https://lars.vilhuber.com/s

(and choose UW Madison 2019) .

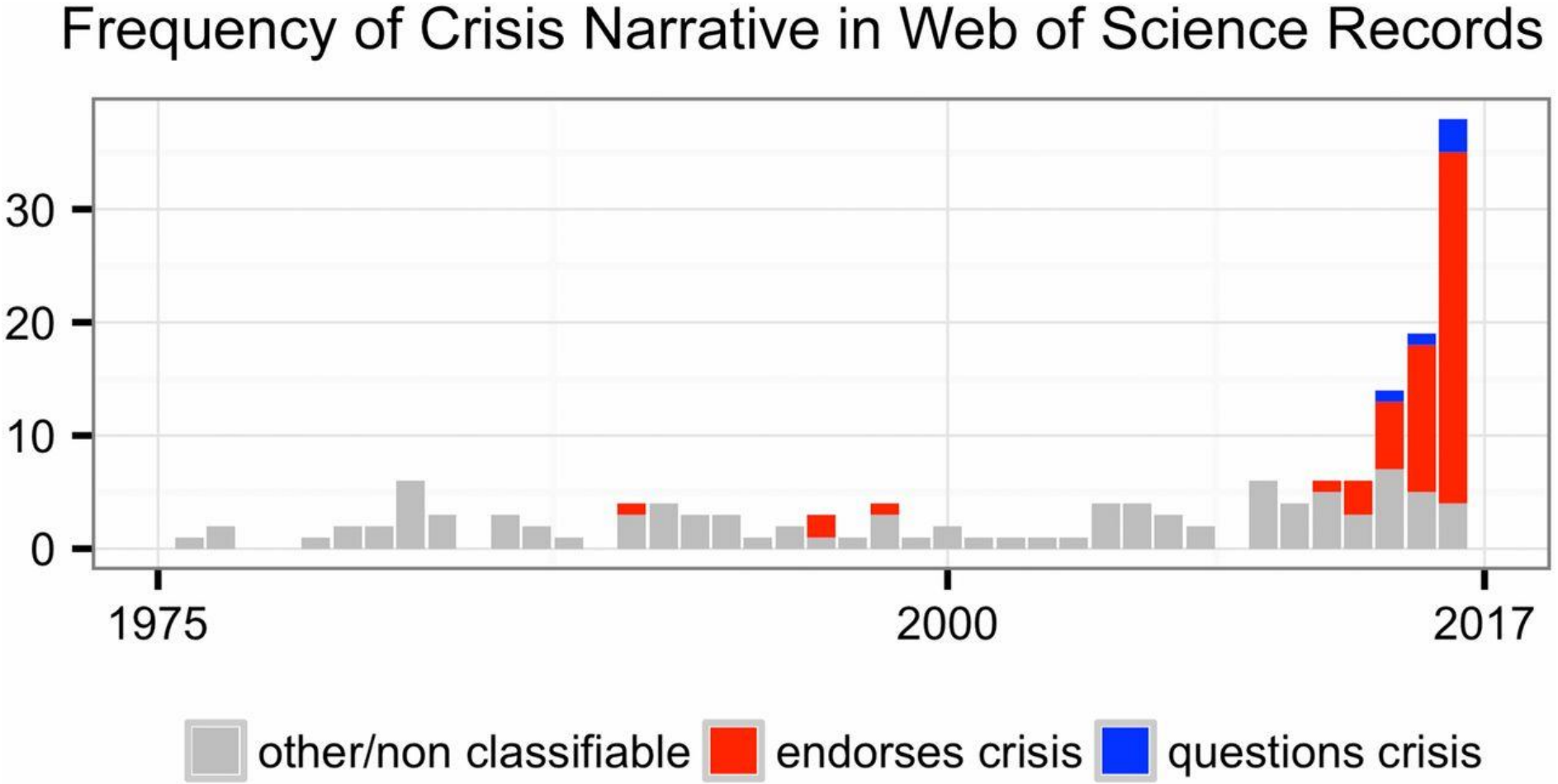
Replication and Reproducibility in Social Sciences and Statistics: Context, Concerns, and Concrete Measures

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….



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

The “crisis” in the 60s and 70s

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

Low power

Flexibility in analysis

Selective reporting

Ignoring nulls

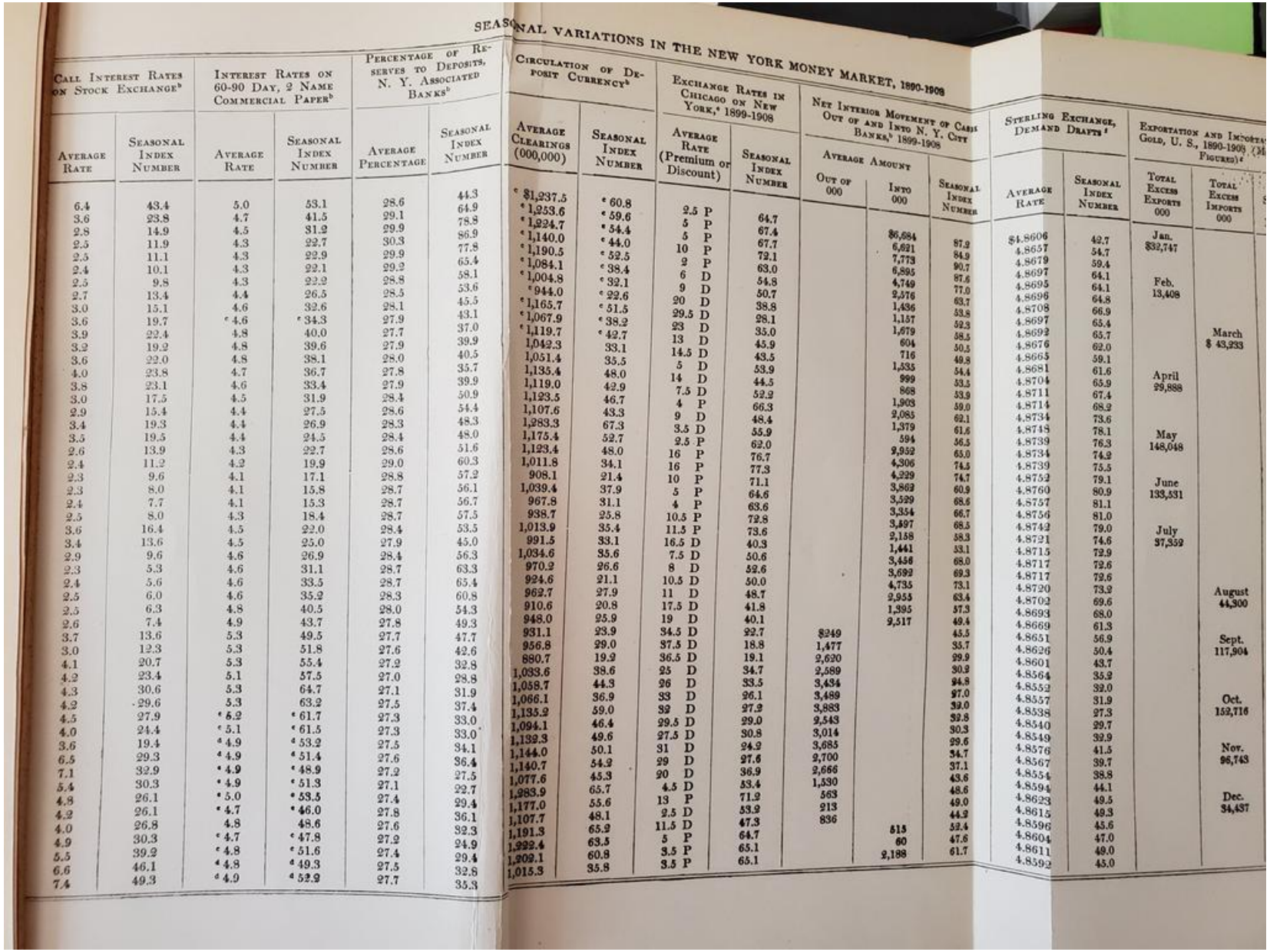
Lack of replication

Misuse of statistics

Source: Nosek Sackler talk 2017

Efficiency of scholary 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 Cash Out of and Into N. Y. City Banks,[b] 1899-1908			Sterling Exchange, Demand Drafts[d]		Exportation and Importation of Gold, U. S., 1890-1908 (Minus 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	Average Amount Into 000	Seasonal Index Number	Average Rate	Seasonal Index Number	Total Excess Exports 000	Total Excess Imports 000
6.4	43.4	5.0	53.1	28.6	44.3	[c] $1,237.5	[c] 60.8	2.5 P	64.7		$6,684	87.2	$4.8606	42.7	Jan. $32,747	
3.6	23.8	4.7	41.5	29.1	64.9	[c] 1,253.6	[c] 59.6	5 P	67.4		6,621	84.9	4.8657	54.7		
2.8	14.9	4.5	31.2	29.9	78.8	[c] 1,224.7	[c] 54.4	5 P	67.7		7,773	90.7	4.8679	59.4		
2.5	11.9	4.3	22.7	30.3	86.9	[c] 1,140.0	[c] 44.0	10 P	72.1		6,895	87.6	4.8697	64.1	Feb. 13,408	
2.5	11.1	4.3	22.9	29.9	77.8	[c] 1,190.5	[c] 52.5	2 P	63.0		4,749	77.0	4.8695	64.1		
2.4	10.1	4.3	22.1	29.2	65.4	[c] 1,084.1	[c] 38.4	6 D	54.8		2,576	63.7	4.8696	64.8		
2.5	9.8	4.3	22.2	28.8	58.1	[c] 1,004.8	[c] 32.1	9 D	50.7		1,436	53.8	4.8708	66.9		
2.7	13.4	4.4	26.5	28.5	53.6	[c] 944.0	[c] 22.6	20 D	38.8		1,157	52.3	4.8697	65.4		
3.0	15.1	4.6	32.6	28.1	45.5	[c] 1,165.7	[c] 51.5	29.5 D	28.1		1,679	58.5	4.8692	65.7		March $ 43,233
3.6	19.7	[c] 4.6	[c] 34.3	27.9	43.1	[c] 1,067.9	[c] 38.2	23 D	35.0		604	50.5	4.8676	62.0		
3.9	22.4	4.8	40.0	27.7	37.0	[c] 1,119.7	[c] 42.7	13 D	45.9		716	49.8	4.8665	59.1		
3.2	19.2	4.8	39.6	27.9	39.9	1,042.3	33.1	14.5 D	43.5		1,535	54.4	4.8681	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	4.8704	65.9	April 29,888	
4.0	23.8	4.7	36.7	27.8	35.7	1,135.4	48.0	14 D	44.5		868	53.9	4.8711	67.4		
3.8	23.1	4.6	33.4	27.9	39.9	1,119.0	42.9	7.5 D	52.2		1,903	59.0	4.8714	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	4.8734	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	4.8748	78.1	May 148,048	
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	4.8739	76.3		
3.5	19.5	4.4	24.5	28.4	48.0	1,175.4	52.7	2.5 P	62.0		2,952	65.0	4.8734	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	4.8739	75.5		
2.4	11.2	4.2	19.9	29.0	60.3	1,011.8	34.1	16 P	77.3		4,229	74.7	4.8752	79.1	June 133,531	
2.3	9.6	4.1	17.1	28.8	57.2	908.1	21.4	10 P	71.1		3,863	69.9	4.8760	80.9		
2.3	8.0	4.1	15.8	28.7	56.1	1,039.4	37.9	5 P	64.6		3,529	68.6	4.8757	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	4.8756	81.0		
2.5	8.0	4.3	18.4	28.7	57.5	938.7	25.8	10.5 P	72.8		3,497	68.5	4.8742	79.0	July 37,352	
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	4.8721	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	4.8715	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	4.8717	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	4.8717	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	4.8720	73.2		August 44,300
2.5	6.0	4.6	35.2	28.3	60.8	962.7	27.9	11 D	48.7		2,955	63.4	4.8702	69.6		
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	4.8693	68.0		
2.6	7.4	4.9	43.7	27.8	49.3	948.0	25.9	19 D	40.1		2,517	49.4	4.8669	61.3		
3.7	13.6	5.3	49.5	27.7	47.7	931.1	23.9	34.5 D	22.7	$249		45.5	4.8651	56.9		Sept. 117,904
3.0	12.3	5.3	51.8	27.6	42.6	956.8	29.0	37.5 D	18.8	1,477		35.7	4.8626	50.4		
4.1	20.7	5.3	55.4	27.2	32.8	880.7	19.2	36.5 D	19.1	2,620		29.9	4.8601	43.7		
4.2	23.4	5.1	57.5	27.0	28.8	1,033.6	38.6	25 D	34.7	2,589		30.2	4.8564	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		24.8	4.8552	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		27.0	4.8557	31.9		Oct. 152,716
4.5	27.9	[c] 5.2	[c] 61.7	27.3	33.0	1,135.2	59.0	32 D	27.2	3,883		32.0	4.8538	27.3		
4.0	24.4	[c] 5.1	[c] 61.5	27.3	33.0	1,094.1	46.4	29.5 D	29.0	2,543		32.8	4.8540	29.7		
3.6	19.4	[d] 4.9	[d] 53.2	27.5	34.1	1,132.3	49.6	27.5 D	30.8	3,014		30.3	4.8549	32.9		
6.5	29.3	[d] 4.9	[d] 51.4	27.6	36.4	1,144.0	50.1	31 D	24.2	3,685		29.6	4.8576	41.5		Nov. 96,743
7.1	32.9	[d] 4.9	[d] 48.9	27.2	27.5	1,140.7	54.2	29 D	27.6	2,700		34.7	4.8567	39.7		
5.4	30.3	[d] 4.9	[d] 51.3	27.1	22.7	1,077.6	45.3	20 D	36.9	2,666		37.1	4.8554	38.8		
4.8	26.1	[d] 5.0	[d] 53.5	27.4	29.4	1,283.9	65.7	4.5 D	53.4	1,530		43.6	4.8594	44.1		
4.2	26.1	[d] 4.7	[d] 46.0	27.8	36.1	1,177.0	55.6	13 P	71.2	563		48.6	4.8623	49.5		Dec. 34,437
4.0	26.8	4.8	48.6	27.6	32.3	1,107.7	48.1	2.5 D	53.2	213		49.0	4.8615	49.3		
4.9	30.3	[c] 4.7	[c] 47.8	27.2	24.9	1,191.3	65.2	11.5 D	47.3	836		44.2	4.8596	45.6		
5.5	39.2	[c] 4.8	[c] 51.6	27.4	29.4	1,222.4	63.5	5 P	64.7		513	52.4	4.8604	47.0		
6.6	46.1	[d] 4.8	[d] 49.3	27.5	32.8	1,202.1	60.8	3.5 P	65.1		60	47.6	4.8611	49.0		
7.4	49.3	[d] 4.9	[d] 52.2	27.7	35.3	1,015.3	35.8	3.5 P	65.1		2,188	61.7	4.8592	45.0		

From @sdellavi
AER 1911

Efficiency of scholary discourse!

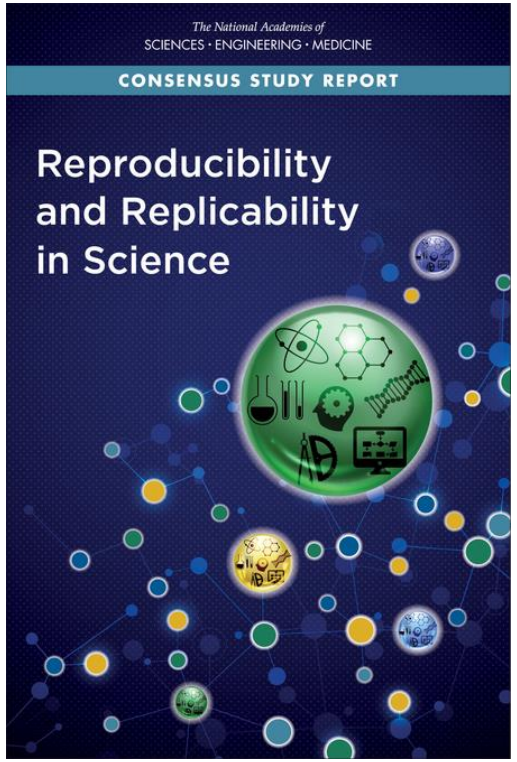
Modern publications thus need
the same transparency and completeness
as in the old days

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

Same data	Same code	Same methods	Same context

Reproducibility

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

Replication continuum

Reproducibility

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

Replicability

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

Replication continuum

Same data	**Different code or software**	Same methods	Same context

Reproducibility

- 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

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

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)

Progress

Progress

- Replication archives and Data (Code) Availability policies

Progress

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

Statistical Software Components

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

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

GAPPORT: Stata module to calculates seats in party-list representation downloads
Ulrich Kohler

GCLSORT: Stata module to sort a single variable via 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

CASD

CRDCN Canadian Research Data Centre Network

FSRDC

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

CIQSS | QICSS
CENTRE INTERUNIVERSITAIRE QUÉBÉCOIS DE STATISTIQUES SOCIALES
QUEBEC INTER-UNIVERSITY CENTER FOR SOCIAL STATISTICS

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

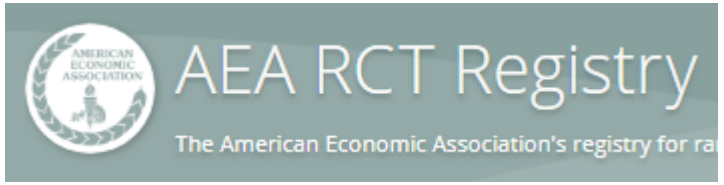
RePEc

Cornell University

arXiv.org

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



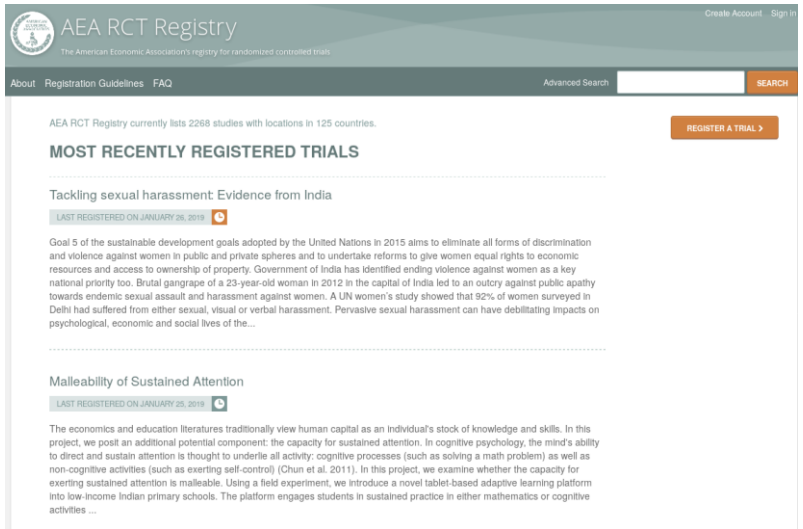
More recently…

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**."



https://www.psychologicalscience.org/publications/psychological_science/preregistration

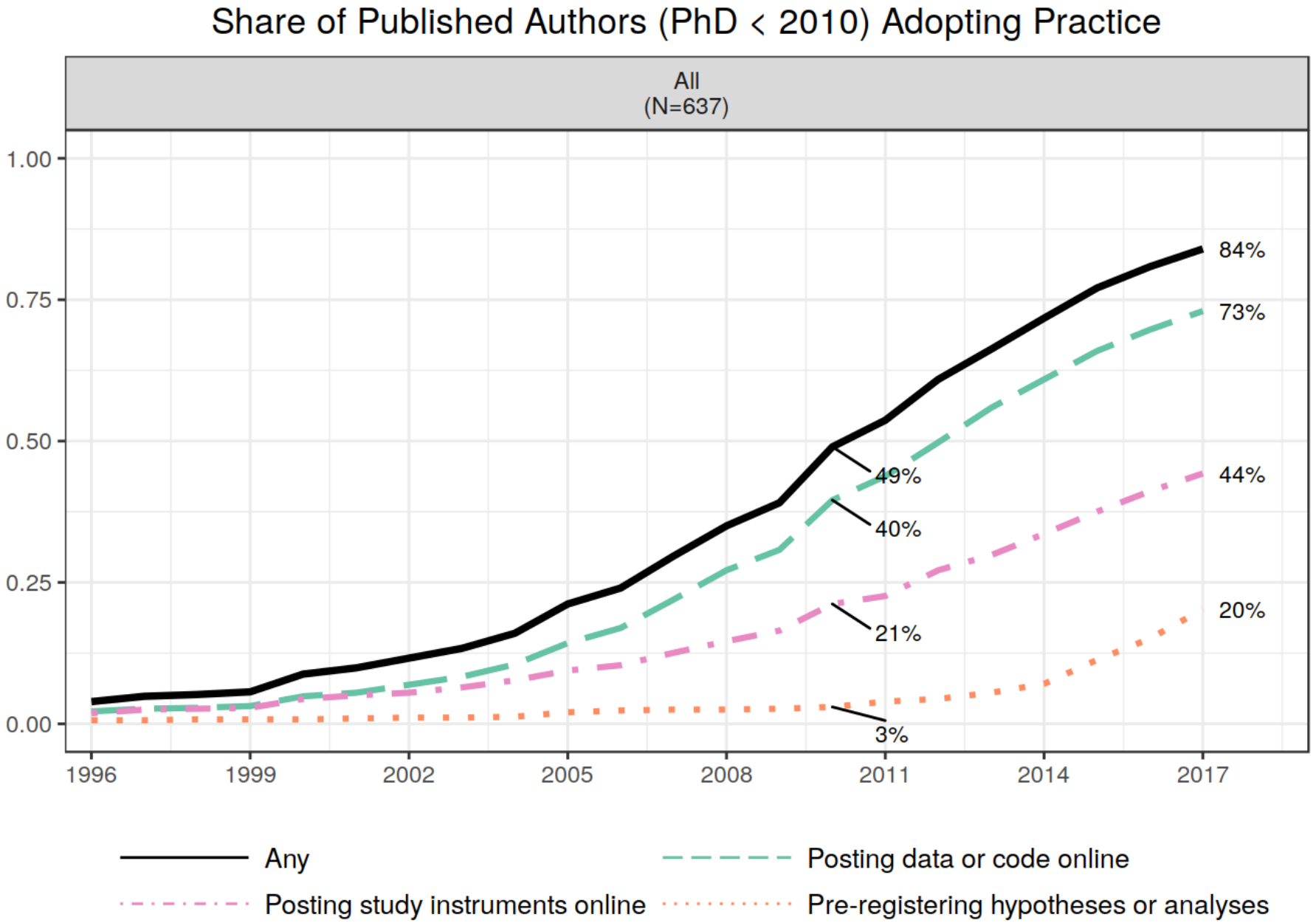
Registered Reports

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

- Close cousin: Results-blind review

Preprints in other sciences

- bioRxiv (2013)
- PsyArXiv (2016)



Paluck (2018) https://osf.io/kvbnh/

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."

Risk

Google

404. That's an error.

The requested URL /a_cool_website was not found on this server. That's all we know.

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

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

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

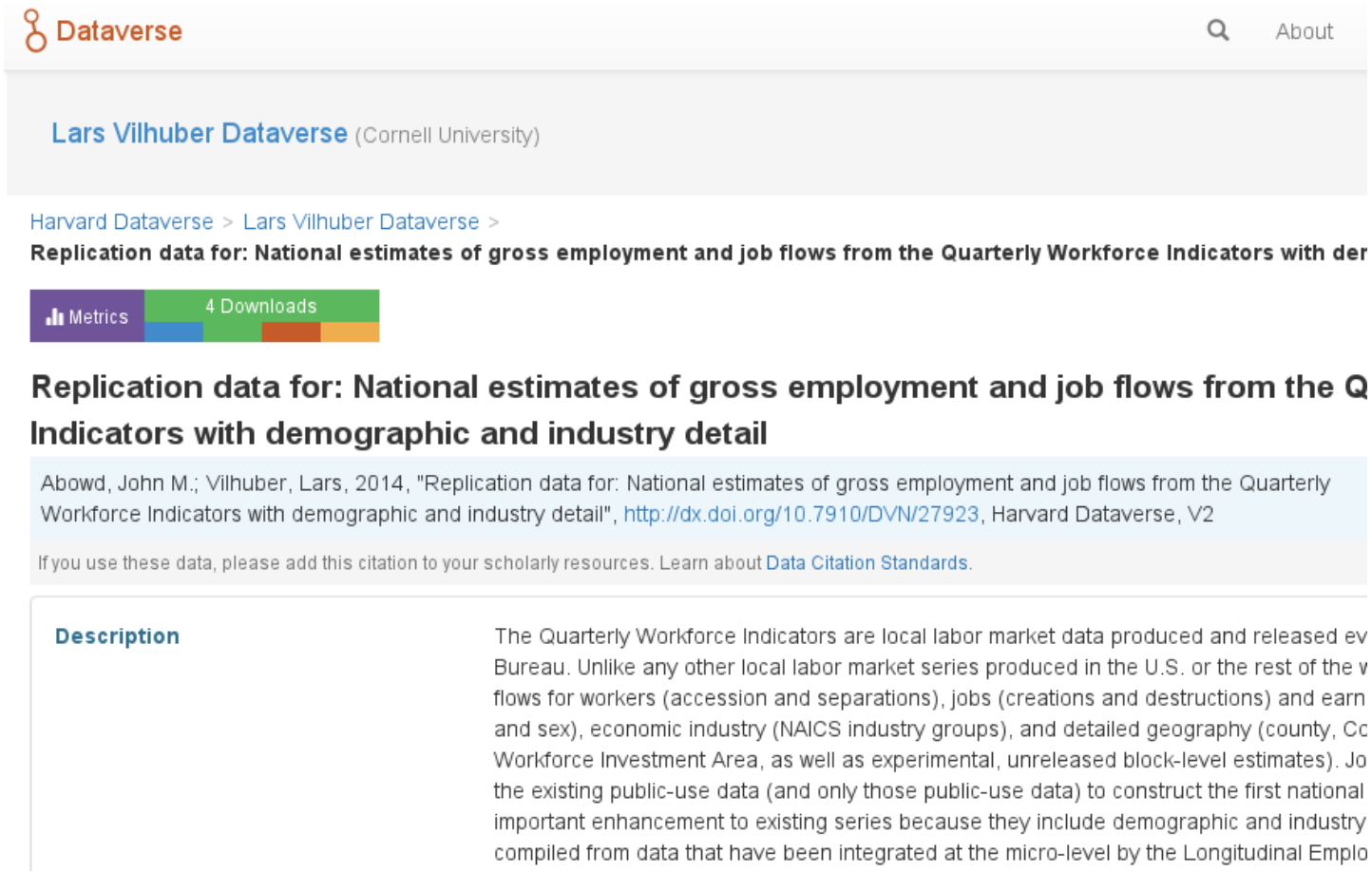
Press for the NBER; 2009. pp. 149–230.

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



We went back, archived it, linked it back

Dataverse

About

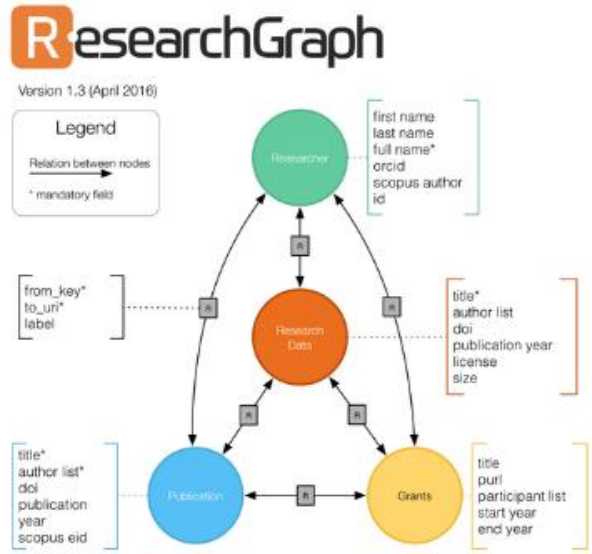
Keyword	Employment Dynamics
Topic Classification	Economics
Related Publication	John M. Abowd and Lars Vilhuber, "National estimates of gross employment and job flows from the Quarterly Wor with demographic and industry detail," Journal of Econometrics, vol. 161, iss. 1, pp. 82-99, 2011. doi: 10.1016/j.jeconom.2010.09.008 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 Wor 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 Wor 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

DataCite

Crossref



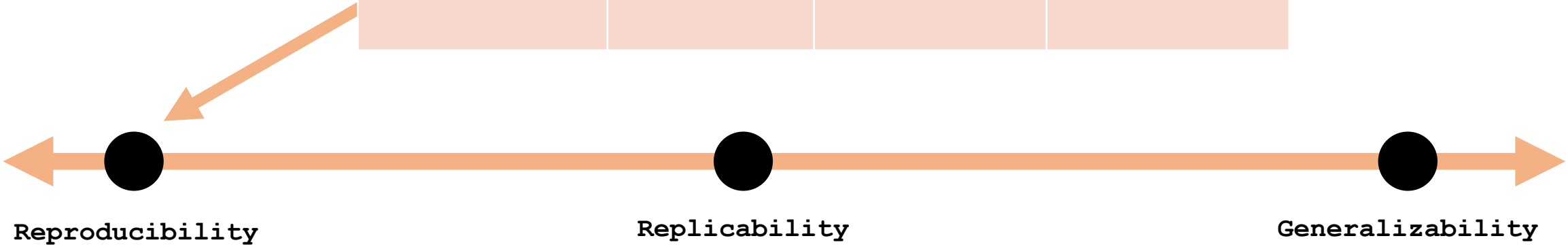
Still true today…

Issues

Not enough articles are reproducible

Replication continuum

Same data	Same code	Same methods	Same context



Reproducibility

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

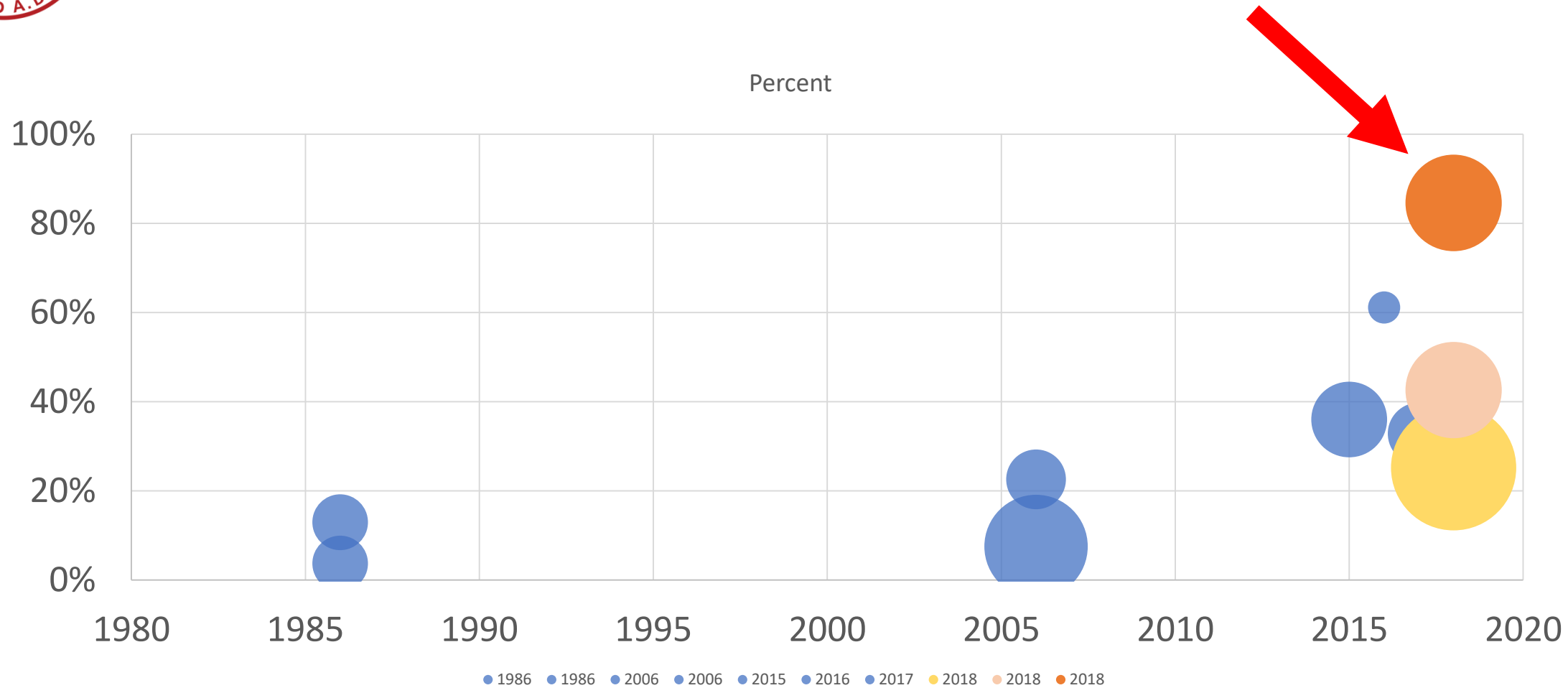
Replicability

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

Generalizability

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

Results?


Percent
100%
80%
60%
40%
20%
0%
1980
1985
1990
1995
2000
2005
2010
2015
2020
1986
1986
2006
2006
2015
2016
2017
2018
2018
2018

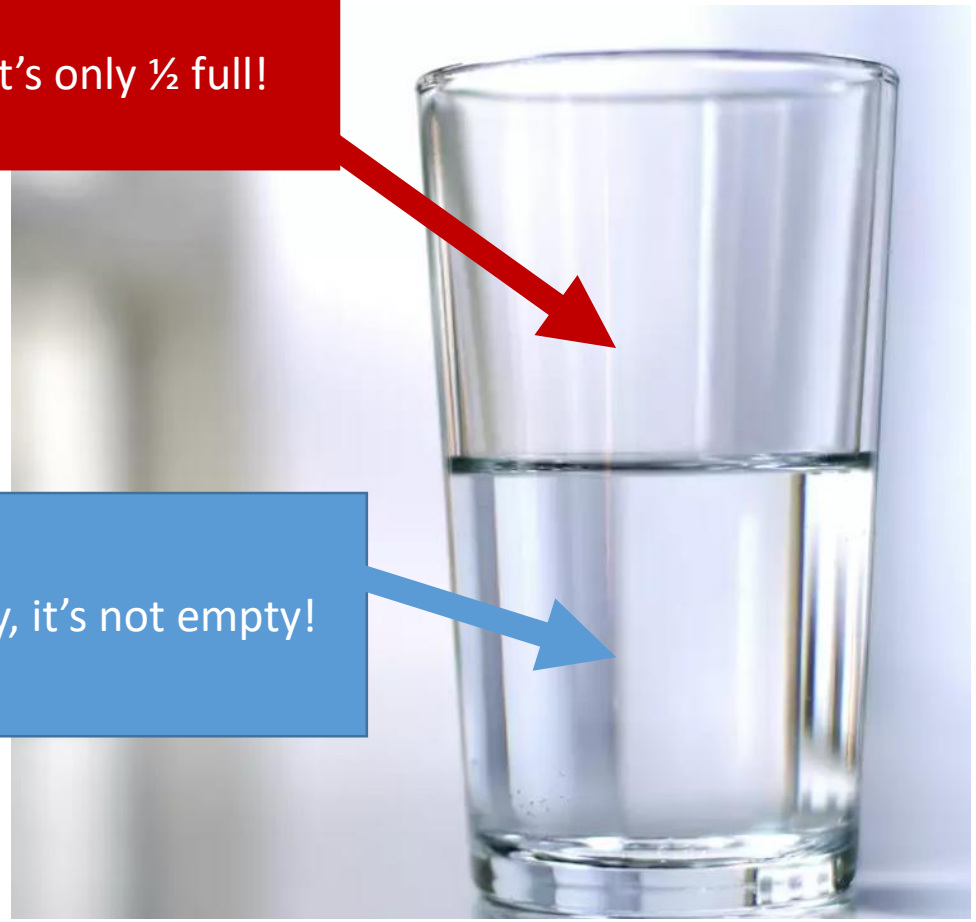
Some key statistics

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								Macroeconomi
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



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?

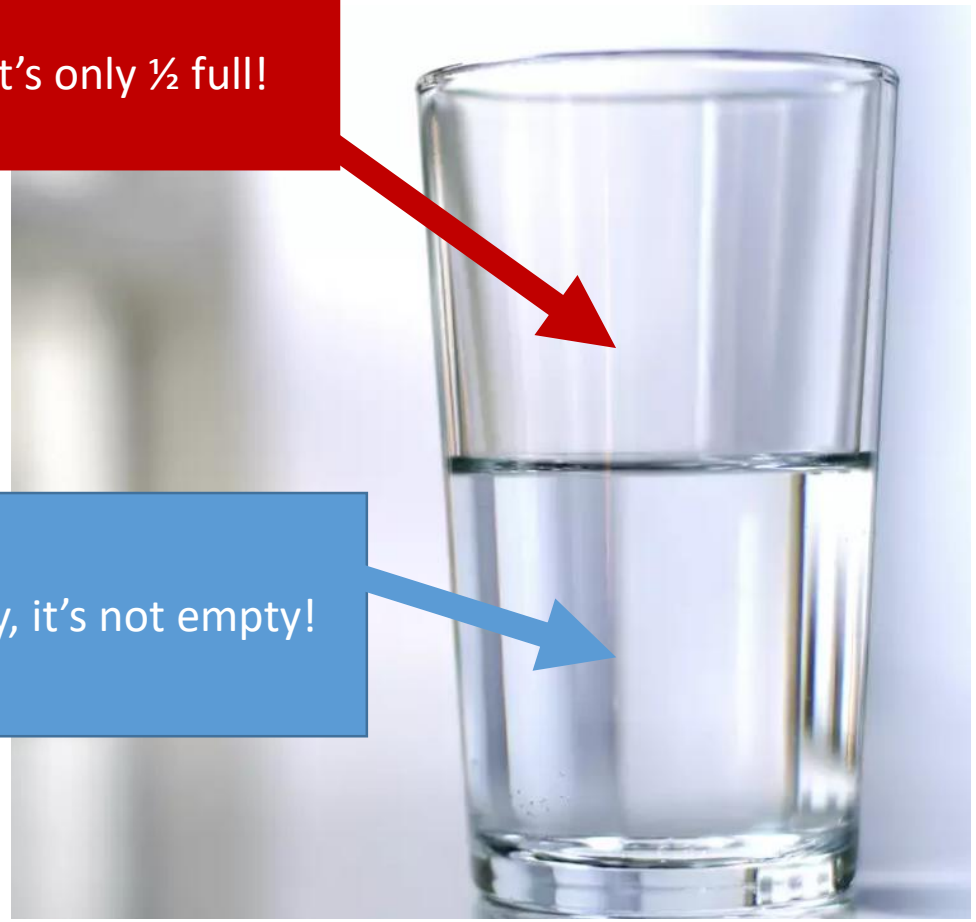
- They DO ensure **longevity/ persistence**
- They DO ensure **access**
- They DO ensure **availability**

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



Evolving Journal and Data Infrastructure

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

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

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 restriced-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" (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 **access to the data and code is clearly and precisely documented and is non-exclusive to the authors.**
- Authors of accepted papers that contain empirical work, simulations, or experimental work must **provide**, **prior to acceptance**, the data, programs, and other details of the computations **sufficient to permit replication**, as well as **information about access to data and programs.**

AEA "Data Availability Policy" (2018)

- 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.

- Data and programs should **be archived in the AEA Data and Code Repository**. Authors will **provide access** to editors and reviewers, if requested**, to both data and programs prior to acceptance**. The Editor should be notified at the time of submission if access to the data used in a paper is restricted or limited, or if, for some other reason, the requirements above cannot be met. **The AEA Data Editor will assess compliance with this policy, and will verify the accuracy of the information prior to acceptance by the Editor.**

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

- Every paper that receives a “conditional acceptance” is verified
  - *Data citations*
  - *Quality of README*
  - *Quality of code*
  - *Reproducibility of code*
  - *Quality of metadata in the repository*

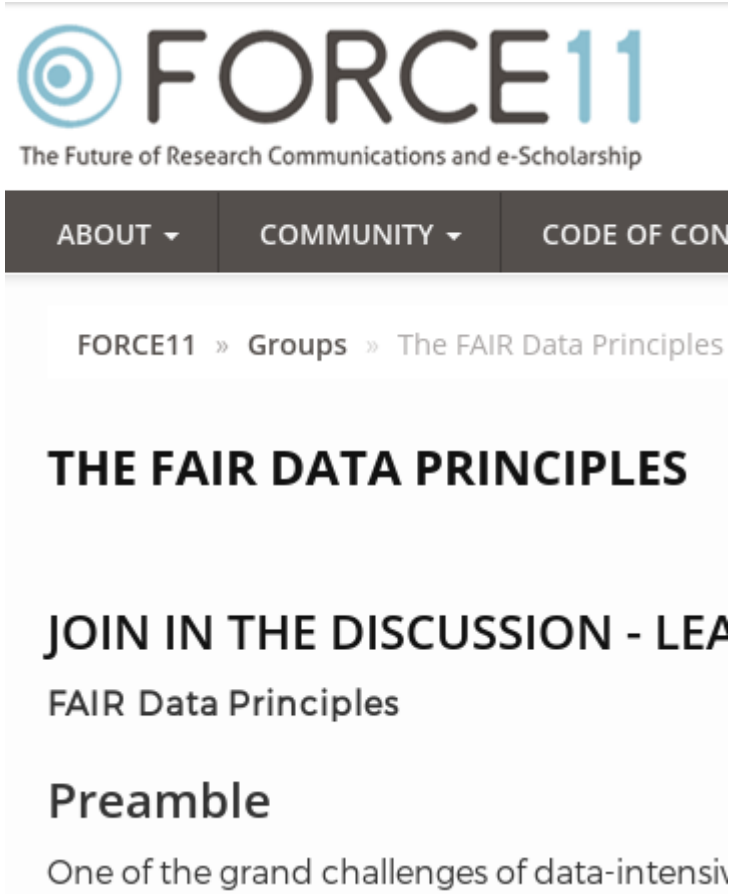
Action: Verifying Data and Code Deposits

- 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?

Action: Data citations and metadata

What is **FAIR**?

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



FAIR principles rely on metadata

Subject Terms
Do not copy/paste multiple terms into this field. Terms must be entered
× Rural roads

JEL Classification
× J43 Agricultural Labor Markets × O12 Microeconomic Analyses of
× O18 Urban, Rural, Regional, and Transportation Analysis • Housing

Manuscript Number
AER-2018-0268.R1 edit remove

Geographic Coverage + add value
India edit remove

Time Period(s) + add value
2000 – 2013 edit remove

Collection Date(s) + add value

Universe
Villages in India without paved roads in 2000. edit remove

Replication continuum

Same data	Same code	Same methods	Same context

Reproducibility

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

Action: Reproducibility Check

Social Science Data Editors

Data and Code Guidance by Data Editors

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

Verification guidance

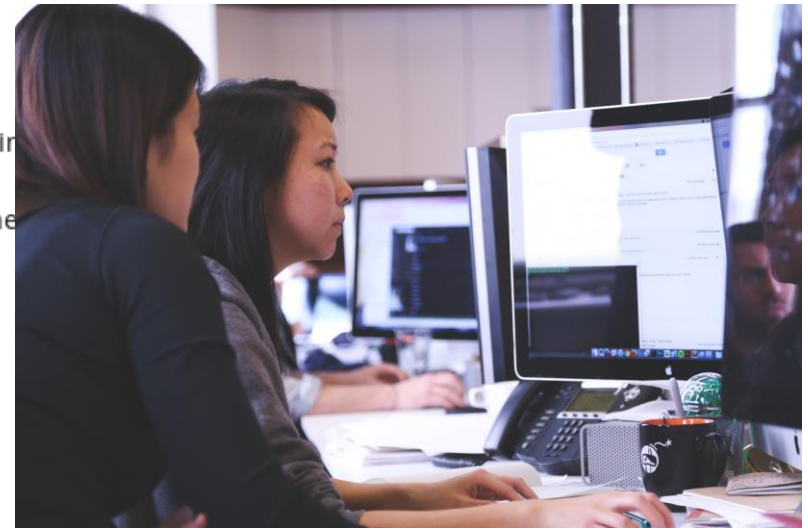
On this page:

- Overview
- Review the README file
- For each listed data source
- For each listed table, figure, in-text number
- Conduct a code verification, if data is available
- Examples

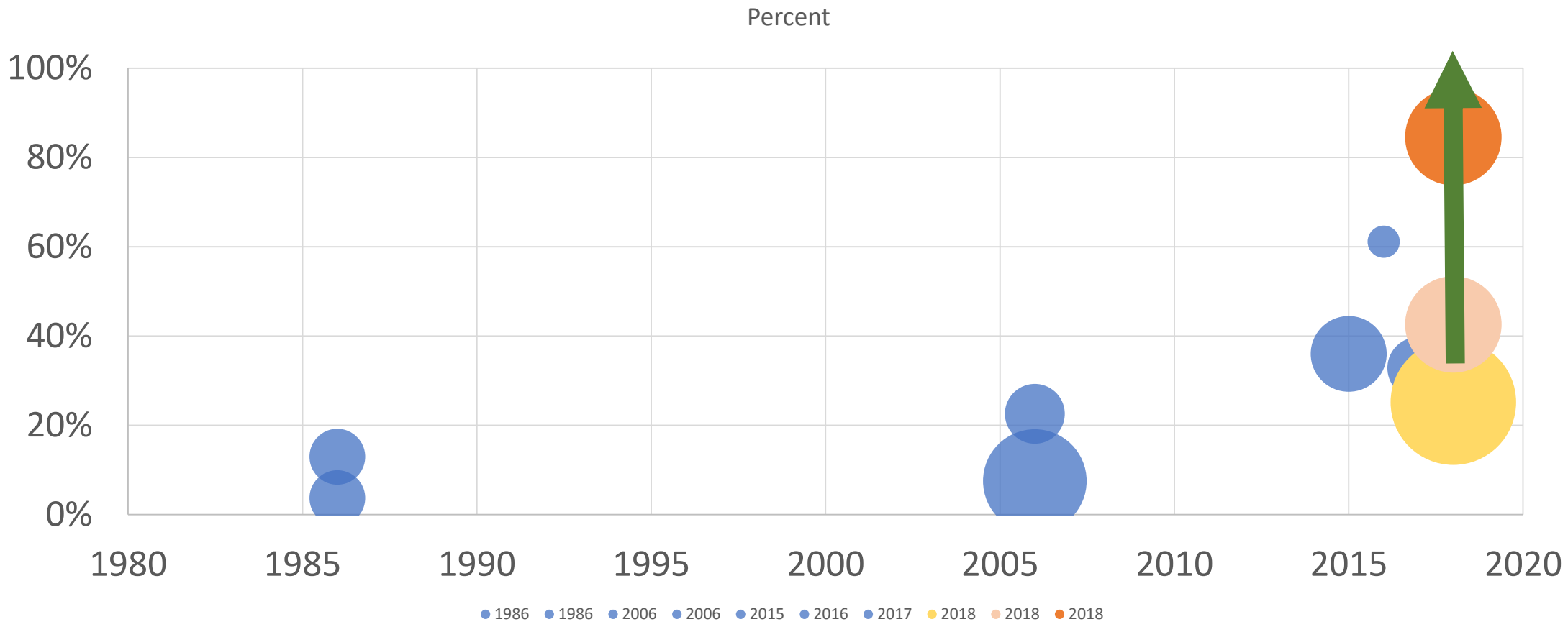
Overview

This document describes

- what authors should check before providir journals
- what verifier teams should check for in the to them for the purpose of verification



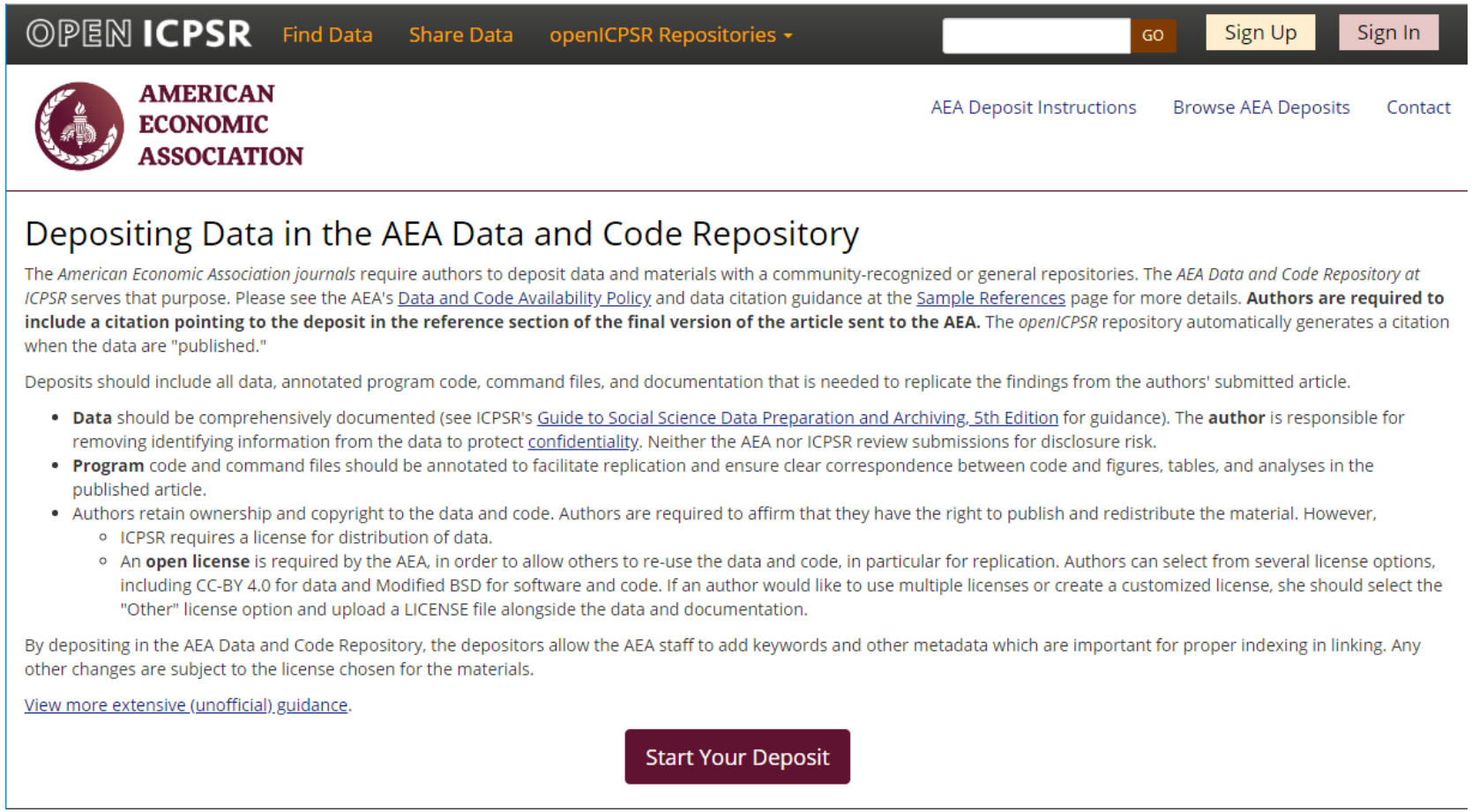
Goal: Improve reproducibility



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

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

Restricted access data

- 40% of Econ articles use restricted access data
  - *Costly or time-intensive to acquire access*
  - *Access may require physical access in California, Australia, Norway, …*
- When collecting own data, informed consent and IRB approval must be obtained to share data
  - *If not done at the start, may not be possible later!*

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

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

  cite!

  - ☑ Data is freely accessible under CC BY-NC 4.0 at 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

  cite!

  - 🔒 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”

Social Science Data Editors

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.

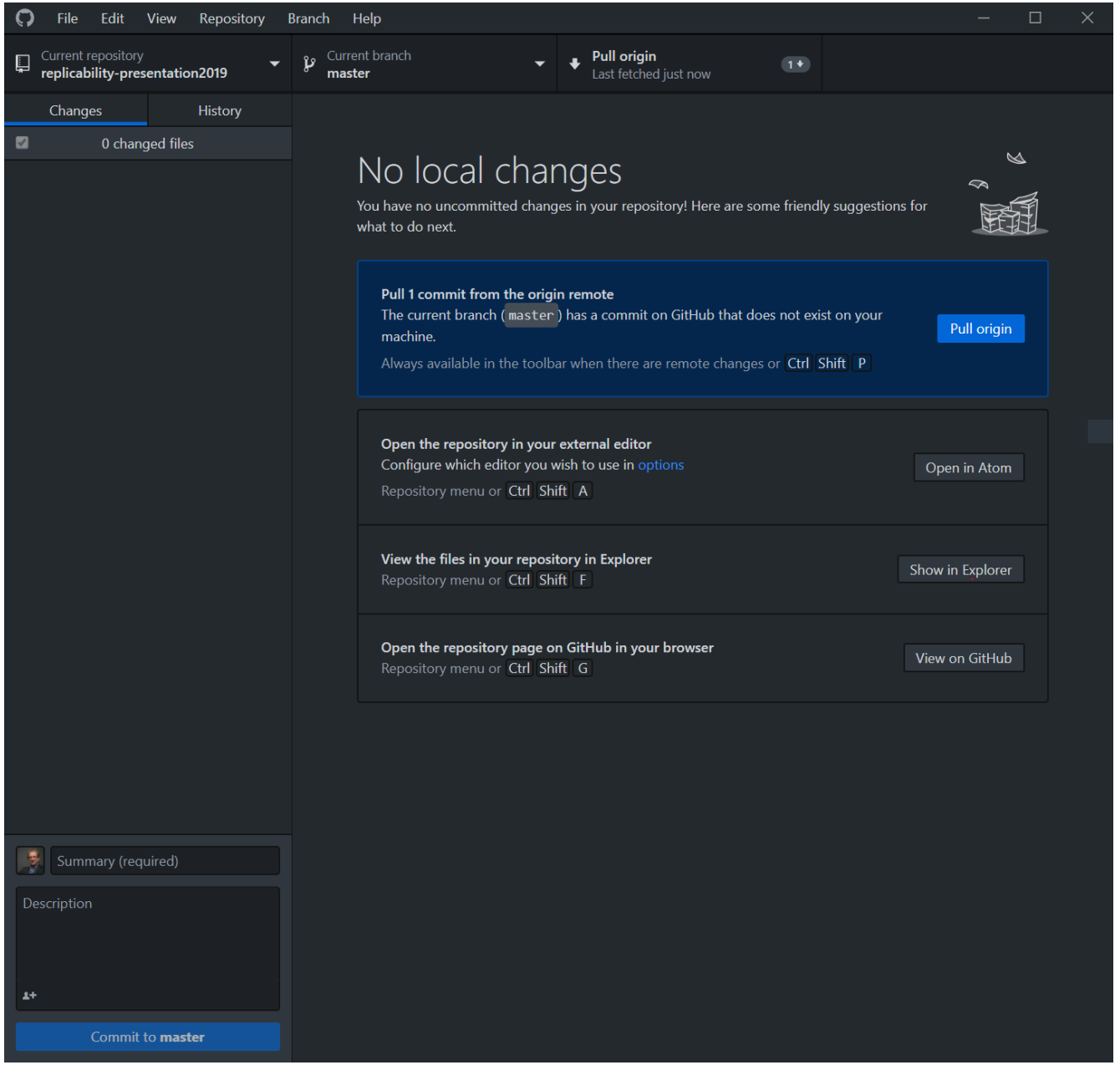
https://social-science-data-editors.github.io/guidance/

Challenges?

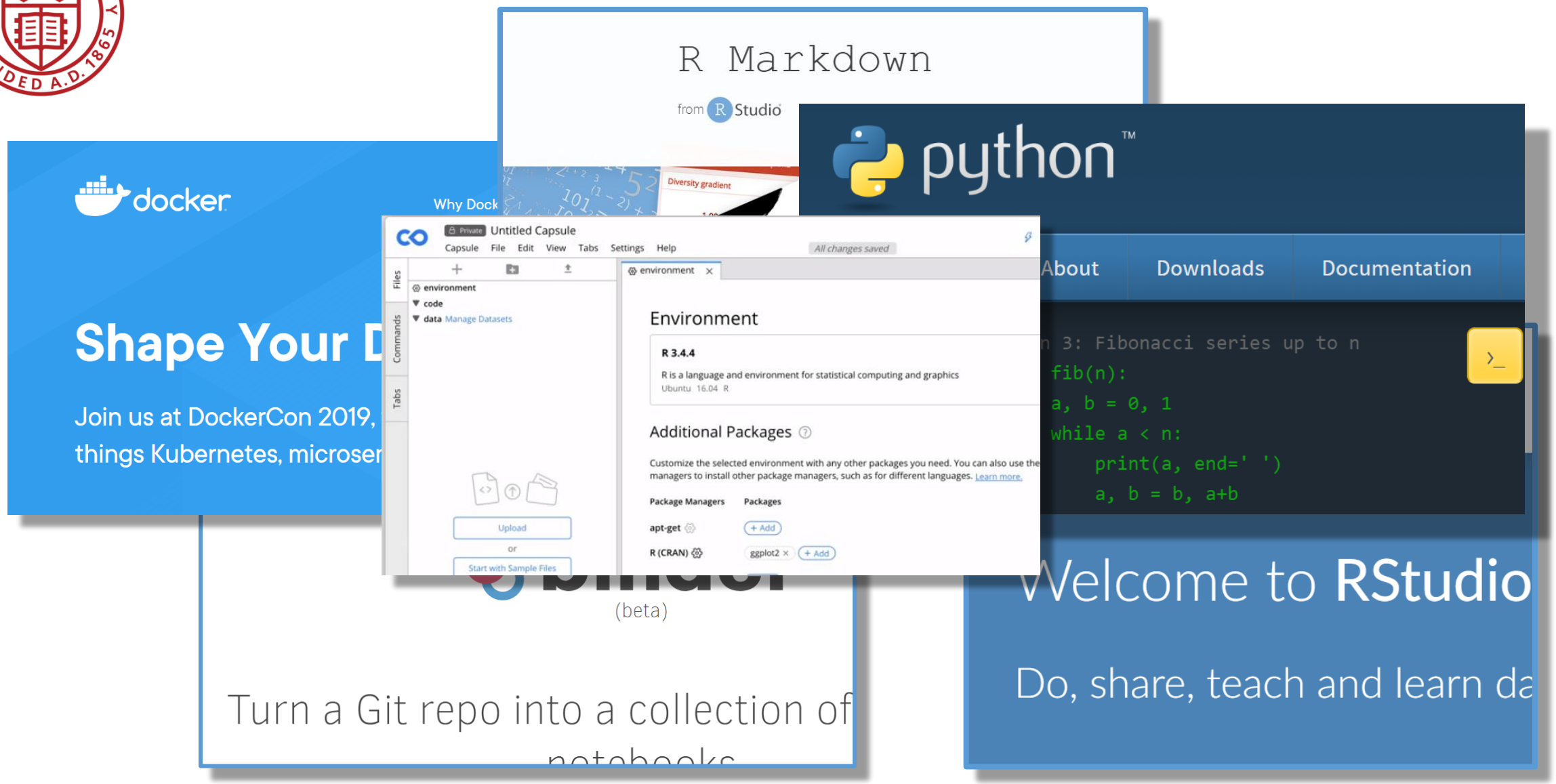
You…

Me…

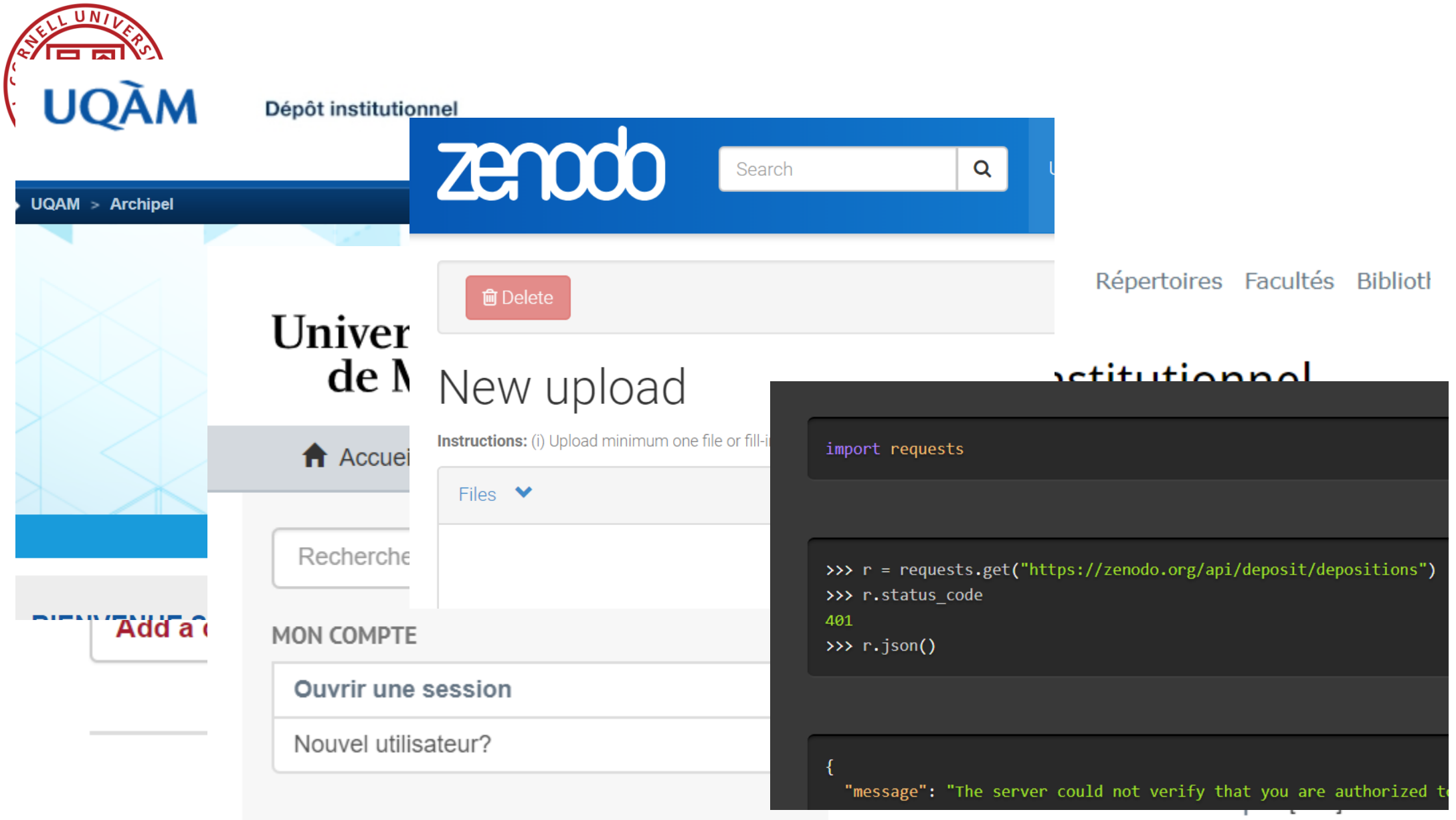
Change ingrained habits…


File
Edit
View
Repository
Branch
Help
Current repository
replicability-presentation2019
Current branch
master
Pull origin
Last fetched just now
1
Changes
History
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.
Pull origin
Always available in the toolbar when there are remote changes or Ctrl Shift P
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
Commit to master

New skills to learn…


R Markdown
from RStudio
Diversity gradient
python™
docker
Why Dock
Shape Your D
Join us at DockerCon 2019,
things Kubernetes, microser
Private Untitled Capsule
Capsule File Edit View Tabs Settings Help
All changes saved
Files
Commands
Tabs
environment
code
data Manage Datasets
Upload
or
Start with Sample Files
environment
Environment
R 3.4.4
R is a language and environment for statistical computing and graphics
Ubuntu 16.04 R
Additional Packages
Customize the selected environment with any other packages you need. You can also use the
managers to install other package managers, such as for different languages. Learn more.
Package Managers
Packages
apt-get
+ Add
R (CRAN)
ggplot2 ×
+ Add
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 …


UQÀM
Dépôt institutionnel
UQAM > Archipel
zenodo
Search
Delete
New upload
Instructions: (i) Upload minimum one file or fill-i
Files
Répertoires
Facultés
Bibliotl
Accuei
Recherche
MON COMPTE
Ouvrir une session
Nouvel utilisateur?
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 t

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 Surrounding Data

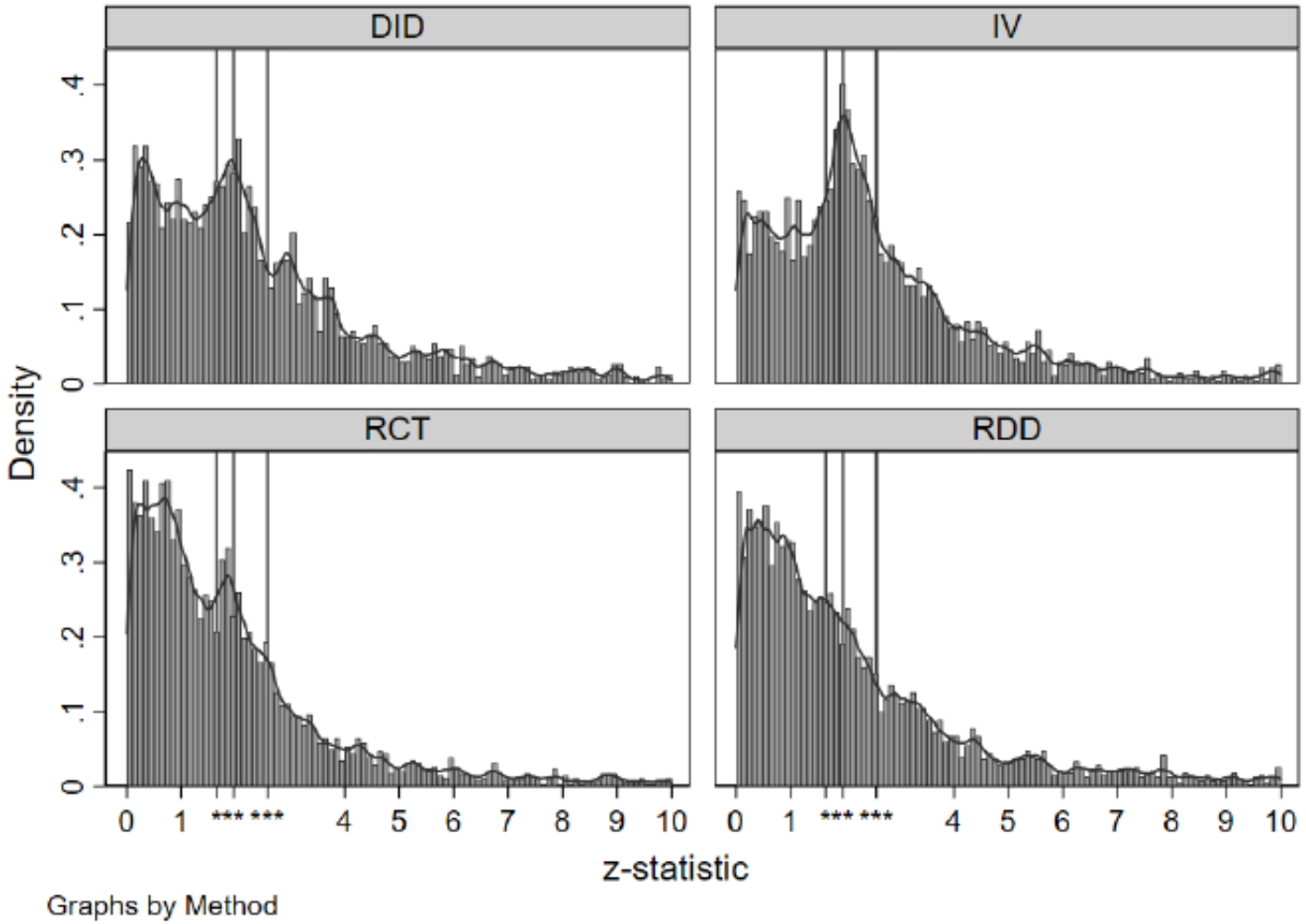
- **Verifiability**
  - How can others obtain access?
- **Documentation**
  - How can others learn about the data?
- **Persistence**
  - How are data and programs preserved?

Thank you!

DOI: 10.5281/zenodo.2573123

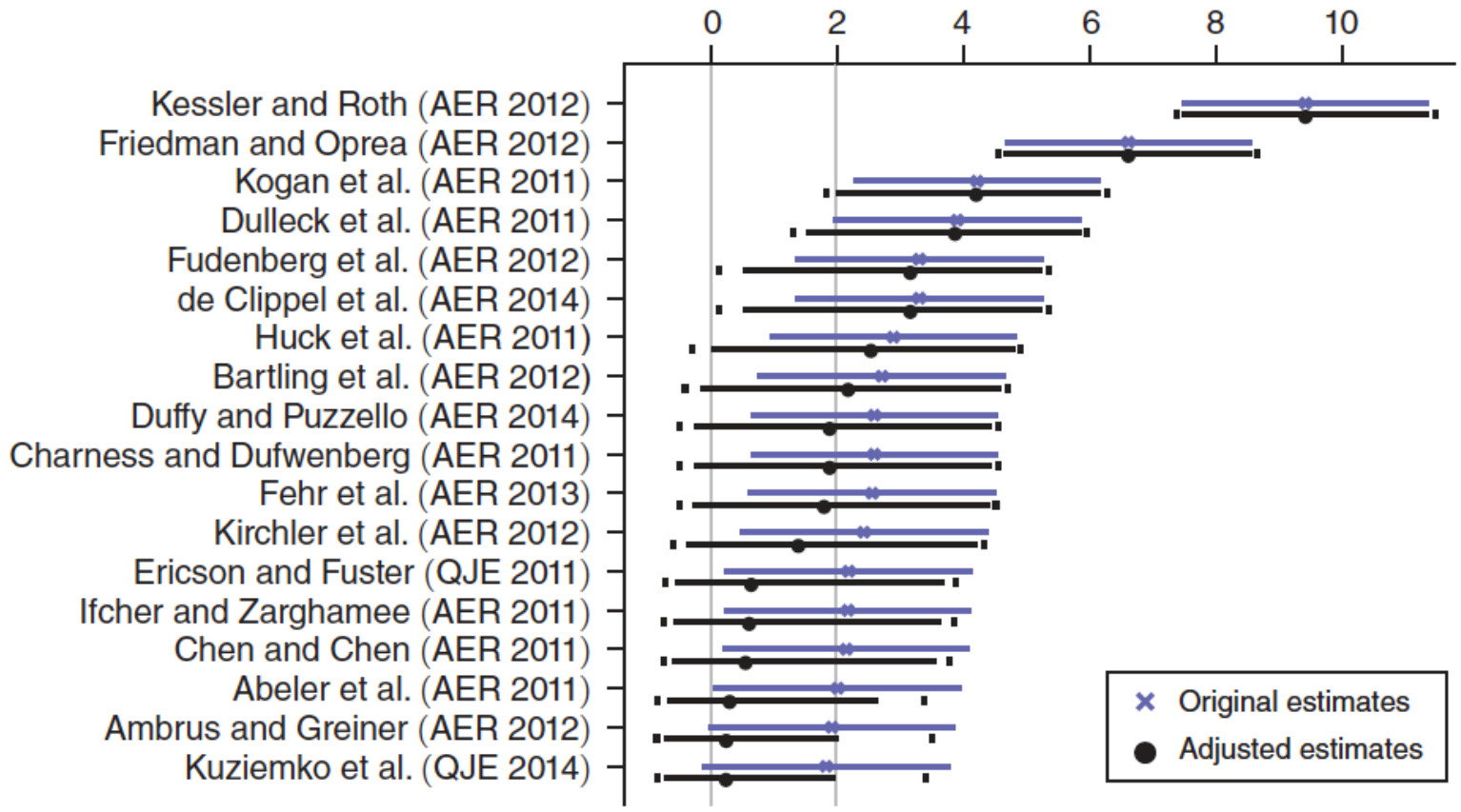
Extra slides

Other challenges remain: Publication bias



Abel Brodeur (2018),
https://osf.io/hg9an/

Correcting for publication bias



Andrews & Kasy (2019), https://doi.org/10.1257/aer.20180310