# Implementing Increased Transparency and Reproducibility in Economics: Lessons learned

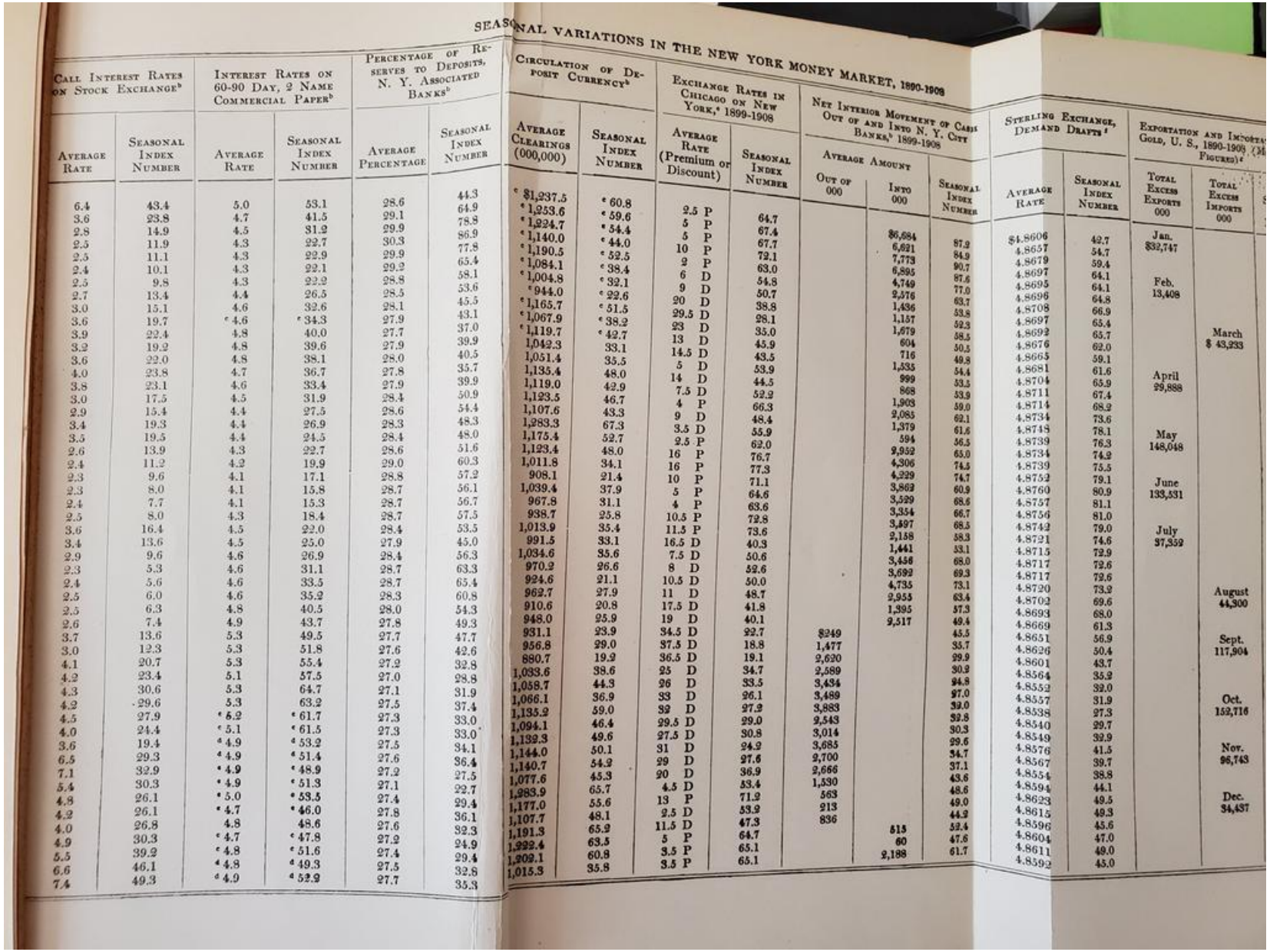
Lars Vilhuber

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[c] | | Exportation and Importation Gold, U. S., 1890-1908 (Minus Figures)[c] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Average Rate | Seasonal Index Number | Average Rate | Seasonal Index Number | Average Percentage | Seasonal Index Number | Average Clearings (000,000) | Seasonal Index Number | Average Rate (Premium or Discount) | Seasonal Index Number | Average Amount Out of 000 | Average Amount Into 000 | Seasonal Index Number | Average Rate | Seasonal Index Number | Total Excess Exports 000 | Total Excess Imports 000 |
| 6.4 | 43.4 | 5.0 | 53.1 | 28.6 | 44.3 | [c] $1,237.5 | [c] 60.8 | 2.5 P | 64.7 | | $6,684 | 87.2 | $4.8606 | 42.7 | Jan. $32,747 | |
| 3.6 | 23.8 | 4.7 | 41.5 | 29.1 | 64.9 | [c] 1,253.6 | [c] 59.6 | 5 P | 67.4 | | 6,621 | 84.9 | 4.8657 | 54.7 | | |
| 2.8 | 14.9 | 4.5 | 31.2 | 29.9 | 78.8 | [c] 1,224.7 | [c] 54.4 | 5 P | 67.7 | | 7,773 | 90.7 | 4.8679 | 59.4 | | |
| 2.5 | 11.9 | 4.3 | 22.7 | 30.3 | 86.9 | [c] 1,140.0 | [c] 44.0 | 10 P | 72.1 | | 6,895 | 87.6 | 4.8697 | 64.1 | Feb. 13,408 | |
| 2.5 | 11.1 | 4.3 | 22.9 | 29.9 | 77.8 | [c] 1,190.5 | [c] 52.5 | 2 P | 63.0 | | 4,749 | 77.0 | 4.8695 | 64.1 | | |
| 2.4 | 10.1 | 4.3 | 22.1 | 29.2 | 65.4 | [c] 1,084.1 | [c] 38.4 | 6 D | 54.8 | | 2,576 | 63.7 | 4.8696 | 64.8 | | |
| 2.5 | 9.8 | 4.3 | 22.2 | 28.8 | 58.1 | [c] 1,004.8 | [c] 32.1 | 9 D | 50.7 | | 1,436 | 53.8 | 4.8708 | 66.9 | | |
| 2.7 | 13.4 | 4.4 | 26.5 | 28.5 | 53.6 | [c] 944.0 | [c] 22.6 | 20 D | 38.8 | | 1,157 | 52.3 | 4.8697 | 65.4 | | |
| 3.0 | 15.1 | 4.6 | 32.6 | 28.1 | 45.5 | [c] 1,165.7 | [c] 51.5 | 29.5 D | 28.1 | | 1,679 | 58.5 | 4.8692 | 65.7 | | March $ 43,233 |
| 3.6 | 19.7 | [c] 4.6 | [c] 34.3 | 27.9 | 43.1 | [c] 1,067.9 | [c] 38.2 | 23 D | 35.0 | | 604 | 50.5 | 4.8676 | 62.0 | | |
| 3.9 | 22.4 | 4.8 | 40.0 | 27.7 | 37.0 | [c] 1,119.7 | [c] 42.7 | 13 D | 45.9 | | 716 | 49.8 | 4.8665 | 59.1 | | |
| 3.2 | 19.2 | 4.8 | 39.6 | 27.9 | 39.9 | 1,042.3 | 33.1 | 14.5 D | 43.5 | | 1,535 | 54.4 | 4.8681 | 61.6 | | |
| 3.6 | 22.0 | 4.8 | 38.1 | 28.0 | 40.5 | 1,051.4 | 35.5 | 5 D | 53.9 | | 999 | 53.5 | 4.8704 | 65.9 | April 29,888 | |
| 4.0 | 23.8 | 4.7 | 36.7 | 27.8 | 35.7 | 1,135.4 | 48.0 | 14 D | 44.5 | | 868 | 53.9 | 4.8711 | 67.4 | | |
| 3.8 | 23.1 | 4.6 | 33.4 | 27.9 | 39.9 | 1,119.0 | 42.9 | 7.5 D | 52.2 | | 1,903 | 59.0 | 4.8714 | 68.2 | | |
| 3.0 | 17.5 | 4.5 | 31.9 | 28.4 | 50.9 | 1,123.5 | 46.7 | 4 P | 66.3 | | 2,085 | 62.1 | 4.8734 | 73.6 | | |
| 2.9 | 15.4 | 4.4 | 27.5 | 28.6 | 54.4 | 1,107.6 | 43.3 | 9 D | 48.4 | | 1,379 | 61.6 | 4.8748 | 78.1 | May 148,048 | |
| 3.4 | 19.3 | 4.4 | 26.9 | 28.3 | 48.3 | 1,283.3 | 67.3 | 3.5 D | 55.9 | | 594 | 56.5 | 4.8739 | 76.3 | | |
| 3.5 | 19.5 | 4.4 | 24.5 | 28.4 | 48.0 | 1,175.4 | 52.7 | 2.5 P | 62.0 | | 2,952 | 65.0 | 4.8734 | 74.2 | | |
| 2.6 | 13.9 | 4.3 | 22.7 | 28.6 | 51.6 | 1,123.4 | 48.0 | 16 P | 76.7 | | 4,306 | 74.5 | 4.8739 | 75.5 | | |
| 2.4 | 11.2 | 4.2 | 19.9 | 29.0 | 60.3 | 1,011.8 | 34.1 | 16 P | 77.3 | | 4,229 | 74.7 | 4.8752 | 79.1 | June 133,531 | |
| 2.3 | 9.6 | 4.1 | 17.1 | 28.8 | 57.2 | 908.1 | 21.4 | 10 P | 71.1 | | 3,863 | 69.9 | 4.8760 | 80.9 | | |
| 2.3 | 8.0 | 4.1 | 15.8 | 28.7 | 56.1 | 1,039.4 | 37.9 | 5 P | 64.6 | | 3,529 | 68.6 | 4.8757 | 81.1 | | |
| 2.4 | 7.7 | 4.1 | 15.3 | 28.7 | 56.7 | 967.8 | 31.1 | 4 P | 63.6 | | 3,354 | 66.7 | 4.8756 | 81.0 | | |
| 2.5 | 8.0 | 4.3 | 18.4 | 28.7 | 57.5 | 938.7 | 25.8 | 10.5 P | 72.8 | | 3,497 | 68.5 | 4.8742 | 79.0 | July 37,352 | |
| 3.6 | 16.4 | 4.5 | 22.0 | 28.4 | 53.5 | 1,013.9 | 35.4 | 11.5 P | 73.6 | | 2,158 | 58.3 | 4.8721 | 74.6 | | |
| 3.4 | 13.6 | 4.5 | 25.0 | 27.9 | 45.0 | 991.5 | 33.1 | 16.5 D | 40.3 | | 1,441 | 53.1 | 4.8715 | 72.9 | | |
| 2.9 | 9.6 | 4.6 | 26.9 | 28.4 | 56.3 | 1,034.6 | 35.6 | 7.5 D | 50.6 | | 3,456 | 68.0 | 4.8717 | 72.6 | | |
| 2.3 | 5.3 | 4.6 | 31.1 | 28.7 | 63.3 | 970.2 | 26.6 | 8 D | 52.6 | | 3,692 | 69.3 | 4.8717 | 72.6 | | |
| 2.4 | 5.6 | 4.6 | 33.5 | 28.7 | 65.4 | 924.6 | 21.1 | 10.5 D | 50.0 | | 4,735 | 73.1 | 4.8720 | 73.2 | | August 44,300 |
| 2.5 | 6.0 | 4.6 | 35.2 | 28.3 | 60.8 | 962.7 | 27.9 | 11 D | 48.7 | | 2,955 | 63.4 | 4.8702 | 69.6 | | |
| 2.5 | 6.3 | 4.8 | 40.5 | 28.0 | 54.3 | 910.6 | 20.8 | 17.5 D | 41.8 | | 1,395 | 57.3 | 4.8693 | 68.0 | | |
| 2.6 | 7.4 | 4.9 | 43.7 | 27.8 | 49.3 | 948.0 | 25.9 | 19 D | 40.1 | | 2,517 | 49.4 | 4.8669 | 61.3 | | |
| 3.7 | 13.6 | 5.3 | 49.5 | 27.7 | 47.7 | 931.1 | 23.9 | 34.5 D | 22.7 | $249 | | 45.5 | 4.8651 | 56.9 | | Sept. 117,904 |
| 3.0 | 12.3 | 5.3 | 51.8 | 27.6 | 42.6 | 956.8 | 29.0 | 37.5 D | 18.8 | 1,477 | | 35.7 | 4.8626 | 50.4 | | |
| 4.1 | 20.7 | 5.3 | 55.4 | 27.2 | 32.8 | 880.7 | 19.2 | 36.5 D | 19.1 | 2,620 | | 29.9 | 4.8601 | 43.7 | | |
| 4.2 | 23.4 | 5.1 | 57.5 | 27.0 | 28.8 | 1,033.6 | 38.6 | 25 D | 34.7 | 2,589 | | 30.2 | 4.8564 | 35.2 | | |
| 4.3 | 30.6 | 5.3 | 64.7 | 27.1 | 31.9 | 1,058.7 | 44.3 | 26 D | 33.5 | 3,434 | | 24.8 | 4.8552 | 32.0 | | |
| 4.2 | 29.6 | 5.3 | 63.2 | 27.5 | 37.4 | 1,066.1 | 36.9 | 33 D | 26.1 | 3,489 | | 27.0 | 4.8557 | 31.9 | | Oct. 152,716 |
| 4.5 | 27.9 | [c] 5.2 | [c] 61.7 | 27.3 | 33.0 | 1,135.2 | 59.0 | 32 D | 27.2 | 3,883 | | 32.0 | 4.8538 | 27.3 | | |
| 4.0 | 24.4 | [c] 5.1 | [c] 61.5 | 27.3 | 33.0 | 1,094.1 | 46.4 | 29.5 D | 29.0 | 2,543 | | 32.8 | 4.8540 | 29.7 | | |
| 3.6 | 19.4 | [c] 4.9 | [c] 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 | [c] 4.9 | [c] 51.4 | 27.6 | 36.4 | 1,144.0 | 50.1 | 31 D | 24.2 | 3,685 | | 29.6 | 4.8576 | 41.5 | | Nov. 96,743 |
| 7.1 | 32.9 | [c] 4.9 | [c] 48.9 | 27.2 | 27.5 | 1,140.7 | 54.2 | 29 D | 27.6 | 2,700 | | 34.7 | 4.8567 | 39.7 | | |
| 5.4 | 30.3 | [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 | [c] 5.0 | [c] 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 | [c] 4.7 | [c] 46.0 | 27.8 | 36.1 | 1,177.0 | 55.6 | 13 P | 71.2 | 563 | | 48.6 | 4.8623 | 49.5 | | Dec. 34,437 |
| 4.0 | 26.8 | 4.8 | 48.6 | 27.6 | 32.3 | 1,107.7 | 48.1 | 2.5 D | 53.2 | 213 | | 49.0 | 4.8615 | 49.3 | | |
| 4.9 | 30.3 | [c] 4.7 | [c] 47.8 | 27.2 | 24.9 | 1,191.3 | 65.2 | 11.5 D | 47.3 | 836 | | 44.2 | 4.8596 | 45.6 | | |
| 5.5 | 39.2 | [c] 4.8 | [c] 51.6 | 27.4 | 29.4 | 1,222.4 | 63.5 | 5 P | 64.7 | | 515 | 52.4 | 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 | 47.6 | 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 | 61.7 | 4.8592 | 45.0 | | |

Via @sdellavi

E. W. Kemmerer. 2011. "Seasonal Variations in the New York Money Market."

*The American Economic Review,* Vol. 1, No. 1 (March 1911), pp. 33-49

# Efficiency of scholary discourse!

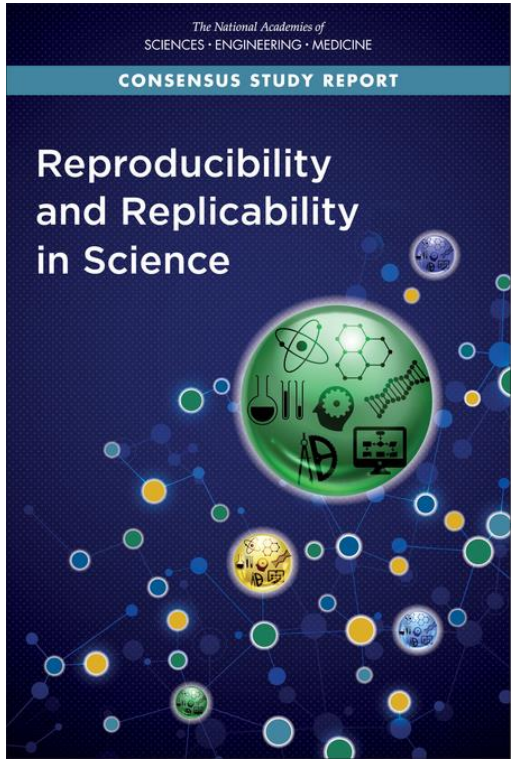
**Modern publications need**
**the same transparency and completeness**
**as in the old days**

**to facilitate replicability**

# Replicability?

# 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 confidential data







# Progress

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

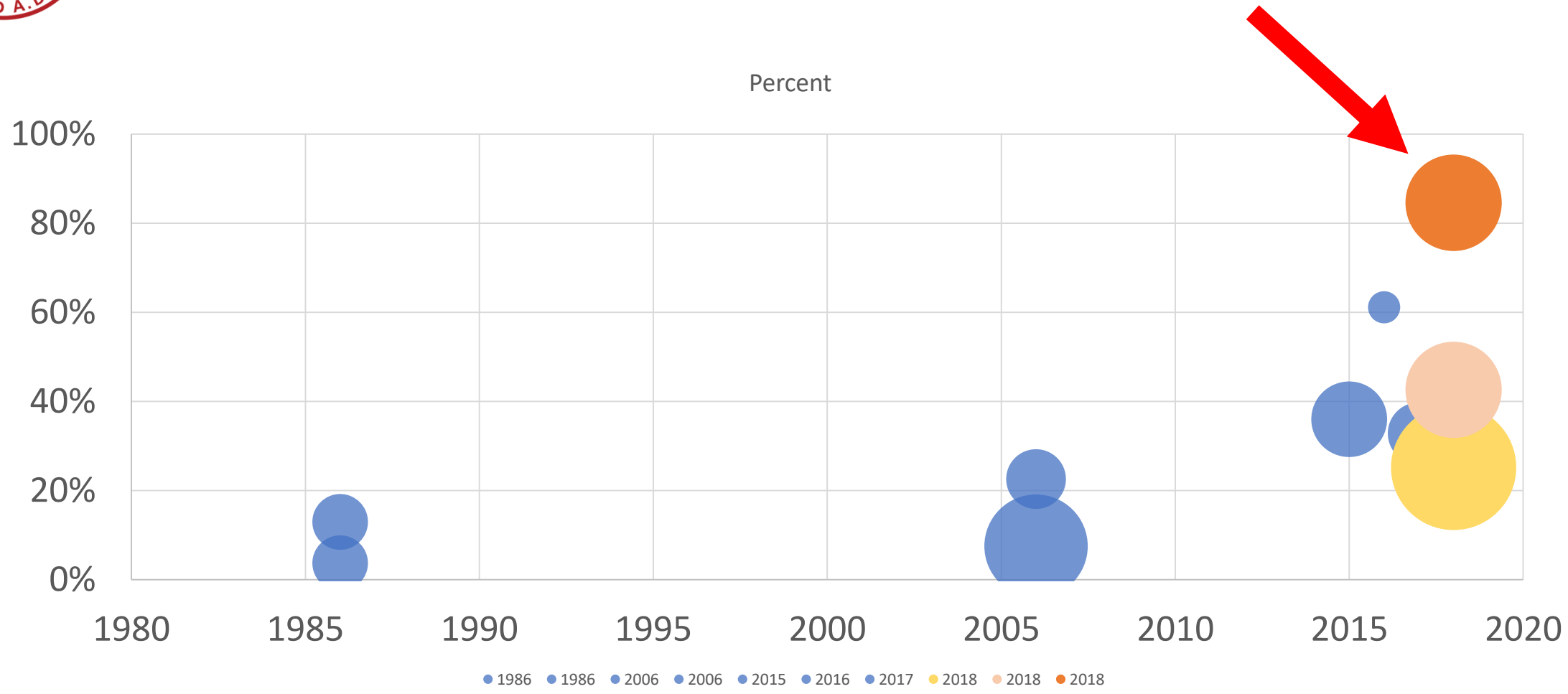
**Replicability**

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

**Generalizability**

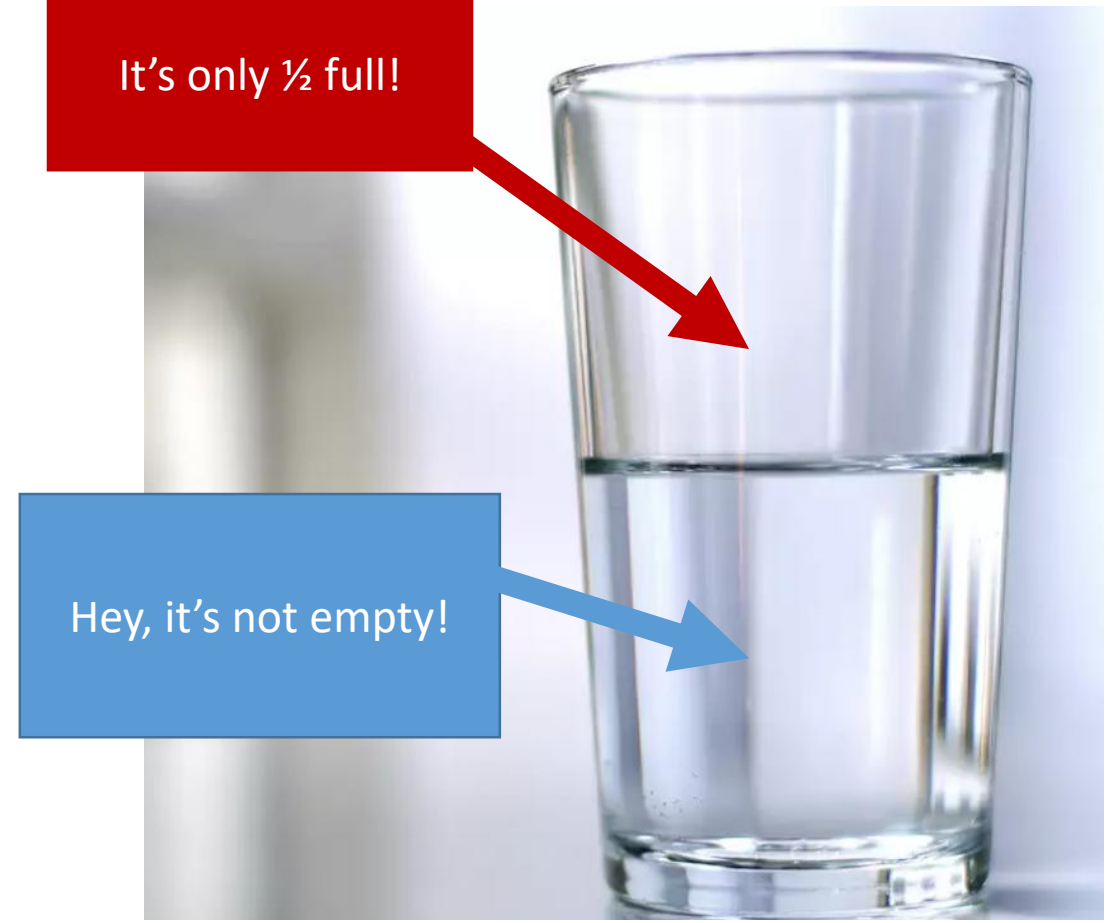
- Wider Replication (Pesaran 2003)
- Scientific Replication (Hamermesh 2007)
- Reanalysis/Robustness (Clemens 2015)

# 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



# Why?

# 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 coding practices

- **Manual/non-automation**

Code produces no meaningful output

- **Lack of robustness**:

Bugs in the code

# 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

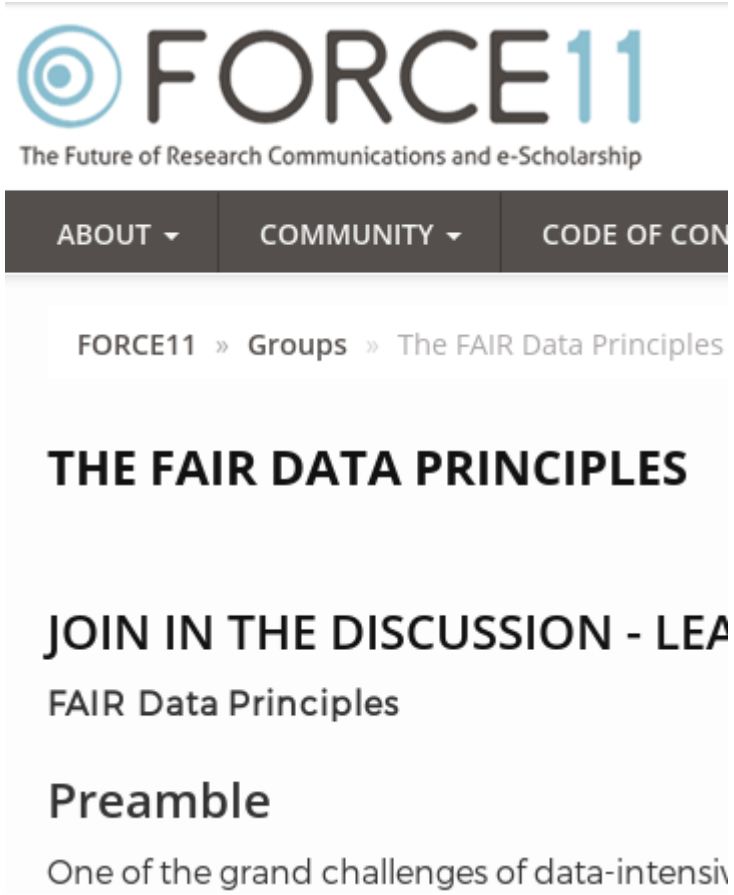
Many datasets

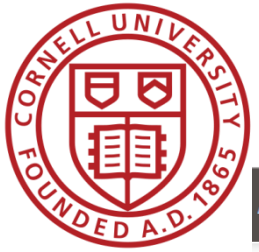
- Are imperfectly described
  - Very few data citations
- Are badly documented
- Have no (permanent) location defined
  - Even for data from high-profile organizations!
- All of the above

# 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].

# 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)

# 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

# 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.**

# 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

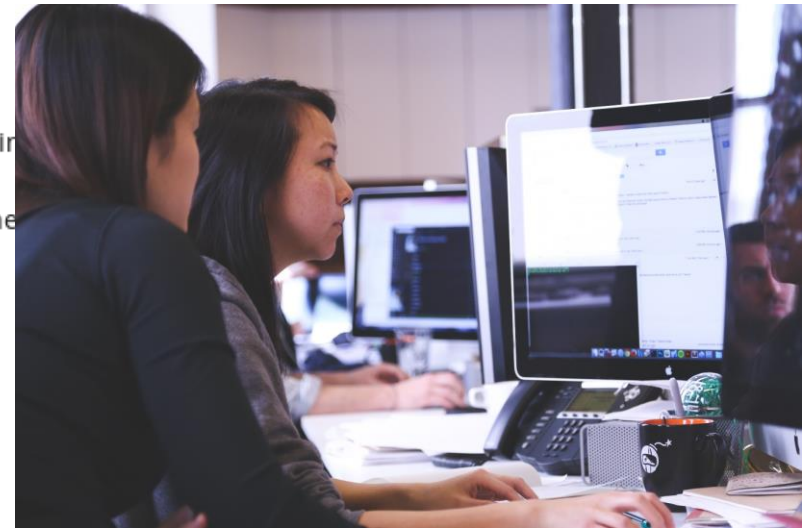
On this page:

- 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

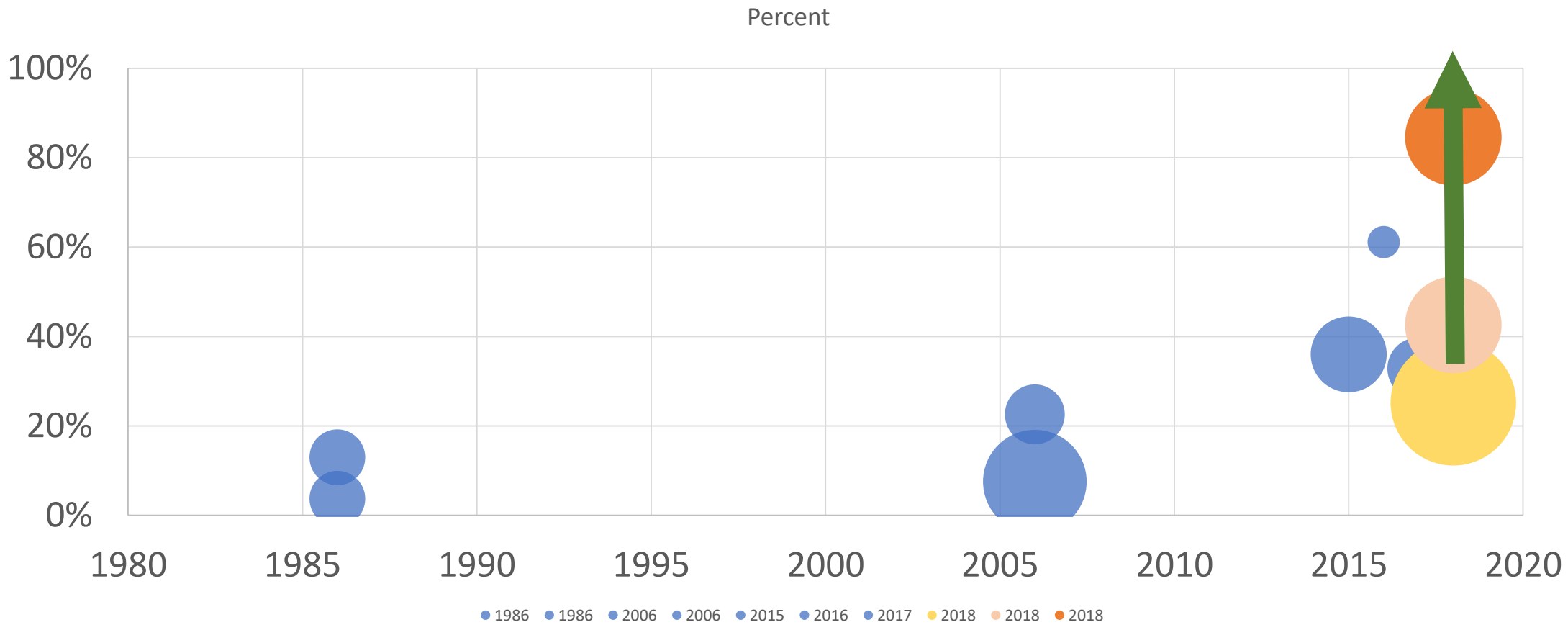
- what authors should check before providir journals
- what verifier teams should check for in the to them for the purpose of verification



# 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, Franklin Omullo, Liam P. Cushen, Ololade Omotoba, Lydia Reiner, Xiangru Li, Melanie Chen, Peter Rafael Sanchez, Jill Crosby, Matthew H. Wang
- Other graduate students: Aviv Caspi, Leah Kim

# Goal: Improve reproducibility



# 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

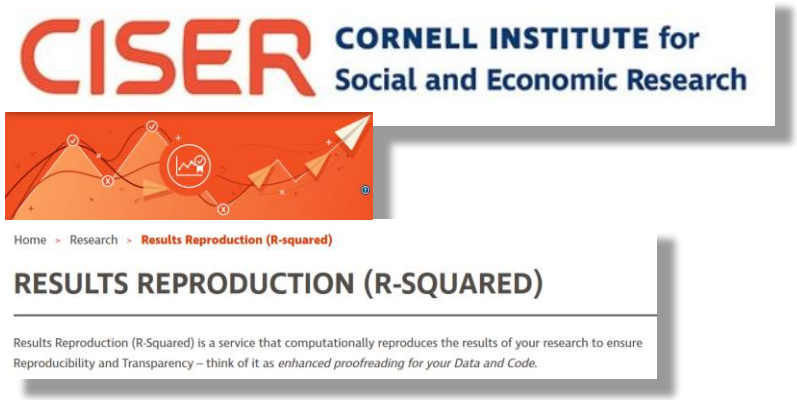
# Stats on reproduced articles

Between July 16, 2019, and June 26, 2020 (11 mths), the AEA Data Editor team conducted

- **636 assessments**
- Of which **233 manuscripts** have been “accepted”

# Also work with 3rd parties

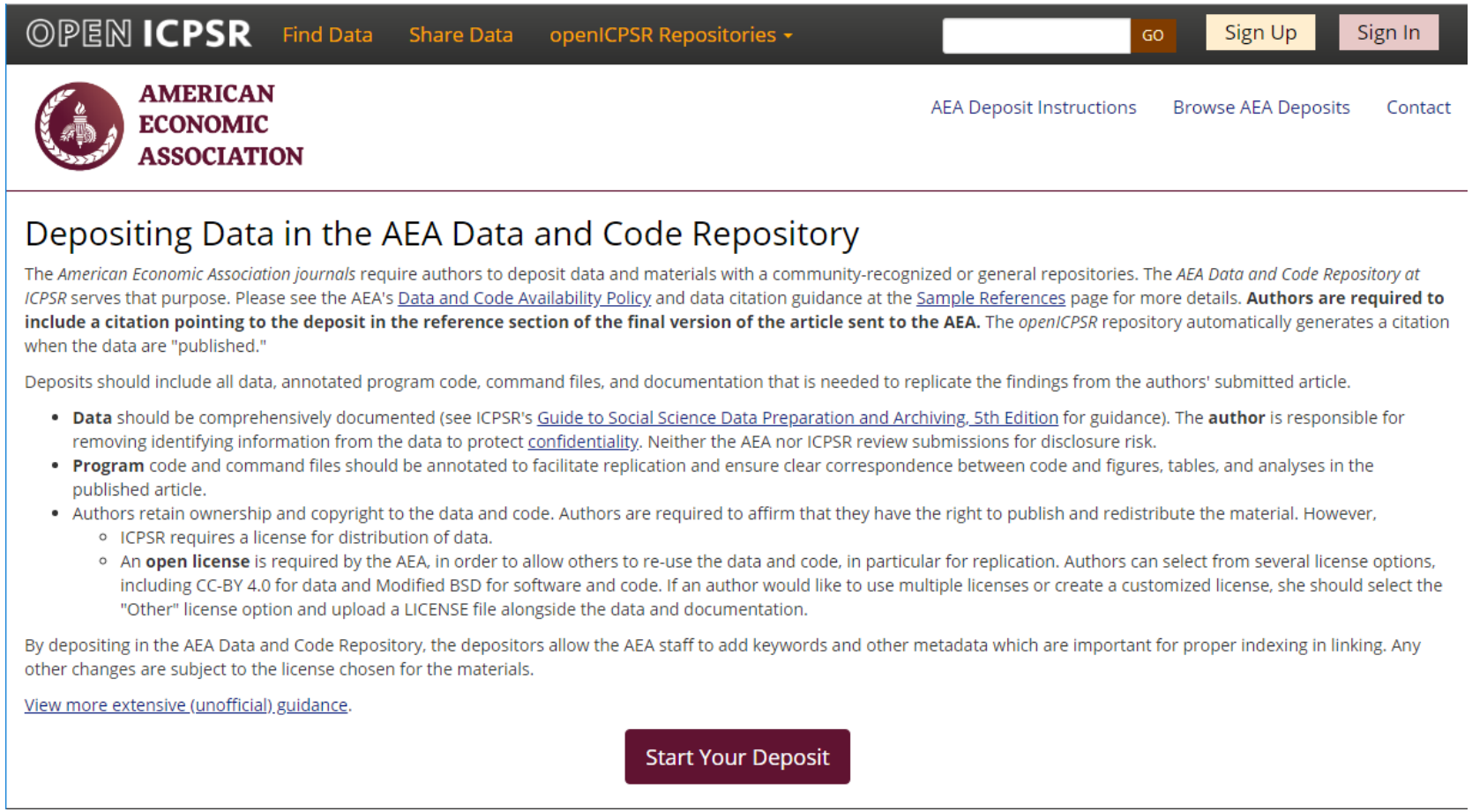
- cascad – Using confidential data!
- CISER (R-Squared)
- Various contributors with access to confidential data


# 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.

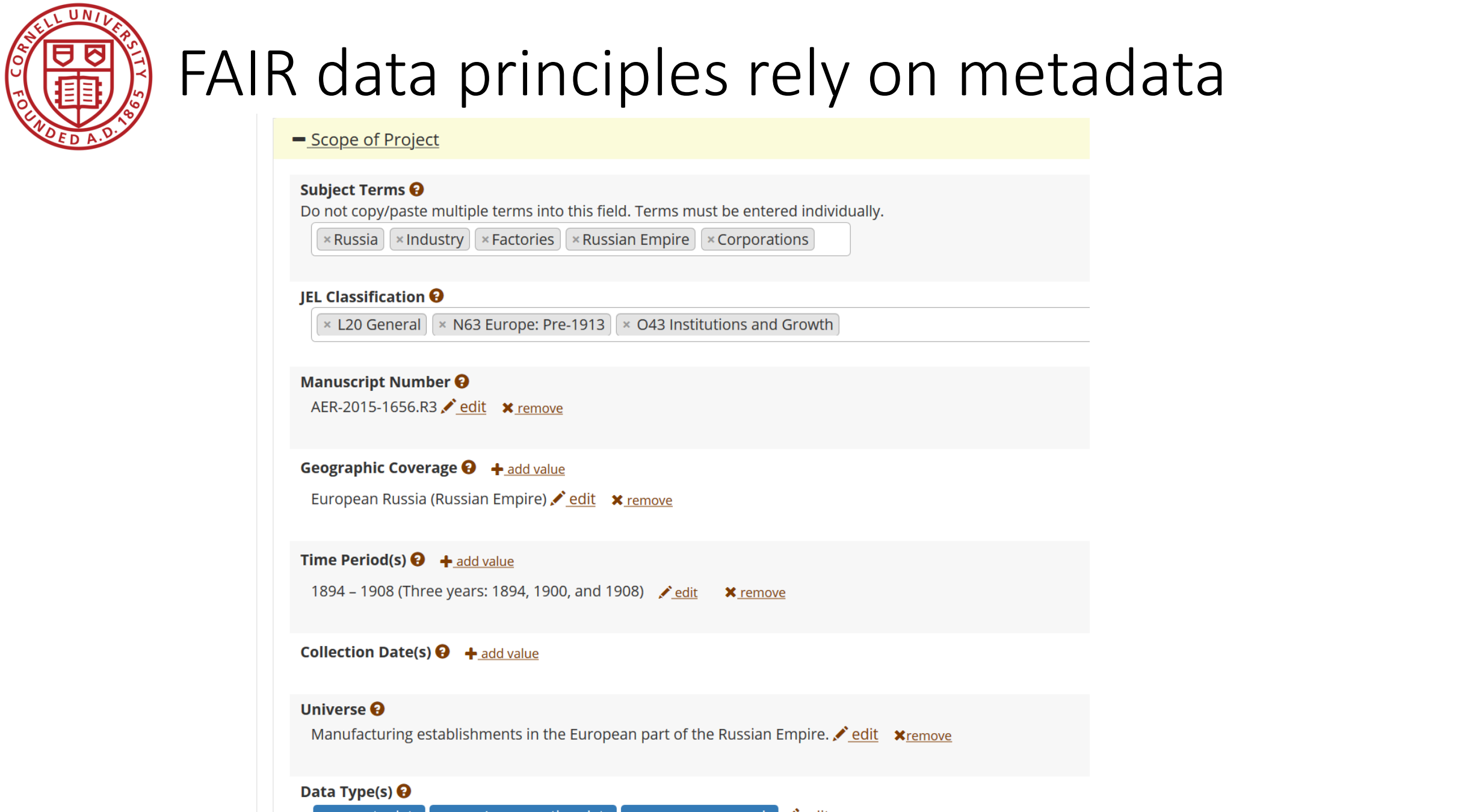
- **Data** should be comprehensively documented (see ICPSR's Guide to Social Science Data Preparation and Archiving, 5th Edition for guidance). The **author** is responsible for removing identifying information from the data to protect confidentiality. Neither the AEA nor ICPSR review submissions for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

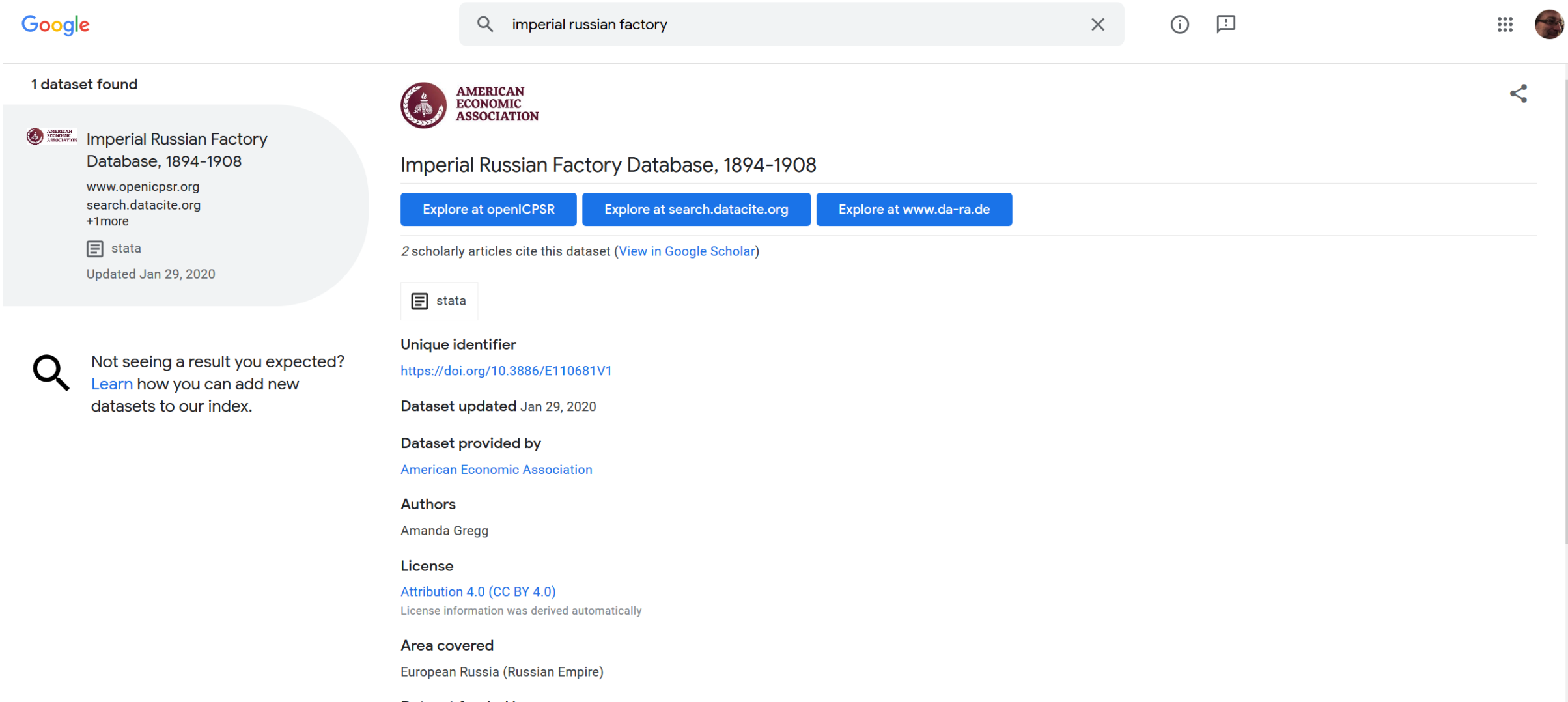
**Version:** V1

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>
```

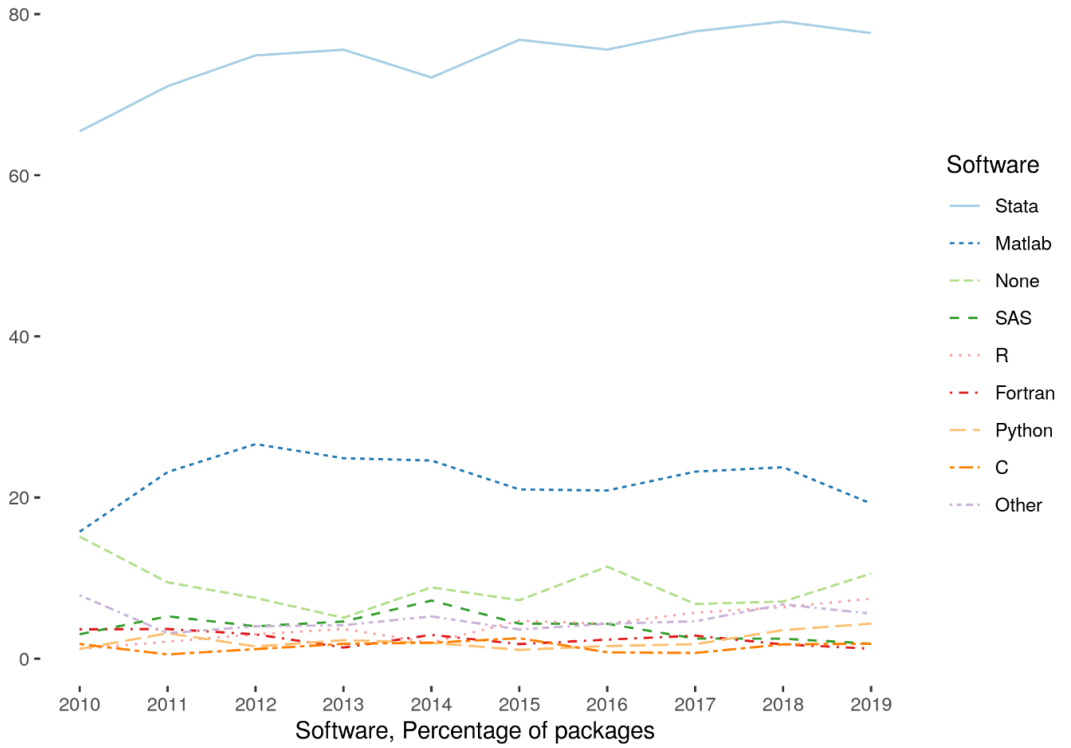

| | | | |
|---|---|---|---|
| AG_Corp_CleaningandDatabaseCompiler.do | | KB | 11:02:AM |
| AG_Corp_Prod_AppendixCode.do | text/x-stata-syntax | 42.2 KB | 12/09/2019 09:19:AM |
| AG_Corp_Prod_Code.do | text/x-stata-syntax | 26.6 KB | 12/12/2019 03:01:AM |
| AG_Corp_Prod_Database.dta | application/x-stata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

# … and findability relies on metadata



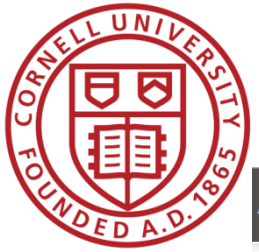
# Very little diversity in software

- **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].

# Some practical tips (based on 700+ assessments)

# Would you buy a car from this guy?

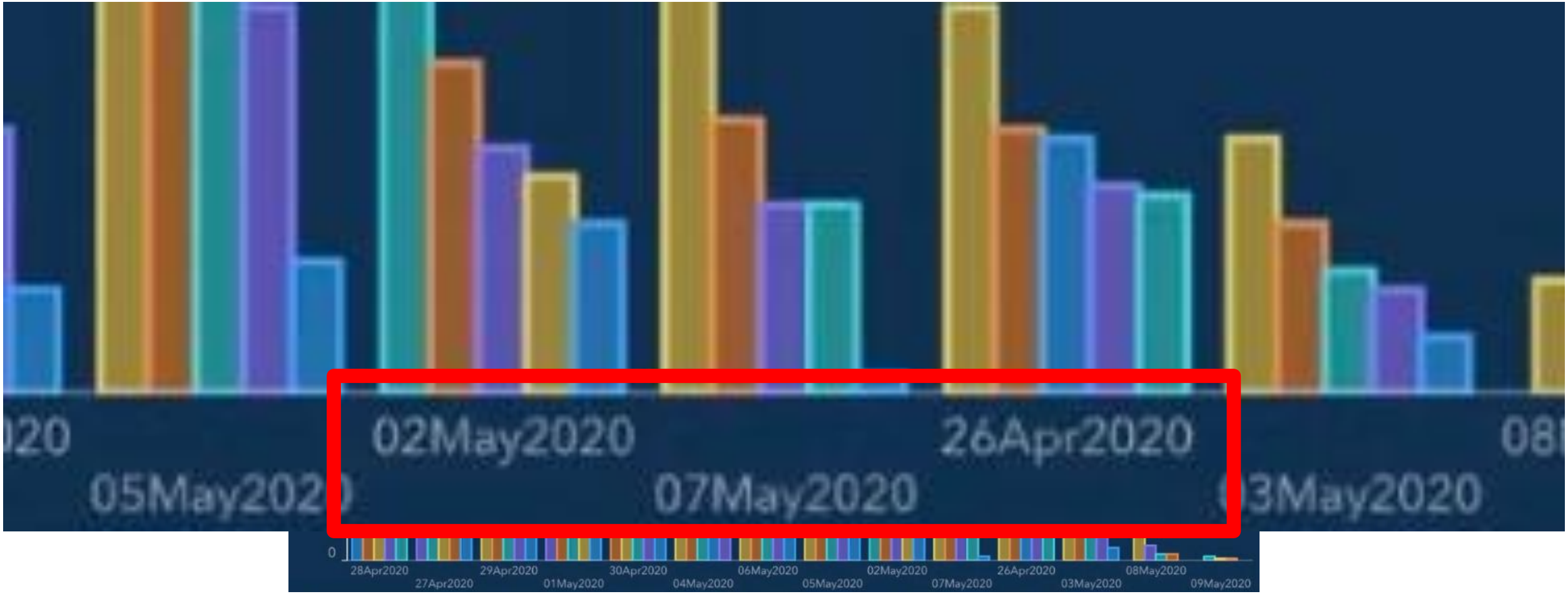

# Provenance!

- Does the sales person have a good record?
- Where does the car come from?
- What do we know about the car?


# Would you use this data?

```
0000000 cfd0 e011 b1a1 e11a 0000 0000 0000 0000
0000010 0000 0000 0000 0000 003e 0003 fffe 0009
0000020 0006 0000 0000 0000 0000 0000 0004 0000
0000030 008f 0000 0000 0000 1000 0000 fffe ffff
0000040 0000 0000 fffe ffff 0000 0000 008b 0000
0000050 008c 0000 008d 0000 008e 0000 ffff ffff
0000060 ffff ffff ffff ffff ffff ffff ffff ffff
*
0000200 0809 0010 0600 0005 209a 07cd c0c9 0000
0000210 0306 0000 00e1 0002 04b0 00c1 0002 0000
0000220 00e2 0000 005c 0070 0001 4c00 2020 2020
0000230 2020 2020 2020 2020 2020 2020 2020 2020
*
0000290 2020 2020 2020 2020 0042 0002 04b0 0161
00002a0 0002 0000 013d 0002 0001 009c 0002 000e
00002b0 0019 0002 0000 0012 0002 0000 0013 0002
00002c0 0000 01af 0002 0000 01bc 0002 0000 003d
00002d0 0012 0000 000f 3f1b 27f6 0038 0000 0000
00002e0 0001 0258 0040 0002 0000 008d 0002 0000
00002f0 0022 0002 0000 000e 0002 0001 01b7 0002
```

# Or would you trust this data?



# Provenance!

- Does the provider have a good record?
- Where do the data come from?
- What do we know about the data?

# Metadata!

# “It’s a file called stockmarket.xlsx”

2101.49
2057.64
2063.11
2077.42
2076.78
0
2068.76
2081.34
2046.68
2051.31
2076.62
2099.60
2108.95
2107.40
2124.29
2126.64
2128.28
2119.21
2114.15
2102.15
2079.65
2067.64
2093.25
2108.57
2108.63
2103.84

# “It’s a file called SP500.xlsx”

| | |
|---|---|
| SP500 | S&P 500, Index, Daily, Not Seasonally Adjusted |

Frequency: Daily, Close

| observation_date | SP500 |
|---|---|
| 2015-06-26 | 2101.49 |
| 2015-06-29 | 2057.64 |
| 2015-06-30 | 2063.11 |
| 2015-07-01 | 2077.42 |
| 2015-07-02 | 2076.78 |
| 2015-07-03 | 0 |
| 2015-07-06 | 2068.76 |
| 2015-07-07 | 2081.34 |
| 2015-07-08 | 2046.68 |
| 2015-07-09 | 2051.31 |
| 2015-07-10 | 2076.62 |
| 2015-07-13 | 2099.60 |
| 2015-07-14 | 2108.95 |
| 2015-07-15 | 2107.40 |
| 2015-07-16 | 2124.29 |
| 2015-07-17 | 2126.64 |
| 2015-07-20 | 2128.28 |

# “It’s a file called SP500.xlsx, downloaded from FRED.”

| | |
|---|---|
| SP500 | S&P 500, Index, Daily, Not Seasonally Adjusted |

Frequency: Daily, Close

| observation_date | SP500 |
|---|---|
| 2015-06-26 | 2101.49 |
| 2015-06-29 | 2057.64 |
| 2015-06-30 | 2063.11 |
| 2015-07-01 | 2077.42 |
| 2015-07-02 | 2076.78 |
| 2015-07-03 | 0 |
| 2015-07-06 | 2068.76 |
| 2015-07-07 | 2081.34 |
| 2015-07-08 | 2046.68 |
| 2015-07-09 | 2051.31 |
| 2015-07-10 | 2076.62 |
| 2015-07-13 | 2099.60 |
| 2015-07-14 | 2108.95 |
| 2015-07-15 | 2107.40 |
| 2015-07-16 | 2124.29 |
| 2015-07-17 | 2126.64 |
| 2015-07-20 | 2128.28 |

# “It’s a file called SP500.xlsx, downloaded from FRED.”

| | |
|---|---|
| SP500 | S&P 500, Index, Daily, Not Seasonally Adjusted |

Frequency: Daily, Close

| observation_date | SP500 |
|---|---|
| 2015-06-26 | 2101.49 |
| 2015-06-29 | 2057.64 |
| 2015-06-30 | 2063.11 |
| 2015-07-01 | 2077.42 |
| 2015-07-02 | 2076.78 |
| 2015-07-03 | 0 |
| 2015-07-06 | 2068.76 |
| 2015-07-07 | 2081.34 |
| 2015-07-08 | 2046.68 |
| 2015-07-09 | 2051.31 |
| 2015-07-10 | 2076.62 |
| 2015-07-13 | 2099.60 |
| 2015-07-14 | 2108.95 |
| 2015-07-15 | 2107.40 |
| 2015-07-16 | 2124.29 |
| 2015-07-17 | 2126.64 |
| 2015-07-20 | 2128.28 |

S&P Dow Jones Indices LLC. 2020. “*S&P 500 [SP500] [dataset]*”, retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SP500, June 26, 2020.

# “SP500.xlsx, from S&P (2020). Not provided as part of replication package because © S&P. ”

| | |
|---|---|
| SP500 | S&P 500, Index, Daily, Not Seasonally Adjusted |

Frequency: Daily, Close

| observation_date | SP500 |
|---|---|
| 2015-06-26 | 2101.49 |
| 2015-06-29 | 2057.64 |
| 2015-06-30 | 2063.11 |
| 2015-07-01 | 2077.42 |
| 2015-07-02 | 2076.78 |
| 2015-07-03 | 0 |
| 2015-07-06 | 2068.76 |
| 2015-07-07 | 2081.34 |
| 2015-07-08 | 2046.68 |
| 2015-07-09 | 2051.31 |
| 2015-07-10 | 2076.62 |
| 2015-07-13 | 2099.60 |
| 2015-07-14 | 2108.95 |
| 2015-07-15 | 2107.40 |
| 2015-07-16 | 2124.29 |
| 2015-07-17 | 2126.64 |
| 2015-07-20 | 2128.28 |

S&P Dow Jones Indices LLC. 2020. “*S&P 500 [SP500] [dataset]”*, retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SP500, June 26, 2020.

# Data Availability Statements

"SP500.xlsx, from S&P (2020). Not provided as part of replication package because © S&P. "

Describes data file, where to get it, how to get it, and any conditions of obtaining it

S&P 500, Index, Daily,
Not Seasonally Adjusted

| | |
|---|---|
| 2015-07-15 | 2107.40 |
| 2015-07-16 | 2124.29 |
| 2015-07-17 | 2126.64 |
| 2015-07-20 | 2128.28 |

S&P Dow Jones Indices LLC. 2020. "*S&P 500 [SP500] [dataset]*", retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SP500, June 26, 2020.

# Data Citation

"SP500.xlsx, from S&P (2020). Not provided as part of replication package because © S&P. "

Attributes the file to the proper source

| | SP500 |
|---|---|
| | S&P 500, Index, Daily, Not Seasonally Adjusted |
| | 2101.49 |
| | 2057.64 |
| | 2063.11 |
| | 2076.78 |
| | 0 |
| | 2068.76 |
| | 2081.34 |
| 2015-07-08 | 2046.68 |
| 2015-07-09 | 2051.31 |
| 2015-07-10 | 2076.62 |
| 2015-07-13 | 2099.60 |
| 2015-07-14 | 2108.95 |
| 2015-07-15 | 2107.40 |
| 2015-07-16 | 2124.29 |
| 2015-07-17 | 2126.64 |
| 2015-07-20 | 2128.28 |

S&P Dow Jones Indices LLC. 2020. "*S&P 500 [SP500] [dataset]*", retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SP500, June 26, 2020.

# 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

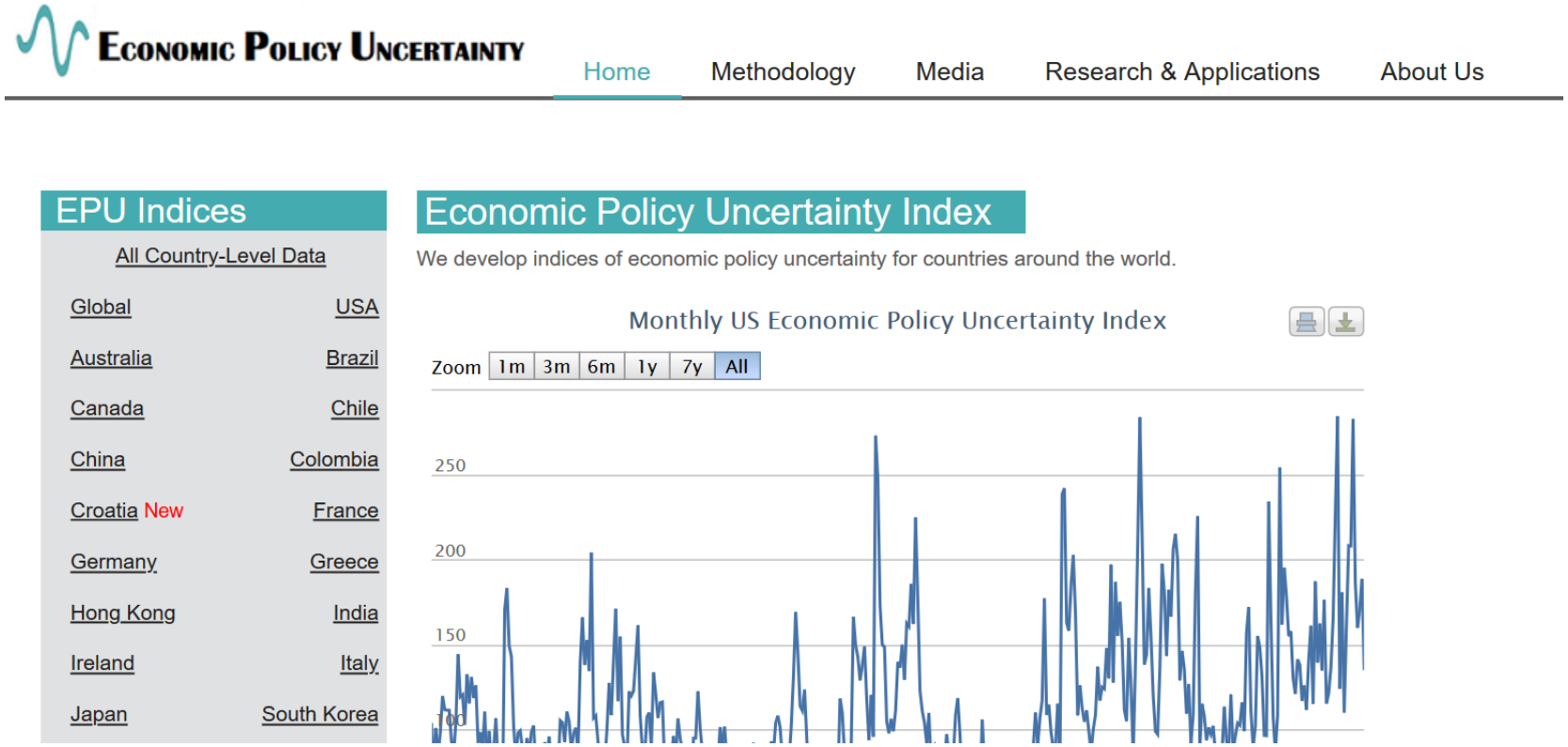
- What datasets are used
- Are they cited?
- Is there additional information access?
  - → URL leads to exact data?
  - → URL leads to application procedure?
  - → other access procedure is described?


Le labyrinthe des juges, par Guillaume Beck

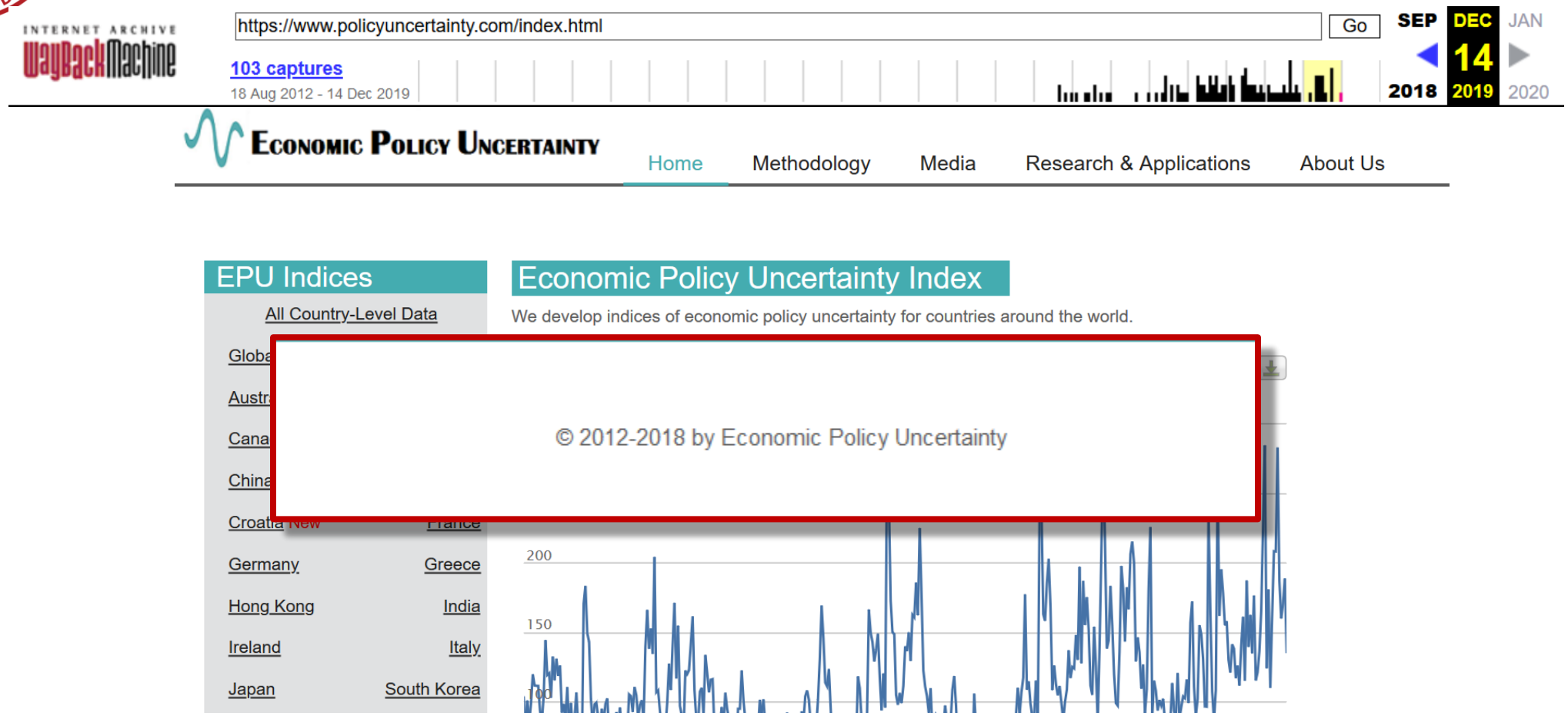
# Every manuscript is checked

- 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?

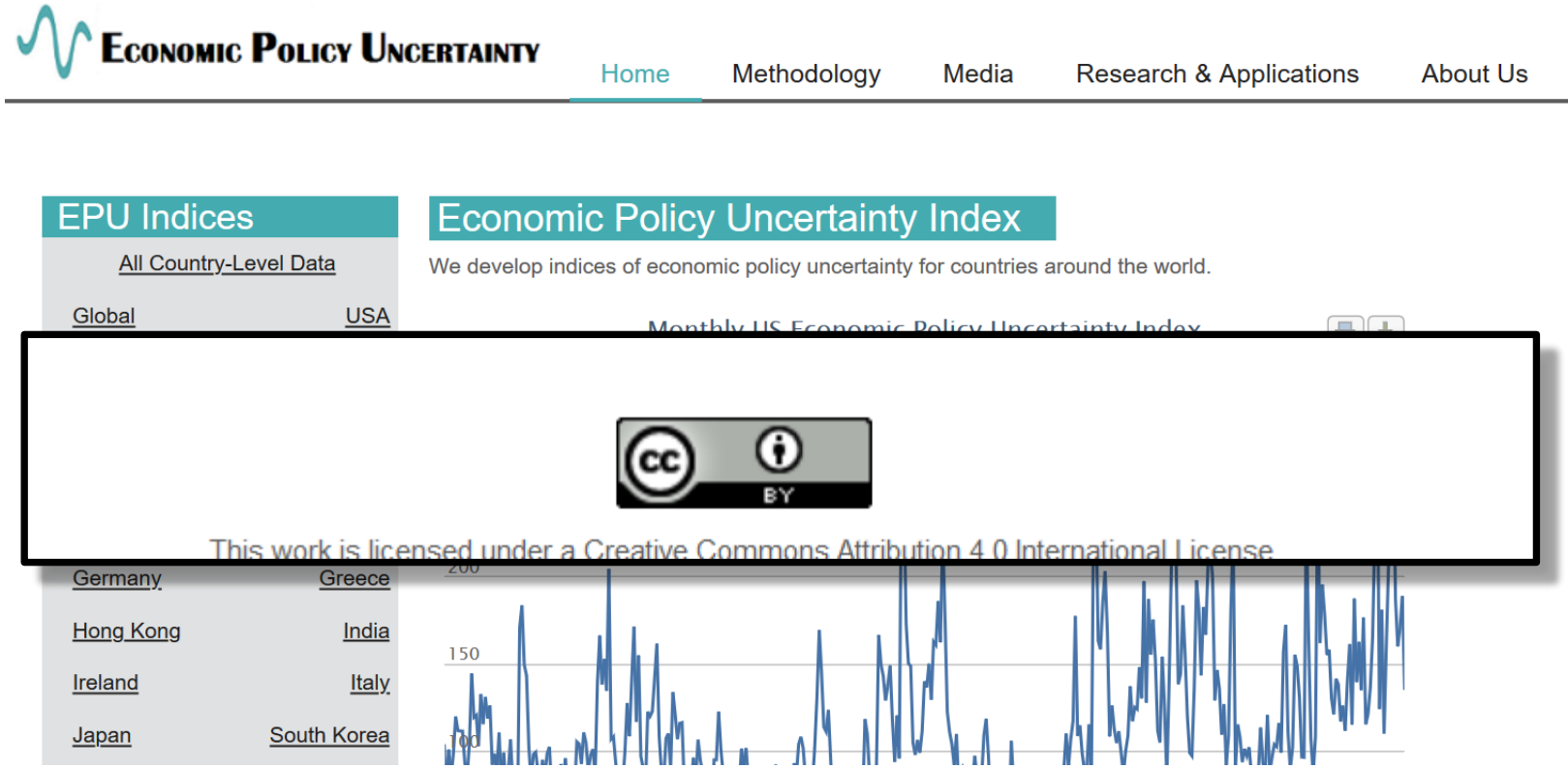
# Example 2: Academic data publisher



# Example 2: Academic data publisher



# Example 2: Academic data publisher-new!



# Rights to **use** data

- You browsed a website
- You purchased the data
- You signed a data use agreement
- You created the data (lab experiment)
- You had survey respondents consent to use (IRB approval!)

# Rights to **distribute** the data

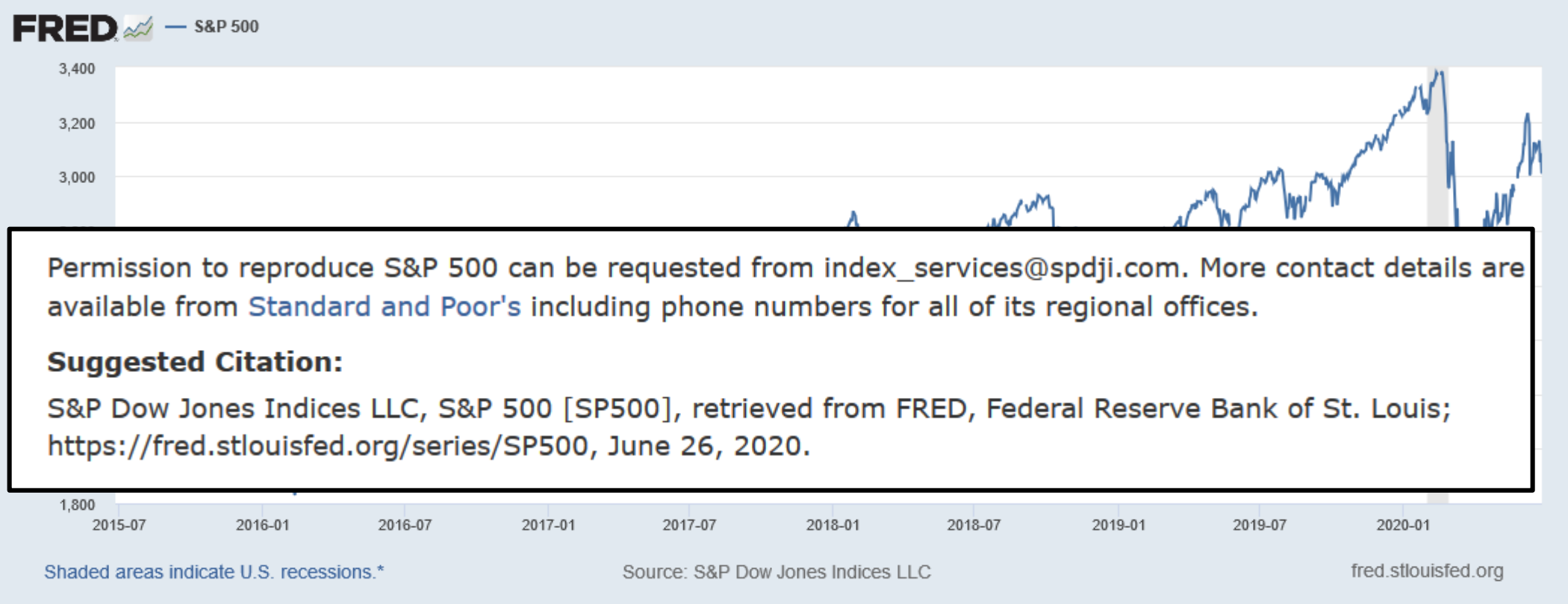
- If you created the data, you decide.
- If you got it from somewhere else:

READ THE TERMS OF USE / DATA USE AGREEMENT / CLICK-THROUGH / ETC.

## 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) and SP500



# Restricted-access data

- Often, you do not have the data on your laptop
- Examples: FSRDC, German, Swedish, French, Canadian, etc. administrative data
- Examples: Uber, Ebay data

# Citing restricted-access data

"Well, I can't download the data, so I can't cite it."

# 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

# But this is not always so easy

## Restricted–Use Economic Census* Microdata

| Data Set | Frequency | Unit of Enumeration | Availability |
|---|---|---|---|
| Census of Auxiliary Establishments (AUX) | Every 5 Years | Establishment | 1977–2012 |
| Census of Construction Industries (CCN) | Every 5 Years | Establishment | 1972–2012 |
| Census of Finance, Insurance, and Real Estate (CFI) | Every 5 Years | Establishment | 1992–2012 |
| Census of Island Areas – Puerto Rico (CIAPR) | Every 5 Years | Establishment | 1992–2007 |
| Census of Manufactures (CMF) | Every 5 Years | Establishment | 1963, 1967–2012 |
| Census of Mining (CMI) | Every 5 Years | Establishment | 1987–2012 |
| Census of Retail Trade (CRT) | Every 5 Years | Establishment | 1977–2012 |

# Element of a (data) citation

ICPSR notes that a citation should include the following items:

- Author
- Title
- Distributor
- Date
- Version
- Persistent identifier

# Element of a (data) citation

ICPSR notes that a citation should include the following items:

- Author
- Title
- Distributor
- Date
- Version
- Persistent identifier

**Suggested Citation:**

S&P Dow Jones Indices LLC, *S&P 500 [SP500]*, retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SP500, June 26, 2020.

# Element of a (data) citation

ICPSR notes that a citation should include the following items:

- Author
- Title
- Distributor
- Date
- Version
- Persistent identifier

**Constructed Citation:**

Institute for Employment Research (IAB), Establishment History Panel 1975-2018. Accessed via the Research Data Centre (FDZ) of the German Federal Employment Agency DOI: 10.5164/IAB.BHP7518.de.en.v1 June 26, 2020.

# Element of a (data) citation

ICPSR notes that a citation should include the following items:

- Author
- Title
- Distributor
- Date
- Version
- Persistent identifier

**Constructed Citation:**

US Census Bureau, Longitudinal Business Database (LBD) 1975-2018. Last accessed via the Federal Statistical Research Data Centre (FSRDC) June 26, 2020.

# Data Availability Statement

- FRED data: URL is sufficient
- IPUMS data: Registration is required, query system requires users to provide details about variables, samples.
- IAB:

"This study uses the weakly anonymous Establishment History Panel 1975-2018. Data access was provided via on-site use at the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) and/or remote data access. DOI: 10.5164/IAB.BHP7518.de.en.v1"

# And we check them!

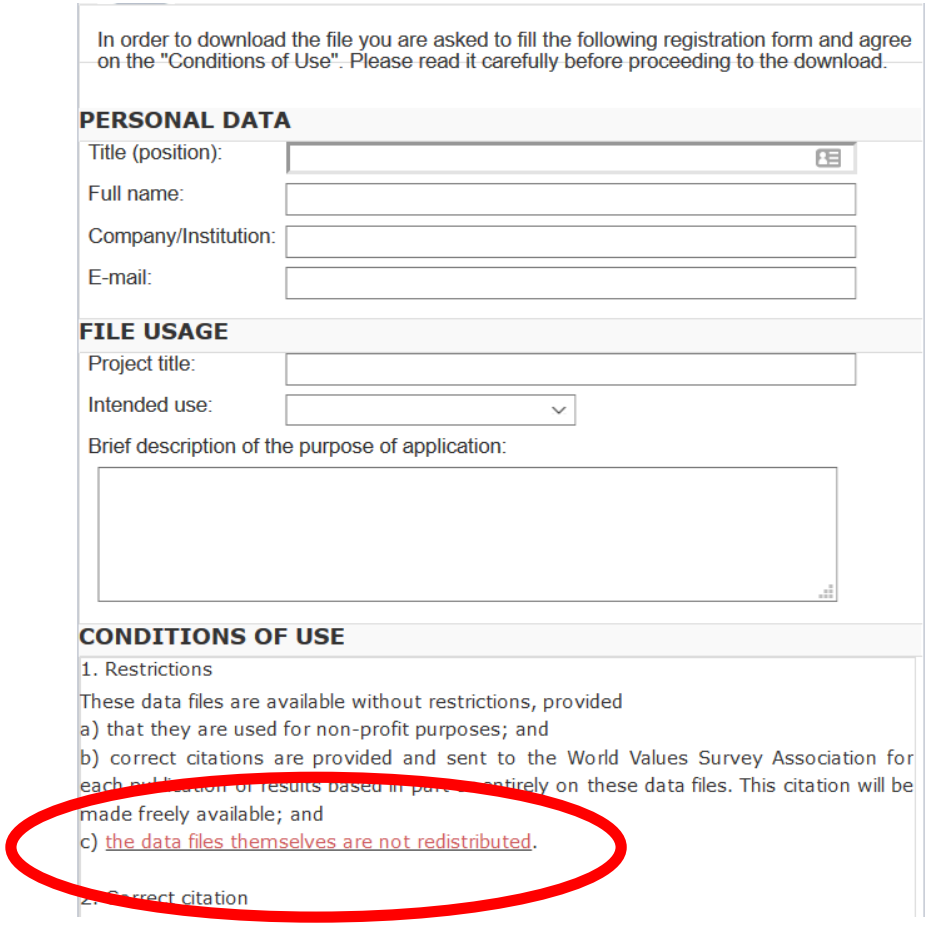
- If the URL does not work, we make a note.
- If the site requires registration, we try it out.
  - How long?
  - Any requirements?

- What does the site say?

Please use the following citation when referring to this file in the different versions:
Inglehart, R., C. Haerpfer, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin & B. Puranen et al. (eds.). 2014. World Values Survey: Round Six - Country-Pooled Datafile Version: www.worldvaluessurvey.org/WVSDocumentationWV6.jsp. Madrid: JD Systems Institute.

- Is that in the README / Paper/ Appendix?

# And we check them!



In order to download the file you are asked to fill the following registration form and agree on the "Conditions of Use". Please read it carefully before proceeding to the download.

**PERSONAL DATA**

Title (position):

Full name:

Company/Institution:

E-mail:

**FILE USAGE**

Project title:

Intended use:

Brief description of the purpose of application:

**CONDITIONS OF USE**

1. Restrictions

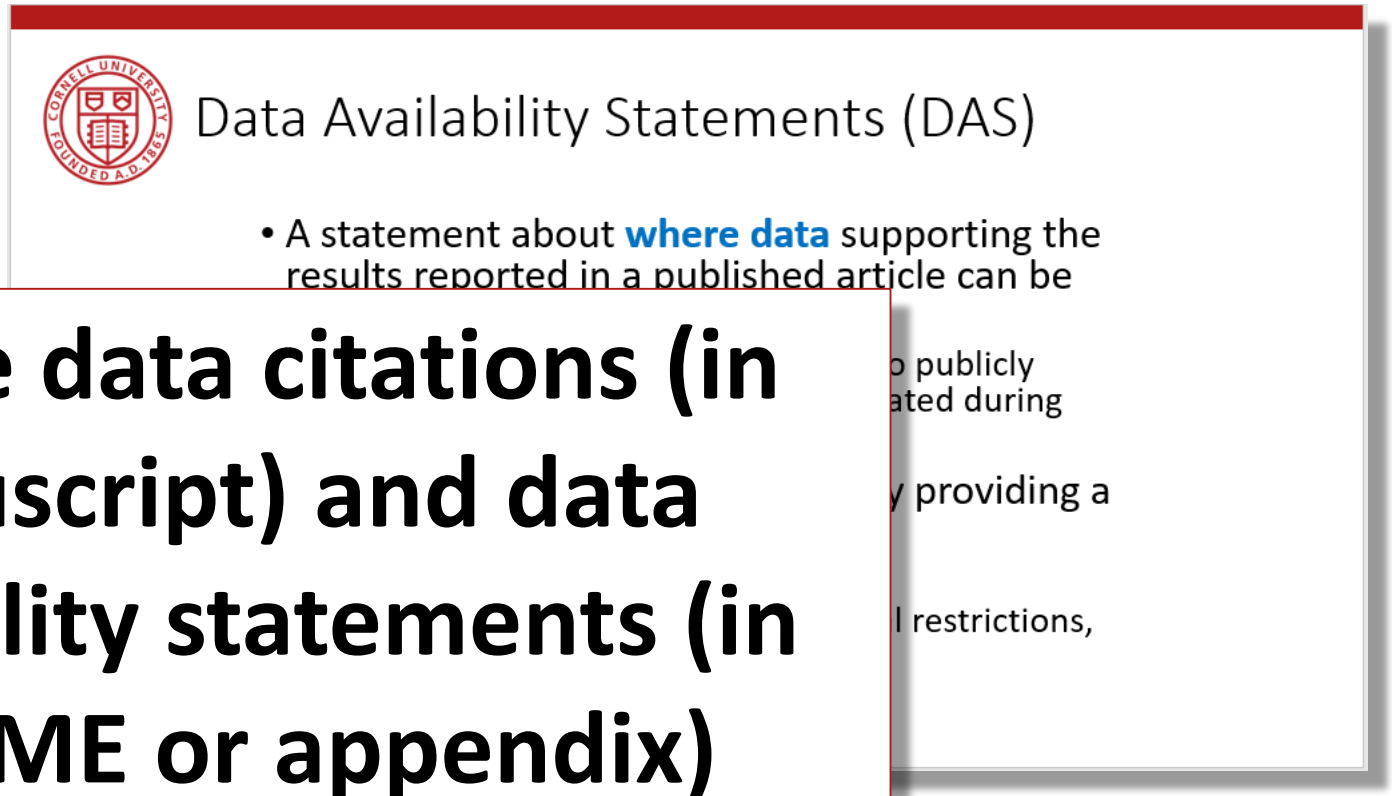
These data files are available without restrictions, provided
a) that they are used for non-profit purposes; and
b) correct citations are provided and sent to the World Values Survey Association for each publication of results based in part or entirely on these data files. This citation will be made freely available; and
c) the data files themselves are not redistributed.

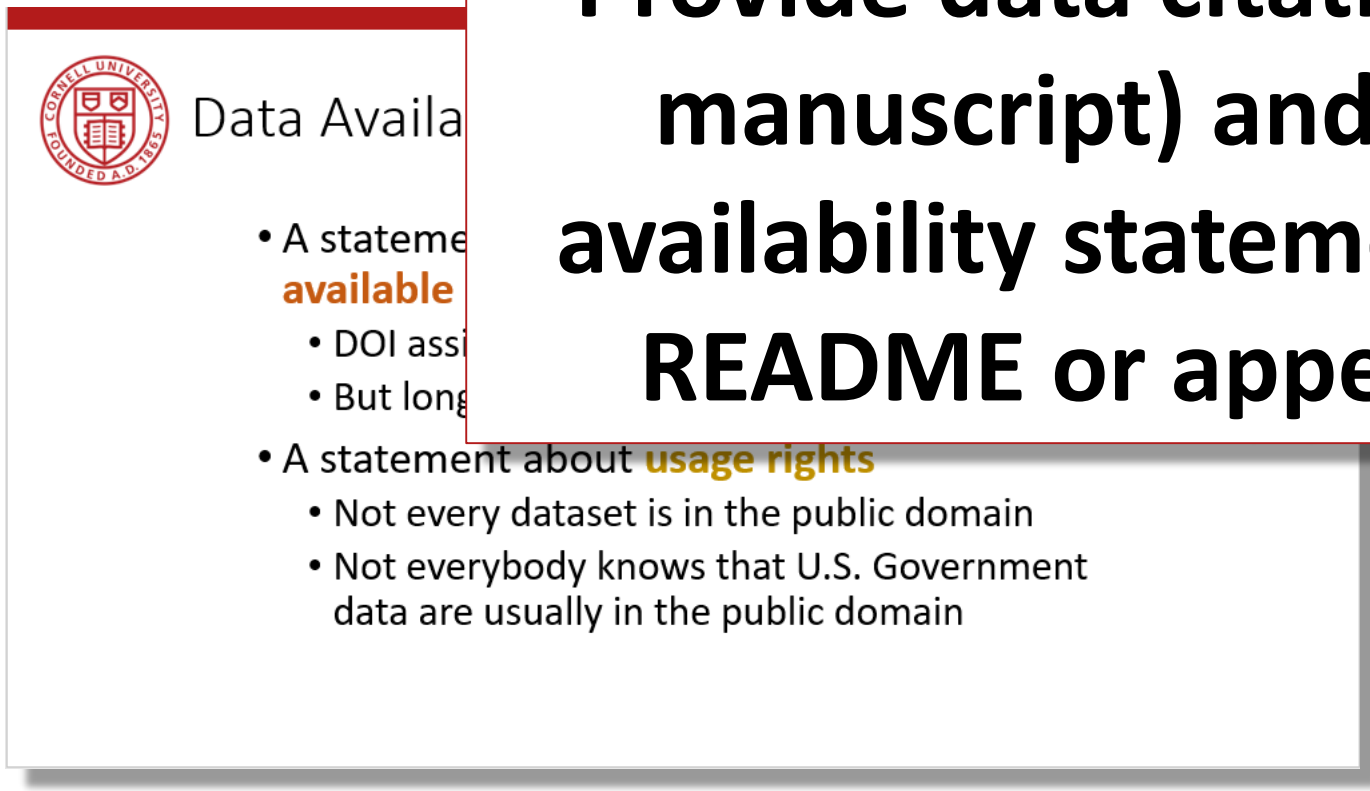
2. Correct citation

- What does the site say?

Please use the following citation when referring to this file in the different versions:
Inglehart, R., C. Haerpfer, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin & B. Puranen et al. (eds.). 2014. World Values Survey: Round Six - Country-Pooled Datafile Version: www.worldvaluessurvey.org/WVSDocumentationWV6.jsp. Madrid: JD Systems Institute.

- Is that in the README / Paper/ Appendix?
- Are all the conditions met/described?




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

# Take-away

# Data: Citations, Access, Rights

- Any data can be cited – even if you can't download it
- Any data that you accessed … can have that access be described
  - But caution: It should be such that others can also repeat the access!
- Just because you "have" the data does not mean you can give it to others
  - Also: distinguish between "sharing" and "publishing"
  - Know your terms of use!

# Coding for Reproducibility

# Follow 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)*


# Mindset

- **Assume nothing*** (fresh laptop, new software install, etc.)
- **Assume your new RA will run the analysis** (she knows Stata/Matlab, but doesn't know your directory setup, secret command-line-fu, which line to find the contents of Table 15, …)

# 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!

# Extreme examples

- Matlab-based simulation
- Real example, 10 figures, 4 panels each

- For Figure 5a, comment line 52, uncomment line 151, run the code, then copy the figure into your document.
- For Figure 5b, comment line 151 again, leave line 52 commented, and change the parameter on line 75 to “3”
- ….

# Extreme examples

- Stata-based estimation
- 4 variants

- Run the data creation programs, then copy the data to Folder A
- Copy programs “b.do” and “c.do” from Folder A to Folder B, but modify “c.do” on line 20
- Once done, convert the output from “d.do” to a Matlab file, and run the simulation in Folder B/C
- ….


# Ideal setup

- 1 program to prepare the setup
  - Installs all packages
  - Creates all directories
- 1 program (or a very small number) that creates the rest
  - Possibly with macros/ ado files/ subroutines
  - Possibly with parameter files that might differ per directory
- All tables and figures are output programmatically

- Setting up can be done in all languages
  - Matlab, Stata, R, Python, Fortran
- Subroutines exist in all languages
  - You might need to learn how!
- Ability to output figures and tables (Excel, LaTeX) exist in all languages

# How to prepare the replication package

- README
- Now ask an RA/ colleague/

## AEA Data and Code Guidance

AMERICAN ECONOMIC ASSOCIATION

Guidance for autho
data and code sup
replicators.

### Steps for the Third-party Replicator

- Download the author's replication archive(s) from the designated URL (public, or privately shared)
- Ensure access to any confidential files that are described in the replication archive's README
  - The replicator should consider whether a third-party person not
    ironment could reasonably
    olely on the instructions in the
- Follow the checklist to conduct the reproducibility exercise, relying exclusively on the README for instructions and guidance.
- Write a report
- Send the report to the AEA Data Editor
- Report any interactions with the author in the course of conducting the reproducibility exercise (help, assistance, clarifications)

That's our Protocol!

But… you just said I can’t give you the data – so … you can’t check my code..

# Tension between access and reproducibility

## 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

## 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

# 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

# Assume we can access the data

Sometimes we cannot

- We will still check if the code seems complete
- We will still verify that all data that *can* be provided have been provided

Sometimes we can:

- In the past 6 months, we have worked ourselves or with 3rd parties on
  - French, Brazilian, and US confidential admin data
  - Purchased commercial data
  - Proprietary data under NDA/DUA
  - Remotely or locally

# The role 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!

https://doi.org/10.5281/zenodo.3662906


# Some resources

- https://social-science-data-editors.github.io/guidance/
- https://aeadataeditor.github.io/aea-de-guidance/
  - **template README**
  - discussion of licensing
  - data citation guidance
- German example:
  - Establishment History Panel (BHP)  DOI: 10.5164/IAB.BHP7516.de.en.v1
- French verification service “cascad” within French RDC CASD
  - https://www.casd.eu/en/le-centre-dacces-securise-aux-donnees-casd/certification-de-resultats-cascad-casd/