Implementing Increased Transparency, and Reproducibility in Economics

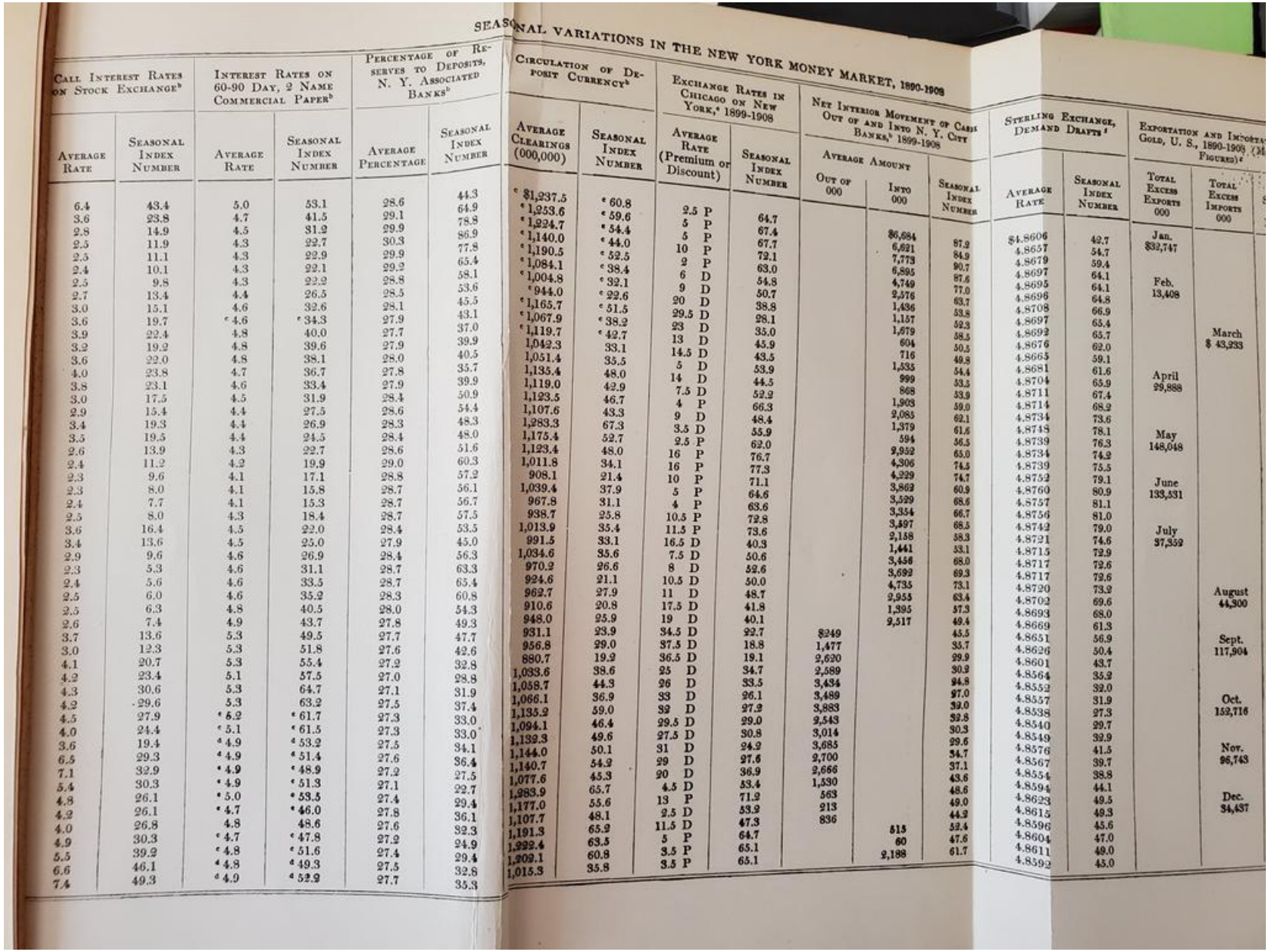
Lars Vilhuber

Cornell University

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.

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				$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		36,684	87.2	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		6,621	84.9	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		7,773	90.7	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		6,895	87.6	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		4,749	77.0	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		2,576	63.7	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		1,436	53.8	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		1,157	52.3	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		1,679	58.5	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		604	50.5	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		716	49.8	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		1,535	54.4	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		999	53.5	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		868	53.9	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		1,903	59.0	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		2,085	62.1	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		1,379	61.6	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		594	56.5	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		2,952	65.0	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		4,306	74.5	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		4,229	74.7	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		3,863	69.9	4.8757	81.1		
2.4	7.7	4.1	15.3	28.7	56.7	967.8	31.1	4 P	63.6		3,529	68.6	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		3,354	66.7	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		3,497	68.5	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		2,158	58.3	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		1,441	53.1	4.8717	72.6		
2.3	5.3	4.6	31.1	28.7	63.3	970.2	26.6	8 D	52.6		3,456	68.0	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		3,692	69.3	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		4,735	73.1	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		2,955	63.4	4.8693	68.0		
2.6	7.4	4.9	43.7	27.8	49.3	948.0	25.9	19 D	40.1		1,395	57.3	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	9,517	49.4	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		45.5	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		33.7	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		29.9	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		30.2	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		34.8	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		37.0	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		32.0	4.8540	29.7		
3.6	19.4	[c] 4.9	[c] 53.2	27.5	34.1	1,132.3	49.6	27.5 D	30.8	3,014		32.8	4.8549	32.9		
6.5	29.3	[c] 4.9	[c] 51.4	27.6	36.4	1,144.0	50.1	31 D	24.2	3,685		30.3	4.8576	41.5		Nov. 96,743
7.1	32.9	[c] 4.9	[c] 48.9	27.2	27.5	1,140.7	54.2	29 D	27.6	2,700		29.6	4.8567	39.7		
5.4	30.3	[c] 4.9	[c] 51.3	27.1	22.7	1,077.6	45.3	20 D	36.9	2,666		34.7	4.8554	38.8		
4.8	26.1	[c] 5.0	[c] 53.5	27.4	29.4	1,283.9	65.7	4.5 D	53.4	1,530		37.1	4.8594	44.1		
4.2	26.1	[c] 4.7	[c] 46.0	27.8	36.1	1,177.0	55.6	13 P	71.2	563		43.6	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		48.6	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		49.0	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		515	44.2	4.8604	47.0		
6.6	46.1	[c] 4.8	[c] 49.3	27.5	32.8	1,202.1	60.8	3.5 P	65.1		60	52.4	4.8611	49.0		
7.4	49.3	[c] 4.9	[c] 52.2	27.7	35.3	1,015.3	35.8	3.5 P	65.1		2,188	47.6	4.8592	45.0		
												61.7				

From @sdellavi
AER 1911

Efficiency of scholary discourse!

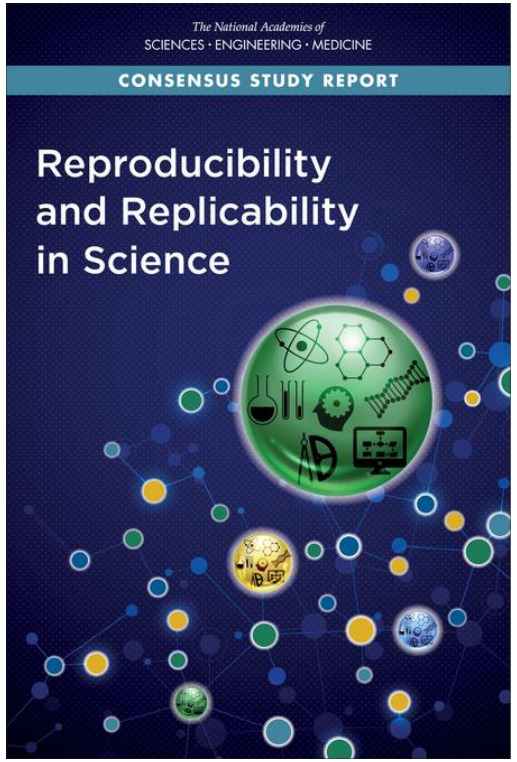
Modern publications thus need
the same transparency and completeness
as in the old days

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	Same code	Same methods	Same context
collection			

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

ICPSR
Sharing data to advance science

CASD

CRDCN Canadian Research Data Centre Network

FSRDC

IPUMS USA

U.S. Census and American Community Survey microdata from 1850 to the present.

VISIT SITE

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

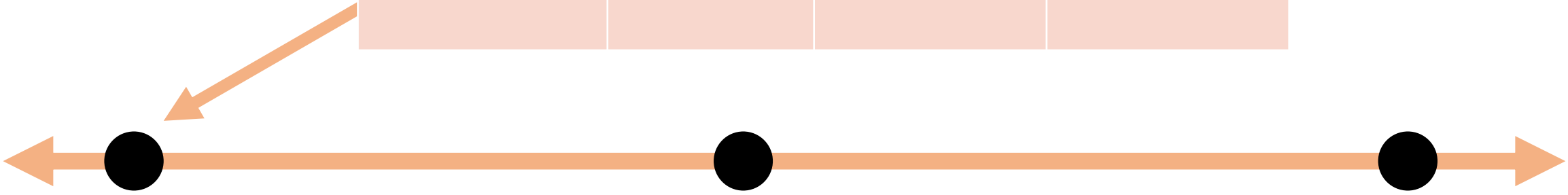
Cornell University

arXiv.org

Issues

Replication continuum

Same data	Same code	Same methods	Same context



Reproducibility

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

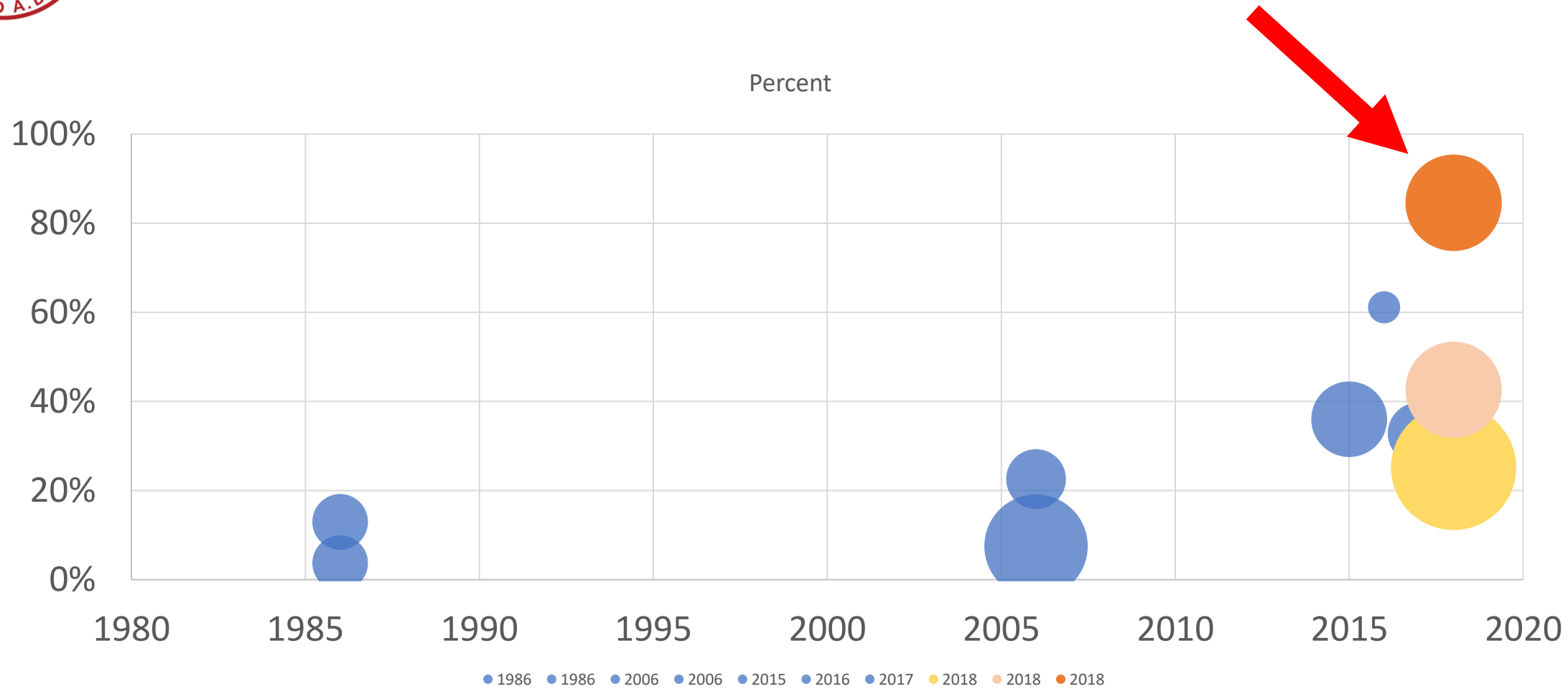
Replicability

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

Generalizability

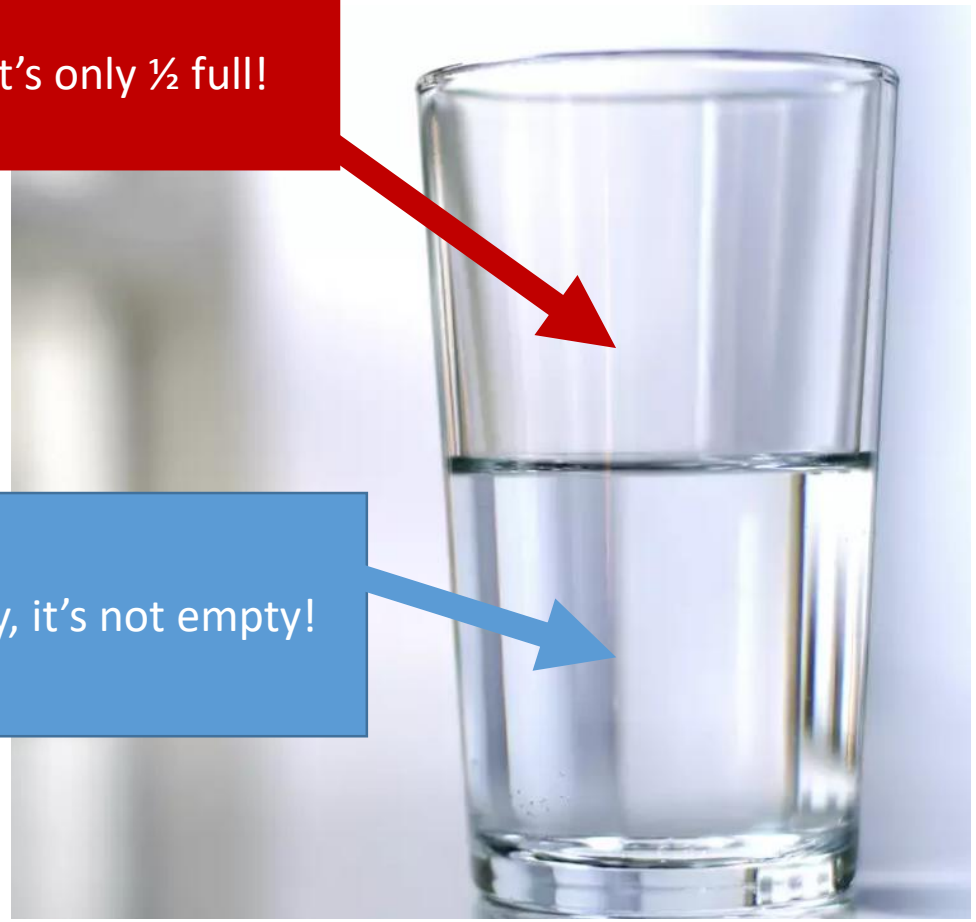
- Wider Replication (Pesaran 2003)
- Scientific Replication (Hamermesh 2007)
- Reanalysis/Robustness (Clemens 2015)

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

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



Failure to curate

Google

404. That's an error.

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

Poor citation practices

- **Macrodata:**

"We use data downloaded from the Bureau of Economic Analysis…"

- **Microdata**:

"… this paper uses data from the Current Population Survey…"

Problems describing 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

What to do?

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)

→ Verifying reproducibility

Current Data Availability Policies are Broken

- If the Data is
  not open-access,

  no systematic information is collected
  (“exemption”)

We asked for “deposits”…

If you used files at
the National Archives,

would we ask you to
“deposit” them?

We asked for “deposits”…

If you used files at
the National Archives,

**you should describe
where they are!**

→Require greater transparency of data/code

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 **trusted repositories** (data archives)?

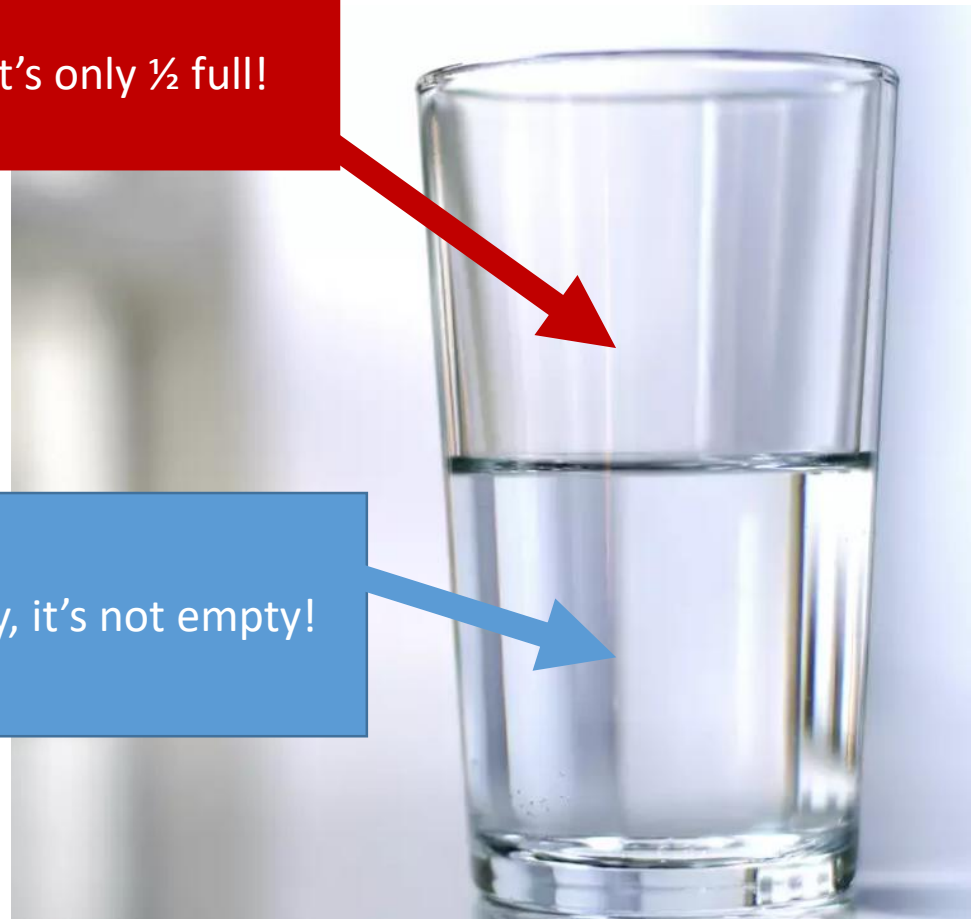
- They DO ensure **longevity/ persistence**
- They DO ensure **access**
- They DO ensure **availability**

Evolving Journal and Data Infrastructure

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

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

→Use trusted repositories where possible

Evolving Journal and Data Infrastructure

Goal: Use any repository!

(subject to conditions)

Problems with that?

Here are the problems…

Failure to curate

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…"

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

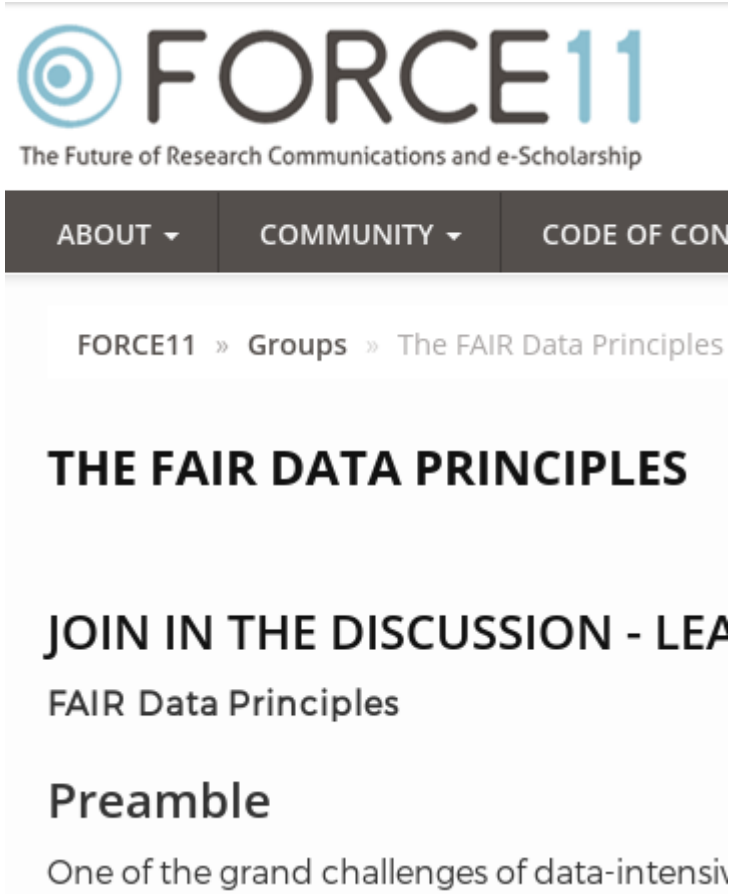
So: Describe them!

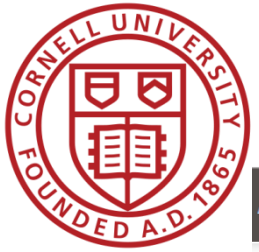
(cite them!)

Action: Data citations and metadata

What is **FAIR**?

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





FORCE11

The Future of Research Communications and e-Scholarship

Search

English

ABOUT | COMMUNITY | CODE OF CONDUCT | GROUPS | RESOURCES | NEWS + BLOGS | EVENTS | PUBLIC

perceived criteria of importance.

1. Importance

Data should be considered legitimate, citable products of research. Data should be accorded the same importance in the scholarly record as citat research objects, such as publications[1].

2. Credit and Attribution

Data citations should facilitate giving scholarly credit and normative and le attribution to all contributors to the data, recognizing that a single style or of attribution may not be applicable to all data[2].

3. Evidence

In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited[3].

4. Unique Identification

A data citation should include a persistent method for identification that i actionable, globally unique, and widely used by a community[4].

5. Access

Data citations should facilitate access to the data themselves and to such

DC1

Data Citation Principles

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [https://www.force11.org/group/joint-declaration-data-citation-principles-final].

Evolving Journal and Data Infrastructure

Data Citations are not enough!

Why are data citations not enough?

- They tell you “where”
- But most do not
  - “who can access”
  - “for how long”
  - “under what conditions”

(Though in theory, these are covered by the Data Citation Principles)

Data Availability Statements (DAS)

- 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

Data Availability Statements

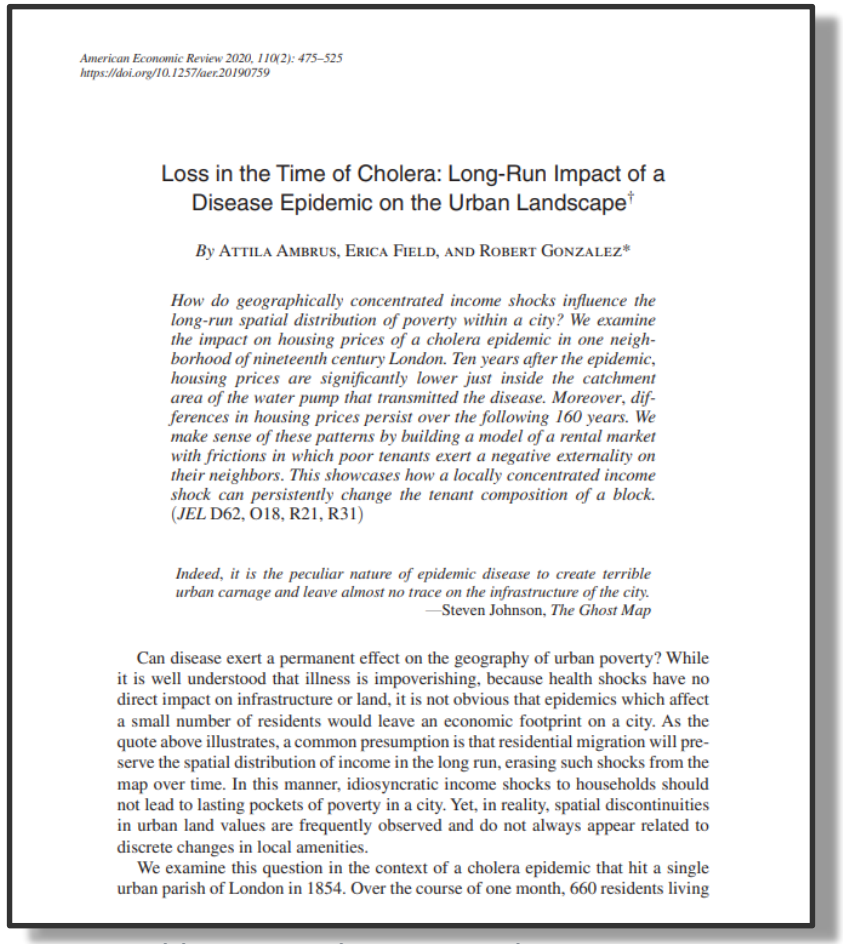
- A statement about **how long** data will be **available** (policy)
  - DOI assignments implies **long-term curation**
  - But long-term curation **does not require DOI**!
- A statement about **usage rights**
  - Not every dataset is in the public domain
  - Not everybody knows that U.S. Government data are usually in the public domain

→Improve provenance documentation

Why Reproducibility, Provenance?

Credibility

Credibility



American Economic Review 2020, 110(2): 475–525
https://doi.org/10.1257/aer.20190759

Loss in the Time of Cholera: Long-Run Impact of a Disease Epidemic on the Urban Landscape†

By Attila Ambrus, Erica Field, and Robert Gonzalez*

How do geographically concentrated income shocks influence the long-run spatial distribution of poverty within a city? We examine the impact on housing prices of a cholera epidemic in one neighborhood of nineteenth century London. Ten years after the epidemic, housing prices are significantly lower just inside the catchment area of the water pump that transmitted the disease. Moreover, differences in housing prices persist over the following 160 years. We make sense of these patterns by building a model of a rental market with frictions in which poor tenants exert a negative externality on their neighbors. This showcases how a locally concentrated income shock can persistently change the tenant composition of a block. (*JEL* D62, O18, R21, R31)

Indeed, it is the peculiar nature of epidemic disease to create terrible urban carnage and leave almost no trace on the infrastructure of the city.
—Steven Johnson, *The Ghost Map*

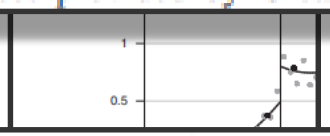
Can disease exert a permanent effect on the geography of urban poverty? While it is well understood that illness is impoverishing, because health shocks have no direct impact on infrastructure or land, it is not obvious that epidemics which affect a small number of residents would leave an economic footprint on a city. As the quote above illustrates, a common presumption is that residential migration will preserve the spatial distribution of income in the long run, erasing such shocks from the map over time. In this manner, idiosyncratic income shocks to households should not lead to lasting pockets of poverty in a city. Yet, in reality, spatial discontinuities in urban land values are frequently observed and do not always appear related to discrete changes in local amenities.

We examine this question in the context of a cholera epidemic that hit a single urban parish of London in 1854. Over the course of one month, 660 residents living

https://doi.org/10.1257/aer.20190759

ics 126 (1): 145–205.
Lagunoff, Roger, and Akihiko Matsui. 1997. "Asynchronous Choice in Repeated Coordination Games." *Econometrica* 65 (6): 1467–77.
Lalive, Rafael. 2008. "How Do Extended Benefits Affect Unemployment Duration? A Regression Discontinuity Approach." *Journal of Econometrics* 142 (2): 785–806.
Land Registry. 2014. "Price Paid Data." http://bit.ly/1HNQAiA (accessed December 19, 2014).
Lee, David S. 2008. "Randomized Experiments from Non-Random Selection in U.S. House Elections." *Journal of Econometrics* 142 (2): 675–97.
Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48 (2): 281–355.
Lee, Sanghoon, and Jeffrey Lin. 2018. "Natural Amenities, Neighbourhood Dynamics, and Persistence in the Spatial Distribution of Income." *Review of Economic Studies* 85 (1): 663–94.
LonRes. 2015. "LonRes: Rental Price Archives." Access provided by Greater London Properties.

long-run spatial distribution of poverty within a city? We examine the impact on housing prices of a cholera epidemic in one neighborhood of nineteenth century London. Ten years after the epidemic, housing prices are significantly lower just inside the catchment area of the water pump that transmitted the disease. Moreover, dif-



the names of the primary occupant at each address
rds.
ended in 1963. Hence, for the years 1995–2013 we
from the Land Registry of England (Land Registry
property address as well as the sale price and date of
ntal prices of all properties rented within the Soho area
May 2015 from the *LonRes* data archives, the primary
ntral London and only available to verified real estate
2015, we obtain house value estimates from Zoopla,
website.[15] We digitized all valuations and addresses
d addresses above by matching them to housing maps
from the relevant time period (for historic records) or using Google's geocoder tool
(for current house records).

To assess the spatial distribution of cholera deaths, we map the total number of
deaths by house using the Cholera Inquiry Committee's 1855 map (Cholera Inquiry
of a scientific
local chaplain
h all residences
eal disease that
resulting map,
of deaths per
eath certificates
which records
om individuals

ure from maps

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik. 2015. "Optimal Data-Driven Regression Discontinuity Plots." *Journal of the American Statistical Association* 110 (512): 1753–69.
Card, David, Alexandre Mas, and Jesse Rothstein. 2008. "Tipping and the Dynamics of Segregation." *Quarterly Journal of Economics* 123 (1): 177–218.
Cholera Inquiry Committee. 1855. ***Report on the Cholera Outbreak in the Parish of St. James, Westminster, during the Autumn of 1854.*** London: J. Churchill.
Conley, T. G. 1999. "GMM Estimation with Cross Sectional Dependence." *Journal of Econometrics* 92 (1): 1–45.
Conley, Timothy G. 2008. "Spatial Econometrics." Unpublished.
Dell, Melissa. 2010. "The Persistent Effects of Peru's Mining Mita." *Econometrica* 78 (6): 1863–1903.

https://doi.org/10.1257/aer.20190759

Data and Code for: Loss in the Time of Cholera: Long-run Impact of a Disease Epidemic on the Urban Landscape

Principal Investigator(s): Attila Ambrus, Duke University; Erica Field, Duke University; Robert Gonzalez, University of South Carolina

Version: V2

AMERICAN ECONOMIC ASSOCIATION

Do-files, input Data, and Output Figures and Tables

NOTE: Master do-file (Master.do) provides all Tables and Figures

Do-file	Input datasets	Output
Table_summary_stats.do	houses_1853_final.dta	Table 1 Table B1
Table_deaths.do	Merged_1853_1864_data.dta	Table 2
Table_main_results.do	Merged_1853_1864_data.dta Merged_1846_1894_data.dta houses_1936_final.dta	Table 3
Table_moved.do	Merged_1853_1864_data.dta	Table 4
Table_migration.do	Merged_1853_1864_data.dta	Table 5
Table_census.do	Data_census.dta	Table 6
Table_Booth_data.do	final_booth_RG.dta	Table 7
Table_current_results.do	houses_current_final.dta current_rentals_final.dta	Table 8
Fig_RD_plots.do	Merged_1853_1864_data.dta houses_1936_final.dta Data_census.dta final_booth_RG.dta houses_current_final.dta current_rentals_final.dta	Figure 2 Figure B1 Figure B2 Figure B3 Figure B4 Figure B5
Fig_variance_grid.do	grid_house_final	Figure 4
Table_fuzzy_iv.do	Merged_1853_1864_data.dta	Table B2

st Modified

/02/2019 02:23:PM

/21/2019 10:47:AM

f Cholera: Long-run Impact of
ation [publisher], 2020. Ann
20-01-31. https://doi.org

un spatial distribution of
one neighborhood of 19th
t inside the catchment area of
ersist over the following 160
tions in which poor tenants

Data and Code for: Loss in the Time of Cholera: Long-run Impact of a Disease Epidemic on the Urban Landscape

Principal Investigator(s): Attila Ambrus, Duke Unive
University; Robert Gonzalez, University of South Carolina

Version: V2

Do-files, input Data, and Output Figures and Tables

NOTE: Master do-file (Master.do) provides all Tables and Figures

Do-file	Input datasets	
Table_summary_stats.do	houses_1853_final.dta	
Table_deaths.do	Merged_1853_1864_data.dta	
Table_main_results.do	Merged_1853_1864_data.dta Merged_1846_1894_data.dta houses_1936_final.dta	
Table_moved.do	Merged_1853_1864_data.dta	
Table_migration.do	Merged_1853_1864_data.dta	
Table_census.do	Data_census.dta	
Table_Booth_data.do	final_booth_RG.dta	
Table_current_results.do	houses_current_final.dta current_rentals_final.dta	Table 8
Fig_RD_plots.do	Merged_1853_1864_data.dta	Figure 2
	houses_1936_final.dta	Figure B1
	Data_census.dta	Figure B2
	final_booth_RG.dta	Figure B3
	houses_current_final.dta	Figure B4
	current_rentals_final.dta	Figure B5
Fig_variance_grid.do	grid_house_final	Figure 4
Table_fuzzy_iv.do	Merged_1853_1864_data.dta	Table B2

Name	File Type
Data_census.dta	application/
Merged_1846_1894_data.dta	application/
Merged_1853_1864_data.dta	application/

Name	File Ty
mccrary-s-ado	
spatial_HAC	
Fig_RD_plots.do	text/x-
Fig_bandwidth_sensitivity.do	text/x-
Fig_pre-trends.do	text/x-

t inside the catchment area of
ersist over the following 160
tions in which poor tenants

Data and Code for: Loss in the Time of Cholera: Long-run Impact of a Disease Epidemic on the Urban Landscape

```
*=====================================================================
* Purpose: Do-file creates RDplots
* Outcome:
* Figure 2:  Cholera Deaths and BSP Boundary (1854)
* Figure 3:  RD plots for Main Outcomes (in logs)
* Figure B1: Covariate RD Plots (1853)
* Figure B2: Histogram and Density of Forcing Variable (Distance to BSP boundary)
* Figure B3: RD Plots for Residential Mobility Outcome
* Figure B4: RD Plots for House Occupancy Outcomes
* Figure B5: RD Plots for Socioeconomic Outcomes
*=====================================================================

clear all
set more off



*********************************************************************************
*  Figure 2a, 2b: Cholera Deaths and BSP Boundary (1854)
*********************************************************************************
* RD Program
capture program drop myrdplot
program define myrdplot
args outcome

    * large sample
    local width = 20
    local hwidth = 10
    local limit = 100 - `width'
    local gr_limit = `limit'+`width'
    local gr_width = `gr_limit'/4
```

Do-files, ir

NOTE: Ma

Do-file		
Table_su		
Table_de		
Table_ma		
Table_mo		
Table_mi		
Table_ce		
Table_Bo		
Table_cu		
Fig_RD_p		
Fig_varia		
Table_fuzzy_iv.do	Merged_1853_1864_data.dta	Table B2

File Type
application/
application/
application/
application/

File Ty
text/x-
text/x-
text/x-

data.dta

data.dta

tivity.do

neighborhood of 19th

inside the catchment area of

persist over the following 160

tions in which poor tenants

Reproducibility

Find Data / Data and Code for: Loss in the Time of Cholera: Long-run Impact of a Disease Epidemic on the Urban Landscape

Data and Code for: Loss in the Time of Cholera: Long-run Impact of a Disease Epidemic on the Urban Landscape

Principal Investigator(s): Attila Ambrus, Duke University; Erica Field, Duke University; Robert Gonzalez, University of South Carolina

Version: V2

Version Title: Corrected author information

AMERICAN ECONOMIC ASSOCIATION

Name	File Type	Size	Last Modified
aer_replication			09/02/2019 02:23:PM
README.pdf	application/pdf	587 KB	08/21/2019 10:47:AM

DOWNLOAD THIS PROJECT

Project Citation:

Ambrus, Attila, Field, Erica, and Gonzalez, Robert. Data and Code for: Loss in the Time of Cholera: Long-run Impact of a Disease Epidemic on the Urban Landscape. Nashville, TN: American Economic Association [publisher], 2020. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020-01-31. https://doi.org/10.3886/E111523V2

Usage Metrics

Overall Project Metrics

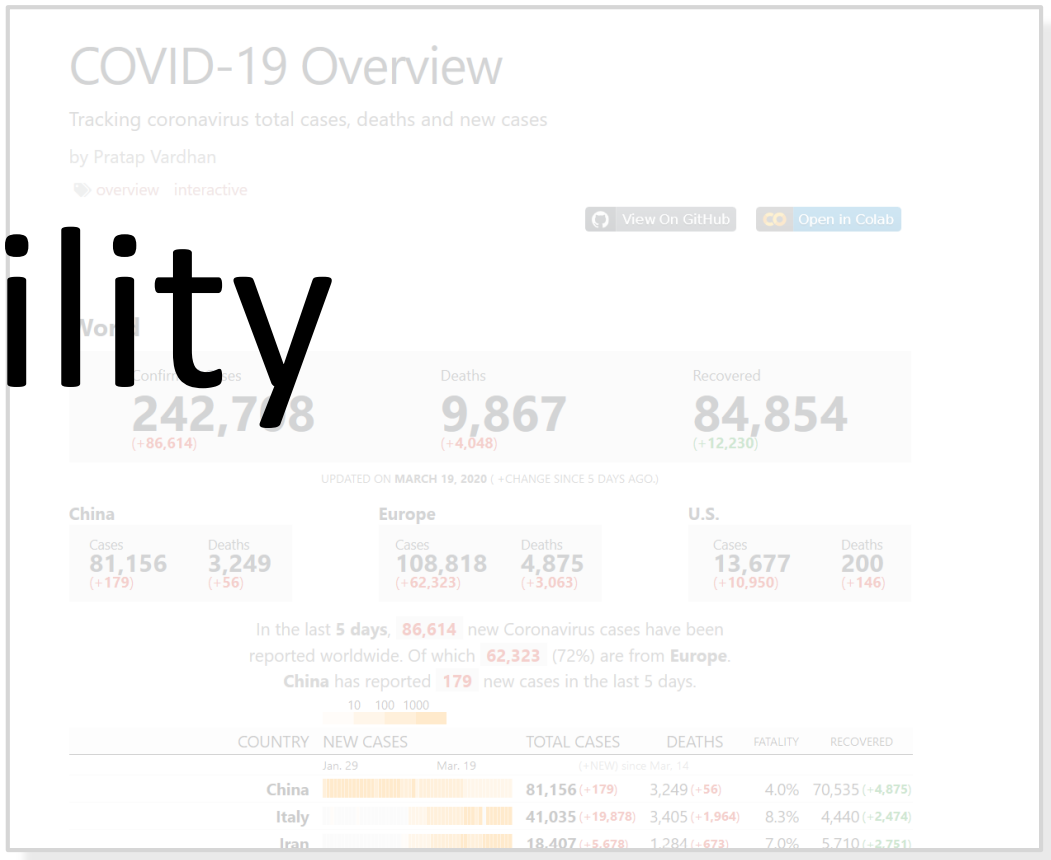
597	155	1
Views	Downloads	Publications

Download Detailed Metrics

Published Versions

Project Description

Summary: How do geographically concentrated income shocks influence the long-run spatial distribution of poverty within a



Credibility

AMERICAN ECONOMIC ASSOCIATION

American Economic Review

The *American Economic Review* is a general-interest economics journal. Established in 1911, the *AER* is among the nation's oldest and most respected scholarly journals in economics.

American Economic Review: Insights

AER: Insights is designed to be a top-tier, general-interest economics journal publishing papers of the same quality and importance as those in the *AER*, but devoted to publishing papers with important insights that can be conveyed succinctly.

Journal of Economic Literature

The *Journal of Economic Literature (JEL)*, first published in 1969, is designed to help economists keep abreast of and synthesize the vast flow of literature.

Journal of Economic Perspectives

The *Journal of Economic Perspectives (JEP)* fills the gap between the general interest press and academic economics journals.

American Economic Journal: Applied Economics

American Economic Journal: Applied Economics publishes papers covering a range of topics in applied economics, with a focus on empirical microeconomic issues.

American Economic Journal: Economic Policy

American Economic Journal: Economic Policy publishes papers covering a range of topics, the common theme being the role of economic policy in economic outcomes.

American Economic Journal: Macroeconomics

American Economic Journal: Macroeconomics focuses on studies of aggregate fluctuations and growth, and the role of policy in that context.

American Economic Journal: Microeconomics

American Economic Journal: Microeconomics publishes papers focusing on microeconomic theory; industrial organization; and the microeconomic aspects of international trade, political economy, and finance.

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

AEA "Data Availability Policy" (2018)

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

AEA Data Availability Policy (2018)

documented
readily available

clearly and precisely

must **provide**, **prior to publication**
details

sufficient to
permit replication

posted on the AEA website

AEA Data Availability Policy (2018)

clearly and precisely documented

readily available

must **provide**, **prior to publication**

details **sufficient to permit replication**

posted on the AEA website.

July 16, 2019

AEA Data & Code 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 DCAP (2018→2019)

- 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.**

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
- **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 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*

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

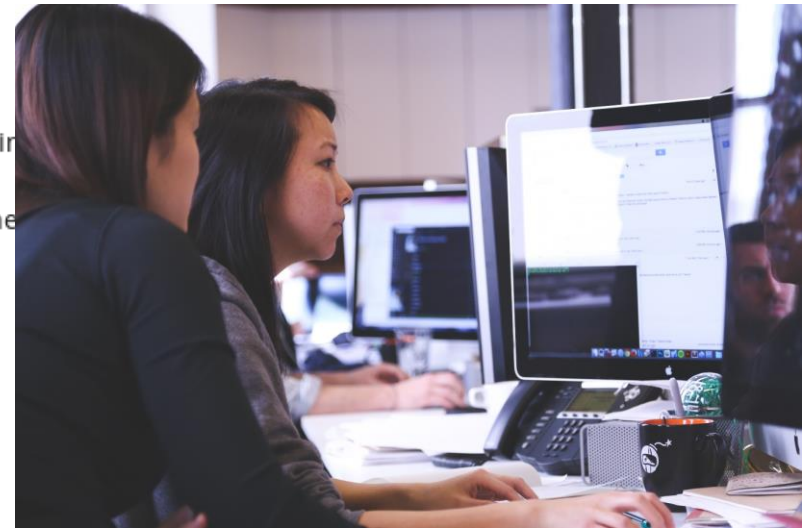
On this page:

- 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

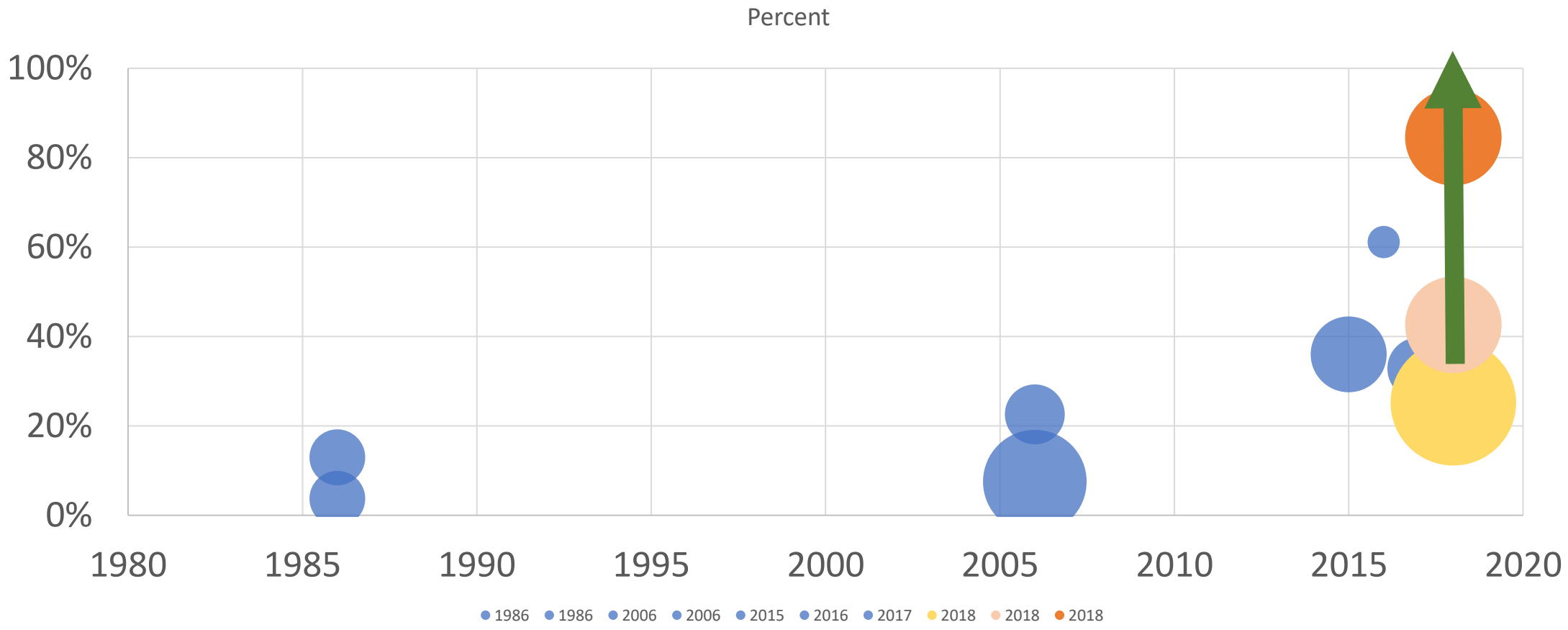
- what authors should check before providir journals
- what verifier teams should check for in the to them for the purpose of verification



Who is doing that?

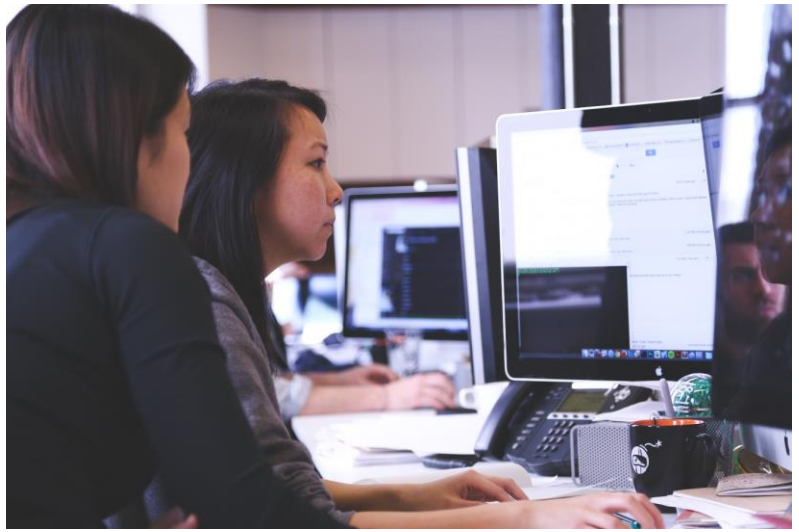
- Earlier reproducibility work: **Flavio Stanchi** (now at AirBnb), **Sylverie Herbert** (on the market), **Hautahi Kingi** (Impaq)
- Current lead graduate students: **David Wasser** (until Dec 2019), **Meredith Welch** (since Jan 2020)
- Current and past undergraduate students: Alexia Ge, Anthony Peraza, Craig Schulman, Elijah B. Ruiz, Gabriel Bond, Jason S. Katz, Jeong Hyun Lee, Jiayin Song, John Park, Joshua Passel, Kirubeal T. Wondimu, Linchen Zhang, Louis Liu, Luis Lopez Cabrera, Luke O'Leary, Mary-Jo Ajiduah, **Naomi Li**, Nicholas Swan, Nishat Peuly, **Ryan Ali**, Samuel Frey, Siyang (Elaine) Yu, **Steve Yeh**, Weilun Shi, William Hernandez, Yanyun (Iris) Chen, Yuan-Hsuan (Sharon) Lin, Zebang Xu, Xing Su, Jiazhen Tan , Xueshi Su, Vendela Norman, Anderson Park, *Nehedin Juarez, Rubal Mistry, Syon Verma, William Silverman, Zechariah Karsana*
- Other graduate students: Aviv Caspi, Leah Kim

Goal: Improve reproducibility



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?



AEA Data and Code Guidance

AMERICAN ECONOMIC ASSOCIATION

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

Unofficial guidance on various topics by the AEA Data Editor

These web pages provide unofficial and developing guidance on the implementation of the American Economic Association (AEA)'s Data and Code Availability Policy. We also provide links to generic guidance being developed by a loose collective ("guild") of data editors and people in a similar role at various social science journals.

Follow @aeadata

Order in which AEA authors should read these resources:

1. Start with the **official Data and Code Availability Policy**
2. Look for general guidance at the **Social Science Data Editors pages**
3. Read the **AEA's FAQ**
4. Look for any guidance specific to the AEA at the **Unofficial AEA Data and Code Guidance**
5. Last but not least, have a look at the **draft FAQ on this site**

Comments are welcome, please file them as issues in our Github repo.

Guidance on creating replicable data and program archives

How should researchers create replicable data and program archives? How

larsvilhuber Minor edits to the report - clarifications about data preparation pro... e9ad1f8 10 days ago

1 contributor

195 lines (135 sloc) 10.3 KB

Raw Blame History

[MC number] [Manuscript Title] Validation and Replication results

> INSTRUCTIONS: Once you've read these instructions, DELETE THESE AND SIMILAR LINES. In the above title, replace [Manuscript Title] with the actual title of the paper, and [MC number] with the Manuscript Central number (e.g., AEJPol-2017-0097) Go through the steps to download and attempt a replication. Document your steps here, the errors generated, and the steps you took to alleviate those errors.

You may want to consult Unofficial Verification Guidance for additional tips and criteria.

SUMMARY

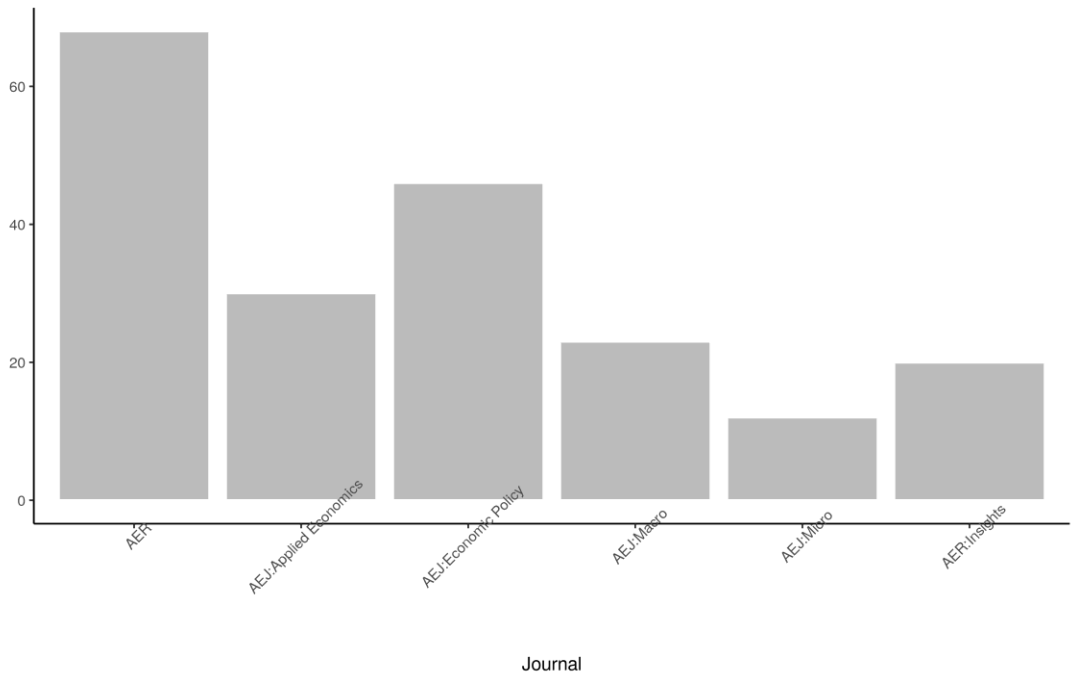
> INSTRUCTION: The Data Editor will fill this part out. It will be based on any [REQUIRED] and [SUGGESTED] action

Stats on reproduced articles

Between July 16, 2019, and November 28, 2019 (4.5 mths), the AEA Data Editor team conducted

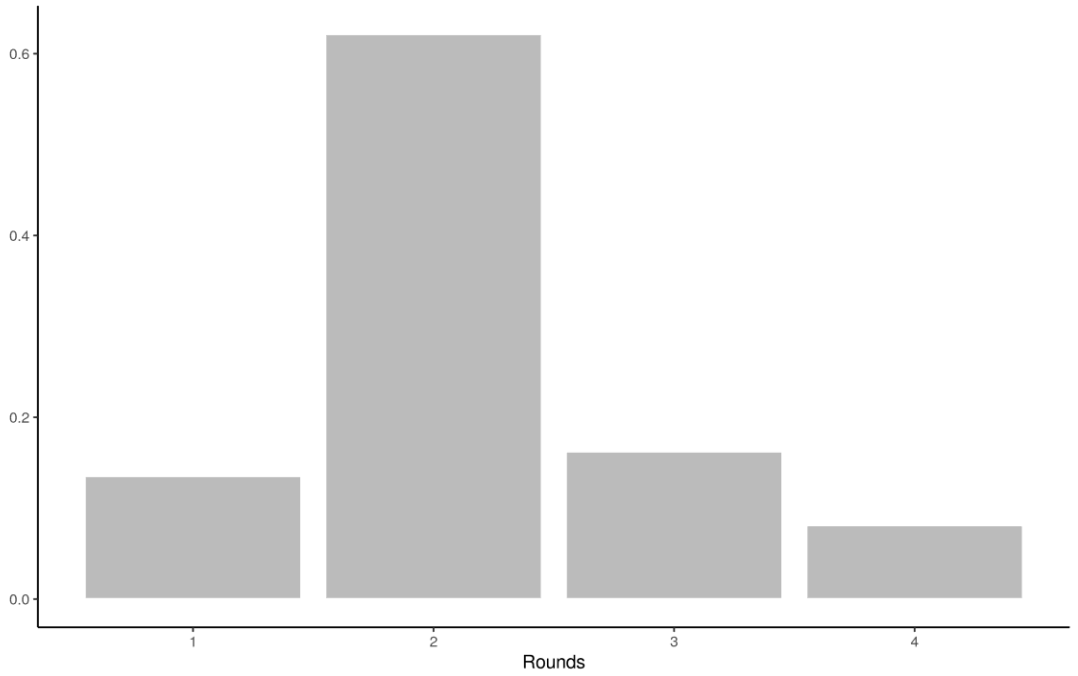
- **216 assessments**
- for **138 manuscripts**.

- (as of today, approx. 600 assessments)



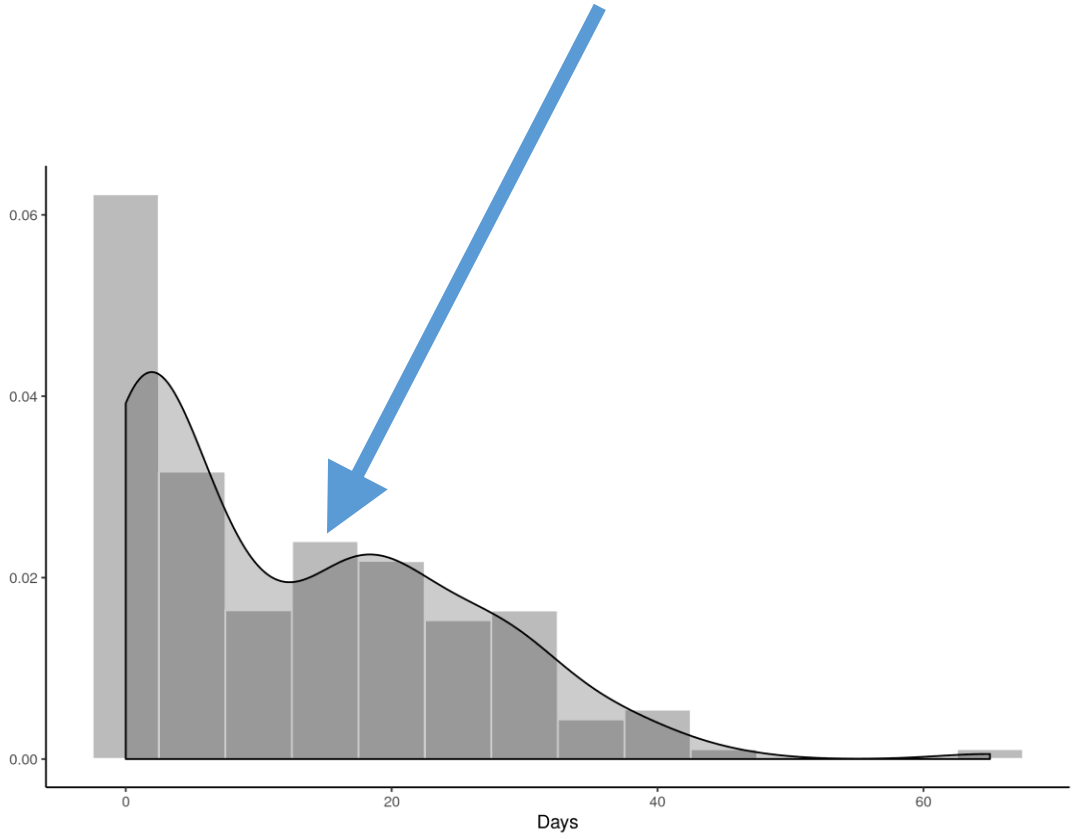
Stats on reproduced articles

- The typical article goes through at least two rounds of assessment (none were perfect)
- Conversely, **not a single study was irreproducible** (not supporting manuscript claims)



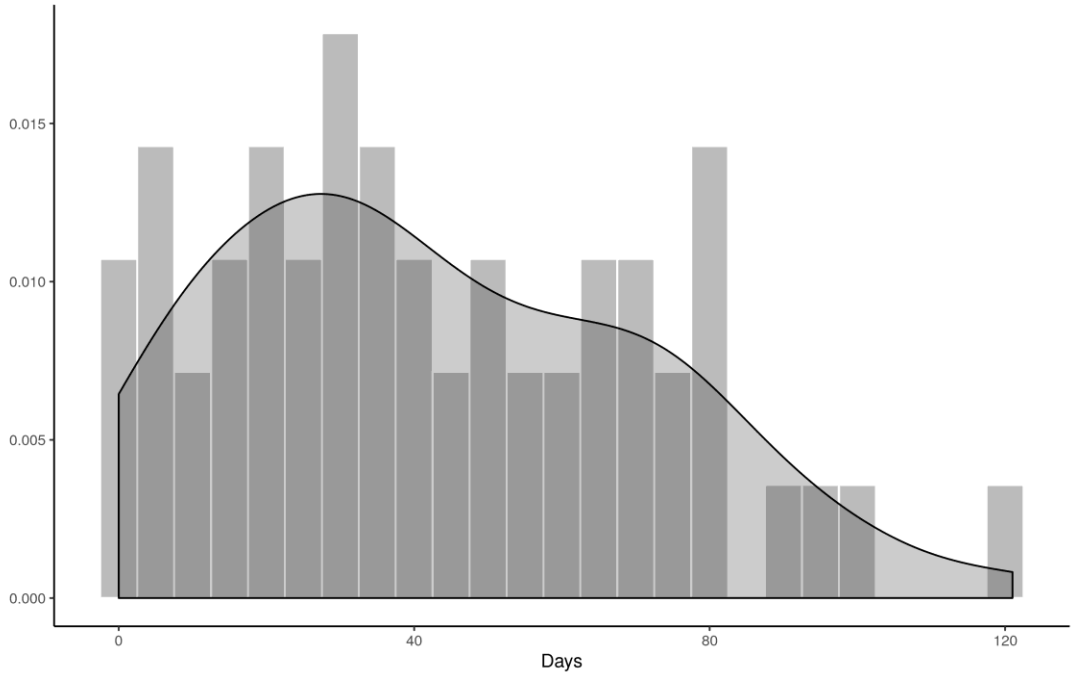
Stats on reproduced articles

- Goal is turnaround of two weeks
- Currently still **too long**



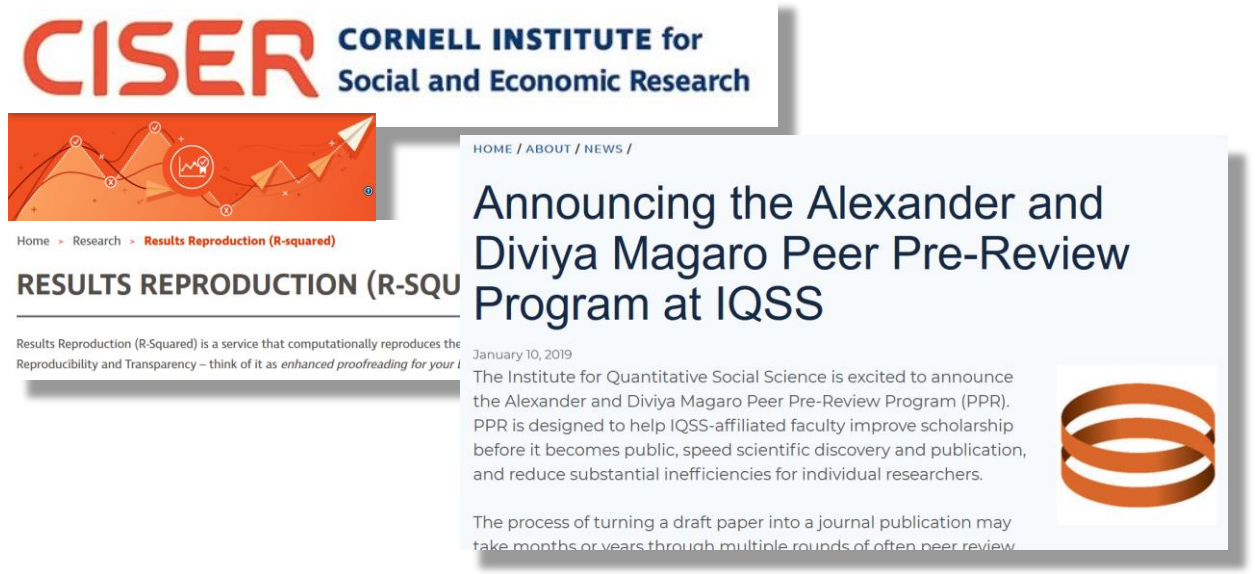
Stats on reproduced articles

- Total time (from first submission to final signoff) is **too long**



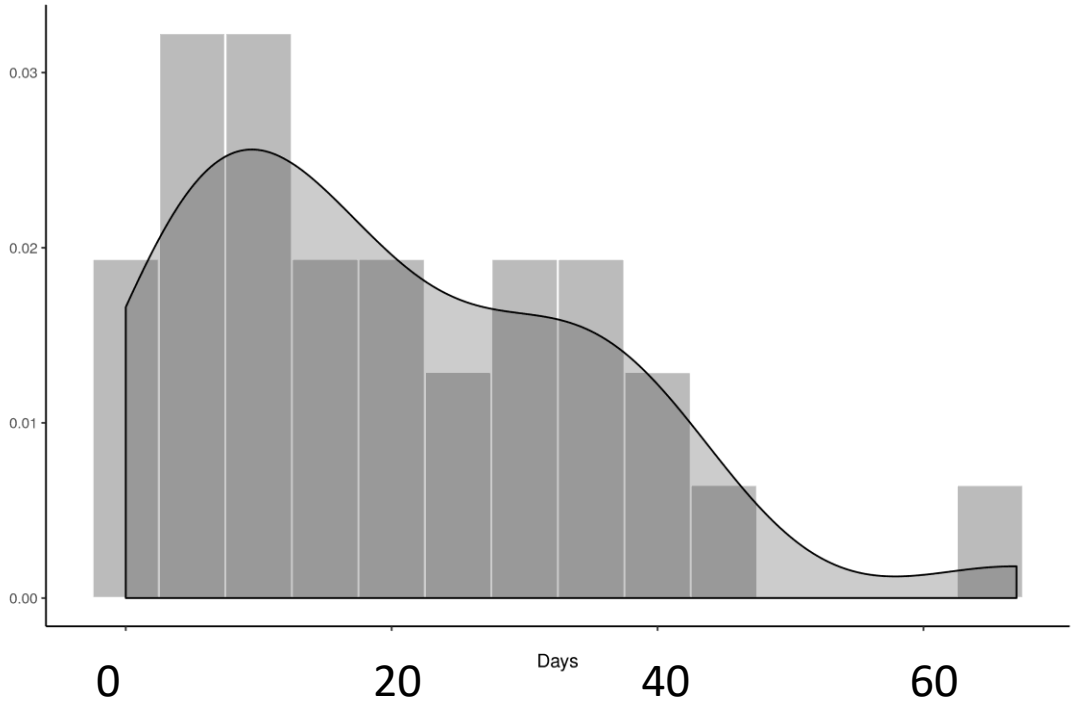
Increasing team size

- Grown from **7 undergrads** + 1 graduate assistant
- To **18 undergraduates**, **15 trainees**, + 1 graduate assistant (+ 1 volunteer)
- And:



Stats on reproduced articles

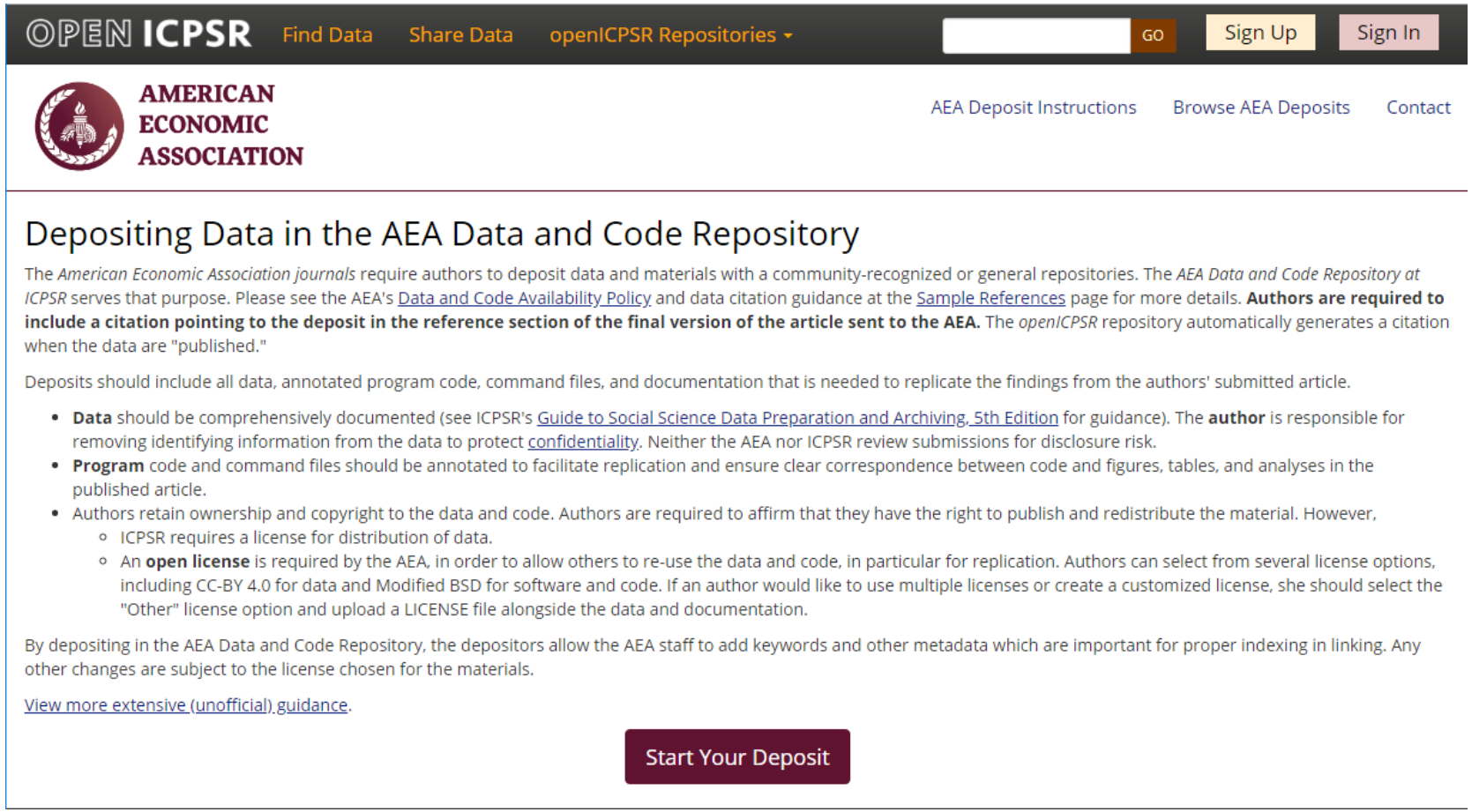
- But author response time is also a contributor



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
- **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

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.

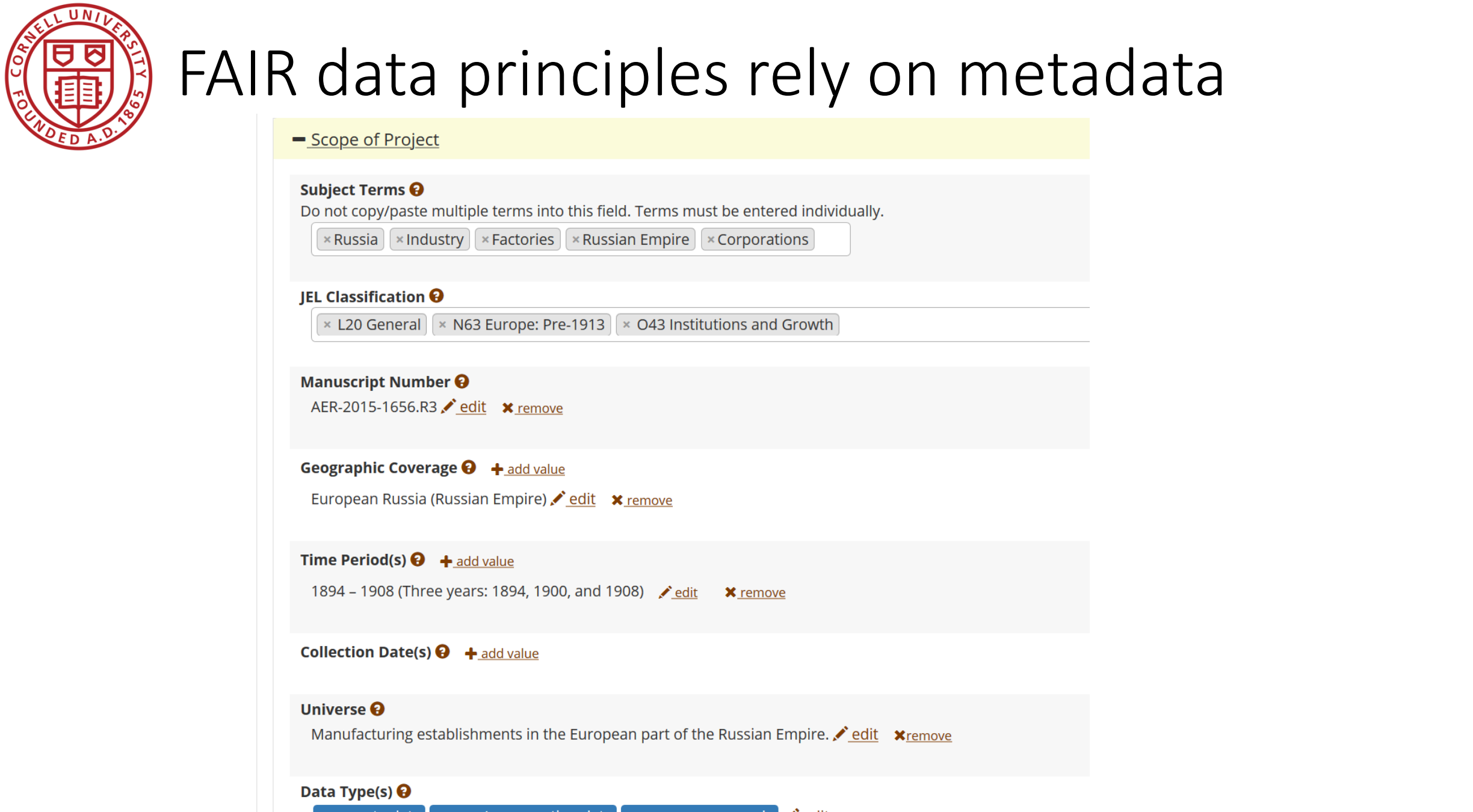
- **Data** should be comprehensively documented (see ICPSR's Guide to Social Science Data Preparation and Archiving, 5th Edition for guidance). The **author** is responsible for removing identifying information from the data to protect confidentiality. Neither the AEA nor ICPSR review submissions for 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

FAIR data principles rely on metadata



OPEN ICPSR Find Data Share Data openICPSR Repositories Management

dataeditor@aeapubs.org

Find Data / Imperial Russian Factory Database, 1894-1908

Imperial Russian Factory Database, 1894-1908

Principal Investigator(s): Amanda Gregg, Middlebury College

Version: V1

AMERICAN ECONOMIC ASSOCIATION

Name	File Type		Last Modified
1894MicroData.xlsx	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet	4.5 MB	08/08/2019 11:01:AM

Project Citation:

Gregg, Amanda. Imperial Russian Factory Database, 1894-1908. Nashville, TN: American Economic Association [publisher], 2020. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020-01-29. https://doi.org/10.3886/E110681V1

AG_Corp_CleaningandDatabaseCompiler.do	text/x-stata-syntax	23.4 KB	08/08/2019 11:02:AM

Related Publications

The following publications are supplemented by the data in this project.

- Gregg, Amanda. "Factory Productivity and the Concession System of Incorporation in Late Imperial Russia, 1894-1908." *American Economic Review* 110, no. 2 (February 2020): 401–27. https://doi.org/10.1257/aer.20151656.

08:55:AM

OPEN ICPSR Find Data Share Data openICPSR Repositories Management

GO dataeditor@aeapubs.org

Find Data / Imperial Russian Factory Database, 1894-1908

Imperial Russian Factory Database, 1894-1908

Principal Investigator(s): Amanda Gregg, Middlebury College

Version: V1

AMERICAN ECONOMIC ASSOCIATION

```
<meta name="DC.identifier" content="10.3886/E110681V1" />
<meta name="DC.title" content="Imperial Russian Factory Database, 1894-1908" />

    <meta name="DC.creator" content="Amanda Gregg, Middlebury College" />

<meta name="DC.publisher" content="Inter-university Consortium for Political and Social Research (ICPSR)" />
<meta name="DC.date" content="2020-01-29" />
<meta name="DC.type" content="Dataset" />
```

	officedocument.spreadsheetml.sheet	MB	08:53:AM
1908MicroData.xlsx	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet	2.3 MB	08/07/2019 11:06:AM
AG_Corp_CleaningandDatabaseCompiler.do	text/x-stata-syntax	23.4 KB	08/08/2019 11:02:AM
AG_Corp_Prod_AppendixCode.do	text/x-stata-syntax	42.2 KB	12/09/2019 09:19:AM
AG_Corp_Prod_Code.do	text/x-stata-syntax	26.6 KB	12/12/2019 03:01:AM
AG_Corp_Prod_Database.dta	application/x-stata	11 MB	08/07/2019 08:55:AM
	application/x-stata	11.9	10/08/2014

OPEN ICPSR Find Data Share Data openICPSR Repositories Management

GO dataeditor@aeapubs.org

Find Data / Imperial Russian Factory Database, 1894-1908

Imperial Russian Factory Database, 1894-1908

Principal Investigator(s): Amanda Gregg, Middlebury College

Version: V1

AMERICAN ECONOMIC ASSOCIATION


```
<script type="application/ld+json">
{"name":"Imperial Russian Factory Database, 1894-1908","identifier":"http://doi.org/10.3886/E110681V1","description":"This database dig
anufacturing censuses. For each factory, the database includes industry, province, enterprise form, total workers, total revenue, and identifiers that
.908 years also include information on the factory's total machine power. The dataset was constructed to study why some Russian firms chose to become c
onsuming concession system. Note that the final analysis files exclude factories located outside of European Russia and, in the main data files, facto
ax.&amp;nbsp;","url":"http://doi.org/10.3886/E110681V1","version":"V1","keywords":["Russia","Industry","Factories","Russian Empire","Corporations"],"spati
mpire)"],"temporalCoverage":["1894-01-01--1908-12-31 (Three years: 1894, 1900, and 1908)"],"creator":[{"name":"Amanda Gregg","affiliation":["Middlebu
"name":"openICPSR Self-Deposit Archive","url":"http://www.openicpsr.org/","@type":"DataCatalog"},"funder":[{"name":"Economic History Association","@ty
irectorate for Social, Behavioral and Economic Sciences","@type":"Organization"},{"name":"Yale Economic Growth Center","@type":"Organization"},{"name"
und","@type":"Organization"},{"name":"Yale Program in Economic History","@type":"Organization"},{"name":"Yale MacMillan Center","@type":"Organization"
{"fileFormat":"stata","contentURL":"https://www.openicpsr.org/openicpsr/project/110681/version/V1/download/terms?path=/openicpsr/110681/fcr:versions/V
tata","encodingFormat":"application/zip"},{"fileFormat":"stata","contentURL":"https://www.openicpsr.org/openicpsr/project/110681/version/V1/download/t
V1/AG_Corp_Prod_Database.dta&type=application/x-stata","encodingFormat":"application/zip"},{"fileFormat":"stata","contentURL":"https://www.openicpsr.o
terms?path=/openicpsr/110681/fcr:versions/V1/AG_Corp_RuscorpMasterFile_Cleaned.dta&type=application/x-stata","encodingFormat":"application/zip"},{"fil
openicpsr/project/110681/version/V1/download/terms?path=/openicpsr/110681/fcr:versions/V1/AG_Corp_Prod_Database_withAktsiz.dta&type=application/x-stat
"fileFormat":"stata","contentURL":"https://www.openicpsr.org/openicpsr/project/110681/version/V1/download/terms?path=/openicpsr/110681/fcr:versions/V1
tata","encodingFormat":"application/zip"}],"license":"https://creativecommons.org/licenses/by/4.0","@context":"http://schema.org","@type":"Dataset"}
</script>
```


AG_Corp_CleaningandDatabaseCompiler.do		KB	11:02:AM
AG_Corp_Prod_AppendixCode.do	text/x-stata-syntax	42.2 KB	12/09/2019 09:19:AM
AG_Corp_Prod_Code.do	text/x-stata-syntax	26.6 KB	12/12/2019 03:01:AM
AG_Corp_Prod_Database.dta	application/x-stata	11 MB	08/07/2019 08:55:AM

… and findability relies on metadata

Google

Imperial Russian Factory

Updated Date | Download Format | Usage Rights | Free

2 datasets found

Imperial Russian Factory Database, 1894-1908
www.openicpsr.org
www.da-ra.de
Updated Jan 29, 2020

Data from: Антиосманские выступления болгар и русско-турецкие войны второй...
explore.openaire.eu
Updated 24 нояб. 2015 г.

Not seeing a result you expected? Learn how you can add new datasets to our index.

Middlebury

Imperial Russian Factory Database, 1894-1908

Explore at openICPSR Self-Deposit Archive | Explore at www.da-ra.de

Unique identifier
https://doi.org/10.3886/E110681V1

Dataset updated Jan 29, 2020

Dataset provided by
Middlebury College

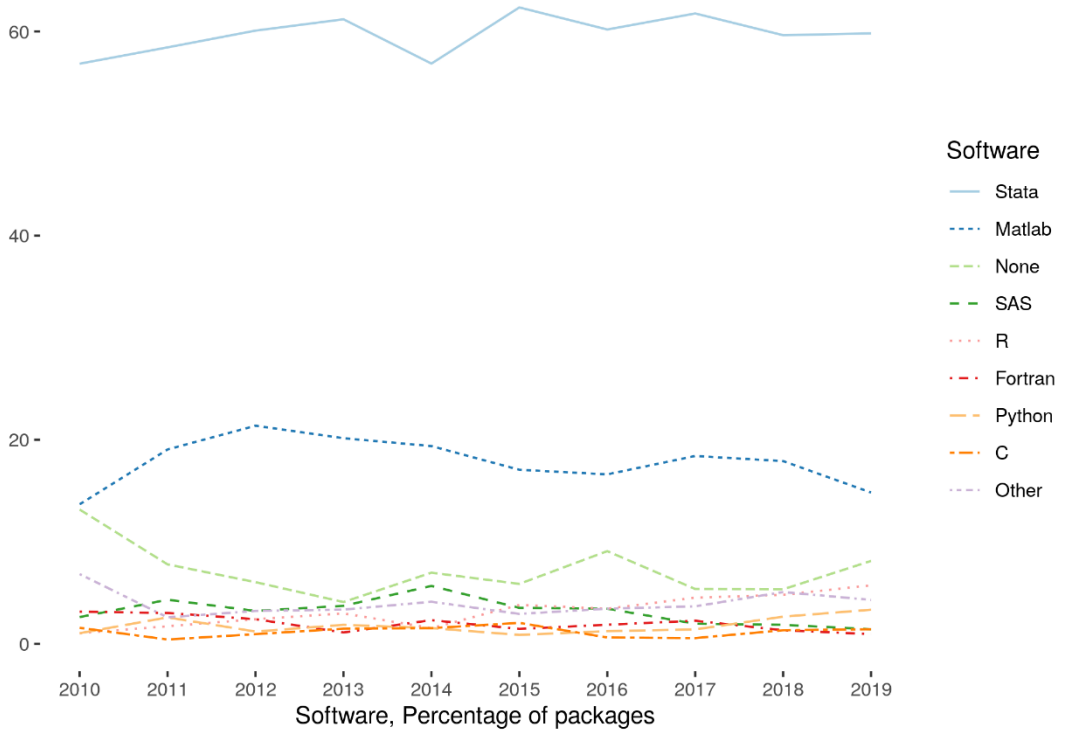
Authors
Amanda Gregg

License
Attribution 4.0 (CC BY 4.0)
License information was derived automatically

Area covered
European Russia (Russian Empire)

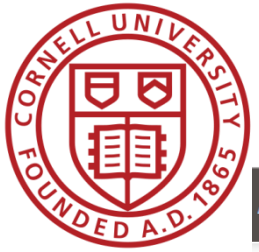
Some statistics

- **Stata** is the most popular statistical software in the journals of the AEA (**72.96%** of all supplements)
- followed by **Matlab** (**22.45%**)



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
- **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



FORCE11

The Future of Research Communications and e-Scholarship

Search

English

ABOUT | COMMUNITY | CODE OF CONDUCT | GROUPS | RESOURCES | NEWS + BLOGS | EVENTS | PUBLIC

perceived criteria of importance.

1. Importance

Data should be considered legitimate, citable products of research. Data should be accorded the same importance in the scholarly record as citat research objects, such as publications[1].

DC1

Data Citation Principles

2. Credit and Attribution

1 **Bureau of Labor Statistics.** 2000–2010. "Current Employment Statistics: Colorado, Total Nonfarm, Seasonally adjusted - SMS08000000000000001." United States Department of Labor. http://data.bls.gov/cgi- bin/surveymost?sm+08 (accessed February 9, 2011).

corresponding data should be cited[3].

4. Unique Identification

A data citation should include a persistent method for identification that i actionable, globally unique, and widely used by a community[4].

5. Access

Data citations should facilitate access to the data themselves and to such

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [https://www.force11.org/group/joint-declaration-data-citation-principles-final].

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.**

Every manuscript is checked

- What datasets are used
- Are they cited?
  - → in Article?
  - → in Online Appendix?
  - → in README?

Every manuscript is checked

- What datasets are used
- Are they cited?
- Is there additional information access?
  - → URL leads to exact data?
  - → URL leads to application procedure?
  - → other access procedure is described?

Le labyrinthe des juges, par Guillaume Beck

Every manuscript is checked

- What datasets are used
- Are they cited?
- Is there additional information on access?
- Is there license/ data use information?
  - → **Should** the author provide the data?
  - → Is the author **allowed** to provide data?

Proposed:
Explicit DAS or
Incorporate into README

Evolving Journal and Data Infrastructure

Authors struggle with this!

The author contacted the data provider, and received an answer similar to this one:

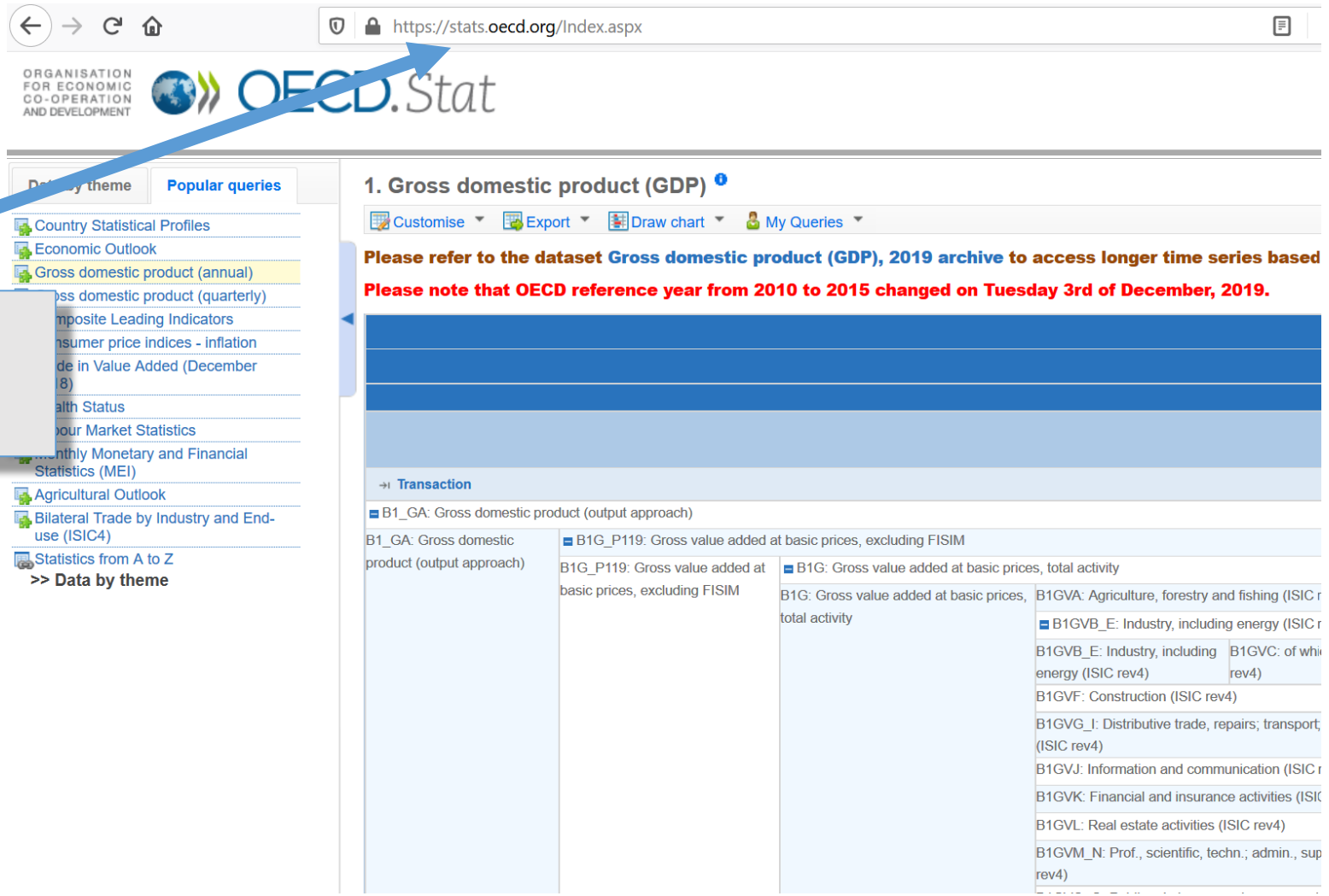
> Hi John Doe, The simple answer is no, we do not want this data to be released publicly. Doing so would require revisiting the data use agreement and another round of Privacy Impact Assessment for internal approval. Regards, Your Data Provider

Evolving Journal and Data Infrastructure

Data Publishers struggle with this!

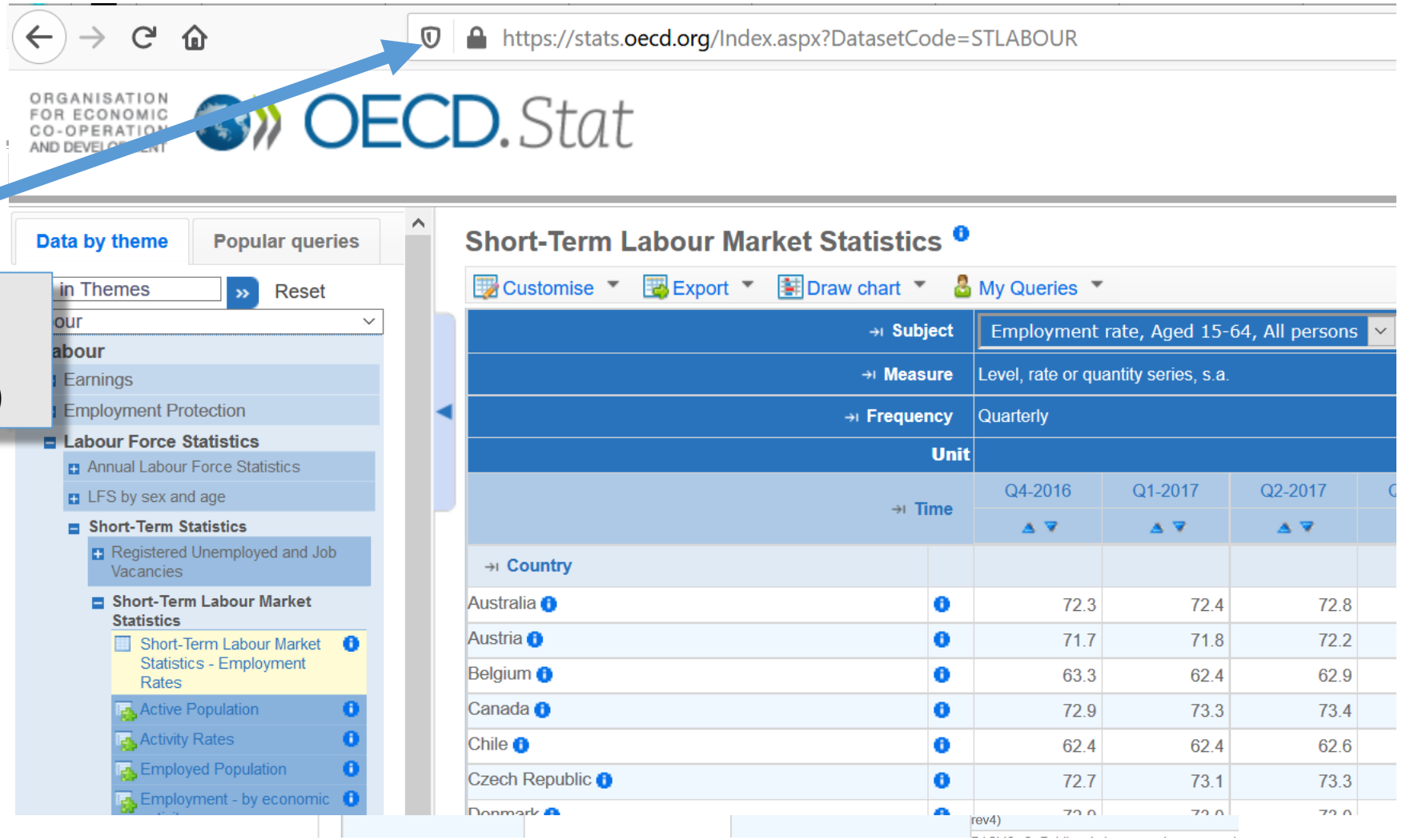
Example 1: OECD

URL does not always change!



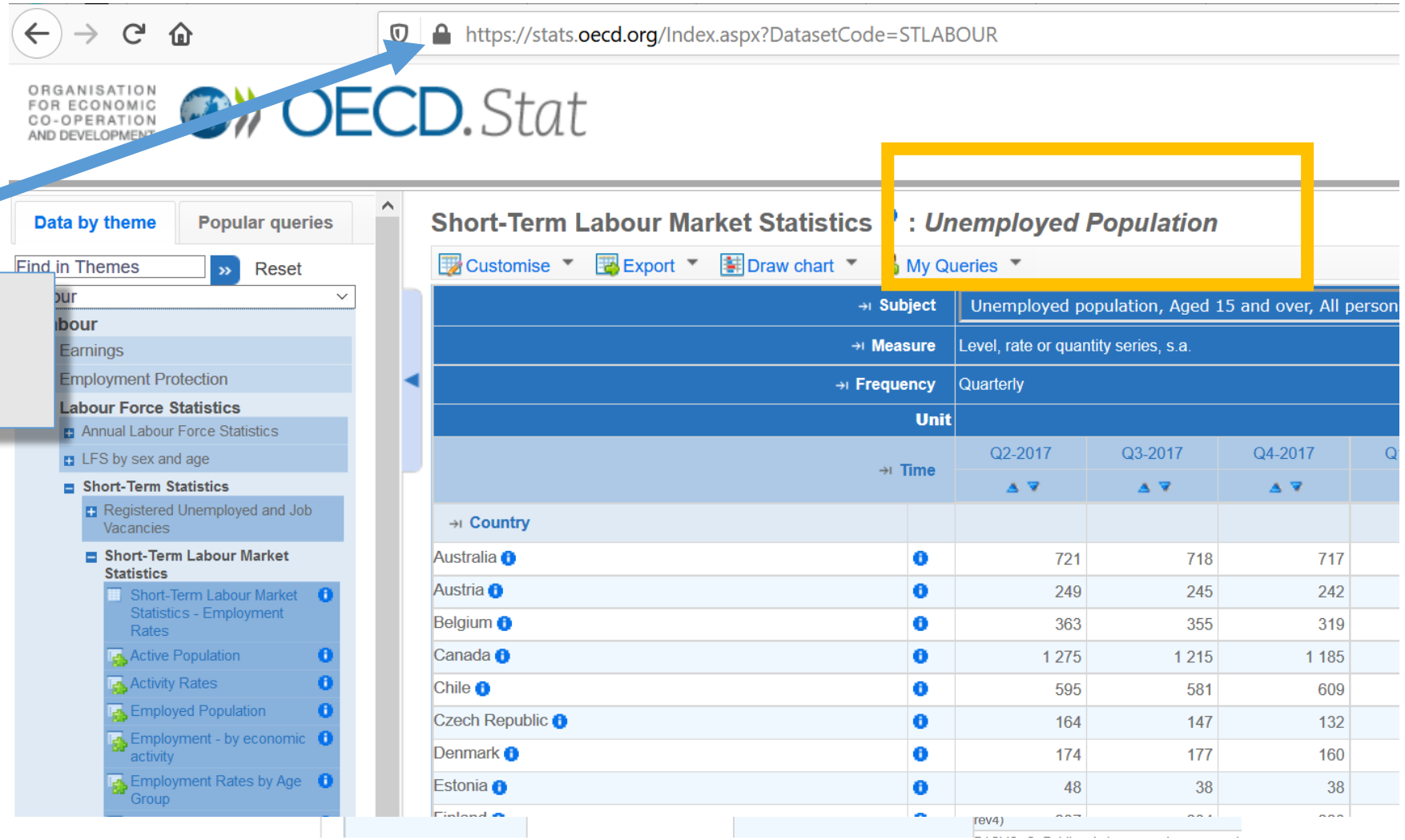
Example 1: OECD

URL does not always change!
(but sometimes it does)

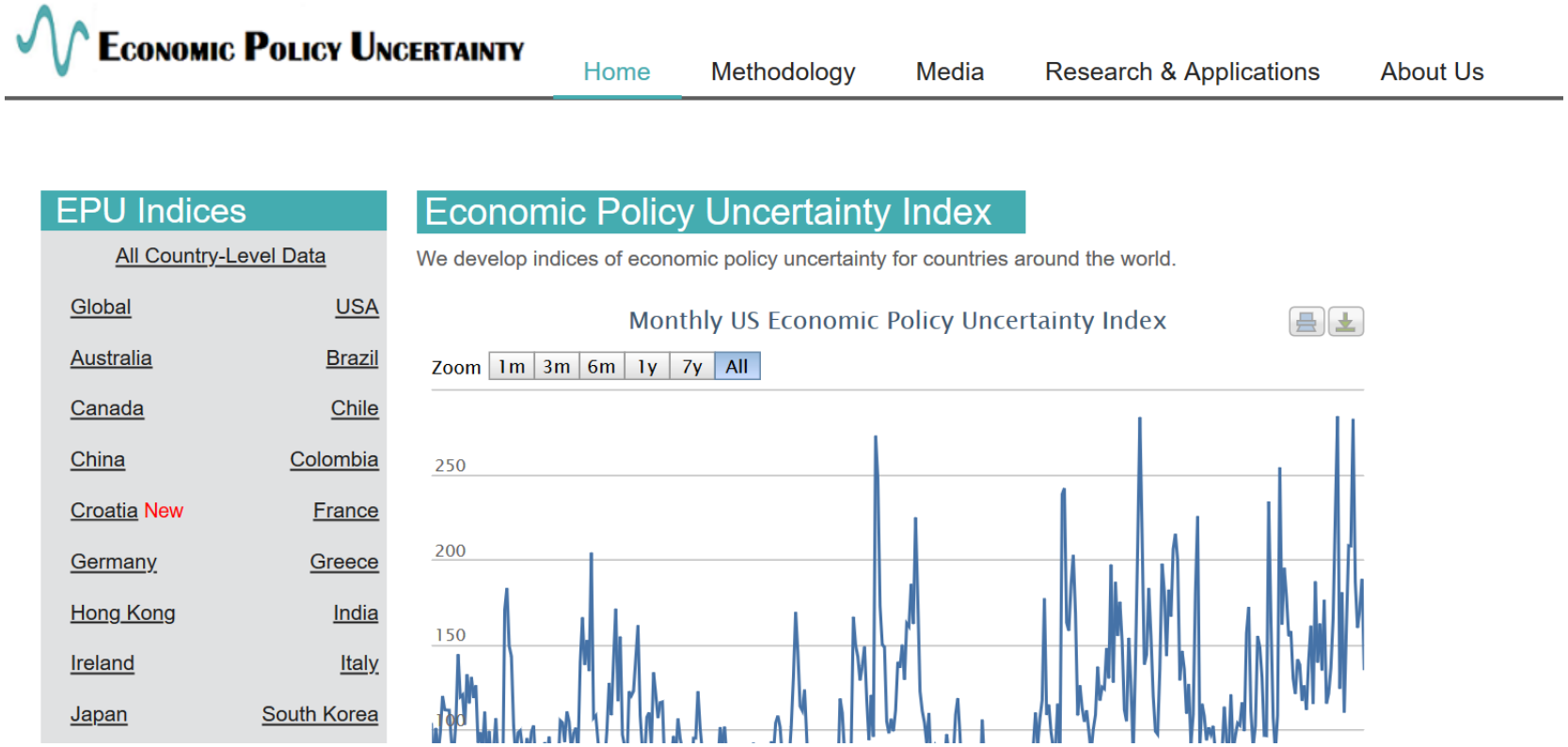


Example 1: OECD

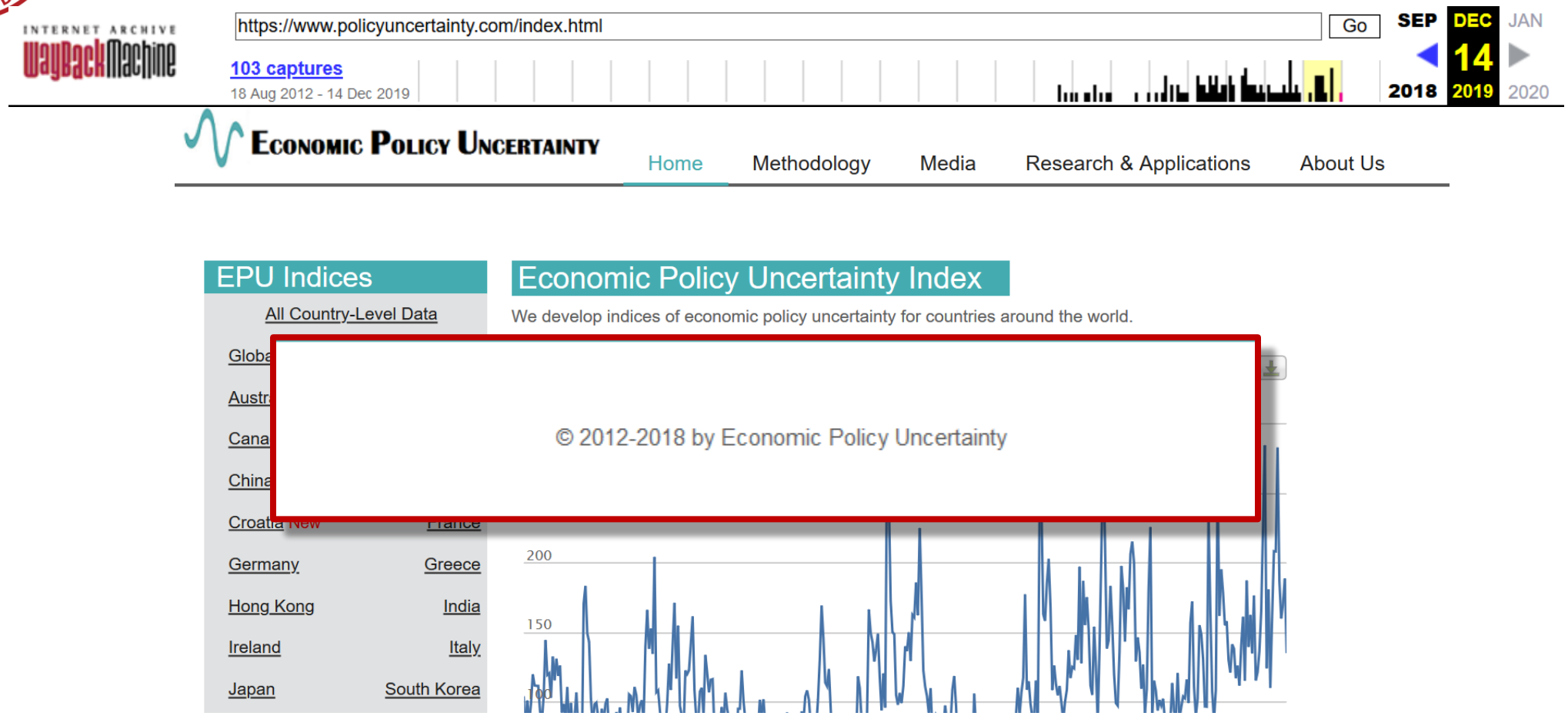
URL does not always change!
(and then it doesn't…)



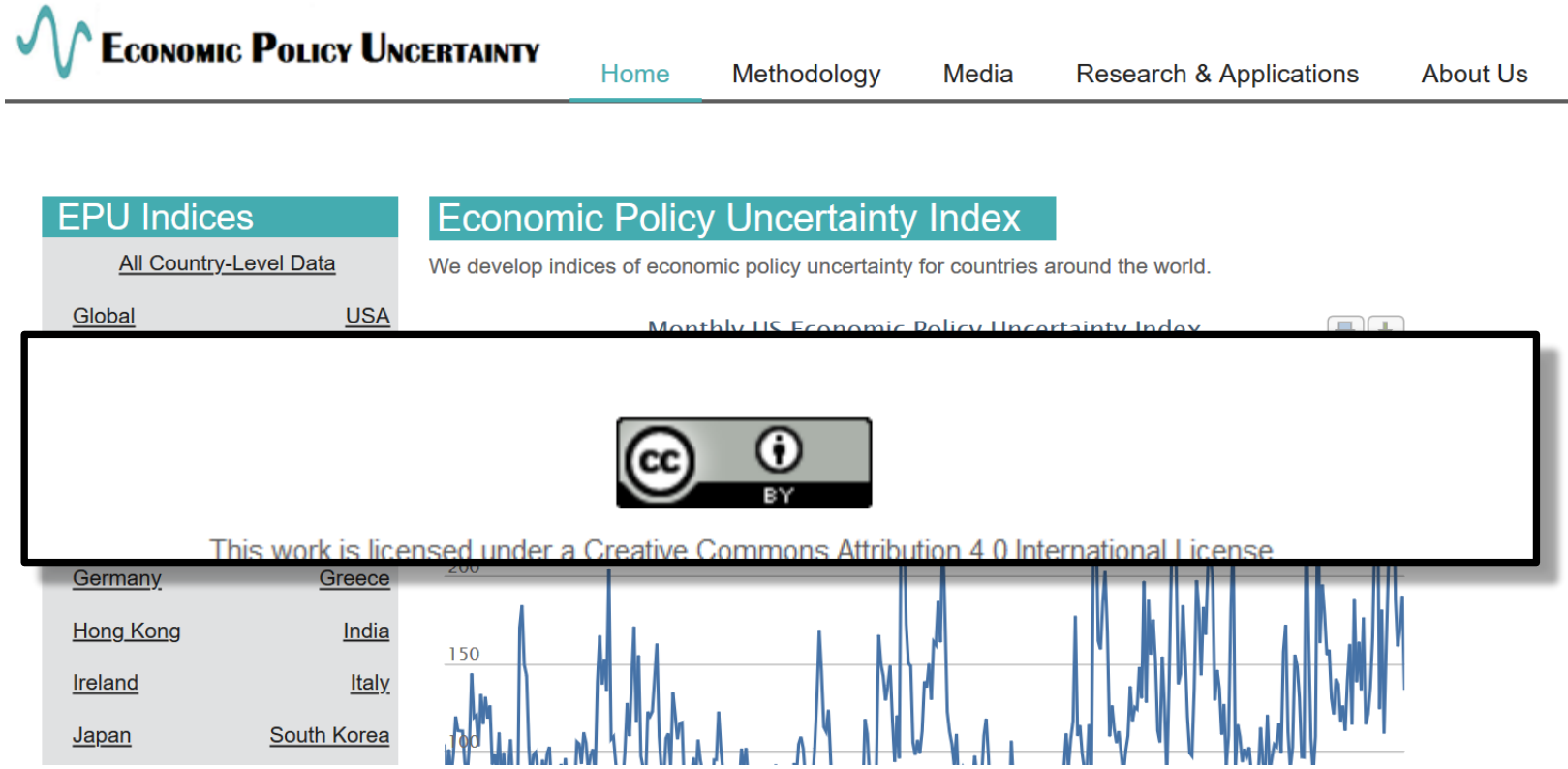
Example 2: Academic data publisher



Example 2: Academic data publisher



Example 2: Academic data publisher-new!



Response

> Thanks, but I'll stick with what I've been doing for at least 20 years. At some point I might figure out the right license, but it's been working so far. And your inference is correct, the authors can use the data but not redistribute it. In this specific case, there is no reason for them to do so because the data are freely available to everyone.

Example 3: FRED (St. Louis Fed)

Example 4: German Restricted-access

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 Version	DOI (Link to Description of Data Version)	Availability (yyyy-mm-dd)
BHP 7518 v1 (current)	10.5164/IAB.BHP7518.de.en.v1	2020-01-13
BHP 7517 v1	10.5164/IAB.BHP7517.de.en.v1	2018-12-12
BHP 7516 v1	10.5164/IAB.BHP7516.de.en.v1	2018-04-11

- External data
- Data Archive
- Data Access
- Campus Files
- Publications
- Events
- Projects of FDZ users
- FDZ Projects
- Complaint point of the RatSWD
- Figures of the FDZ

employees, both in total and broken down by gender, age, occupational status, qualification and nationality. Means and medians of wages for full-time employees are given, too. Additional datasets providing information about (gross) worker flows and about foundations and closures of establishments are available on request.

Data Versions

Old versions are only available for replication studies and only in justified exceptional cases for new Projects.

Data Version	DOI (Link to Description of Data Version)	Availability (yyyy-mm-dd)
BHP 7518 v1 (current)	10.5164/IAB.BHP7518.de.en.v1	2020-01-13

Example 4: German Restricted-access

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 Version	DOI (Link to Description of Data Version)	Availability (yyyy-mm-dd)
BHP 7518 v1 (current)	10.5164/IAB.BHP7518.de.en.v1	2020-01-13
BHP 7517 v1	10.5164/IAB.BHP7517.de.en.v1	2018-12-12
BHP 7516 v1	10.5164/IAB.BHP7516.de.en.v1	2018-04-11

- External data
- Data Archive
- Data Access
- Campus Files
- Publications
- Events
- Projects of FDZ users
- FDZ Projects
- Complaint point of the RatSWD
- Figures of the FDZ

employees, both in total and broken down by gender, age, occupational status, qualification and nationality. Means and medians of wages for full-time employees are given, too. Additional datasets providing information about (gross) worker flows and about foundations and closures of establishments are available on request.

Data Versions

Old versions are only available for replication studies and only in justified exceptional cases for new Projects.

Data Version	DOI (Link to Description of Data Version)	Availability (yyyy-mm-dd)
BHP 7518 v1 (current)	10.5164/IAB.BHP7518.de.en.v1	2020-01-13

Example 4: German Restricted-access

Establishment History Panel (BHP) – Version 7518 v1

DOI: 10.5164/IAB.BHP7518.de.en.v1

Summary

Data source:

Data Access

The IAB Establishment Panel is available via the following ways of access:

- On-site use at the FDZ. Further information on Applying for on-site use.
- Remote data Access. Further information on Applying for remote data access.

nationality. Means and medians of wages for full-time employees are given, too. Additional datasets providing information about (gross) worker flows and about foundations and closures of establishments are available on request.

Dataset Descriptions and Frequencies

German

- DOI: 10.5164/IAB.FDZD.2001.de.v1
- FDZ-Datenreport 01/2020
- Fallzahlen und Labels

English

- DOI: 10.5164/IAB.FDZD.2001.en.v1

Some Suggestions

For authors

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)*

Improving replication packages

- Victoria Stodden's TIER talk 2020-02-07
  - Whole Tale
  - CodeOcean
  - OSF
  - Binder.org
  - Runmycode.org

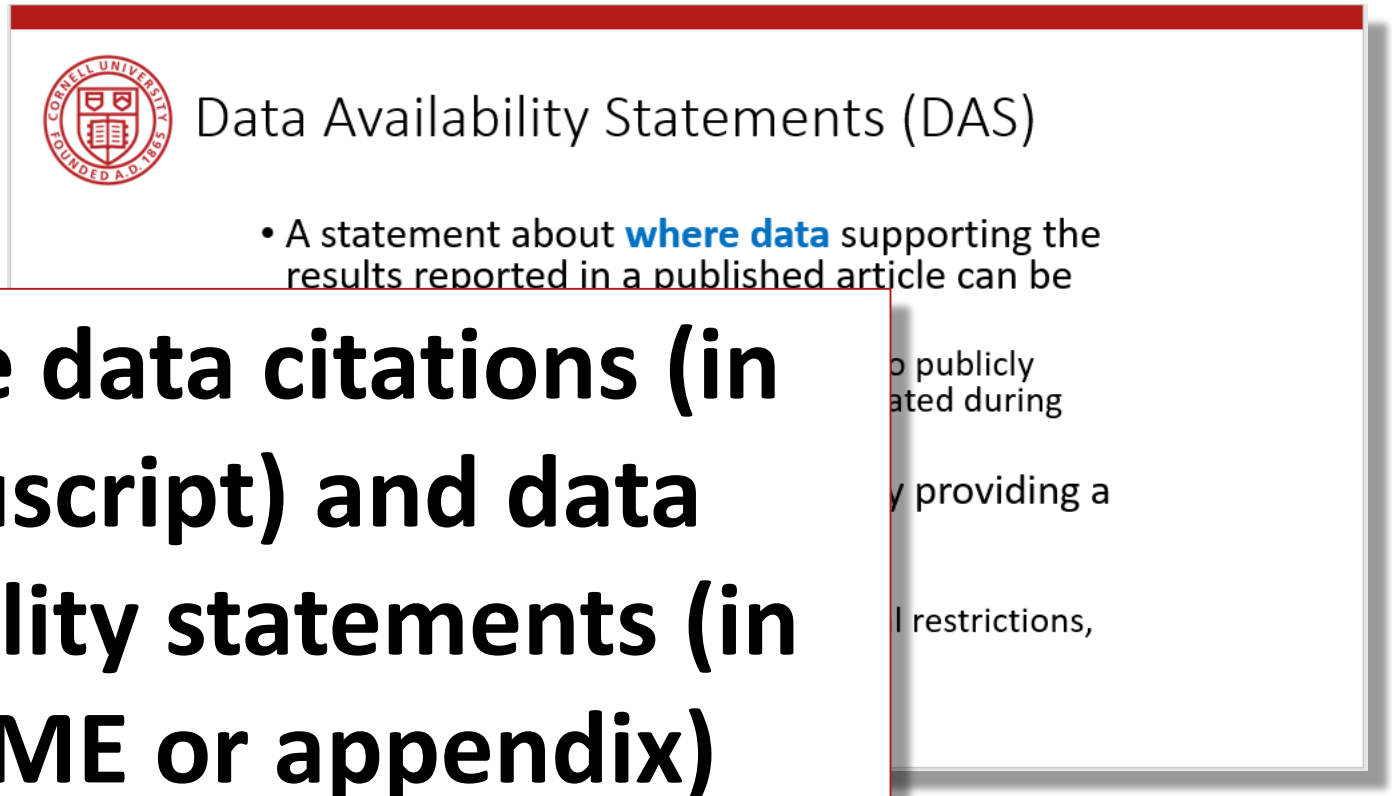
Beware: reproducibility tomorrow, platform gone the day after!

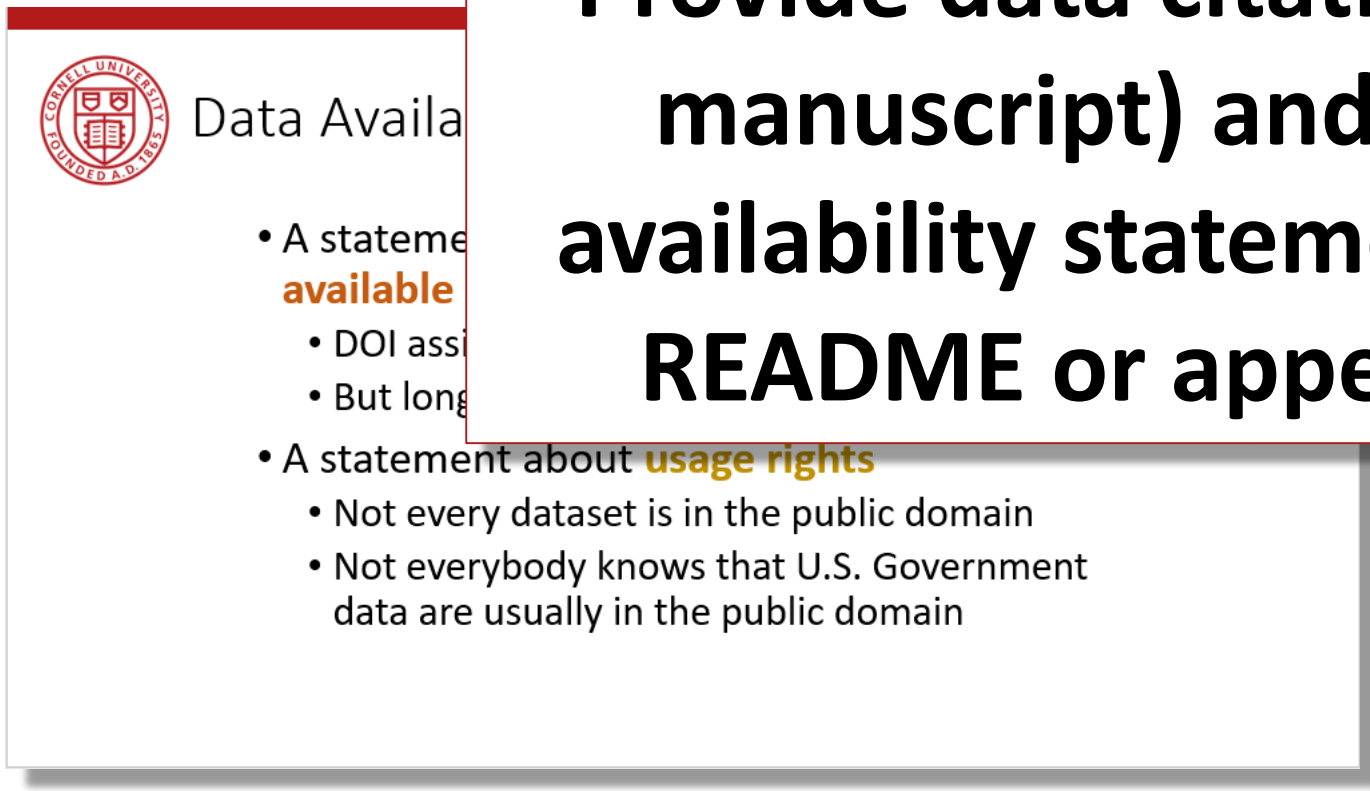
- Scott Long's TIER talk 2020-03-06
  - Dual-workflow
  - Emphasize reproducibility from the start

Beware: document provenance of data!

Streamlining replication packages

- Master script preferred
  - Least amount of manual effort
- No manual manipulation
  - "Change the parameter to 0.2, then run the code again"
- No manual copying of results
  - Write out/save tables and figures using packages
  - Compute all numbers in package
- No manual install of packages
  - Use a script to create all directories, install all necessary packages/requirements/etc.
- Clear instructions!





Provide data citations (in manuscript) and data availability statements (in README or appendix)

For institutions

Better support for researchers

- Training in methods (with various centers, institutions, etc.)
  - For current researchers (Carpentries, custom, etc.)
  - 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

Verification services

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

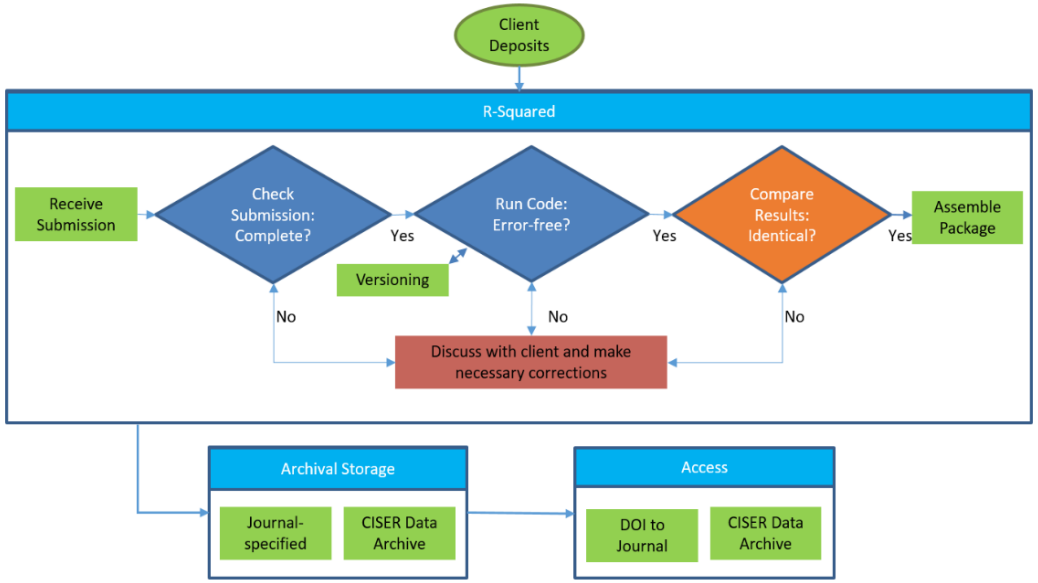
Cornell University

CISER CORNELL INSTITUTE for Social and Economic Research

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

RESULTS REPRODUCTION (R-SQUARED)

Results Reproduction (R-Squared) is a service that computationally reproduces the results of your research to ensure Reproducibility and Transparency – think of it as *enhanced proofreading for your Data and Code*.



HARVARD UNIVERSITY

People

IQSS
The Institute for Quantitative Social Science

About ▾ Programs & Products ▾ Resear

HOME / ABOUT / NEWS /

Announcing the Alexander and Diviya Magaro Peer Pre-Review Program at IQSS

January 10, 2019

The Institute for Quantitative Social Science is excited to announce the Alexander and Diviya Magaro Peer Pre-Review Program (PPR). PPR is designed to help IQSS-affiliated faculty improve scholarship before it becomes public, speed scientific discovery and publication, and reduce substantial inefficiencies for individual researchers.

Verification services

Your students!

Your colleagues!

For journals

Goal: Transportability

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

Social science “guild”

Social Science Data Editors

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

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

Thank you!

DOI: to come