Reproducibility and transparency in academia, and implications for statistical agencies

Lars Vilhuber

2025-10-09

Follow along

larsvilhuber.github.io/transparency-statistical-agencies/ (HTML zipped, PDF)

Disclaimer

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.

Goals of my talk

What is the state of reproducibility and transparency in academic economics?

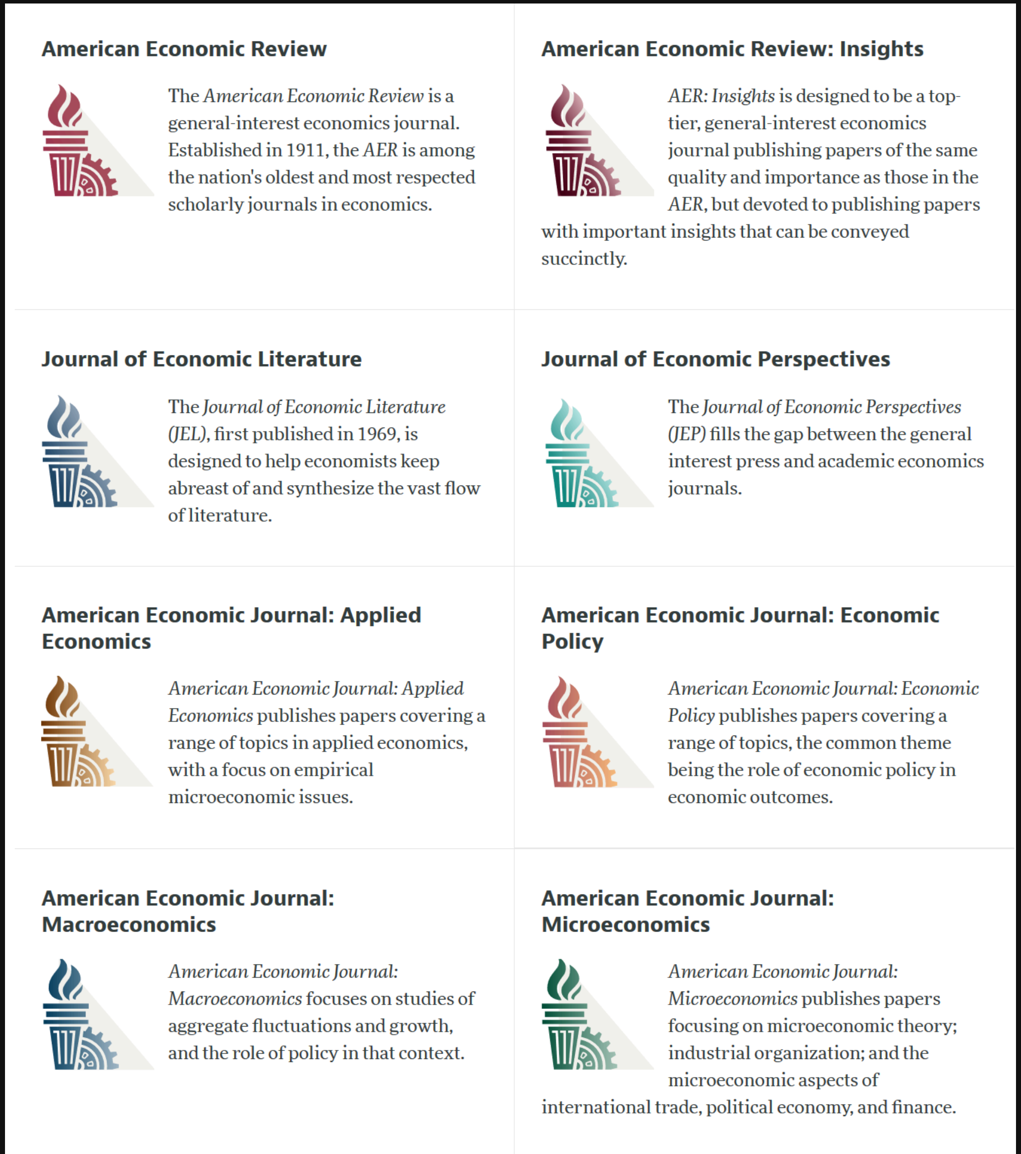
What are the benefits of reproducibility and transparency?

Increasing broad consensus in academia

What are the implications for statistical agencies?

Reproducibility in Economics

AEA Journals



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.

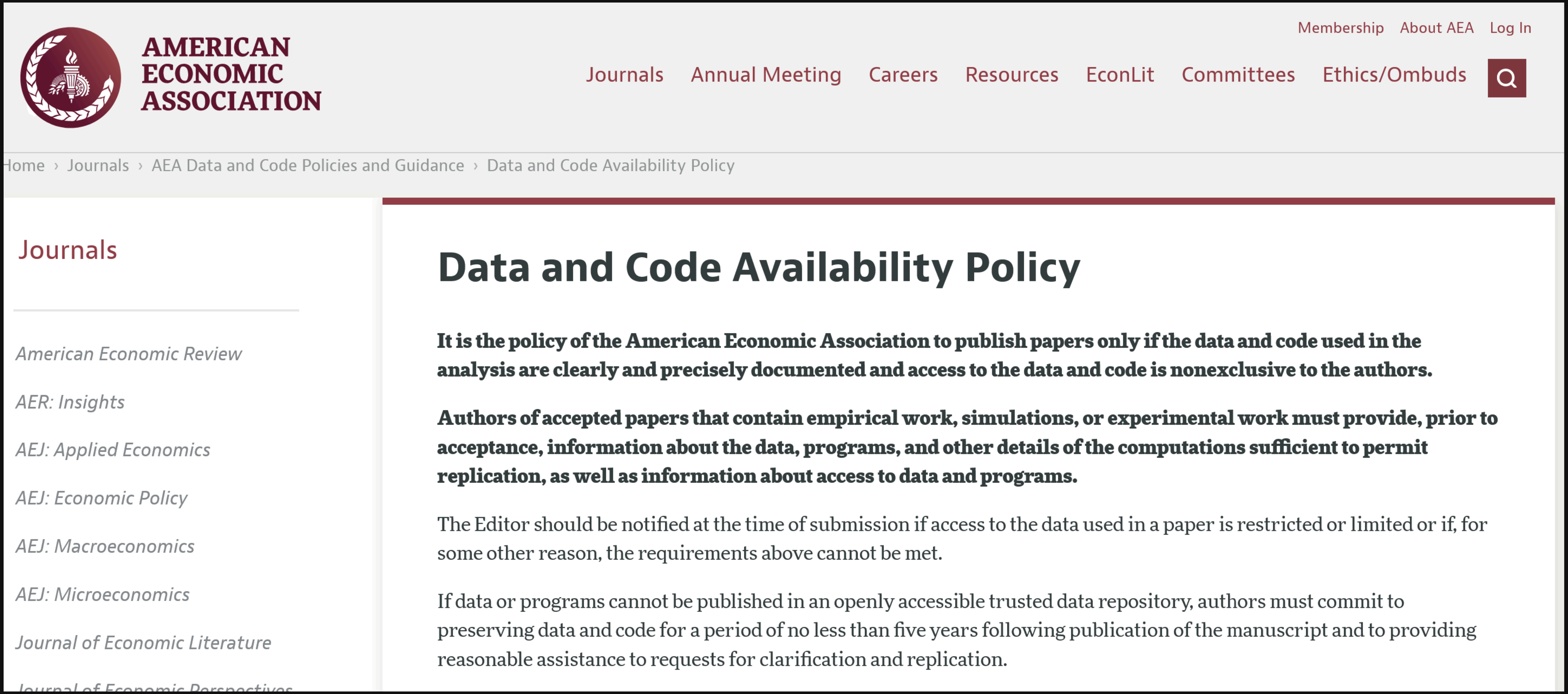
American Economic Journal: Microeconomics

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

What is a replication package?

- AEA Data and Code Availability policy
- Data and Code Availability Standard DCAS v1.0
- AEA Data and Code Repository

AEA policy



Data and Code Availability Policy

It is the policy of the American Economic Association to publish papers only if the data and code used in the analysis are clearly and precisely documented and access to the data and code is nonexclusive to the authors.

Authors of accepted papers that contain empirical work, simulations, or experimental work must provide, prior to acceptance, information about the data, programs, and other details of the computations sufficient to permit replication, as well as information about access to data and programs.

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.

If data or programs cannot be published in an openly accessible trusted data repository, authors must commit to preserving data and code for a period of no less than five years following publication of the manuscript and to providing reasonable assistance to requests for clarification and replication.

Tenets of the Policy

- **Transparency**
- **Completeness**
- **Preservation**

Transparency

- Provenance of the ***data***
- Processing of the data, from raw data to results (code)

> 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.

Completeness

- All data needs to be identified and and access described
- All code needs to be described and provided

> Authors … must provide, prior to acceptance, the **data, programs, and other details** of the computations **sufficient** to permit replication

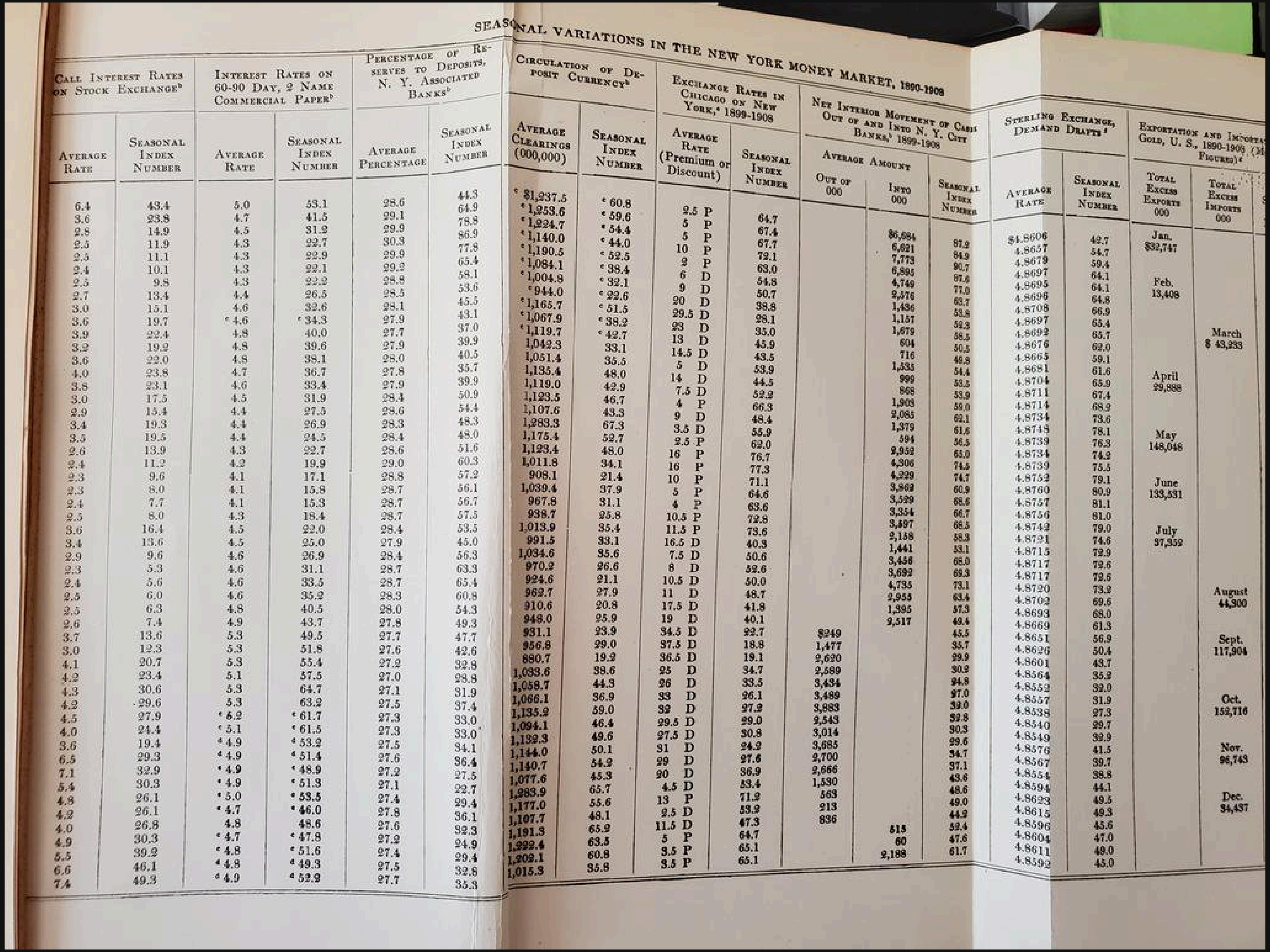
Preservation

- All data needs to be preserved for future replicators
  - Ideally, within the replication package, subject to ToU, for convenience
  - Otherwise, in a **trusted repository**

Preservation

- Code must be in a trusted repository
  - Usually, within the replication package
  - Websites, Github, are ***not acceptable***

Historically

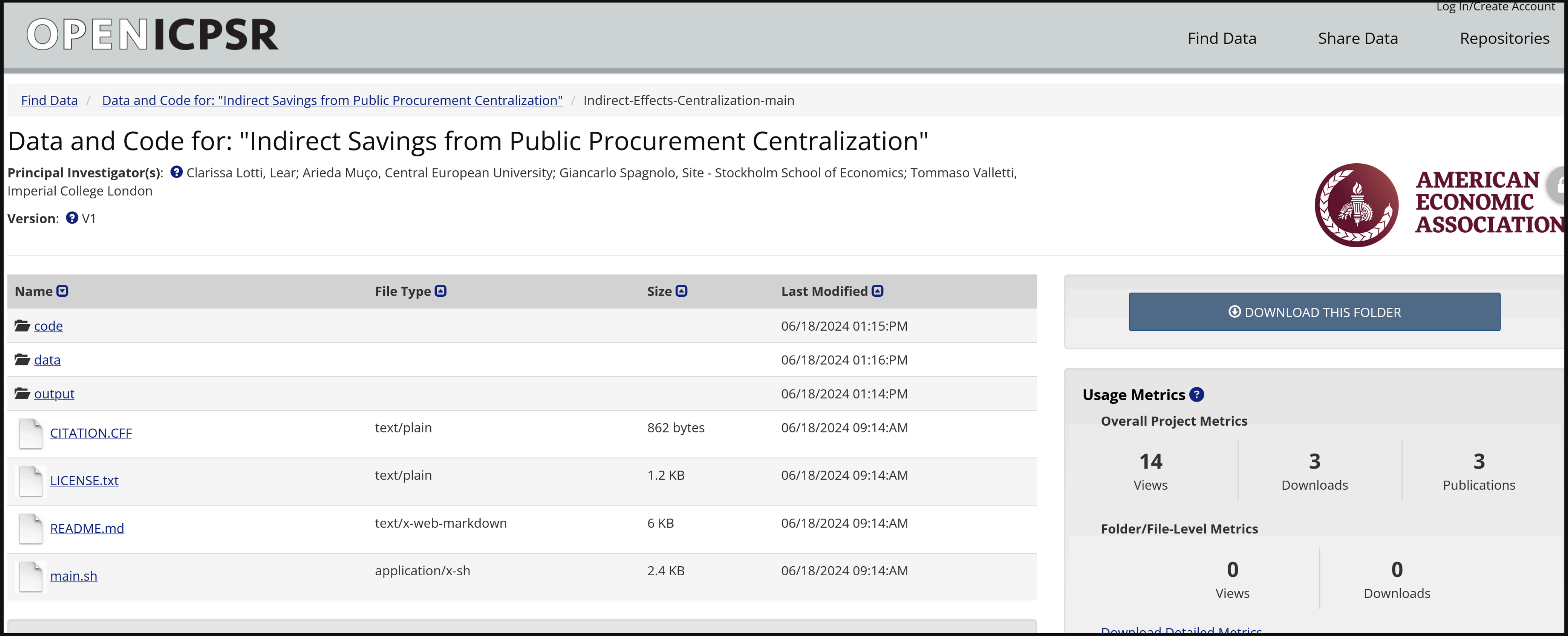


SEASONAL VARIATIONS IN THE NEW YORK MONEY MARKET, 1890-1909

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,[e] 1899-1908		Net Interior Movement of Cash Out of and Into N. Y. City Banks,[b] 1899-1908			Sterling Exchange, Demand Drafts[f]		Exportation and Importa Gold, U. S., 1890-1909 (M Figures)[g]	
										Average Amount						
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	Out of 000	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.	
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	$32,747	
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		
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	Feb.	
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	13,408	
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
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		$ 43,233
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	
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	29,888	
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	68.9		
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		
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	May	
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	148,048	
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	
2.3	9.6	4.1	17.1	28.8	57.2	908.1	21.4	10 P	71.1		3,862	60.9	4.8760	80.9	133,531	
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	
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	37,352	
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
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		44,300
2.5	6.3	4.8	40.5	28.0	54.3	910.6	20.8	17.5 D	41.8		1,395	57.3	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.
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		117,904
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.
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		152,716
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.
7.1	32.9	[e] 4.9	[e] 48.9	27.2	27.5	1,140.7	54.2	29 D	27.6	2,700		34.7	4.8567	39.7		96,743
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		37.1	4.8554	38.8		
4.8	26.1	[e] 5.0	[e] 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	[e] 4.7	[e] 46.0	27.8	36.1	1,177.0	55.6	13 P	71.2	563		48.6	4.8623	49.5		Dec.
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		34,437
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		613	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] 52.9	27.7	35.3	1,015.3	35.8	3.5 P	65.1		2,188	61.7	4.8592	45.0		

AER 2011 thanks to Stefano Dellavigna

Modern preservation



Log In/Create Account

Find Data / Data and Code for: "Indirect Savings from Public Procurement Centralization" / Indirect-Effects-Centralization-main

Data and Code for: "Indirect Savings from Public Procurement Centralization"

Principal Investigator(s): Clarissa Lotti, Lear; Arieda Muço, Central European University; Giancarlo Spagnolo, Site - Stockholm School of Economics; Tommaso Valletti, Imperial College London

Version: V1

Name	File Type	Size	Last Modified
code			06/18/2024 01:15:PM
data			06/18/2024 01:16:PM
output			06/18/2024 01:14:PM
CITATION.CFF	text/plain	862 bytes	06/18/2024 09:14:AM
LICENSE.txt	text/plain	1.2 KB	06/18/2024 09:14:AM
README.md	text/x-web-markdown	6 KB	06/18/2024 09:14:AM
main.sh	application/x-sh	2.4 KB	06/18/2024 09:14:AM

Usage Metrics

Overall Project Metrics

14	3	3
Views	Downloads	Publications

Folder/File-Level Metrics

0	0
Views	Downloads

Download Detailed Metrics

Side note: Government

- Data are often confidential
  - Are they preserved? (**NARA**, otherwise)
  - Are they accessible to others? (**FSRDC, NORC**, etc.)
- Code is sometimes deemed “confidential”
  - We will return to this topic!

Exceptions to the Policy

None

…

… there is a grey zone:

- When data do not belong to researcher, no control over preservation, access!
- Sometimes, ToU prevent researcher from revealing metadata (name of company, location)

Transparency again

- However:
  - No exception for need to **describe** access (own and other)
  - No exception for need to fully **describe** processing (possibly with redacted code)

Enforcement of the AEA Policy

Reproducibility?

Reproducibility

> “Reproducibility” refers to the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator.” [1]

Testing for …

- **Transparency**
- **Completeness**

through **reproducibility**

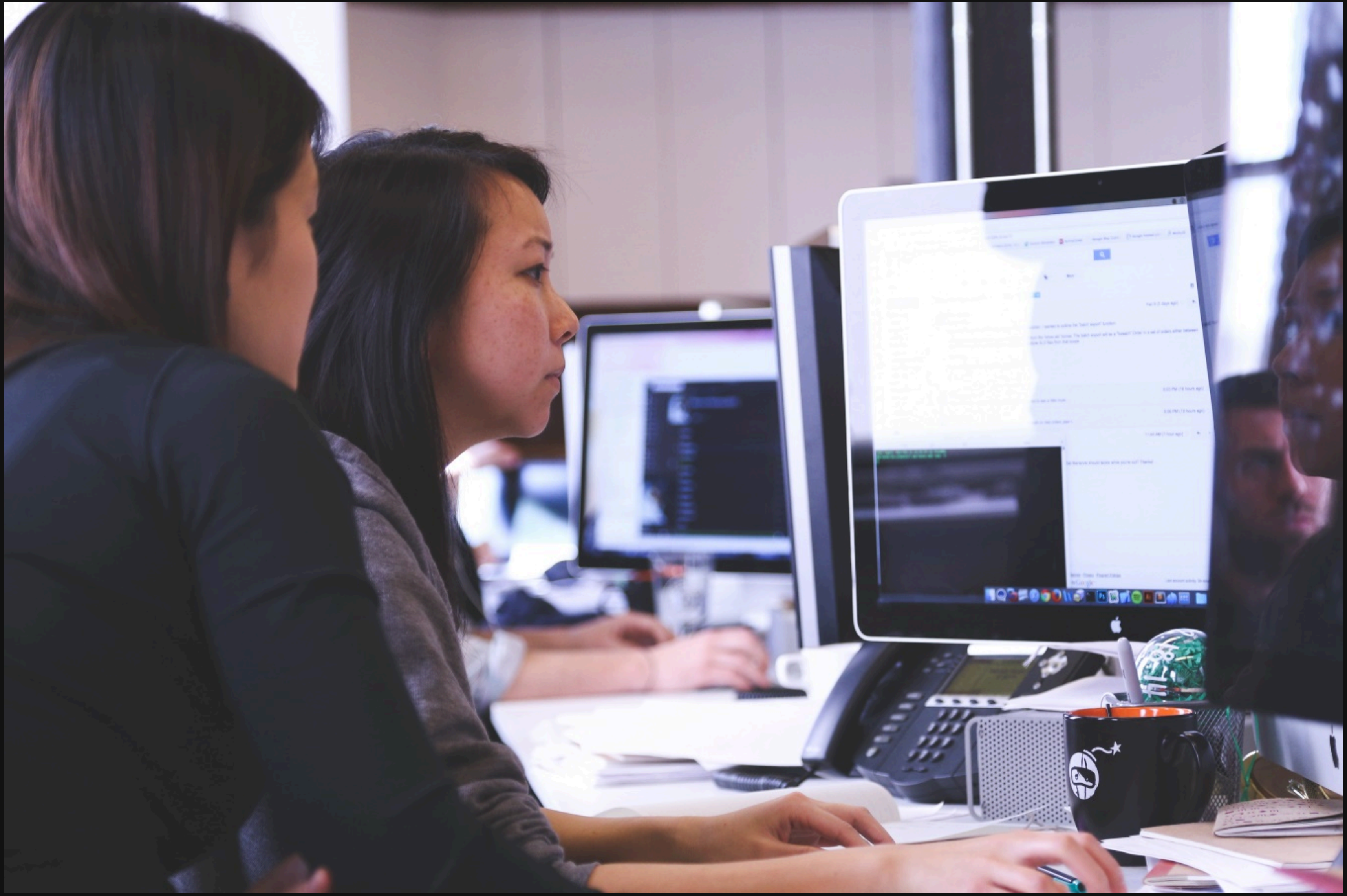
Criteria: Transparency?

- Can a reasonable person understand the description of **acquisition of data** and **processing** via code?

Criteria: Completeness?

- Do the provided materials allow to reproduce all the **tables** and **figures** in the paper?

Who is the target person?

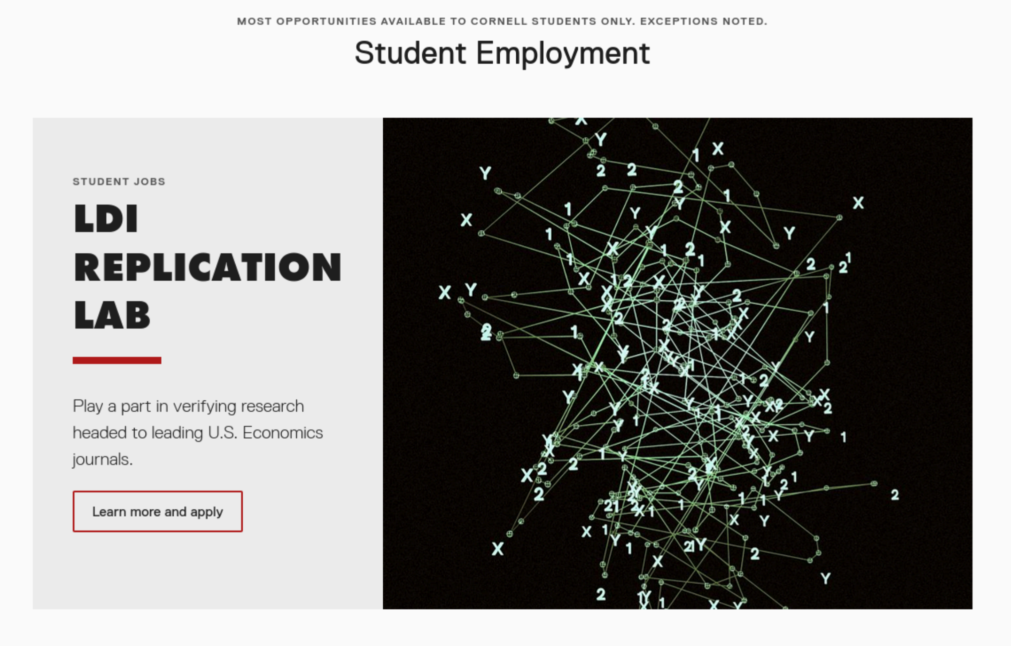


Student replicators

Who is the target person?

Over the past 6 years, over **170** ***undergraduate*** students have been involved in verifying these articles.

- Economics, biostatistics, sociology
- Typically recruited in sophomore or junior year, but will consider freshmen through master’s students



MOST OPPORTUNITIES AVAILABLE TO CORNELL STUDENTS ONLY. EXCEPTIONS NOTED.

Student Employment

STUDENT JOBS

LDI REPLICATION LAB

Play a part in verifying research headed to leading U.S. Economics journals.

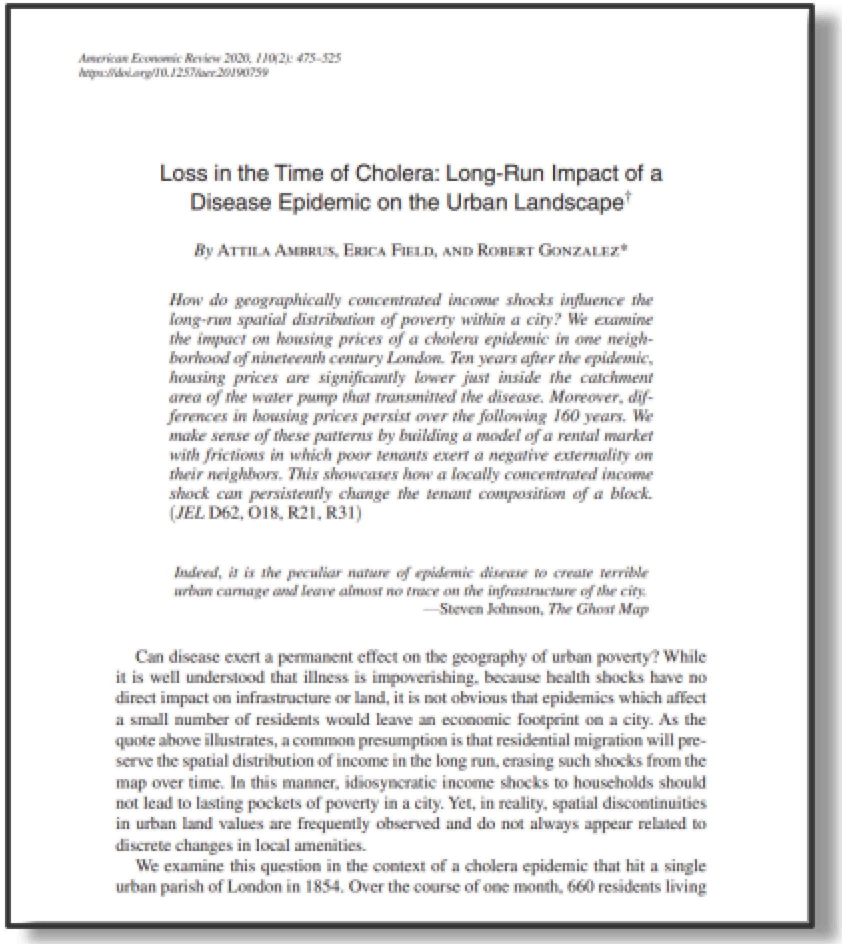
Who is the target person?

- **You** (in 4 years, between prepping 2 new courses, an R&R, a new child, and tenure coming up in 2 years)
- **Your RA** (in 4 years, because you are… see above)
- Your **future readers** who will cite you (in 4-10 years, who may want to extend or replicate your study, but won’t if it is too complex)

Tracing inputs from outputs

Credibility

2



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*

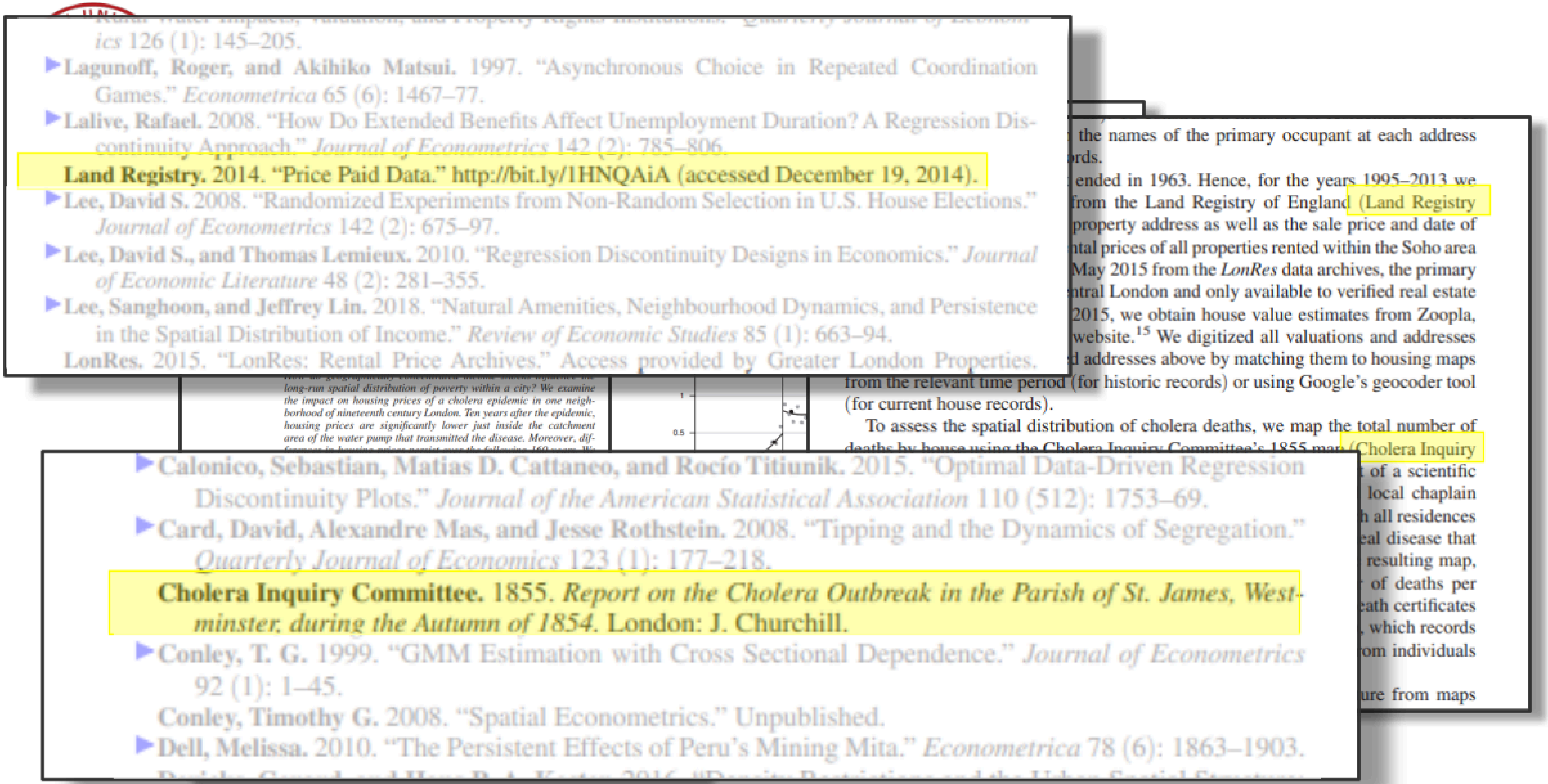
How do geographically concentrated income shocks influence the long-run spatial distribution of poverty within a city? We examine the impact on housing prices of a cholera epidemic in one neighborhood of nineteenth century London. Ten years after the epidemic, housing prices are significantly lower just 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
t 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.

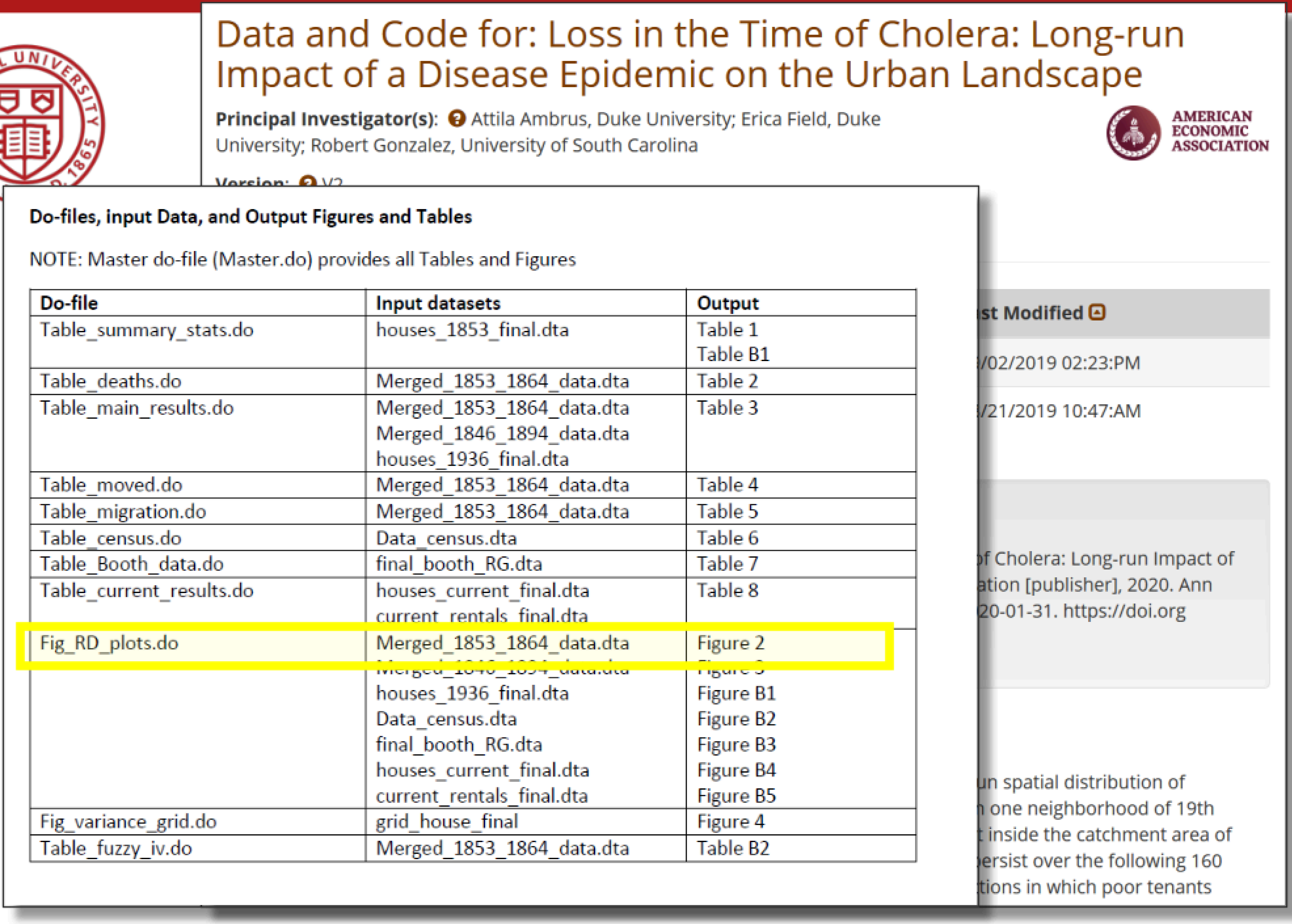
Cholera Inquiry Committee. 1855. *Report on the Cholera Outbreak in the Parish of St. James, Westminster, during the Autumn of 1854*. London: J. Churchill.

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

...st Modified

/02/2019 02:23:PM

/21/2019 10:47:AM

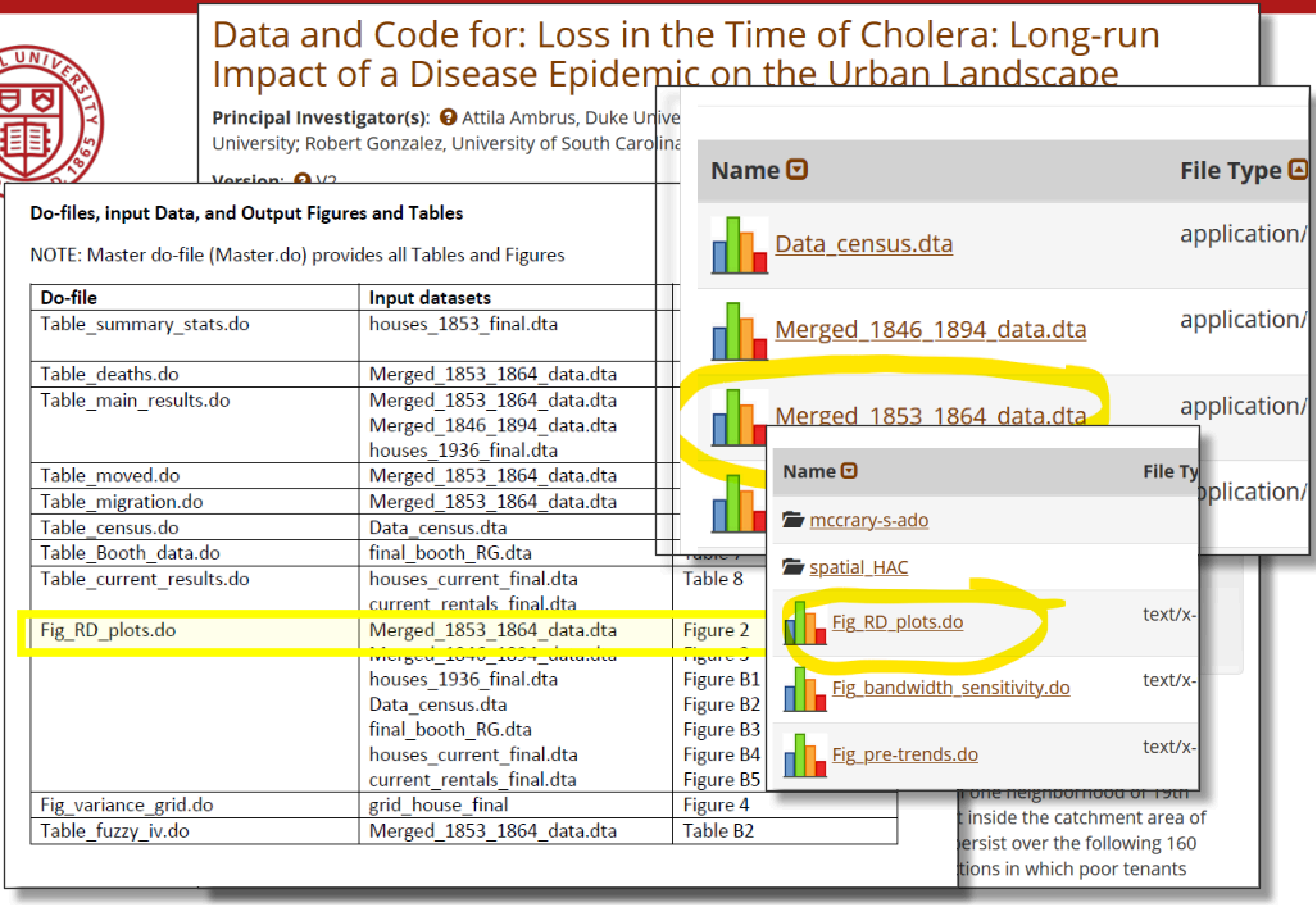
f Cholera: Long-run Impact of
ation [publisher], 2020. Ann
20-01-31. https://doi.org

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

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 Merged_1846_1894_data.dta houses_1936_final.dta Data_census.dta final_booth_RG.dta houses_current_final.dta current_rentals_final.dta	Figure 2 Figure 3 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



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

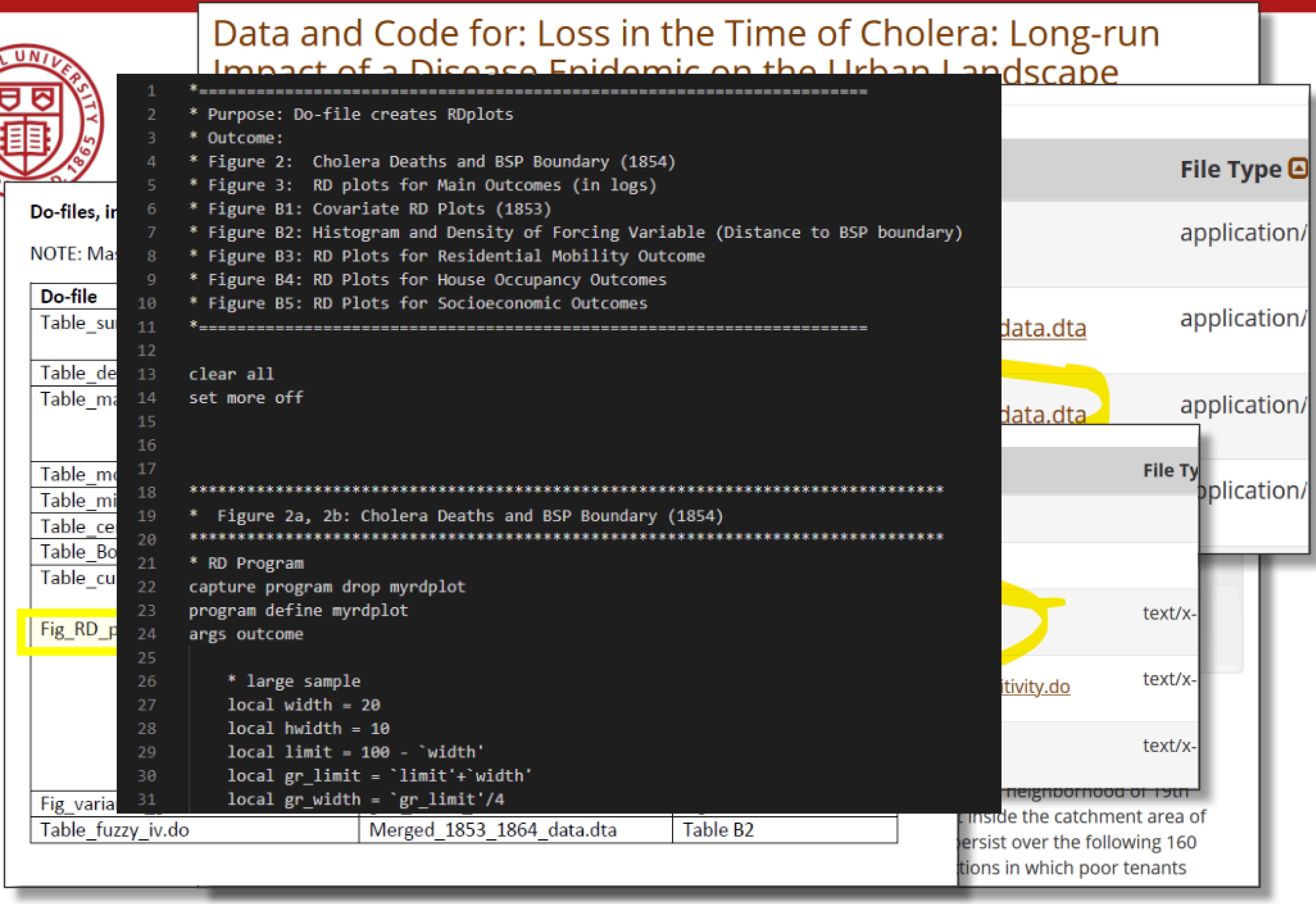
NOTE: Master do-file (Master.do) provides all Tables and Figures

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 7
Table_current_results.do	houses_current_final.dta current_rentals_final.dta	Table 8
Fig_RD_plots.do	Merged_1853_1864_data.dta Merged_1846_1894_data.dta houses_1936_final.dta Data_census.dta final_booth_RG.dta houses_current_final.dta current_rentals_final.dta	Figure 2 Figure 3 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

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

one neighborhood of 19th
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

Do-files, i

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

```
*=======================================================================
* 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
```

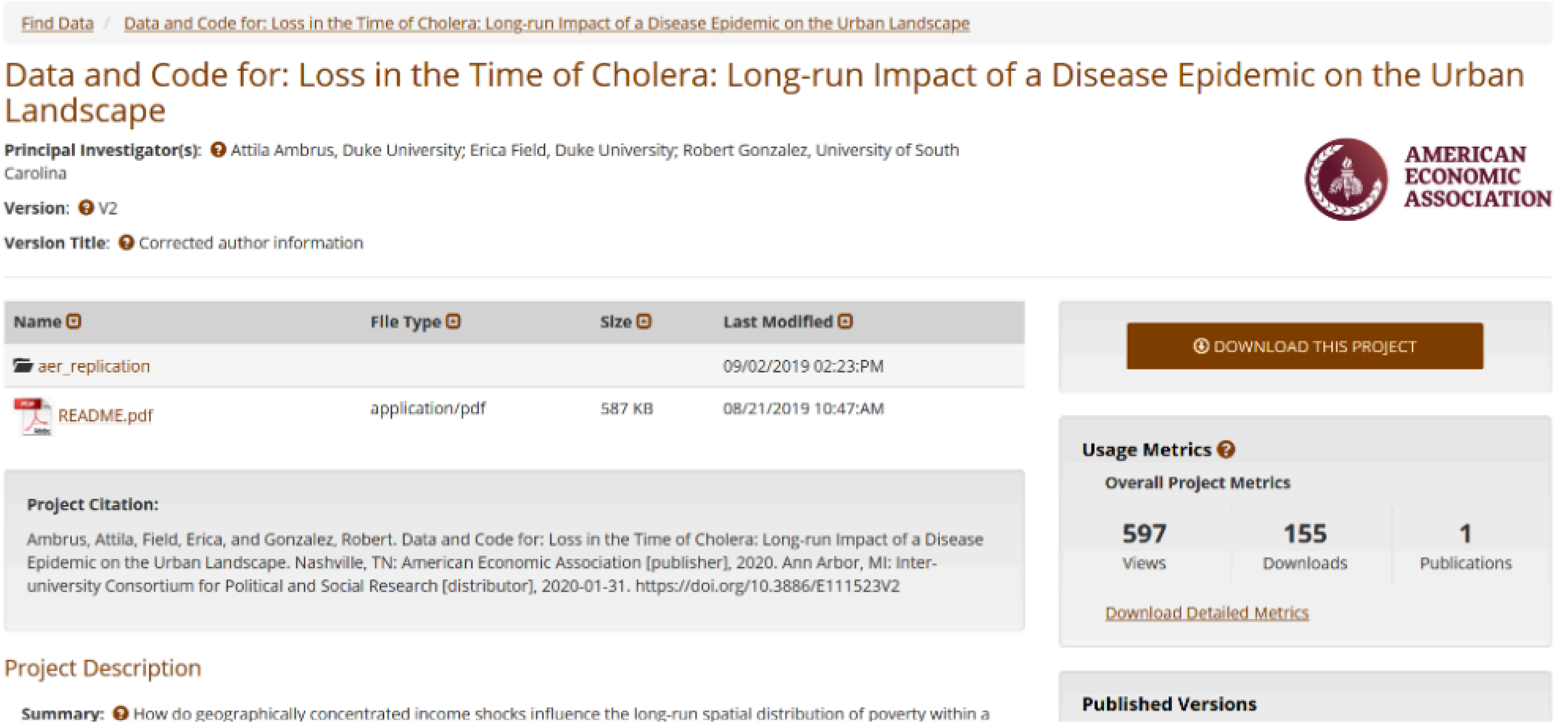
	File Type
	application/
data.dta	application/
data.dta	application/
	pplication/

	File Ty
	text/x-
itivity.do	text/x-
	text/x-

neighborhood of 19th
inside the catchment area of
ersist over the following 160
tions in which poor tenants

Reproducibility

3



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

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

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

Project Description

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

Published Versions

Credibility

Novel Coronavirus (COVID-19) Situation

209,839 confirmed cases

8,778 deaths

168 countries, areas or territories with cases

Last updated:19/03/2020 00:00 CET

Countries, areas or territories with cases

China :81174 cases

Italy :35713 cases

Iran (Islamic Republic of) :17361 cases

Spain :13716 cases

France :9043 cases

Republic of Korea :8413 cases

Germany :8198 cases

United States of America :7087 cases

Switzerland :3010 cases

The United Kingdom :2630 cases

Netherlands :2051 cases

Austria :1646 cases

Belgium :1486 cases

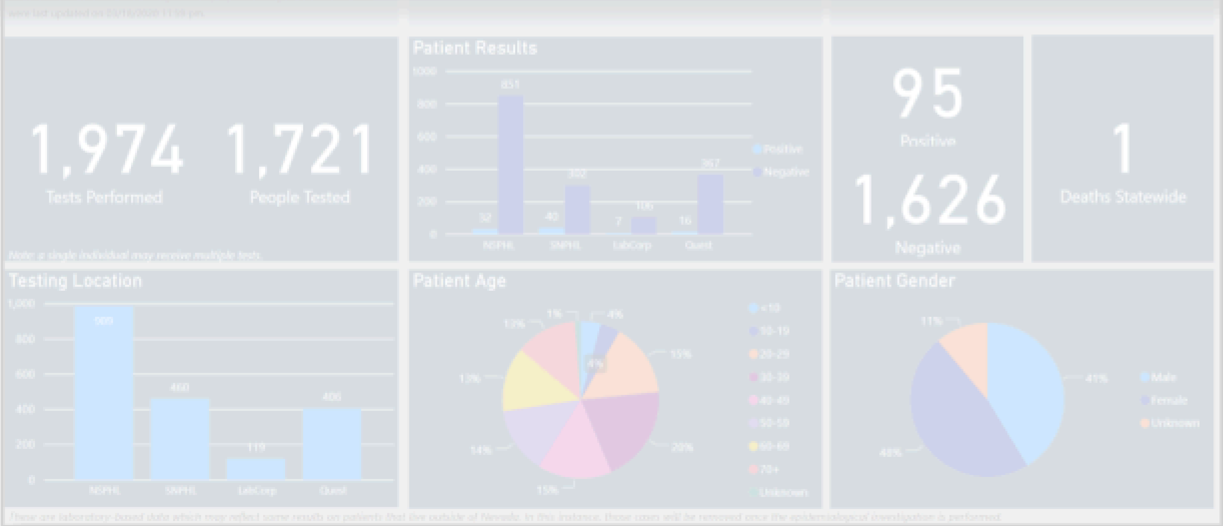
Norway :1423 cases

Sweden :1279 cases

Denmark :1044 cases

Czechia :522 cases

Australia :510 cases



were last updated on 03/18/2020 11:59 pm.

1,974 Tests Performed

1,721 People Tested

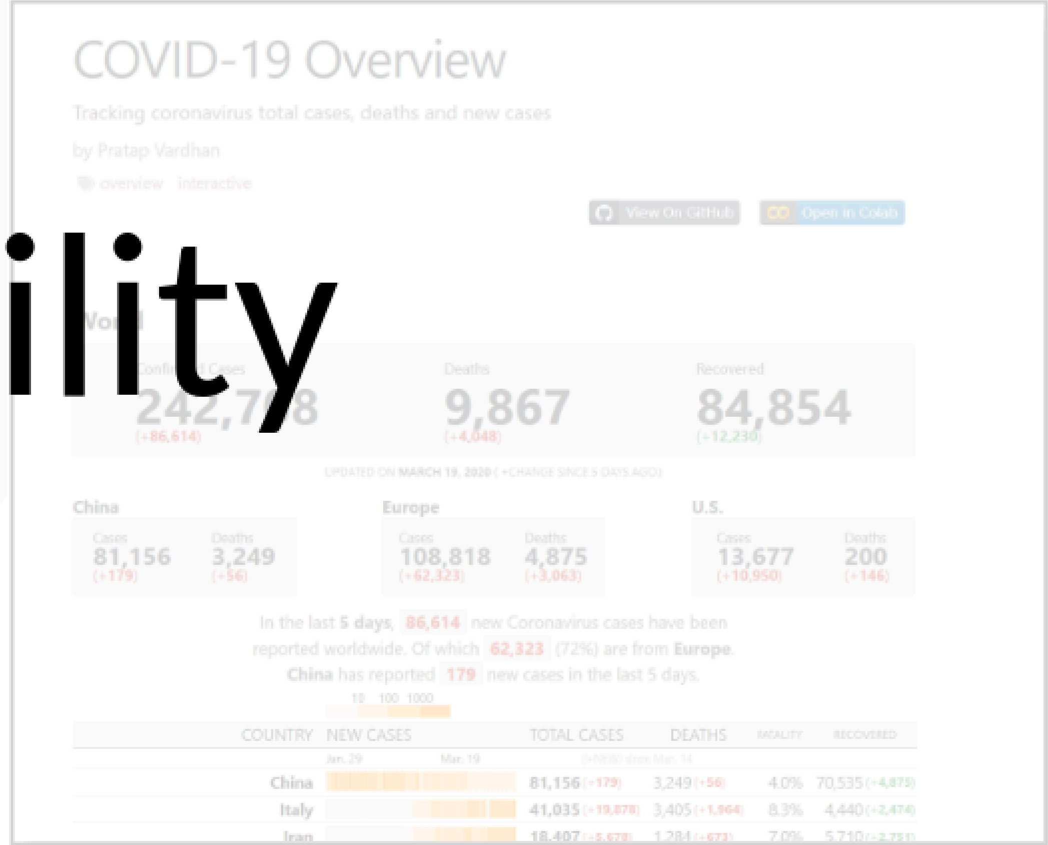
Note: a single individual may receive multiple tests.

95 Positive

1,626 Negative

1 Deaths Statewide

These are laboratory-based data which may reflect some results on patients that live outside of Nevada. In this instance, those cases will be removed once the epidemiological investigation is performed.



COVID-19 Overview

Tracking coronavirus total cases, deaths and new cases

by Pratap Vardhan

overview interactive

View On GitHub | Open in Colab

World

Confirmed Cases	Deaths	Recovered
242,708 (+86,614)	9,867 (+4,048)	84,854 (+12,230)

UPDATED ON MARCH 19, 2020 (+CHANGE SINCE 5 DAYS AGO)

	China		Europe		U.S.
Cases	Deaths	Cases	Deaths	Cases	Deaths
81,156 (+179)	3,249 (+56)	108,818 (+62,323)	4,875 (+3,063)	13,677 (+10,950)	200 (+146)

In the last **5 days**, 86,614 new Coronavirus cases have been reported worldwide. Of which 62,323 (72%) are from **Europe**. **China** has reported 179 new cases in the last 5 days.

COUNTRY	NEW CASES	TOTAL CASES	DEATHS	FATALITY	RECOVERED
	Jan. 29 Mar. 19	(+NEW) since Mar, 14			
China		81,156 (+179)	3,249 (+56)	4.0%	70,535 (+4,875)
Italy		41,035 (+19,878)	3,405 (+1,964)	8.3%	4,440 (+2,474)
Iran		18,407 (+5,678)	1,284 (+673)	7.0%	5,710 (+2,751)

Reproducibility in Economics and beyond

Social Science Data Editors DCAS README Guidance Contact

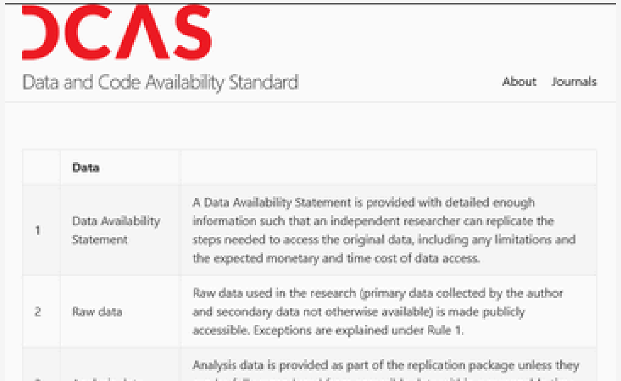
Social Science Data Editors

Improving reproducibility in the social and economic sciences

Data and Code Availability Standard

DCAS The Data and Code Availability Standard (DCAS) is a standard for sharing research code and data, endorsed by leading journals in social sciences. See https://datacodestandard.org/ for more information.

DOI 10.5281/zenodo.7436134



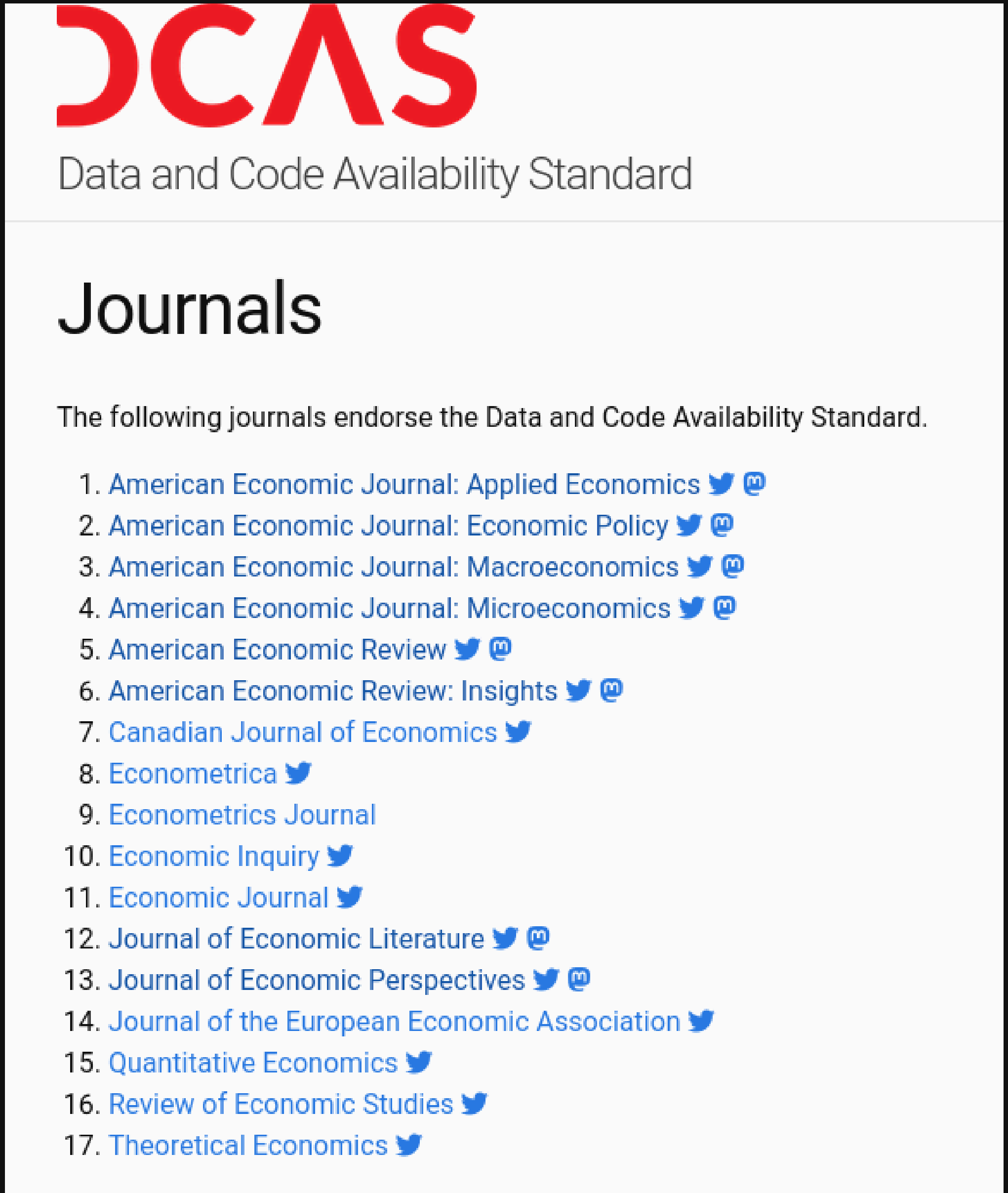
DCAS

Data and Code Availability Standard

	Data	
1	Data Availability Statement	A Data Availability Statement is provided with detailed enough information such that an independent researcher can replicate the steps needed to access the original data, including any limitations and the expected monetary and time cost of data access.
2	Raw data	Raw data used in the research (primary data collected by the author and secondary data not otherwise available) is made publicly accessible. Exceptions are explained under Rule 1.
3	Analysis data	Analysis data is provided as part of the replication package unless they

Data Editors

- American Economic Association (8)
- Econometric Society (3)
- Canadian Journal of Economics (1)
- Royal Economic Society (2)
- Western Economic Association International (1)
- European Economic Association (1)
- Review of Economic Studies (1)
- **Journal of the European Economic Association** (1)
- **Journal of Political Economy** (3)



Data and Code Availability Standard

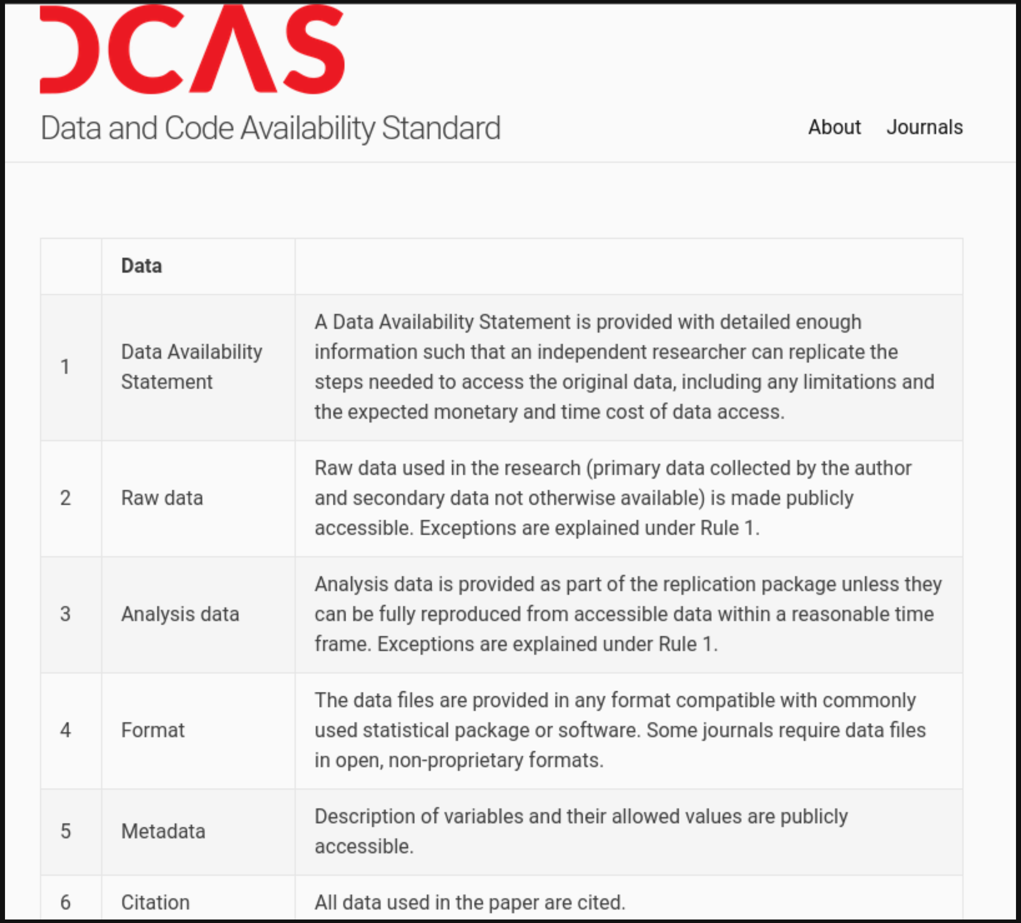
Journals

The following journals endorse the Data and Code Availability Standard.

1. American Economic Journal: Applied Economics
2. American Economic Journal: Economic Policy
3. American Economic Journal: Macroeconomics
4. American Economic Journal: Microeconomics
5. American Economic Review
6. American Economic Review: Insights
7. Canadian Journal of Economics
8. Econometrica
9. Econometrics Journal
10. Economic Inquiry
11. Economic Journal
12. Journal of Economic Literature
13. Journal of Economic Perspectives
14. Journal of the European Economic Association
15. Quantitative Economics
16. Review of Economic Studies
17. Theoretical Economics

Common policies

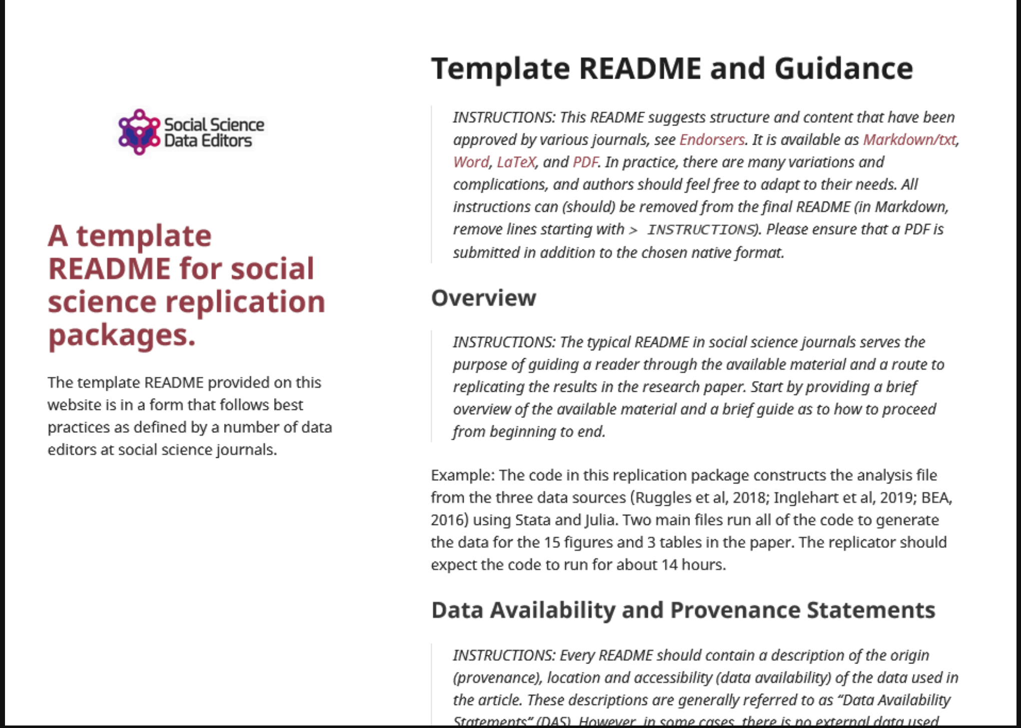
https://social-science-data-editors.github.io/



DCAS

Data and Code Availability Standard

	Data	
1	Data Availability Statement	A Data Availability Statement is provided with detailed enough information such that an independent researcher can replicate the steps needed to access the original data, including any limitations and the expected monetary and time cost of data access.
2	Raw data	Raw data used in the research (primary data collected by the author and secondary data not otherwise available) is made publicly accessible. Exceptions are explained under Rule 1.
3	Analysis data	Analysis data is provided as part of the replication package unless they can be fully reproduced from accessible data within a reasonable time frame. Exceptions are explained under Rule 1.
4	Format	The data files are provided in any format compatible with commonly used statistical package or software. Some journals require data files in open, non-proprietary formats.
5	Metadata	Description of variables and their allowed values are publicly accessible.
6	Citation	All data used in the paper are cited.



A template README for social science replication packages.

The template README provided on this website is in a form that follows best practices as defined by a number of data editors at social science journals.

Template README and Guidance

> *INSTRUCTIONS: This README suggests structure and content that have been approved by various journals, see Endorsers. It is available as Markdown/txt, Word, LaTeX, and PDF. In practice, there are many variations and complications, and authors should feel free to adapt to their needs. All instructions can (should) be removed from the final README (in Markdown, remove lines starting with `> INSTRUCTIONS`). Please ensure that a PDF is submitted in addition to the chosen native format.*

Overview

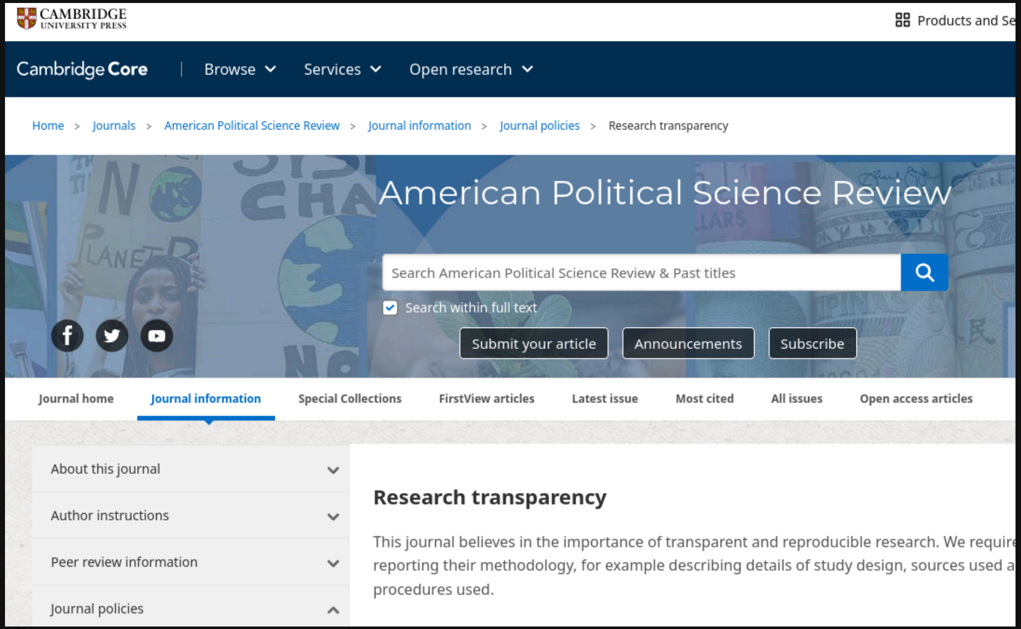
> *INSTRUCTIONS: The typical README in social science journals serves the purpose of guiding a reader through the available material and a route to replicating the results in the research paper. Start by providing a brief overview of the available material and a brief guide as to how to proceed from beginning to end.*

Example: The code in this replication package constructs the analysis file from the three data sources (Ruggles et al, 2018; Inglehart et al, 2019; BEA, 2016) using Stata and Julia. Two main files run all of the code to generate the data for the 15 figures and 3 tables in the paper. The replicator should expect the code to run for about 14 hours.

Data Availability and Provenance Statements

> *INSTRUCTIONS: Every README should contain a description of the origin (provenance), location and accessibility (data availability) of the data used in the article. These descriptions are generally referred to as "Data Availability Statements" (DAS). However, in some cases, there is no external data used*

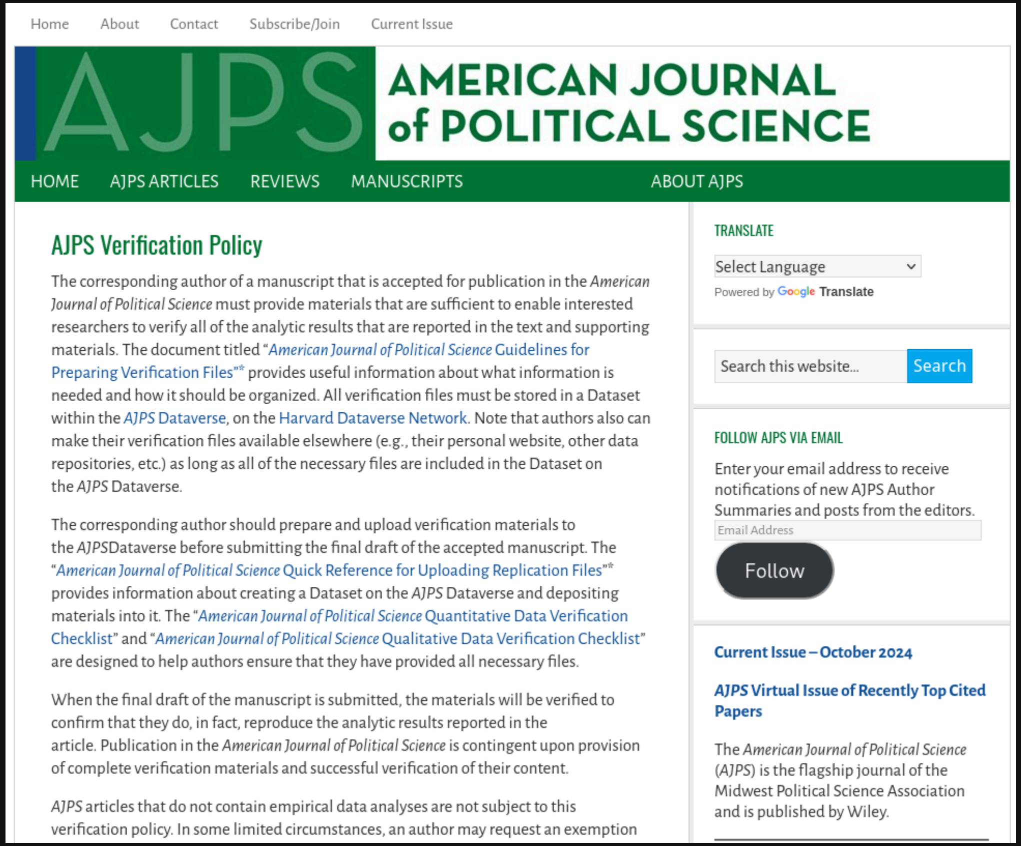
Elsewhere: Political Science



Research transparency

This journal believes in the importance of transparent and reproducible research. We require reporting their methodology, for example describing details of study design, sources used a procedures used.

APSR



AJPS Verification Policy

The corresponding author of a manuscript that is accepted for publication in the *American Journal of Political Science* must provide materials that are sufficient to enable interested researchers to verify all of the analytic results that are reported in the text and supporting materials. The document titled "*American Journal of Political Science* Guidelines for Preparing Verification Files"* provides useful information about what information is needed and how it should be organized. All verification files must be stored in a Dataset within the *AJPS* Dataverse, on the Harvard Dataverse Network. Note that authors also can make their verification files available elsewhere (e.g., their personal website, other data repositories, etc.) as long as all of the necessary files are included in the Dataset on the *AJPS* Dataverse.

The corresponding author should prepare and upload verification materials to the *AJPS*Dataverse before submitting the final draft of the accepted manuscript. The "*American Journal of Political Science* Quick Reference for Uploading Replication Files"* provides information about creating a Dataset on the *AJPS* Dataverse and depositing materials into it. The "*American Journal of Political Science* Quantitative Data Verification Checklist" and "*American Journal of Political Science* Qualitative Data Verification Checklist" are designed to help authors ensure that they have provided all necessary files.

When the final draft of the manuscript is submitted, the materials will be verified to confirm that they do, in fact, reproduce the analytic results reported in the article. Publication in the *American Journal of Political Science* is contingent upon provision of complete verification materials and successful verification of their content.

AJPS articles that do not contain empirical data analyses are not subject to this verification policy. In some limited circumstances, an author may request an exemption

Enter your email address to receive notifications of new AJPS Author Summaries and posts from the editors.

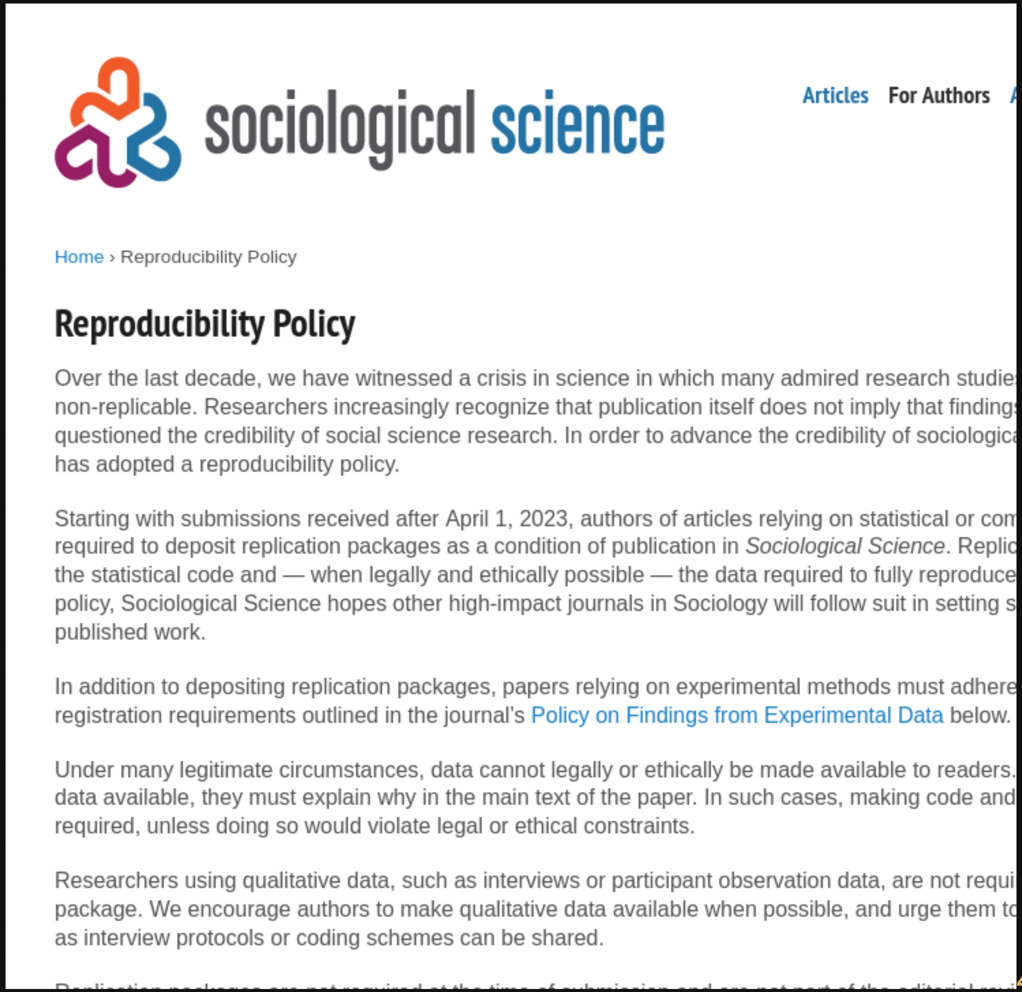
Current Issue – October 2024

***AJPS* Virtual Issue of Recently Top Cited Papers**

The *American Journal of Political Science* (*AJPS*) is the flagship journal of the Midwest Political Science Association and is published by Wiley.

AJPS

Elsewhere: Sociology



sociological science

Home › Reproducibility Policy

Reproducibility Policy

Over the last decade, we have witnessed a crisis in science in which many admired research studies non-replicable. Researchers increasingly recognize that publication itself does not imply that findings questioned the credibility of social science research. In order to advance the credibility of sociologica has adopted a reproducibility policy.

Starting with submissions received after April 1, 2023, authors of articles relying on statistical or com required to deposit replication packages as a condition of publication in *Sociological Science*. Replic the statistical code and — when legally and ethically possible — the data required to fully reproduce policy, Sociological Science hopes other high-impact journals in Sociology will follow suit in setting s published work.

In addition to depositing replication packages, papers relying on experimental methods must adhere registration requirements outlined in the journal's Policy on Findings from Experimental Data below.

Under many legitimate circumstances, data cannot legally or ethically be made available to readers. data available, they must explain why in the main text of the paper. In such cases, making code and required, unless doing so would violate legal or ethical constraints.

Researchers using qualitative data, such as interviews or participant observation data, are not requi package. We encourage authors to make qualitative data available when possible, and urge them to as interview protocols or coding schemes can be shared.

Replication packages are not required at the time of submission and are not part of the editorial revi

4

Trust in Government Statistics

United Nations

Fundamental Principles of Official Statistics, **Principle 3**:

> ***Accountability and Transparency*** To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics [5]

National Academies

Principles and Practices for a Federal Statistical Agency, **Principle 2**:

> Credibility among Data Users A federal statistical agency must have credibility with those who use its data and information [6]

OMB

“… flow of objective, **credible** statistics to support the decisions of individuals, households, governments, businesses, and other organizations.”

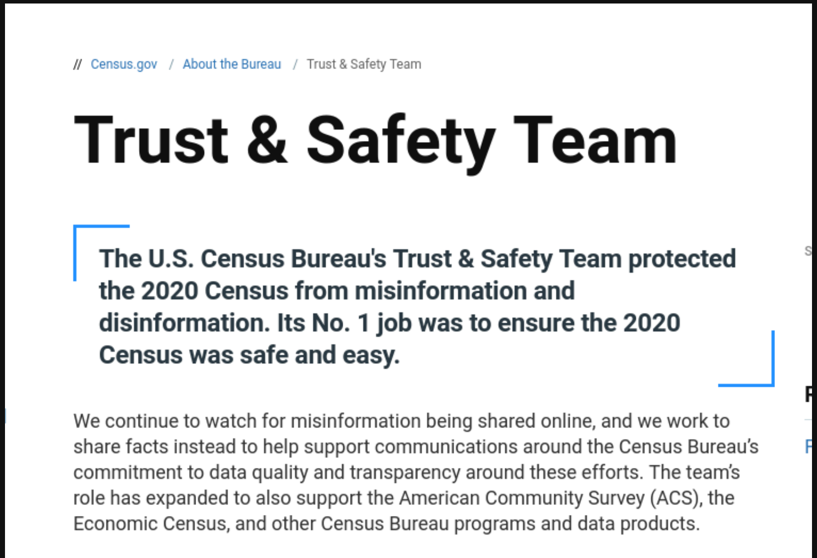
OMB

Statistical Policy Directive No. 1, 4:

“Any **loss of trust** in the integrity of the Federal statistical system and its products could lessen respondent cooperation with Federal statistical surveys, decrease the quality of statistical system products, and foster uncertainty about the validity of measures our Nation uses to monitor and assess its performance and progress.”

Agency efforts

U.S. Census Bureau



// Census.gov / About the Bureau / Trust & Safety Team

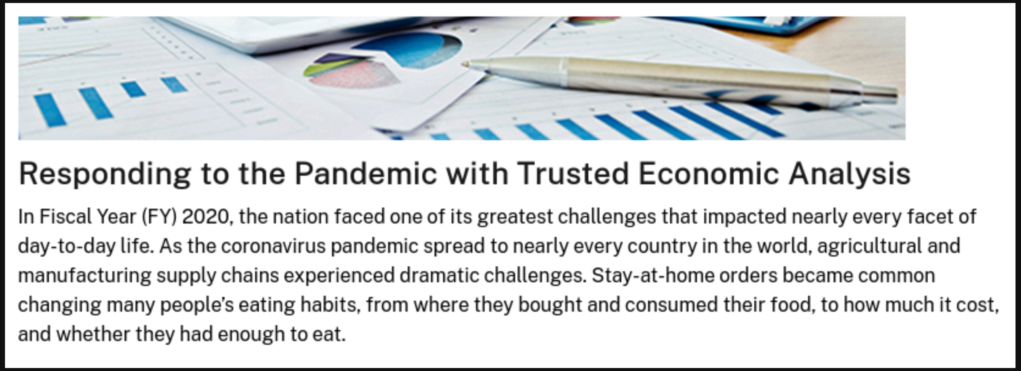
Trust & Safety Team

The U.S. Census Bureau's Trust & Safety Team protected the 2020 Census from misinformation and disinformation. Its No. 1 job was to ensure the 2020 Census was safe and easy.

We continue to watch for misinformation being shared online, and we work to share facts instead to help support communications around the Census Bureau's commitment to data quality and transparency around these efforts. The team's role has expanded to also support the American Community Survey (ACS), the Economic Census, and other Census Bureau programs and data products.

U.S. Census Trust and Safety Center

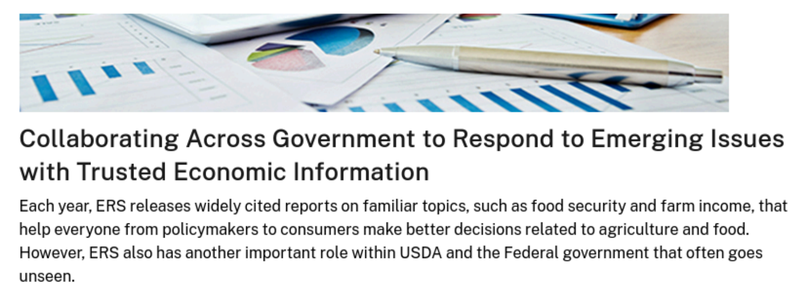
Agency efforts



Responding to the Pandemic with Trusted Economic Analysis

In Fiscal Year (FY) 2020, the nation faced one of its greatest challenges that impacted nearly every facet of day-to-day life. As the coronavirus pandemic spread to nearly every country in the world, agricultural and manufacturing supply chains experienced dramatic challenges. Stay-at-home orders became common changing many people's eating habits, from where they bought and consumed their food, to how much it cost, and whether they had enough to eat.

ERS Annual Report 2020



Collaborating Across Government to Respond to Emerging Issues with Trusted Economic Information

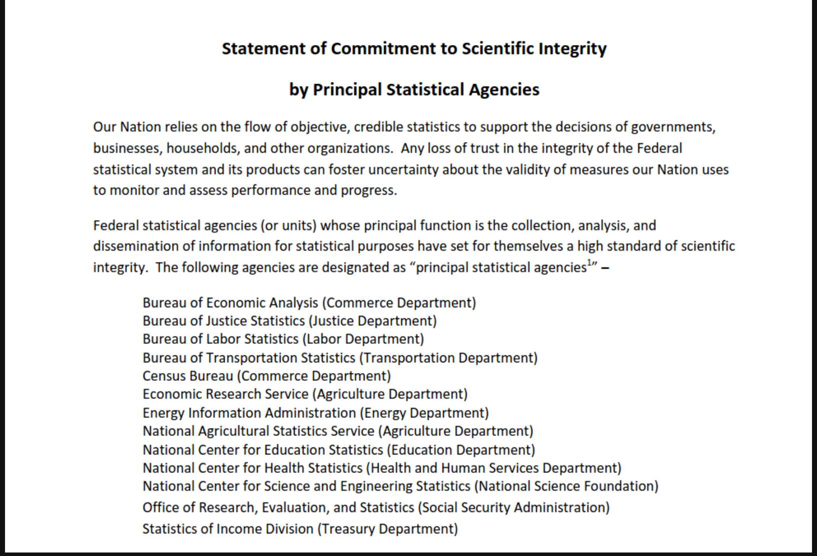
Each year, ERS releases widely cited reports on familiar topics, such as food security and farm income, that help everyone from policymakers to consumers make better decisions related to agriculture and food. However, ERS also has another important role within USDA and the Federal government that often goes unseen.

ERS Annual Report 2021

Joint Statement

Joint Statement on Commitment to Scientific Integrity and Transparency

- **Principle 2**: a Federal statistical agency must have credibility with those who use its data and information;
- **Principle 3**: a Federal statistical agency must have the trust of those whose information it obtains;



Statement of Commitment to Scientific Integrity

by Principal Statistical Agencies

Our Nation relies on the flow of objective, credible statistics to support the decisions of governments, businesses, households, and other organizations. Any loss of trust in the integrity of the Federal statistical system and its products can foster uncertainty about the validity of measures our Nation uses to monitor and assess performance and progress.

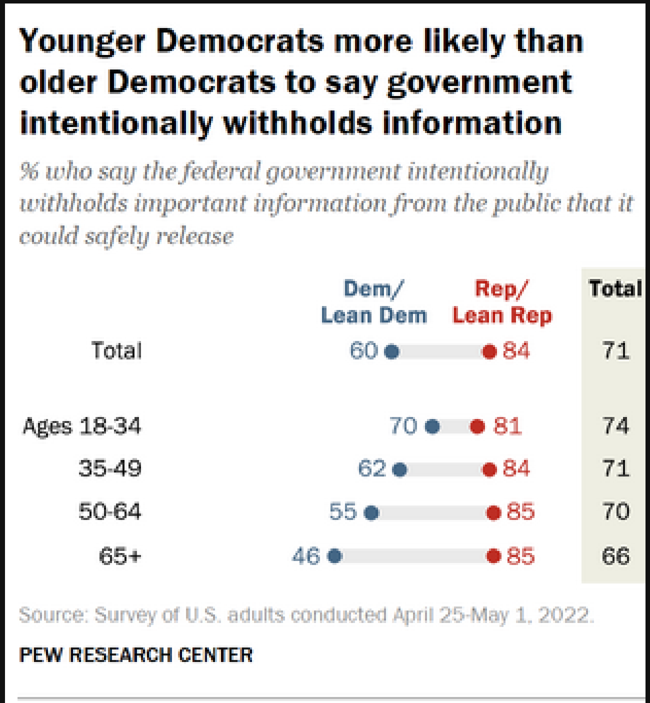
Federal statistical agencies (or units) whose principal function is the collection, analysis, and dissemination of information for statistical purposes have set for themselves a high standard of scientific integrity. The following agencies are designated as "principal statistical agencies[1]" –

Bureau of Economic Analysis (Commerce Department)
Bureau of Justice Statistics (Justice Department)
Bureau of Labor Statistics (Labor Department)
Bureau of Transportation Statistics (Transportation Department)
Census Bureau (Commerce Department)
Economic Research Service (Agriculture Department)
Energy Information Administration (Energy Department)
National Agricultural Statistics Service (Agriculture Department)
National Center for Education Statistics (Education Department)
National Center for Health Statistics (Health and Human Services Department)
National Center for Science and Engineering Statistics (National Science Foundation)
Office of Research, Evaluation, and Statistics (Social Security Administration)
Statistics of Income Division (Treasury Department)

Joint Statement

Waning trust

Pew Research



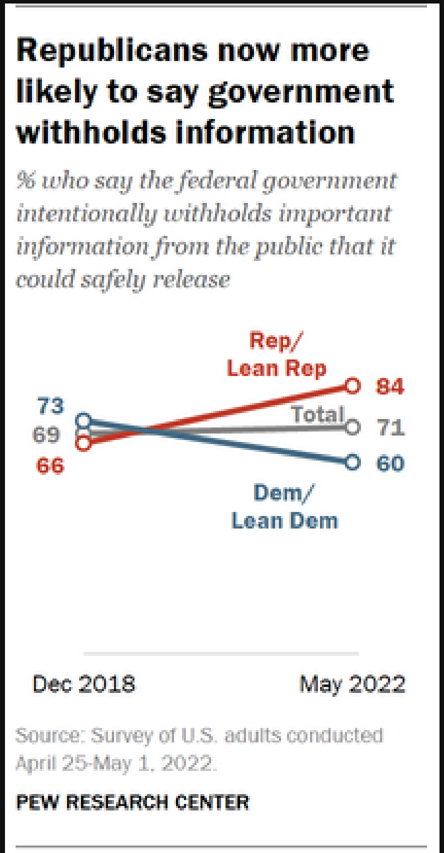
Younger Democrats more likely than older Democrats to say government intentionally withholds information

% who say the federal government intentionally withholds important information from the public that it could safely release

Source: Survey of U.S. adults conducted April 25-May 1, 2022.

PEW RESEARCH CENTER

Democrats on withholding data



Republicans now more likely to say government withholds information

% who say the federal government intentionally withholds important information from the public that it could safely release

Source: Survey of U.S. adults conducted April 25-May 1, 2022.

PEW RESEARCH CENTER

Republicans on withholding data

Computational Reproducibility and Official Statistics

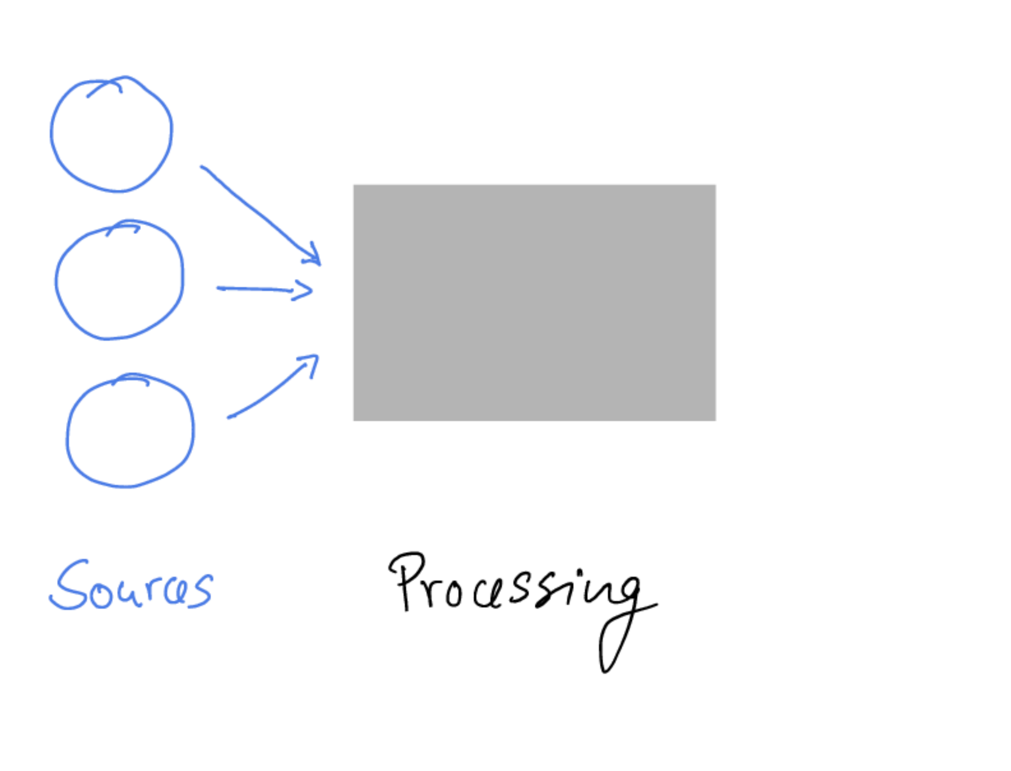
Agencies do provide **detailed information on sources**

- Surveys
- Administrative data

Computational Reproducibility and Official Statistics

But: **Availability of “computing instructions”**?

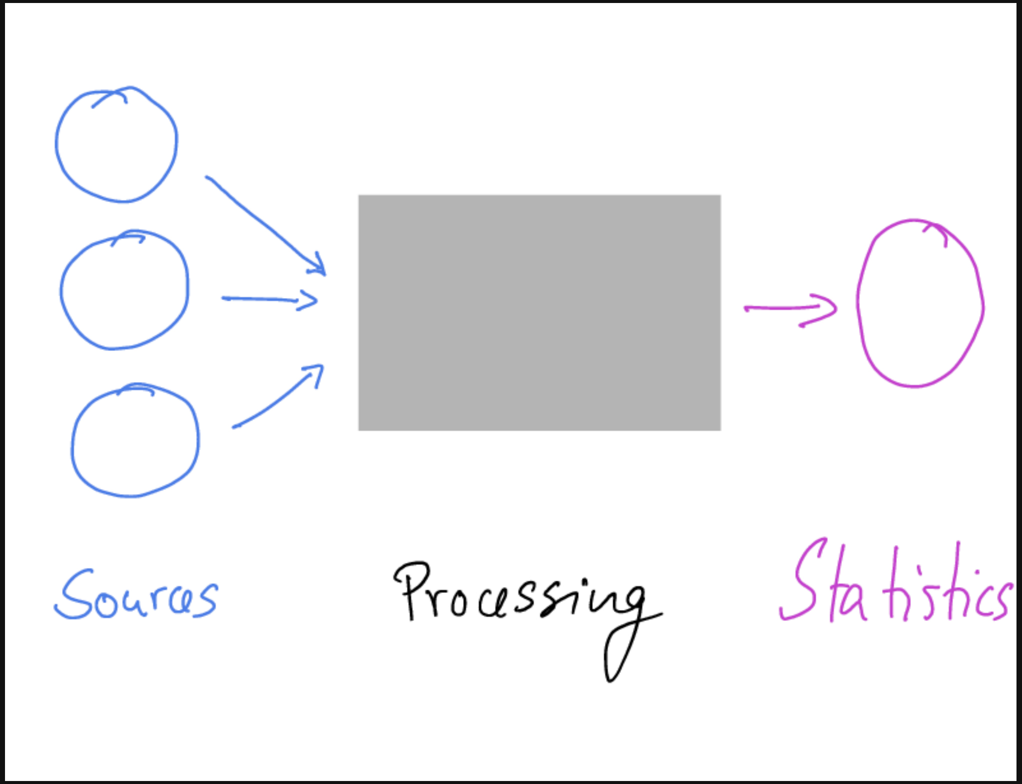
- ***Code*** for cleaning, aggregation, imputation
- Including for ***disclosure avoidance***



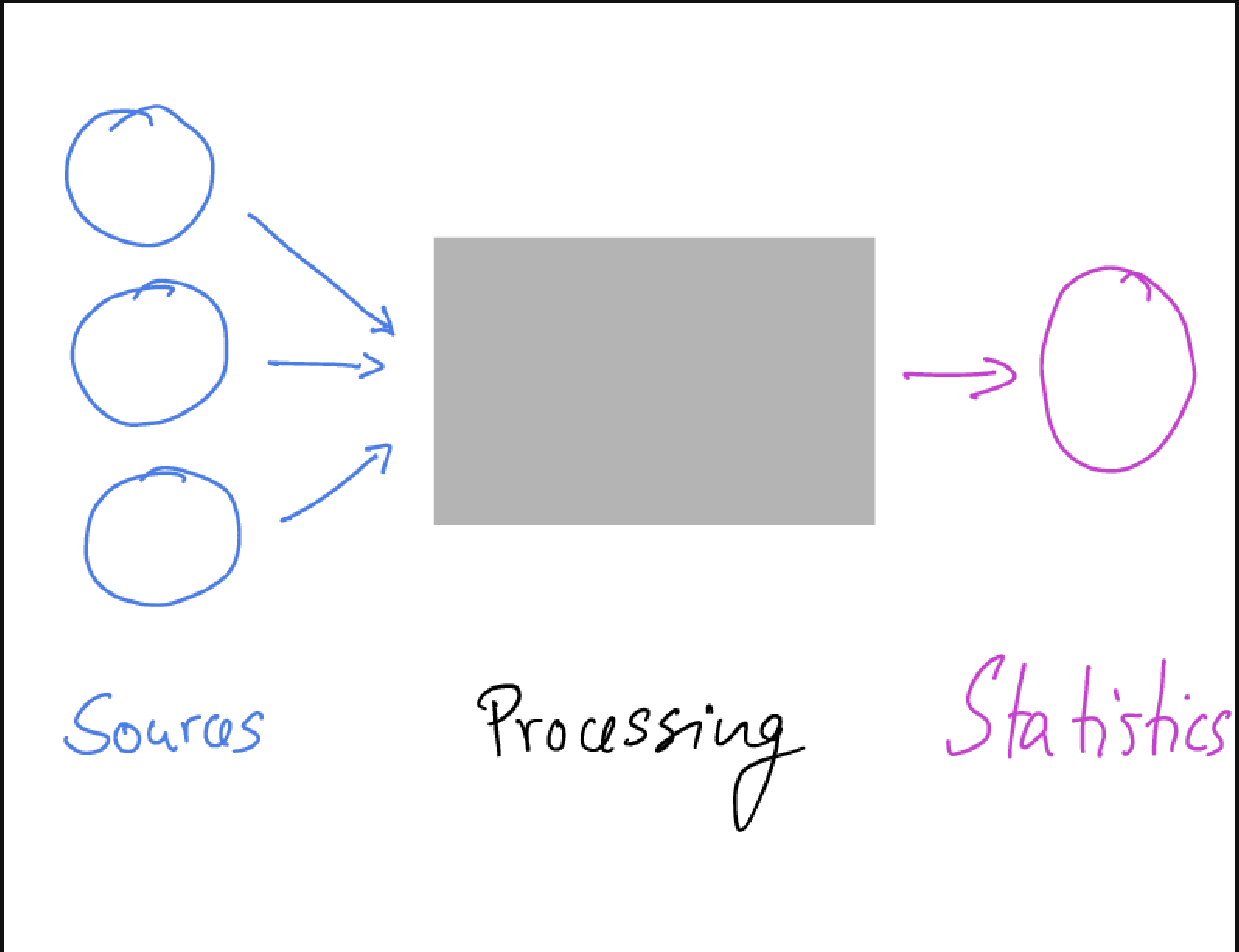
Computational Reproducibility and Official Statistics

But: Availability of **reliable, trusted data archives**

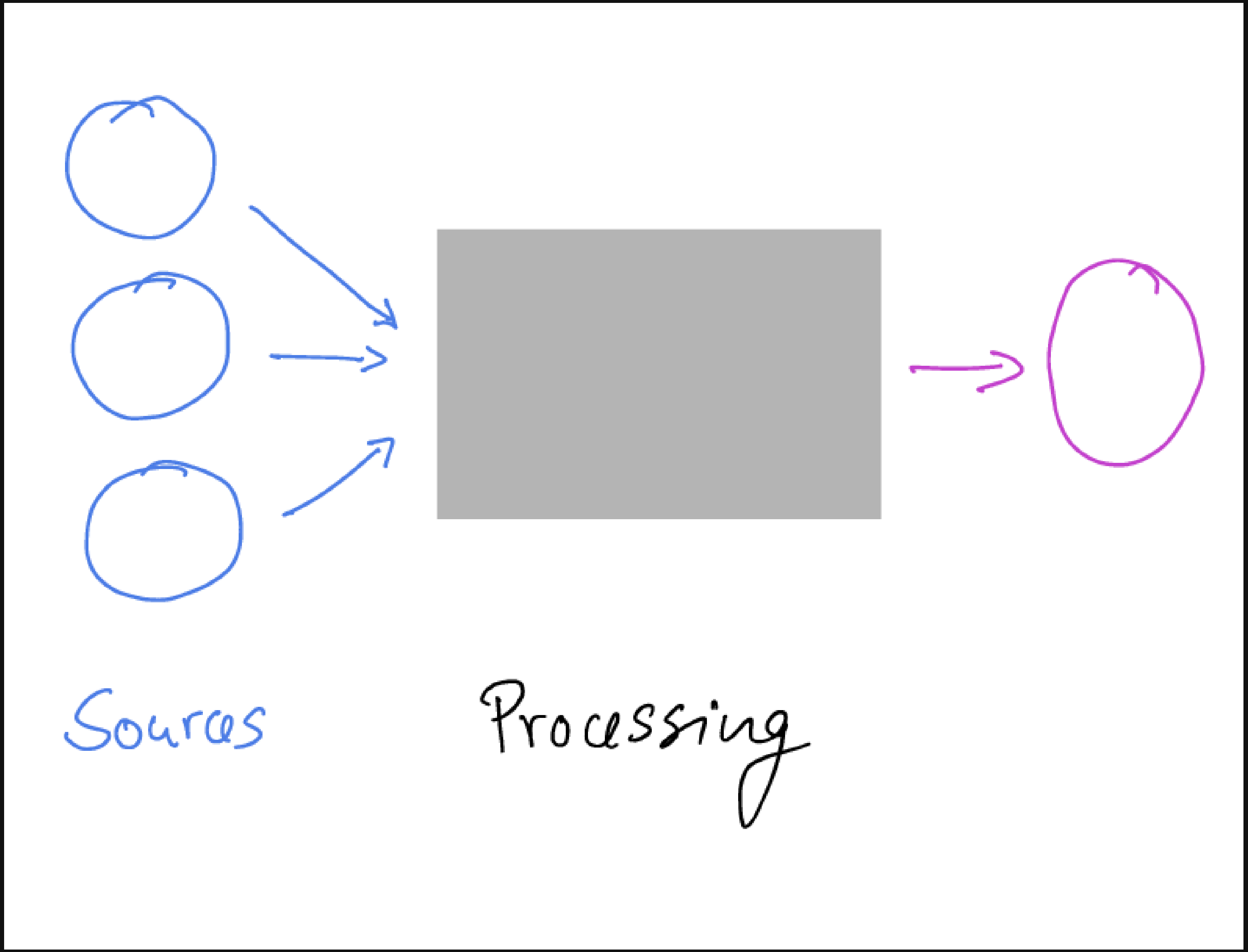
- Of released data – ability to reproduce downstream uses
- Of source data – ability to reproduce released data



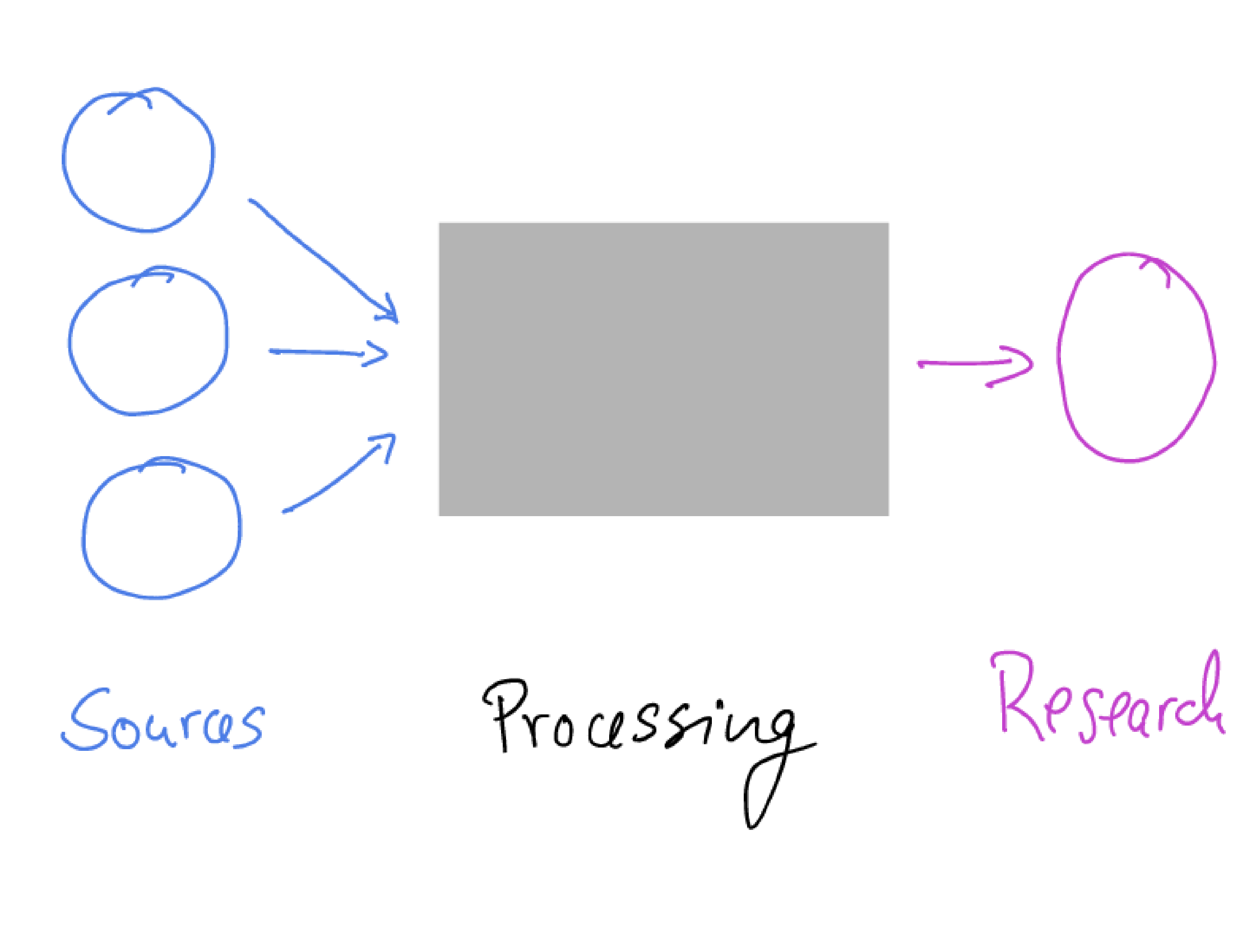
The analogy



The analogy



The analogy



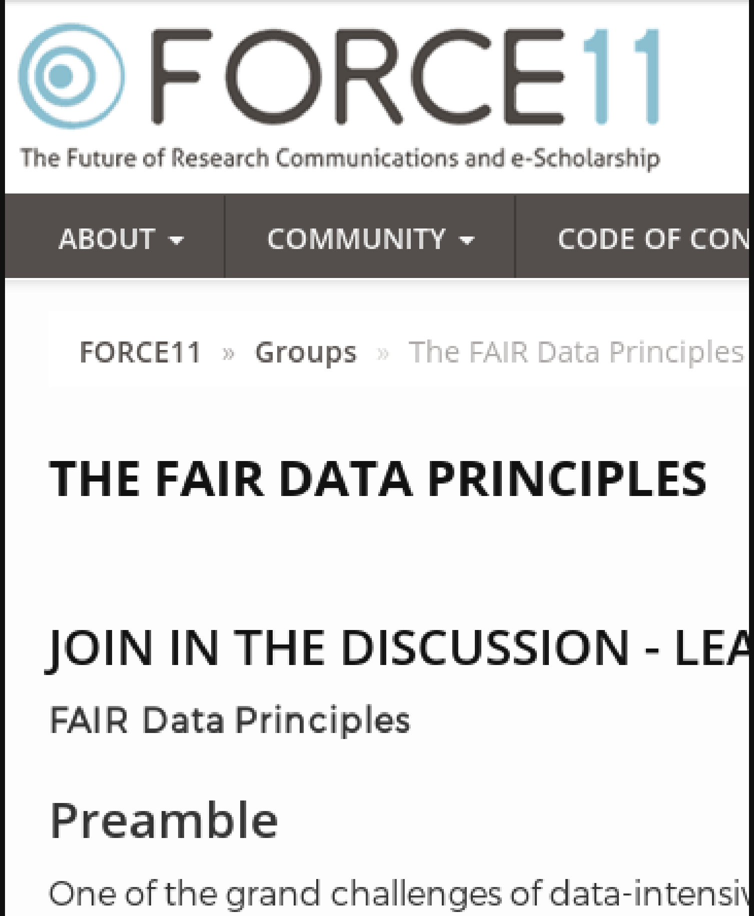
Some principles from the academic world

Which are starting to be infused into the federal system

FAIR Principles

FAIR:

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable



FORCE11 » Groups » The FAIR Data Principles

THE FAIR DATA PRINCIPLES

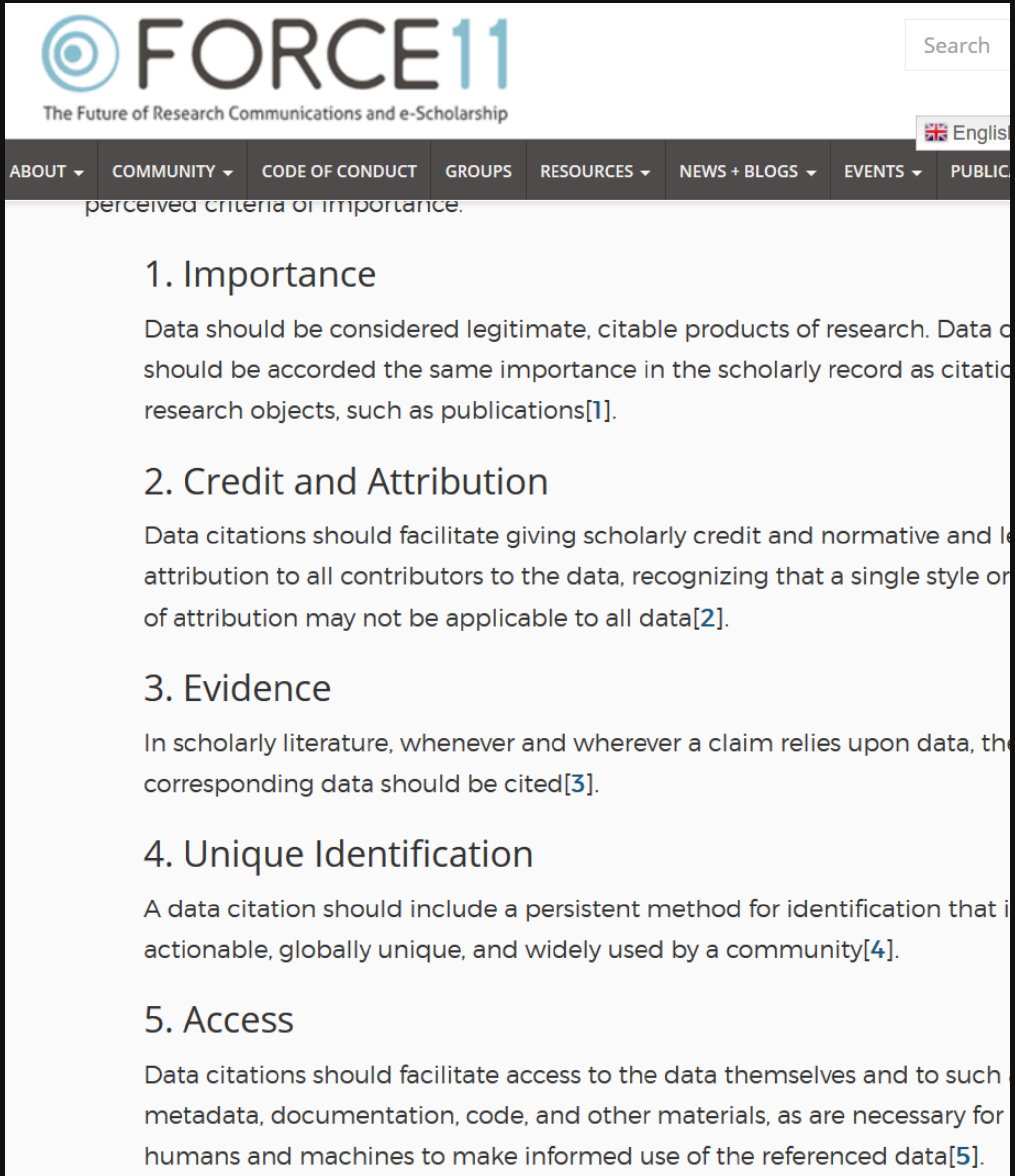
JOIN IN THE DISCUSSION - LEA

FAIR Data Principles

Preamble

One of the grand challenges of data-intensiv

Data Citation Principles



perceived criteria of importance.

1. Importance

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

2. Credit and Attribution

Data citations should facilitate giving scholarly credit and normative and l 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, th 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 metadata, documentation, code, and other materials, as are necessary for humans and machines to make informed use of the referenced data[5].

To make it **findable**,

DC1

Data Citation Principles

7

An example from ERS

USDA Economic Research Service
U.S. DEPARTMENT OF AGRICULTURE

Economic Research Service

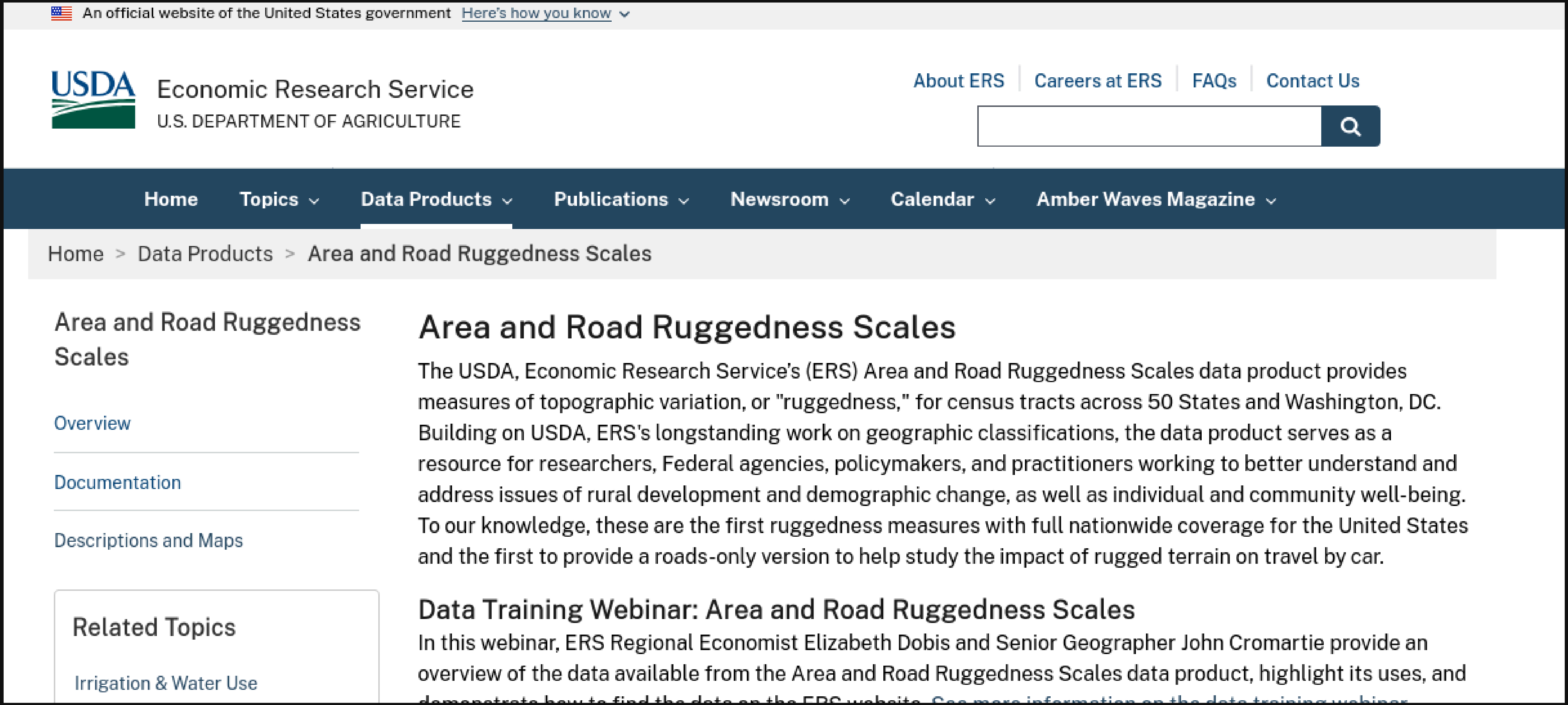
Economic Research Report Number 322

August 2023

Characterizing Rugged Terrain in the United States

Elizabeth A. Dobis, John Cromartie, Ryan Williams, and Kyle Reed

Website



Area and Road Ruggedness Scales

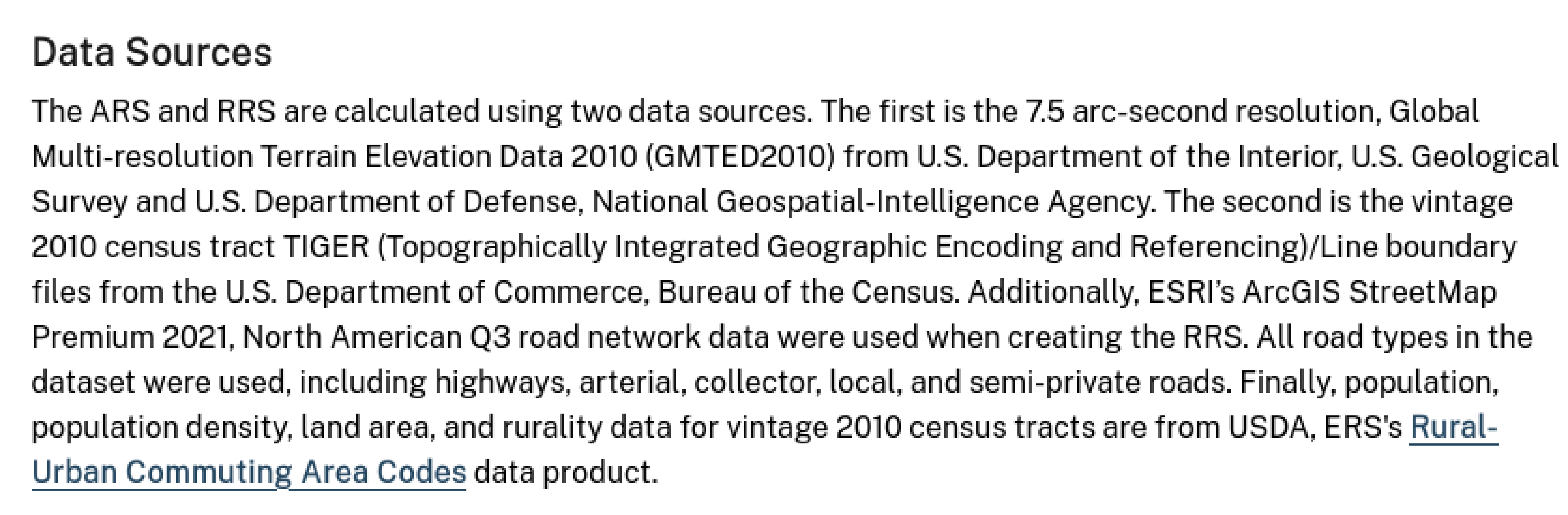
The USDA, Economic Research Service's (ERS) Area and Road Ruggedness Scales data product provides measures of topographic variation, or "ruggedness," for census tracts across 50 States and Washington, DC. Building on USDA, ERS's longstanding work on geographic classifications, the data product serves as a resource for researchers, Federal agencies, policymakers, and practitioners working to better understand and address issues of rural development and demographic change, as well as individual and community well-being. To our knowledge, these are the first ruggedness measures with full nationwide coverage for the United States and the first to provide a roads-only version to help study the impact of rugged terrain on travel by car.

Data Training Webinar: Area and Road Ruggedness Scales

In this webinar, ERS Regional Economist Elizabeth Dobis and Senior Geographer John Cromartie provide an overview of the data available from the Area and Road Ruggedness Scales data product, highlight its uses, and demonstrate how to find the data on the ERS website. See more information on the data training webinar.

Website

Sources

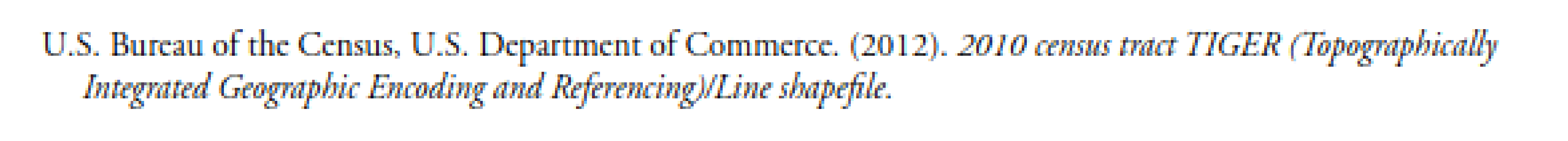


Data Sources

The ARS and RRS are calculated using two data sources. The first is the 7.5 arc-second resolution, Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010) from U.S. Department of the Interior, U.S. Geological Survey and U.S. Department of Defense, National Geospatial-Intelligence Agency. The second is the vintage 2010 census tract TIGER (Topographically Integrated Geographic Encoding and Referencing)/Line boundary files from the U.S. Department of Commerce, Bureau of the Census. Additionally, ESRI's ArcGIS StreetMap Premium 2021, North American Q3 road network data were used when creating the RRS. All road types in the dataset were used, including highways, arterial, collector, local, and semi-private roads. Finally, population, population density, land area, and rurality data for vintage 2010 census tracts are from USDA, ERS's Rural-Urban Commuting Area Codes data product.

High-level description of sources

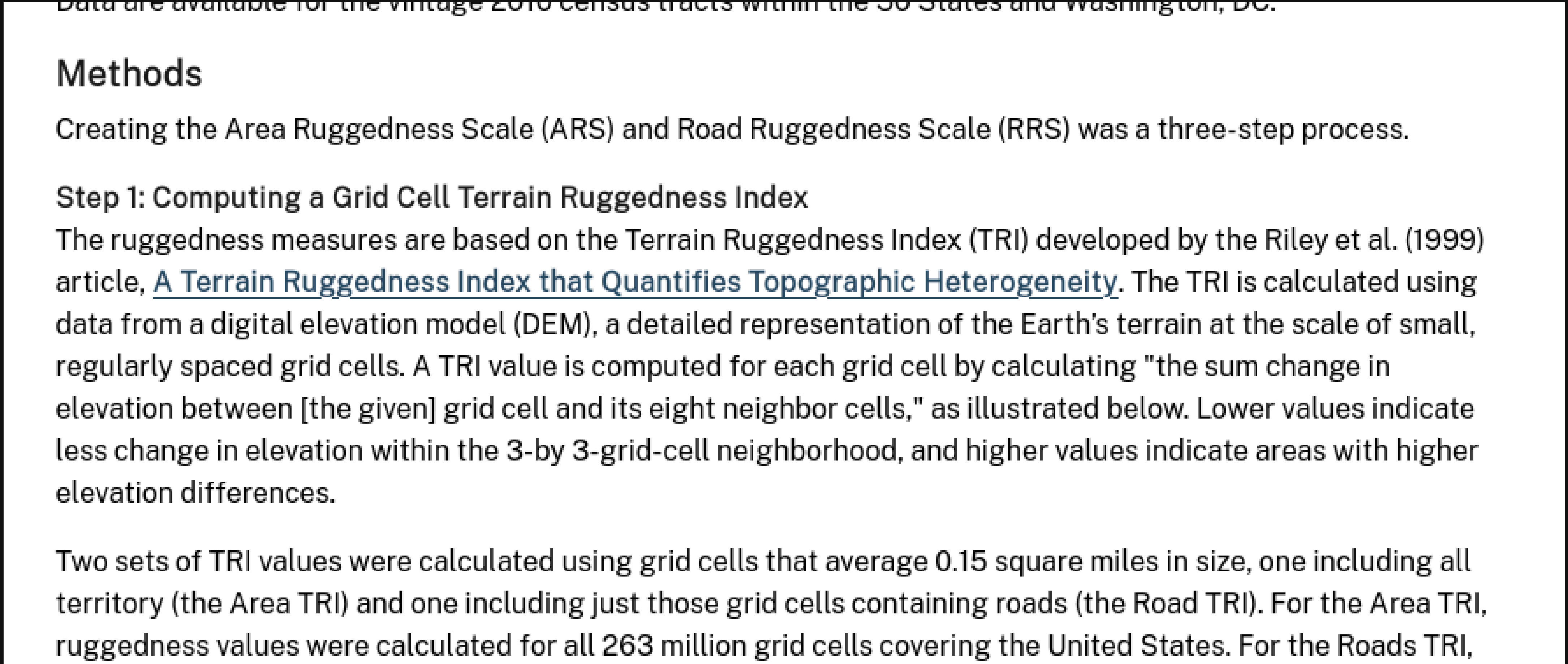
Sources



U.S. Bureau of the Census, U.S. Department of Commerce. (2012). *2010 census tract TIGER (Topographically Integrated Geographic Encoding and Referencing)/Line shapefile.*

Sources are cited!

Methods



Data are available for the vintage 2010 census tracts within the 50 States and Washington, DC.

Methods

Creating the Area Ruggedness Scale (ARS) and Road Ruggedness Scale (RRS) was a three-step process.

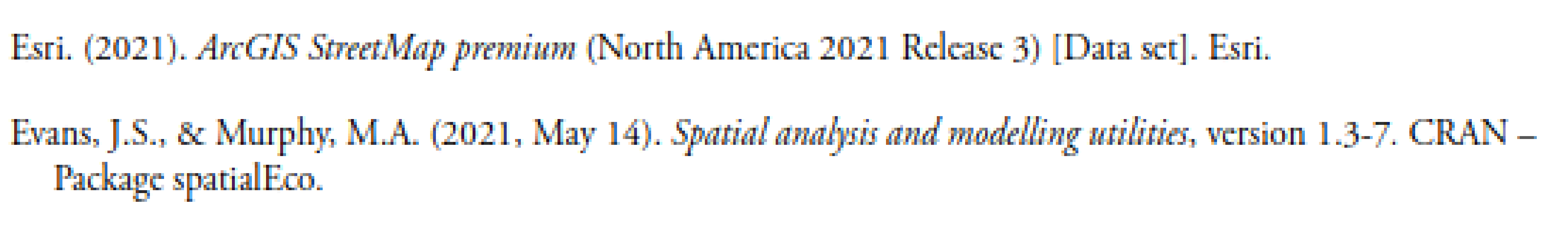
Step 1: Computing a Grid Cell Terrain Ruggedness Index

The ruggedness measures are based on the Terrain Ruggedness Index (TRI) developed by the Riley et al. (1999) article, A Terrain Ruggedness Index that Quantifies Topographic Heterogeneity. The TRI is calculated using data from a digital elevation model (DEM), a detailed representation of the Earth's terrain at the scale of small, regularly spaced grid cells. A TRI value is computed for each grid cell by calculating "the sum change in elevation between [the given] grid cell and its eight neighbor cells," as illustrated below. Lower values indicate less change in elevation within the 3-by 3-grid-cell neighborhood, and higher values indicate areas with higher elevation differences.

Two sets of TRI values were calculated using grid cells that average 0.15 square miles in size, one including all territory (the Area TRI) and one including just those grid cells containing roads (the Road TRI). For the Area TRI, ruggedness values were calculated for all 263 million grid cells covering the United States. For the Roads TRI,

High-level description of methods, but no (obvious) code

Methods

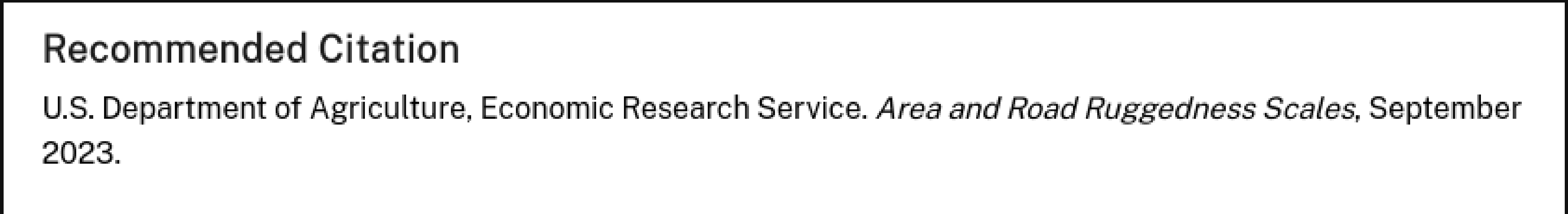


Esri. (2021). *ArcGIS StreetMap premium* (North America 2021 Release 3) [Data set]. Esri.

Evans, J.S., & Murphy, M.A. (2021, May 14). *Spatial analysis and modelling utilities*, version 1.3-7. CRAN – Package spatialEco.

Some methods - R code - is cited

Own citation?

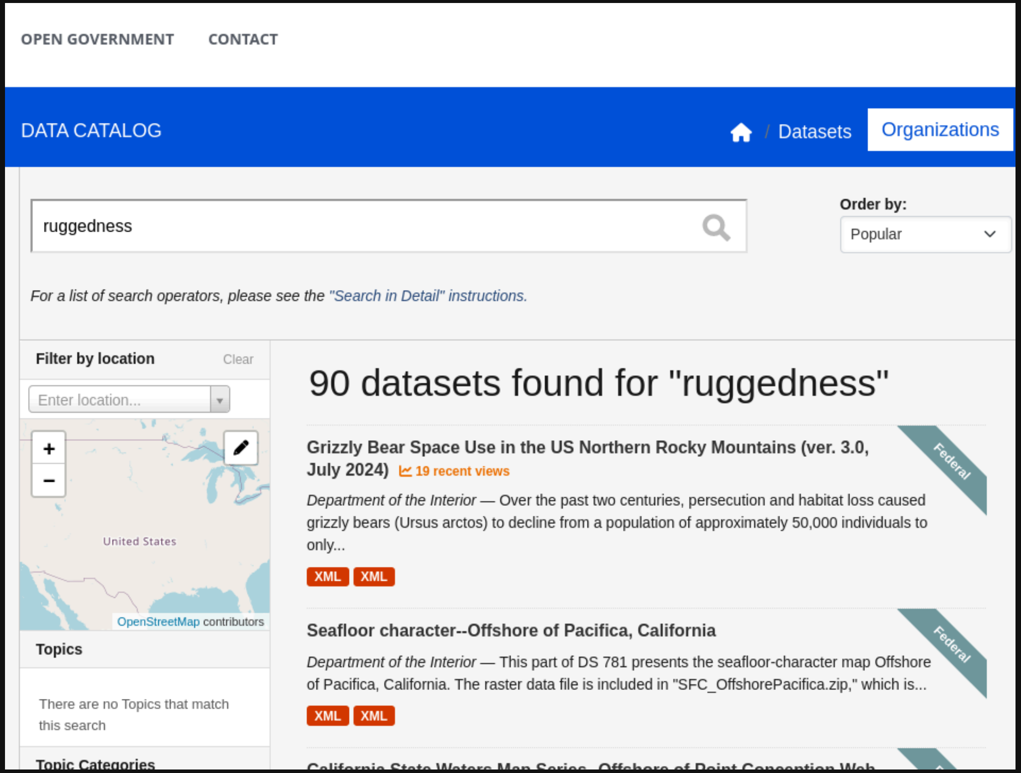


Recommended Citation

U.S. Department of Agriculture, Economic Research Service. *Area and Road Ruggedness Scales*, September 2023.

Own citation does not include a URL

Findability?



OPEN GOVERNMENT CONTACT

DATA CATALOG

Datasets / Organizations

ruggedness

Order by: Popular

For a list of search operators, please see the "Search in Detail" instructions.

Filter by location Clear

Enter location...

Topics

There are no Topics that match this search

Topic Categories

90 datasets found for "ruggedness"

Grizzly Bear Space Use in the US Northern Rocky Mountains (ver. 3.0, July 2024) 19 recent views

Federal

Department of the Interior — Over the past two centuries, persecution and habitat loss caused grizzly bears (Ursus arctos) to decline from a population of approximately 50,000 individuals to only...

XML XML

Seafloor character--Offshore of Pacifica, California

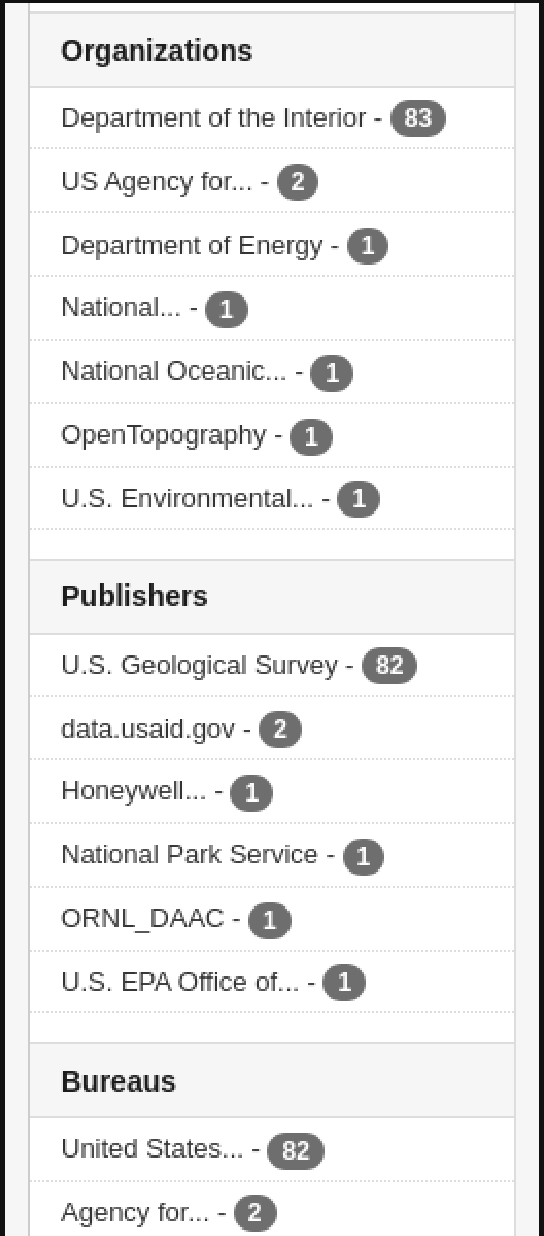
Federal

Department of the Interior — This part of DS 781 presents the seafloor-character map Offshore of Pacifica, California. The raster data file is included in "SFC_OffshorePacifica.zip," which is...

XML XML

California State Waters Map Series--Offshore of Point Conception Web

Data.gov is not great



Organizations

Department of the Interior - 83

US Agency for... - 2

Department of Energy - 1

National... - 1

National Oceanic... - 1

OpenTopography - 1

U.S. Environmental... - 1

Publishers

U.S. Geological Survey - 82

data.usaid.gov - 2

Honeywell... - 1

National Park Service - 1

ORNL_DAAC - 1

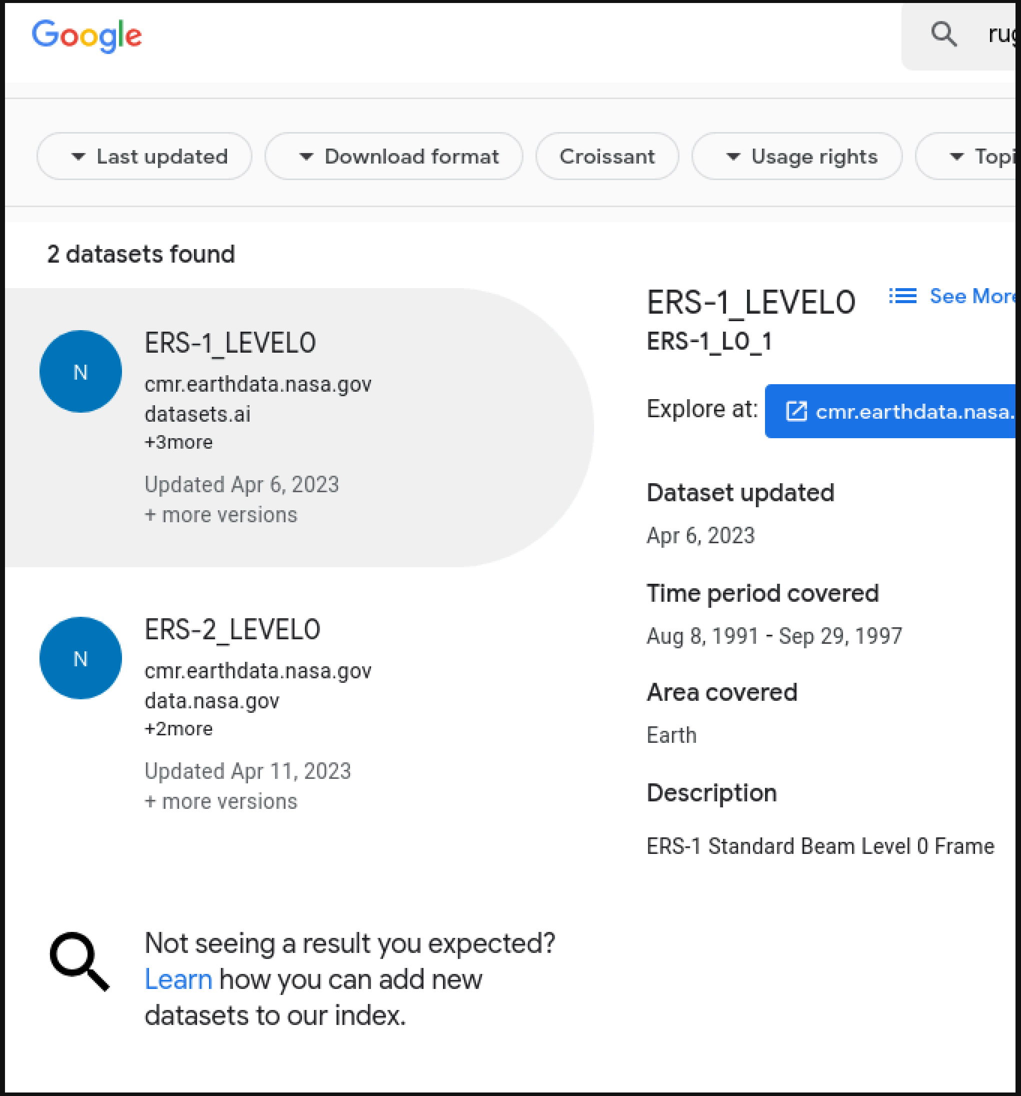
U.S. EPA Office of... - 1

Bureaus

United States... - 82

Agency for... - 2

Findability?



2 datasets found

N ERS-1_LEVEL0
cmr.earthdata.nasa.gov
datasets.ai
+3more
Updated Apr 6, 2023
+ more versions

N ERS-2_LEVEL0
cmr.earthdata.nasa.gov
data.nasa.gov
+2more
Updated Apr 11, 2023
+ more versions

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

ERS-1_LEVEL0 See More
ERS-1_L0_1

Explore at: cmr.earthdata.nasa.

Dataset updated
Apr 6, 2023

Time period covered
Aug 8, 1991 - Sep 29, 1997

Area covered
Earth

Description
ERS-1 Standard Beam Level 0 Frame

Not even close.

Google Dataset Search is worse

Repeatability of downloads?

URL is

https://ers.usda.gov/webdocs/DataFiles/107356/Ruggedne v=6316.8

- What will the 2020 tract-based data URL look like?

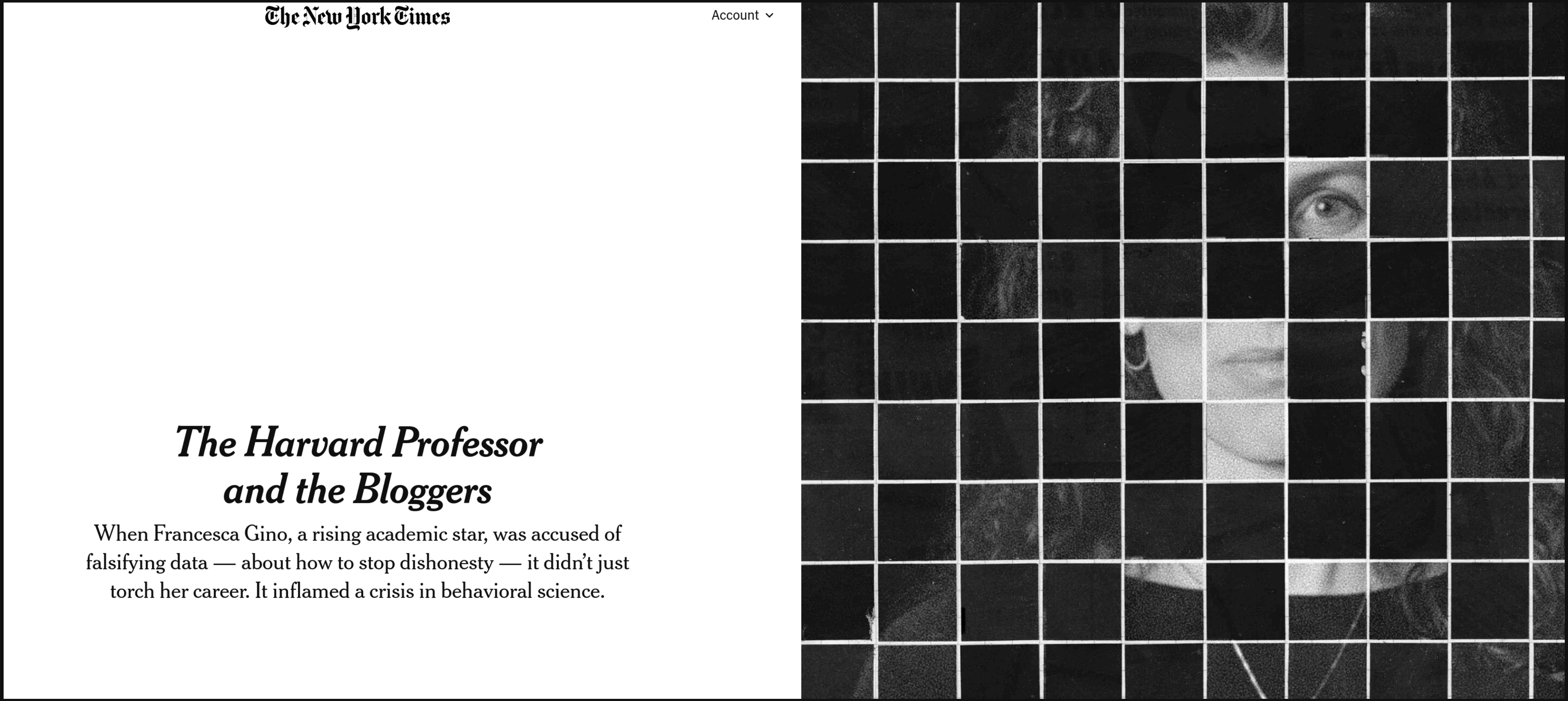
Reproducibility?

- Most of the data inputs seem to be public data, or commercially available (ESRI)
- If the code were provided, others should be able to reproduce the analysis

Opening up technical possibilities

How can we know that a data source is reliably obtained?

Consider the case of Gino



The Harvard Professor and the Bloggers

When Francesca Gino, a rising academic star, was accused of falsifying data — about how to stop dishonesty — it didn't just torch her career. It inflamed a crisis in behavioral science.

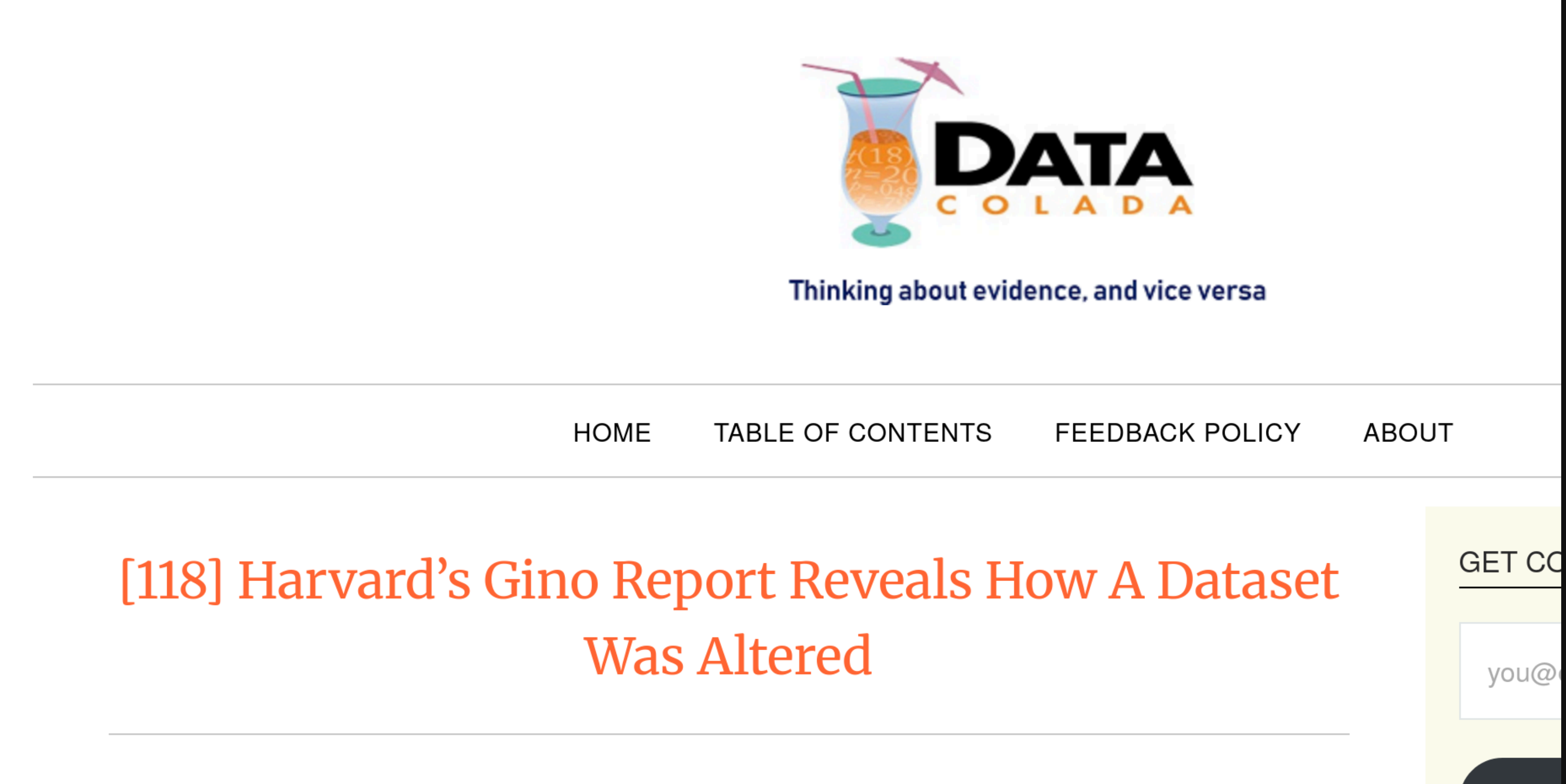
Francesca Gino

The case of Gino

- Francesca Gino was a tenured professor at Harvard Business School, writing on honesty (!)

The case of Gino

- Several articles were investigated by third parties (Data Colada, in particular [9]), and found to be problematic



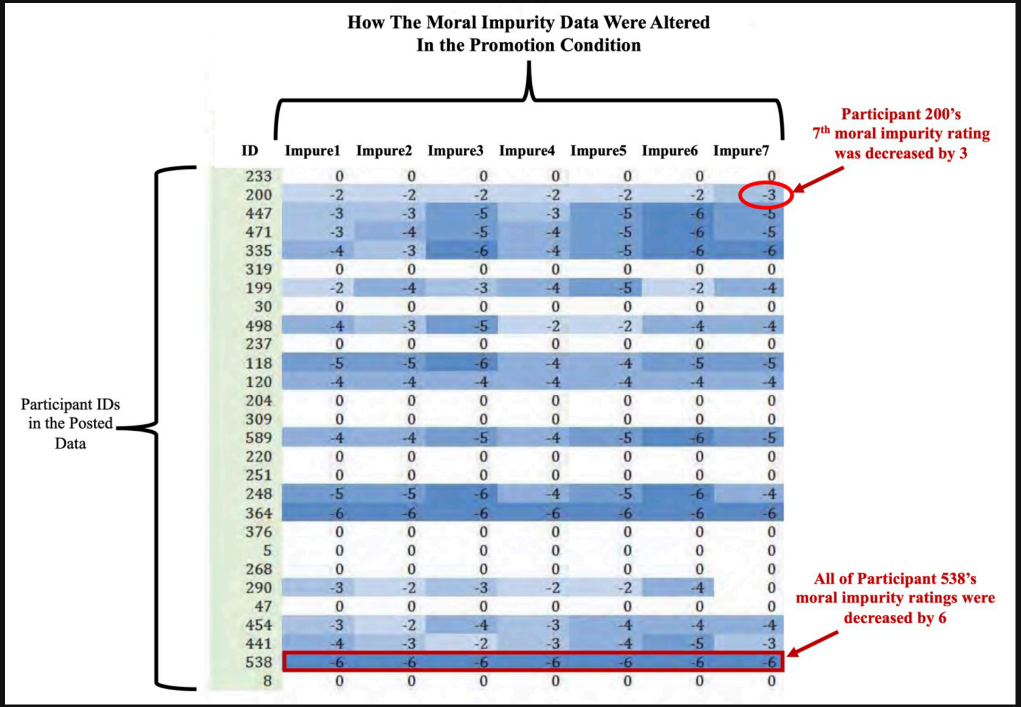
DATA COLADA

Thinking about evidence, and vice versa

[118] Harvard's Gino Report Reveals How A Dataset Was Altered

The case of Gino

- At least one of them had manipulated data **AFTER** it had been collected, **BEFORE** it had been analyzed.



How The Moral Impurity Data Were Altered In the Promotion Condition

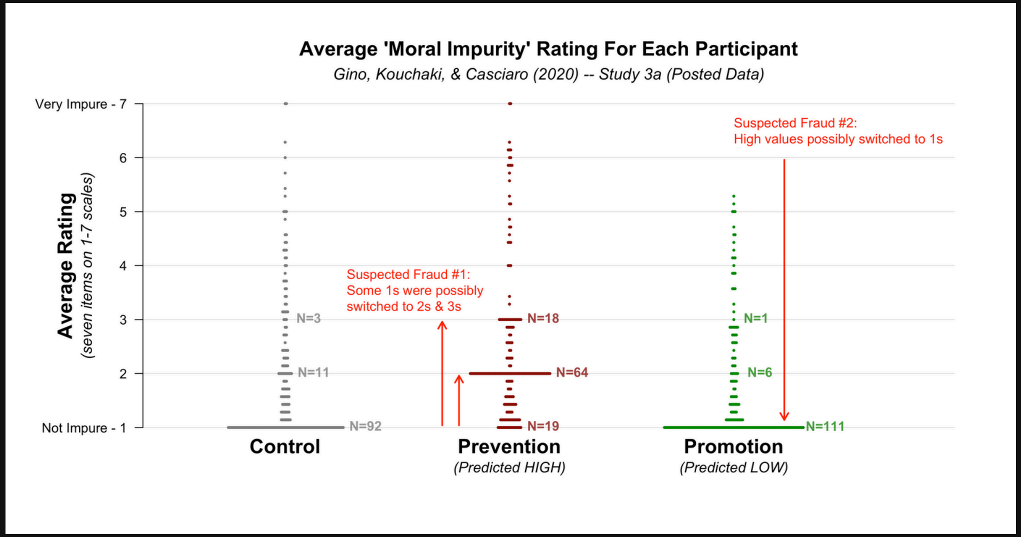
Participant IDs in the Posted Data

ID	Impure1	Impure2	Impure3	Impure4	Impure5	Impure6	Impure7
233	0	0	0	0	0	0	0
200	-2	-2	-2	-2	-2	-2	-3
447	-3	-3	-5	-3	-5	-6	-5
471	-3	-4	-5	-4	-5	-6	-5
335	-4	-3	-6	-4	-5	-6	-6
319	0	0	0	0	0	0	0
199	-2	-4	-3	-4	-5	-2	-4
30	0	0	0	0	0	0	0
498	-4	-3	-5	-2	-2	-4	-4
237	0	0	0	0	0	0	0
118	-5	-5	-6	-4	-4	-5	-5
120	-4	-4	-4	-4	-4	-4	-4
204	0	0	0	0	0	0	0
309	0	0	0	0	0	0	0
589	-4	-4	-5	-4	-5	-6	-5
220	0	0	0	0	0	0	0
251	0	0	0	0	0	0	0
248	-5	-5	-6	-4	-5	-6	-4
364	-6	-6	-6	-6	-6	-6	-6
376	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
268	0	0	0	0	0	0	0
290	-3	-2	-3	-2	-2	-4	0
47	0	0	0	0	0	0	0
454	-3	-2	-4	-3	-4	-4	-4
441	-4	-3	-2	-3	-4	-5	-3
538	-6	-6	-6	-6	-6	-6	-6
8	0	0	0	0	0	0	0

Participant 200's 7th moral impurity rating was decreased by 3

All of Participant 538's moral impurity ratings were decreased by 6

Data manipulation

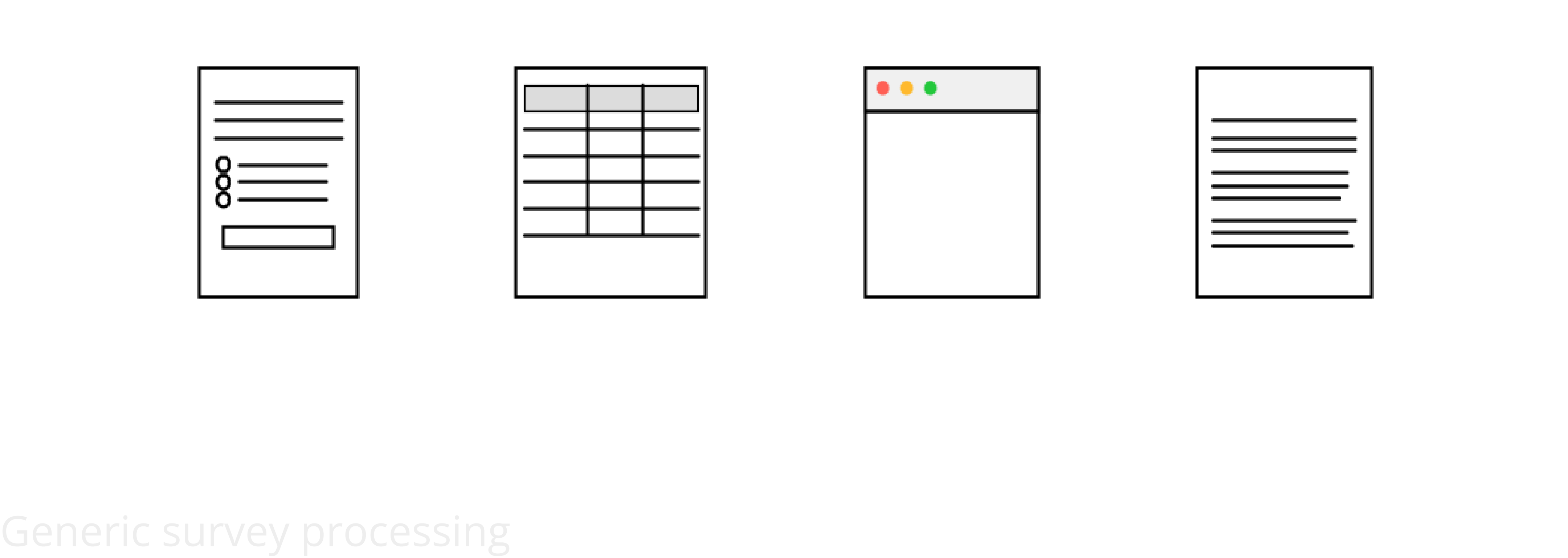


Average 'Moral Impurity' Rating For Each Participant

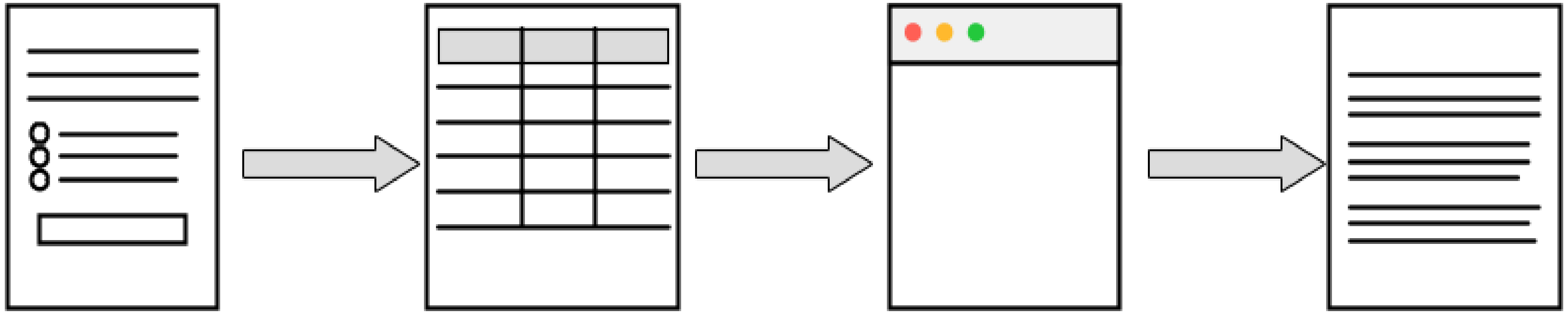
Gino, Kouchaki, & Casciaro (2020) -- Study 3a (Posted Data)

Results of manipulation

Generic survey processing

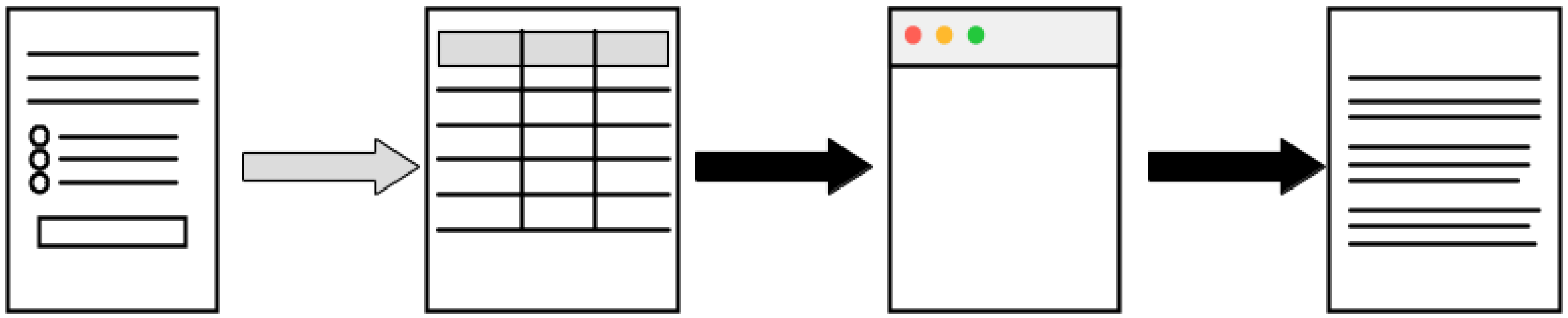


Generic survey processing



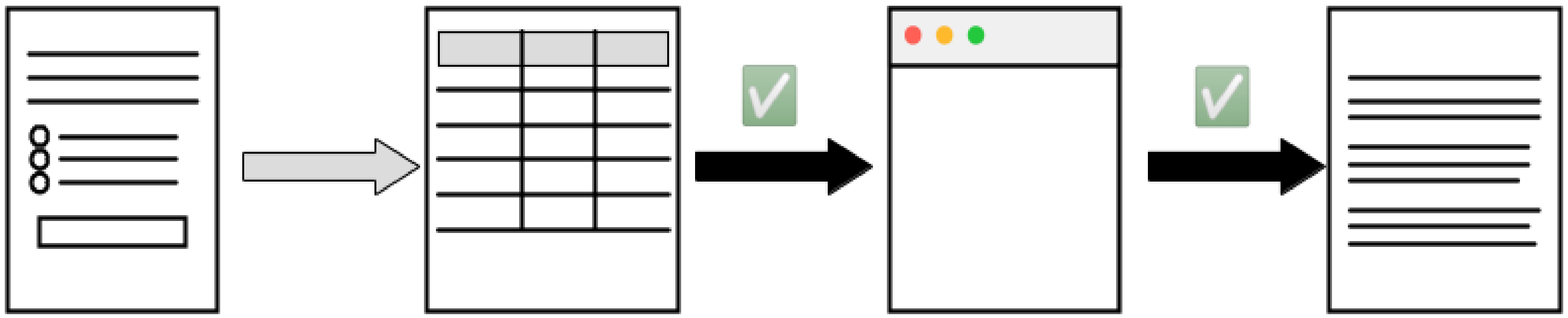
Generic survey processing

Requiring transparency in academia



Generic survey processing

Verifying transparency in academia



Generic survey processing

Verification by journals

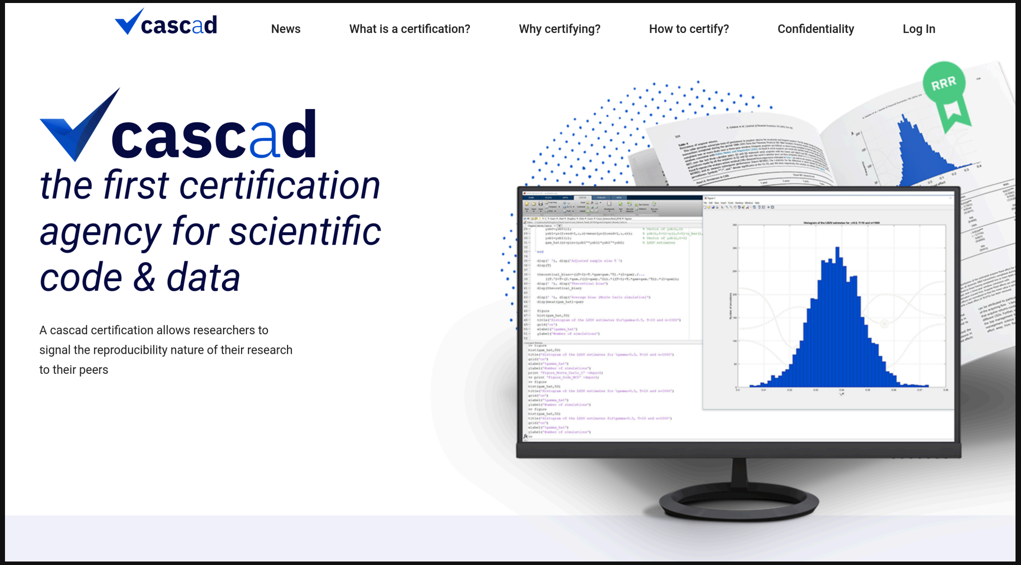
- **Provision** (publication of materials) provides transparency
- **Verification** (running the analysis again - computational reproducibility) compensates for ***mistrust***/***absence of trust***

Which journals again

- American Economic Association (8)
- Econometric Society (3)
- Canadian Journal of Economics (1)
- Royal Economic Society (2)
- Western Economic Association International (1)
- European Economic Association (1)
- Review of Economic Studies (1)
- Journal of the European Economic Association (1)
- Journal of Political Economy (3)
- American Journal of Political Science (1)
- American Political Science Review (1)

Verification by others

- Pre-publication: cascad



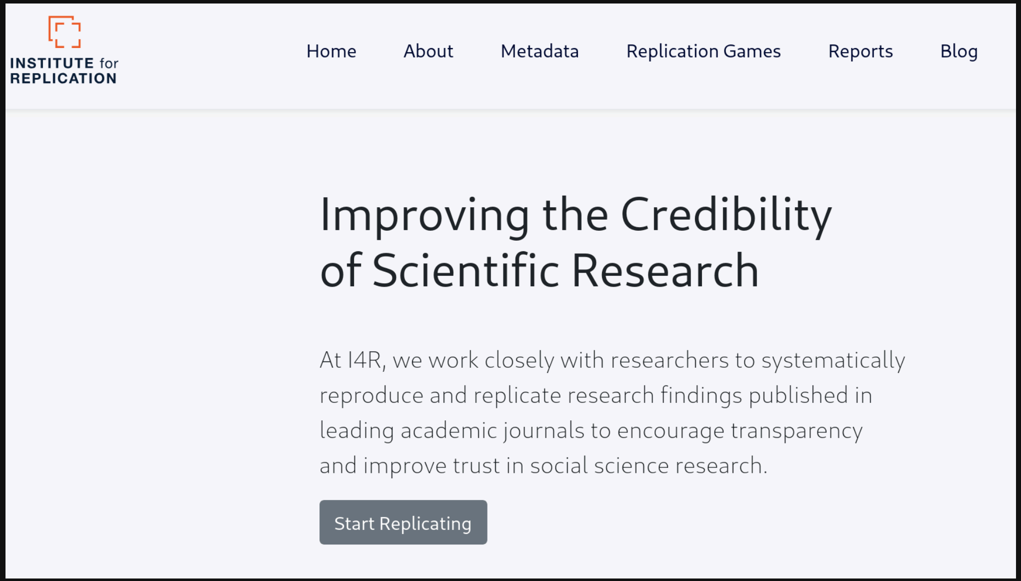
cascad

the first certification agency for scientific code & data

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

cascad

- Post-publication: Data Colada, Institute for Replication



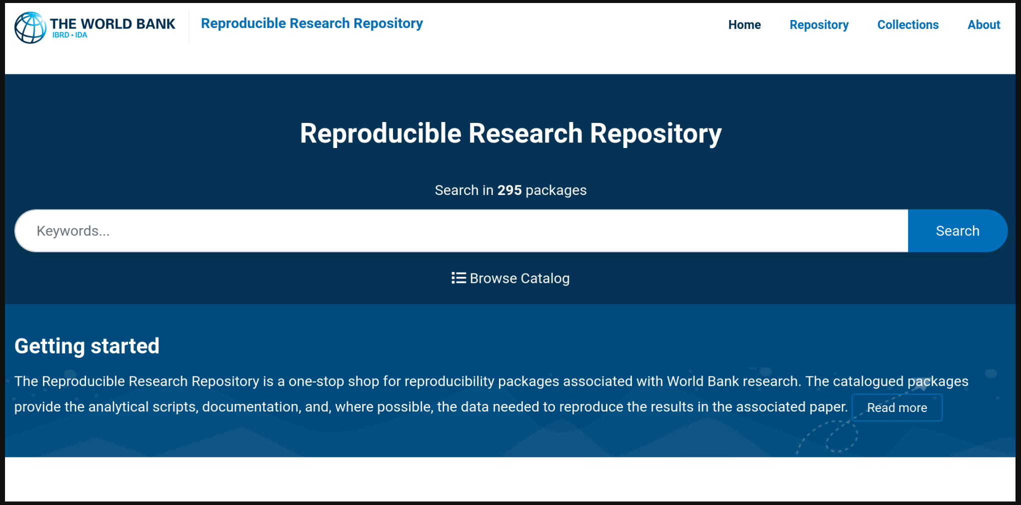
Improving the Credibility of Scientific Research

At I4R, we work closely with researchers to systematically reproduce and replicate research findings published in leading academic journals to encourage transparency and improve trust in social science research.

I4R

Verification by institutions

- World Bank[10]



Reproducible Research Repository

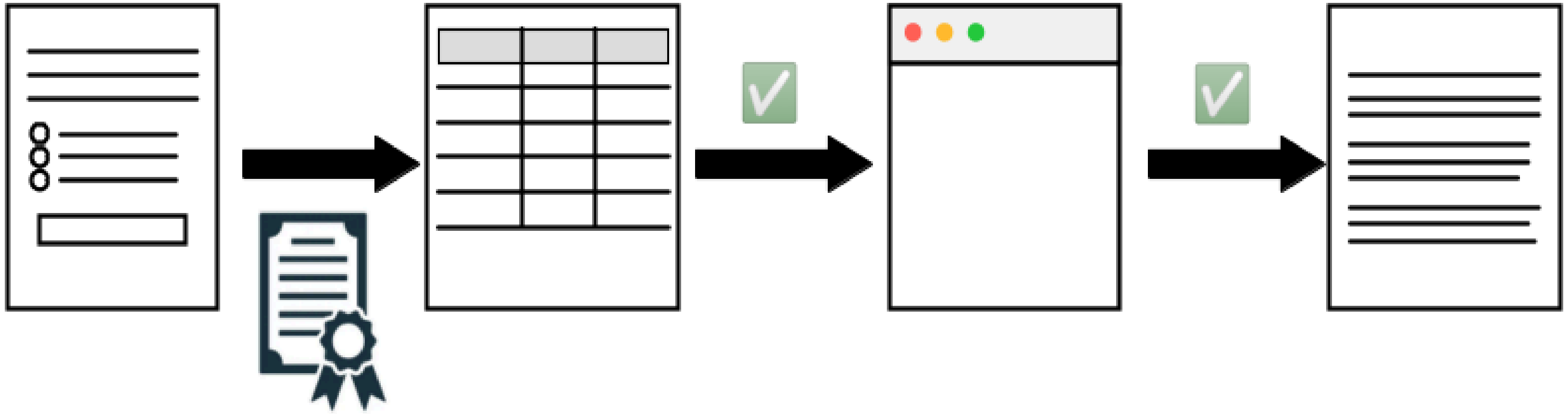
Search in **295** packages

Getting started

The Reproducible Research Repository is a one-stop shop for reproducibility packages associated with World Bank research. The catalogued packages provide the analytical scripts, documentation, and, where possible, the data needed to reproduce the results in the associated paper.

World Bank RRR

Taking it a step further



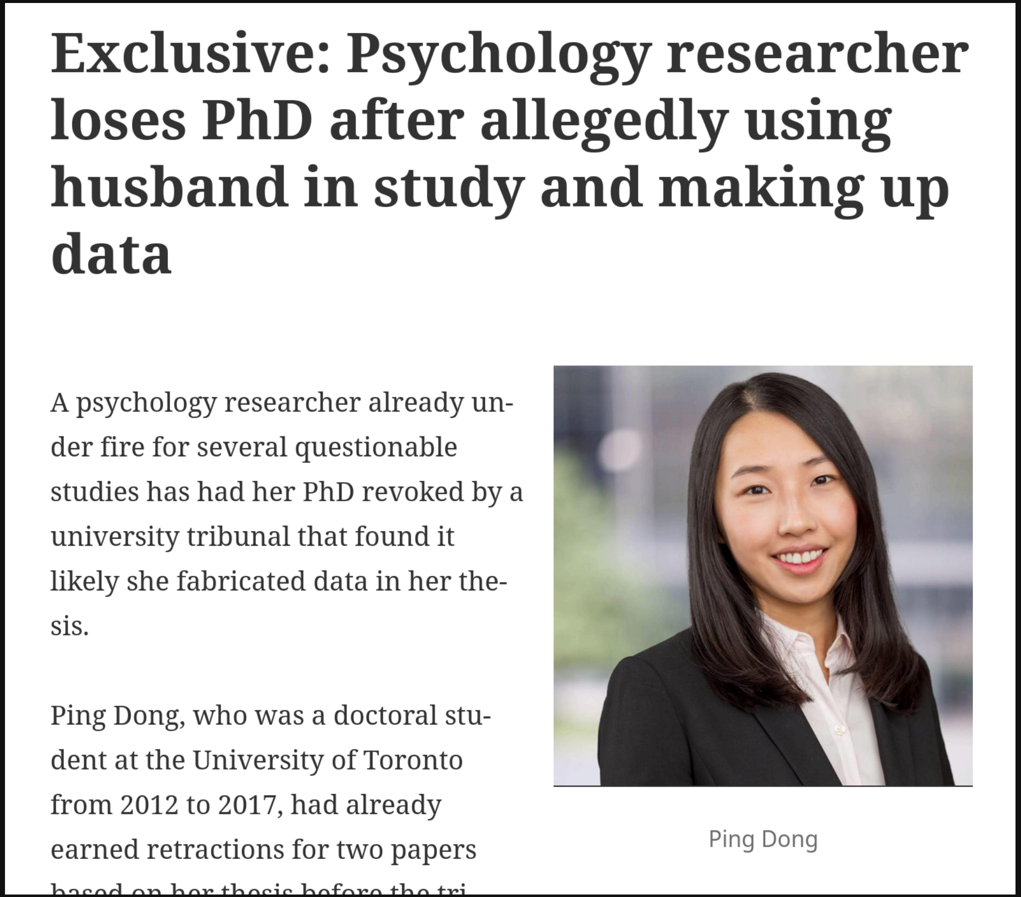
Survey flow

Taking it a step further

- Has been discussed by authors behind Data Colada
- Survey tool provider (Qualtrics, etc.) exports data, posts checksum
- Survey tool provider exports data only to institution directly into trusted repository, researchers obtain data from there (with privacy protections)

Does not prevent all fraud

Toronto researcher loses Ph.D.



Exclusive: Psychology researcher loses PhD after allegedly using husband in study and making up data

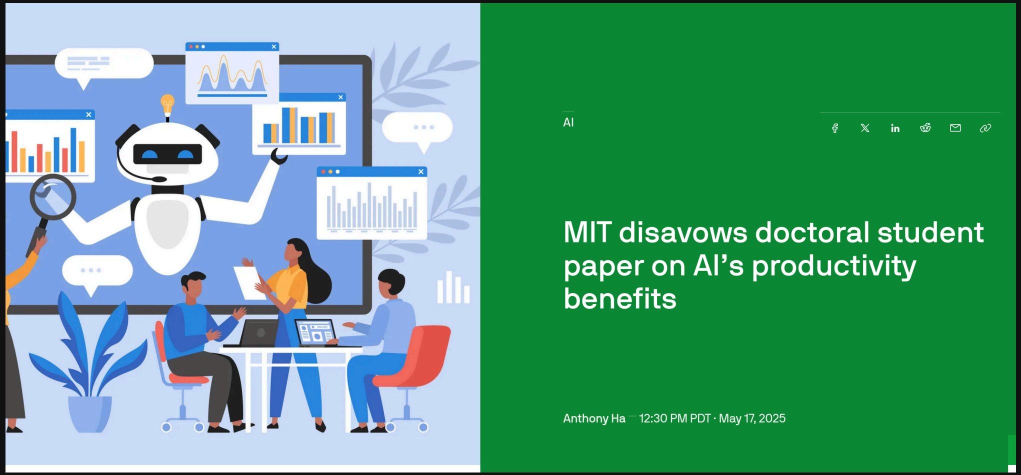
A psychology researcher already under fire for several questionable studies has had her PhD revoked by a university tribunal that found it likely she fabricated data in her thesis.

Ping Dong, who was a doctoral student at the University of Toronto from 2012 to 2017, had already earned retractions for two papers based on her thesis before the tri-

Ping Dong

Toronto case

MIT student makes up firm data



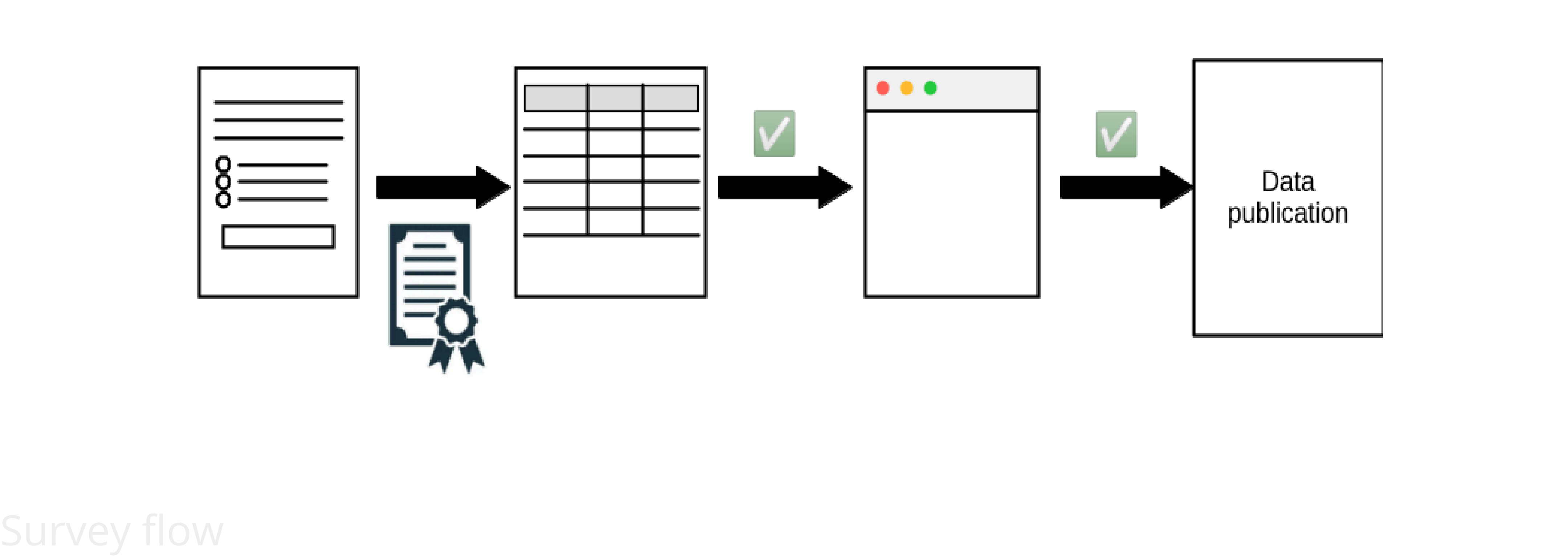
AI

MIT disavows doctoral student paper on AI's productivity benefits

Anthony Ha · 12:30 PM PDT · May 17, 2025

MIT case

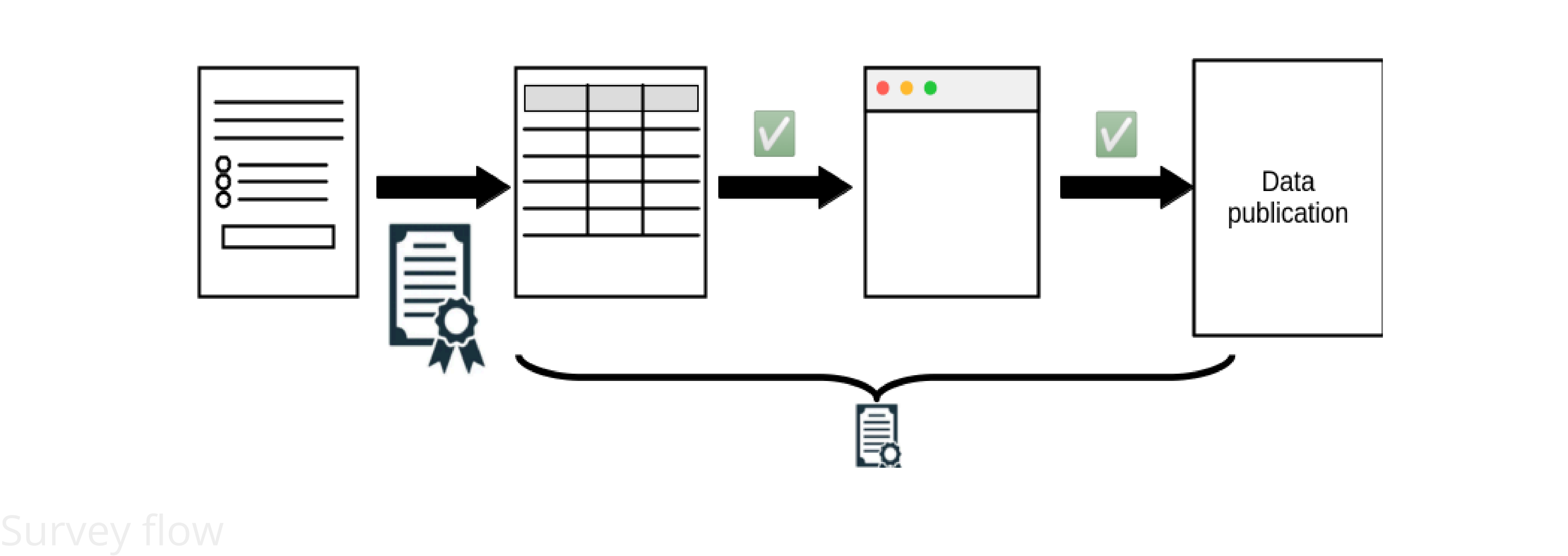
Can Assurances be created for Statistical Agencies?



Challenges for statistical agencies

- Documenting full transparency is hard
- Complex and legacy survey tools make the process harder
- Presence of (legitimate!) manual edits is an issue
- Production processes are long and complex
- Much of the code base is not open source

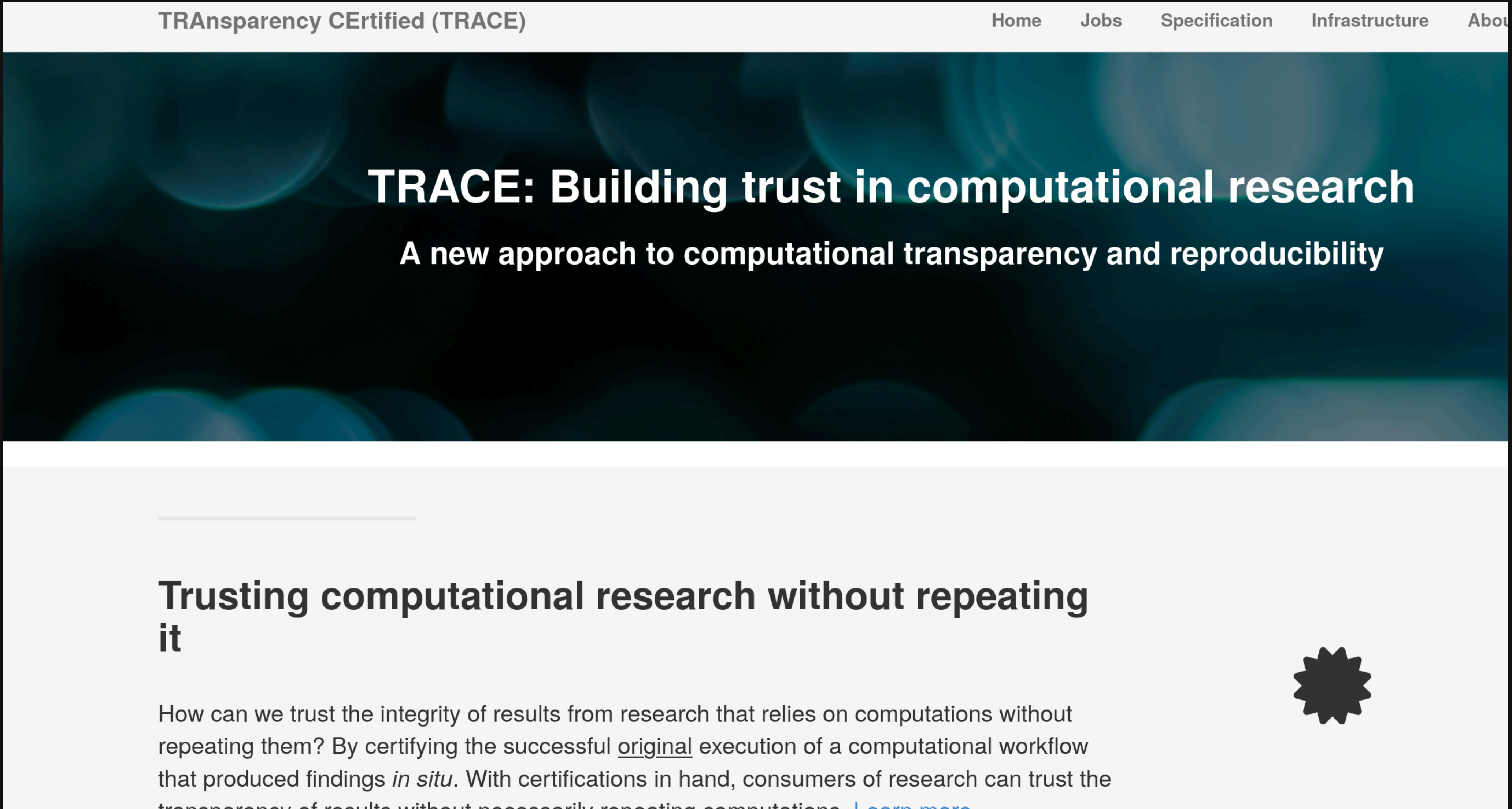
How to document the full process?



Survey flow

A sketch: Transparency Certified

https://transparency-certified.github.io/



TRAnsparency CErtified (TRACE)

TRACE: Building trust in computational research

A new approach to computational transparency and reproducibility

Trusting computational research without repeating it

How can we trust the integrity of results from research that relies on computations without repeating them? By certifying the successful original execution of a computational workflow that produced findings *in situ*. With certifications in hand, consumers of research can trust the transparency of results without necessarily repeating computations. Learn more

Work in progress

- Working with cascad, several INEXDA members, and others
- Relying on external certification of data inputs (data catalogs with metadata, checksums)

Wrapping it all up

What is the state of reproducibility and transparency in academic economics?

- An increasing number of journals are not just **requiring** complete data, code, and transparent description, but also **verifying** that the code and data are correct.
- At the AEA: since 2019, reviewed around **3000** articles, ran code for about **2/3** of them.

What are the benefits of reproducibility and transparency?

- Greater **trust** in the results
- Greater **ease** of building on results
- Greater **transparency** of the process, but also of the **provenance**

Increasing broad consensus in academia

- FAIR principles
- Data Citation Principles
- Computational Reproducibility

What are the implications for statistical agencies?

Producers of statistical products

- May want to provide greater transparency into the **production process**.
- May need to do more for **long-term, unbiased preservation** of input data, output products, and code/software to link the two.
- May want to start with the low-hanging fruit: ***dashboards and fully public processes***.

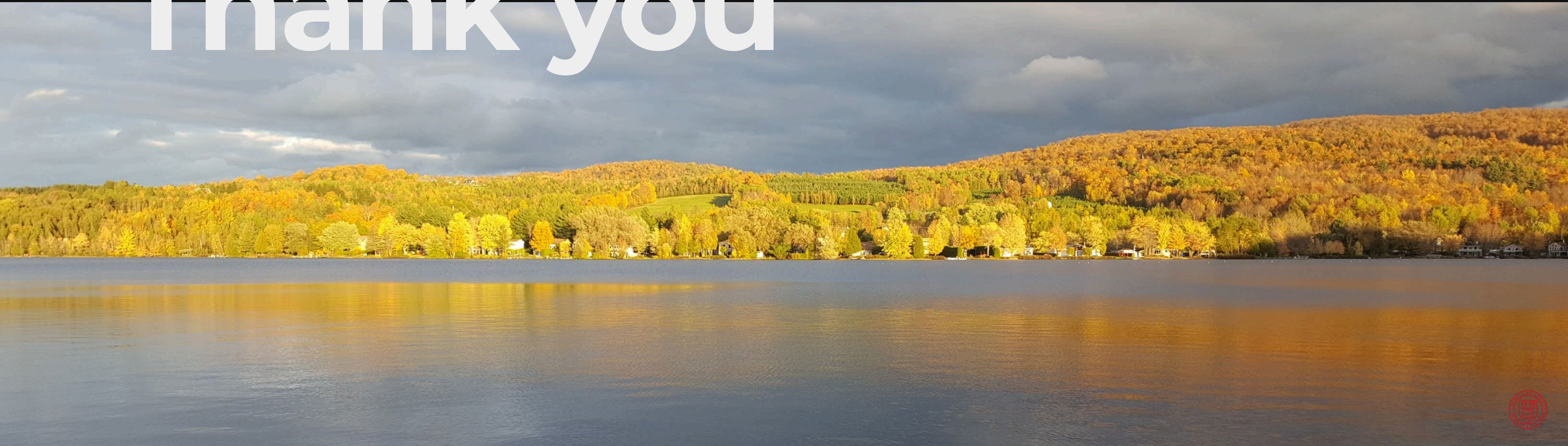
Coherence with stated principles

The emerging consensus is fully in line with the **decades-strong principles** of statistical agencies:

Greater trust by the public?

- **Transparency** should be correlated with greater trust in the work of the statistical agencies
- But: Transparency can also lead to vulnerability through misinterpretation (no panacea)

Thank you



One more thing...

That confidential code thing...

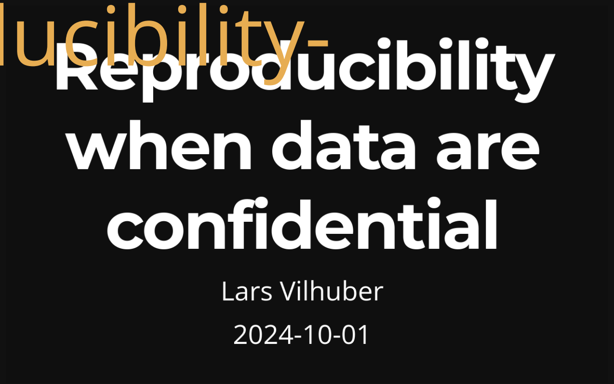
- IRS variable names
- File paths b/c your IT department said so
- Use of confidential **data** in code (`if name="Lars" then confid=2`)

Solution

Don’t do that.

Solution

labordynamicsinstitute.github.io/reproducibility-confidential/



Reproducibility when data are confidential

Lars Vilhuber

2024-10-01

Also here.

Appendix

Secrets in the code

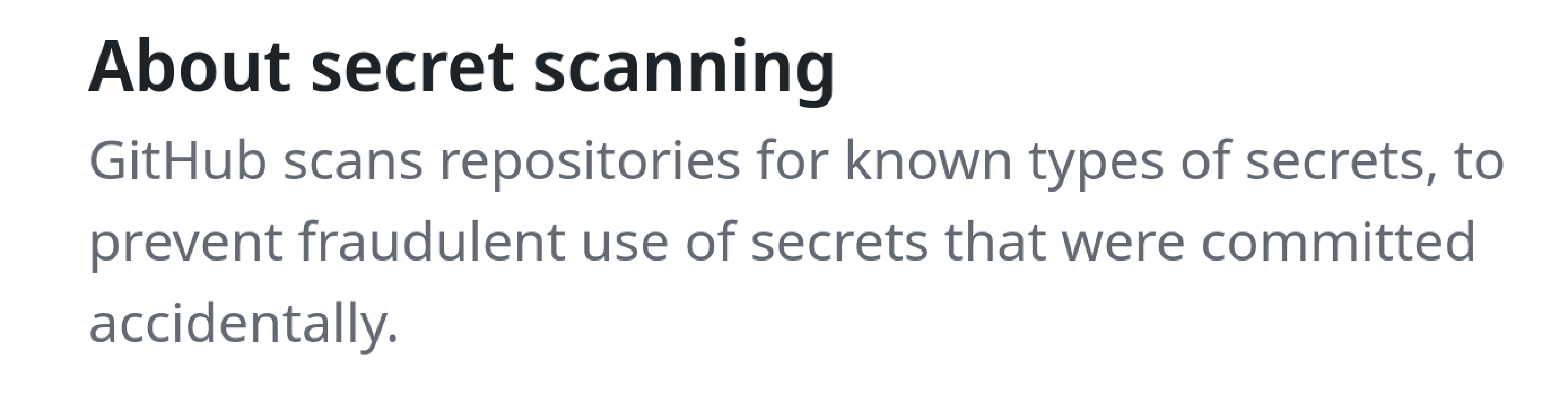
What are secrets?

- API keys
- Login credentials for data access
- File paths (FSRDC!)
- Variable names (IRS!)

Standard practice

Store secrets in environment variables or files that are not published.

Some services are serious about this



About secret scanning

GitHub scans repositories for known types of secrets, to prevent fraudulent use of secrets that were committed accidentally.

Github secret scanning

Where to store secrets

- **environment variables**
- “dot-env” files (Python), “Renviron” files (R)
- or some other clearly identified file in the project or home directory

Environment variables

Typed interactively (here for Linux and Mac)

```
MYSECRET="dfad89ald"
CONFDATALOC="/path/to/irs/files"
```

(this is **not** recommended)

Storing these in files

Same syntax used for contents of “dot-env” or “Renviron” files, and in fact `bash` or `zsh` startup files (`.bash_profile`, `.zshrc`)

Using In R

Edit `.Renviron` (note the dot!) files:

```r
# Edit global (personal) Renviron
usethis::edit_r_environ()
# You can also consider creating project-specific settings:
usethis::edit_r_environ(scope = "project")
```

Use the variables defined in `.Renviron`:

```r
mysecret <- Sys.getenv('MYSECRET')
```

Using In Python

Loading regular environment variables:

```
import os
mysecret = os.getenv("MYSECRET")   # will load environment variables
```

Loading with `dotenv`

```
from dotenv import load_dotenv
load_dotenv()   # take environment variables from project .env.
mysecret = os.getenv("MYSECRET")   # will load environment variables
```

Using in Stata

Yes, this also works in Stata

```
// load from environment
global mysecret : env MYSECRET
display "$mysecret"  // don't actually do this in code
```

and via (what else) a user-written package for loading from files:

```
net install doenv, from(https://github.com/vikjam/doenv/raw/master/)
doenv using ".env"
global mysecret "`r(MYSECRET)'"
display "$mysecret"
```

Simplest solution

```
//============ non-confidential parameters =========
include "config.do"

//============ confidential parameters =============
capture confirm file "$code/confidential/confparms.do"
if _rc == 0 {
    // file exists
    include "$code/confidential/confparms.do"
} else {
    di in red "No confidential parameters found"
}
//============ end confidential parameters =========
```

Confidential code?

What is confidential code, you say?

- In the United States, some **variables on IRS databases** are considered super-top-secret. So you can’t name that-variable-that-you-filled-out-on-your-Form-1040 in your analysis code of same data. (They are often referred to in jargon as “Title 26 variables”).

What is confidential code, you say?

- Your code contains the **random seed you used to anonymize** the sensitive identifiers. This might allow to reverse-engineer the anonymization, and is not a good idea to publish.

What is confidential code, you say?

- You used a **look-up table hard-coded** in your Stata code to anonymize the sensitive identifiers (`replace anoncounty=1 if county="Tompkins, NY"`).

A **really bad idea**, but yes, you probably want to hide that.

What is confidential code, you say?

- Your IT specialist or disclosure officer thinks publishing the **exact path** to your copy of the confidential 2010 Census data, e.g., “/data/census/2010”, is a security risk and refuses to let that code through.

What is confidential code, you say?

- You have adhered to disclosure rules, but for some reason, the precise minimum cell size is a confidential parameter.

What is confidential code, you say?

So whether reasonable or not, **this is an issue**. How do you do that, without messing up the code, or spending hours redacting your code?

Example

- This will serve as an example. None of this is specific to Stata, and the solutions for R, Python, Julia, Matlab, etc. are all quite similar.
- Assume that variables `q2f` and `q3e` are considered confidential by some rule, and that the minimum cell size `10` is also confidential.

```
set seed 12345
use q2f q3e county using "/data/economic/cmf2012/extract.dta", clear
gen logprofit = log(q2f)
by county: collapse (count)  n=q3e (mean) logprofit
drop if n<10
graph twoway n logprofit
```

Example

Only one line that does not contain “confidential” information.

```
set seed 12345
use q2f q3e county using "/data/economic/cmf2012/extract.dta", clear
gen logprofit = log(q2f)
by county: collapse (count)  n=q3e (mean) logprofit
drop if n<10
graph twoway n logprofit
```

Do not do this

A bad example, because literally making more work for you and for future replicators, is to manually redact the confidential information with text that is not legitimate code:

```
set seed NNNNN
use <removed vars> county using "<removed path>", clear
gen logprofit = log(XXXX)
by county: collapse (count)  n=XXXX (mean) logprofit
drop if n<XXXX
graph twoway n logprofit
```

The redacted program above will no longer run, and will be very tedious to un-redact if a subsequent replicator obtains legitimate access to the confidential data.

Better

Simply replacing the confidential data with replacement that are valid placeholders in the programming language of your choice is already better. Here’s the confidential version of the file:

```
//============ confidential parameters =============
global confseed    12345
global confpath    "/data/economic/cmf2012"
global confprofit  q2f
global confemploy  q3e
global confmincell 10
//============ end confidential parameters =========
set seed $confseed
use $confprofit county using "${confpath}/extract.dta", clear
gen logprofit = log($confprofit)
by county: collapse (count)  n=$confemploy (mean) logprofit
drop if n<$confmincell
graph twoway n logprofit
```

Better

and this could be the released file, part of the replication package:

```
//============ confidential parameters =============
global confseed    XXXX    // a number
global confpath    "XXXX"  // a path that will be communicated to you
global confprofit  XXX     // Variable name for profit T26
global confemploy  XXX     // Variable name for employment T26
global confmincell XXX     // a number
//============ end confidential parameters =========
set seed $confseed
use $confprofit county using "${confpath}/extract.dta", clear
gen logprofit = log($confprofit)
by county: collapse (count)  n=$confemploy (mean) logprofit
drop if n<$confmincell
graph twoway n logprofit
```

While the code won’t run as-is, it is easy to un-redact, regardless of how many times you reference the confidential values, e.g., `q2f`, anywhere in the code.

Best

- Main file
- Conditional processing
- Separate file for confidential parameters which can simply be excluded from disclosure request

Best

Main file `main.do`:

```
//============ confidential parameters =============
capture confirm file "$code/confidential/confparms.do"
if _rc == 0 {
    // file exists
    include "$code/confidential/confparms.do""
} else {
    di in red "No confidential parameters found"
}
//============ end confidential parameters =========

//============ non-confidential parameters =========
global safepath "$rootdir/releasable"
cap mkdir "$safepath"

//============ end parameters ======================
```

Best

Main file `main.do` (continued)

```
// ::::  Process only if confidential data is present

capture confirm  file "${confpath}/extract.dta"
if _rc == 0 {
   set seed $confseed
   use $confprofit county using "${confpath}/extract.dta", clear
   gen logprofit = log($confprofit)
   by county: collapse (count)  n=$confemploy (mean) logprofit
   drop if n<$confmincell
   save "${safepath}/figure1.dta", replace
} else { di in red "Skipping processing of confidential data" }

//============ at this point, the data is releasable ======
// ::::  Process always

use "${safepath}/figure1.dta", clear
graph twoway n logprofit
graph export "${safepath}/figure1.pdf", replace
```

Best

Auxiliary file `$code/confidential/confparms.do"` (not released)

```
//============ confidential parameters =============
global confseed    12345
global confpath    "/data/economic/cmf2012"
global confprofit  q2f
global confemploy  q3e
global confmincell 10
//============ end confidential parameters =========
```

Best

Auxiliary file `$code/include/confparms_template.do` (this is released)

```
//============ confidential parameters =============
// Copy this file to $code/confidential/confparms.do and edit
global confseed    XXXX    // a number
global confpath    "XXXX"  // a path that will be communicated to you
global confprofit  XXX     // Variable name for profit T26
global confemploy  XXX     // Variable name for employment T26
global confmincell XXX     // a number
//============ end confidential parameters =========
```

Best replication package

Thus, the replication package would have:

```
...
code/main.do
README.md
include/confparms_template.do
releasable/figure1.dta
releasable/figure1.pdf
```

Keeping on top of provenance

- Licenses
- Streamlining for reproducibility

Licenses

Where does the file come from?

- How can we describe this later to somebody?
  - Point and click is long to describe
  - What are the rights we have?

What is a license?

> A license (licence) is an official permission or permit to do, use, or own something (as well as the document of that permission or permit).[11] [12]

Examples

- Creative Commons licenses, used for artistic products and data
- Open Source licenses (BSD, GPL, MIT, etc.), used for software (code)

License applying to Geodist data

- CEPII GeoDist is under an “Etalab 2.0 license”

Can we re-publish the file?

Downloading via code

Easiest:

Stata

```
use "$URL" , clear
```

Why not?

- will it be there in two months? in 6 years?
- what if the internet connection is down?

Easy:

Stata

```
global URL "https://www.cepii.fr/distance/dist_cepii.dta"
copy "$URL" (outputfile), replace
```

R

```
download.file(url="$URL",destfile="(outputfile)")
```

We will get to even better methods a bit later

Creating a README

- Template README
  - Cite both dataset and working paper
  - Add data URL and time accessed (can you think of a way to automate this?)
  - Add a link to license (also: download and store the license)

Link

Step 1: Stata, R[13]

Wrapping it all up

Wrapping up

- Public replication package contains intelligible code, omits confidential details (but provides template code), has detailed data provenance statements
- Confidential replication package contains all the same, plus the confidential code, is archived in the FSRDC

Things to remember

- Use code to save figures and tables (`estout`, `graph export`, `regsave`)
- Create log files for each run (`stata -b do file.do` not fine-grained enough) link

Things to remember

Run it all again, top to bottom!

Things to remember

- When doing a disclosure review request, remember to request the **code**
- When outputting statistics, ***consider the disclosure rules*** - the less changes, the faster the output (in theory), but in particular fewer surprises
- Do not think "***nobody will ever read this code***" - somebody is very likely to!

End

Now you wait for the replicators to show up!

Footnotes

1.

Bollen et al. 2015. “Social, Behavioral, and Economic Sciences Perspectives on F
and Reliable Science.” National Science Foundation.
https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Rep

2.

Ambrus, Attila, Erica Field, and Robert Gonzalez. 2020. “Loss in the Time of Cho
Long-Run Impact of a Disease Epidemic on the Urban Landscape.” American Ec
Review, 110 (2): 475–525. https://doi.org/10.1257/aer.20190759

3.

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

4.