



Implementing Increased Transparency and Reproducibility in Economics: Lessons learned

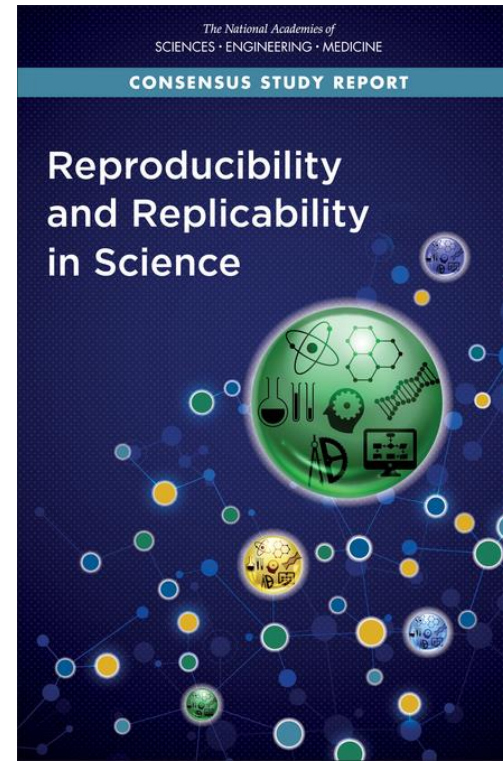
Lars Vilhuber
Cornell University

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.



Replication continuum

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



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)

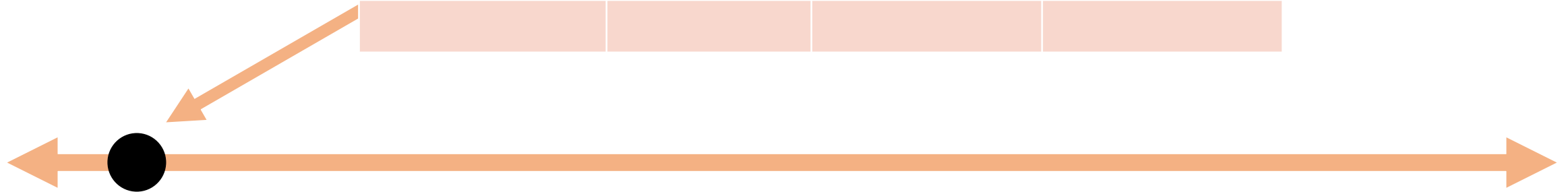
Generalizability

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



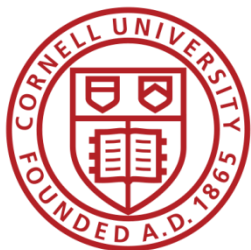
Replication continuum

Same data	Same code	Same methods	Same context



Reproducibility

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



Progress

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

The logo for IPUMS USA, featuring a stylized American flag icon to the left of the text "IPUMS USA".

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

[VISIT SITE](#)



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

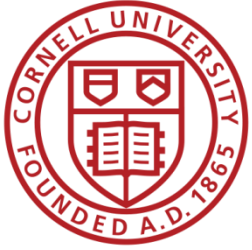




Progress

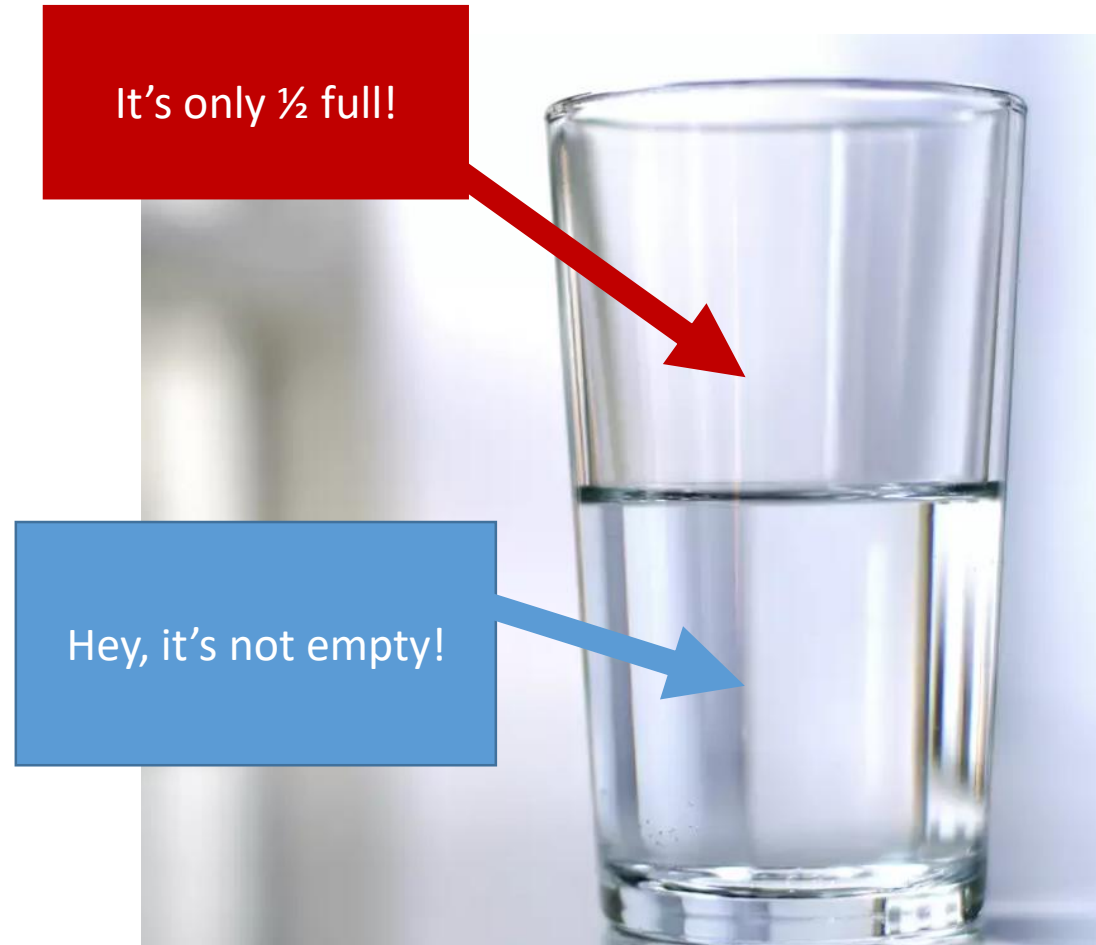
- Replication archives and Data (Code) Availability policies
- Shared open source software
- Better public-use and shared confidential data
- History of accessing preprints/ grey literature

RePEc

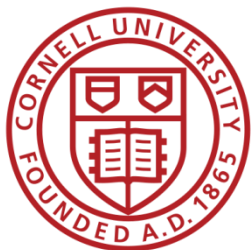


In a nutshell

- **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



404. That's an error.

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





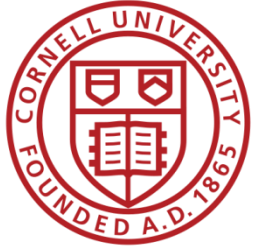
Poor coding practices

- **Manual/non-automation**

Code produces no meaningful output

- **Lack of robustness:**

Bugs in the code



Poor citation practices

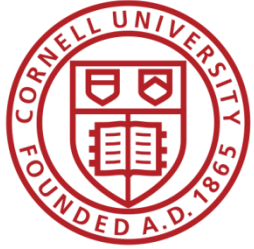
- **Macrodata:**

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

- **Microdata:**

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

Actions



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

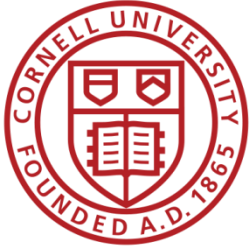


American Economic Journal: Macroeconomics focuses on studies of aggregate fluctuations and growth, and the role of policy in that context.

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.



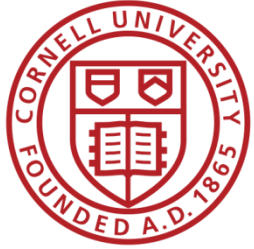
AEA Data & Code Availability Policy (2019)

- 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*



Action: Reproducibility Check



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:

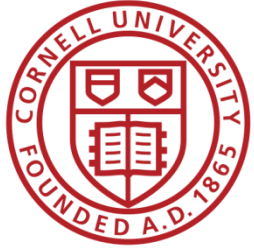
- [Overview](#)
- [Review the README file](#)
- [For each listed data source](#)
- [For each listed table, figure, in-text number](#)
- [Conduct a code verification, if data is available](#)
- [Examples](#)

Overview

This document describes

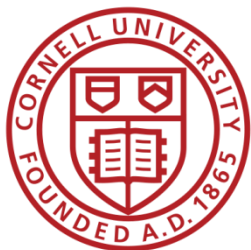
- what authors should check before providing data to journals
- what verifier teams should check for in the data submitted to them for the purpose of verification





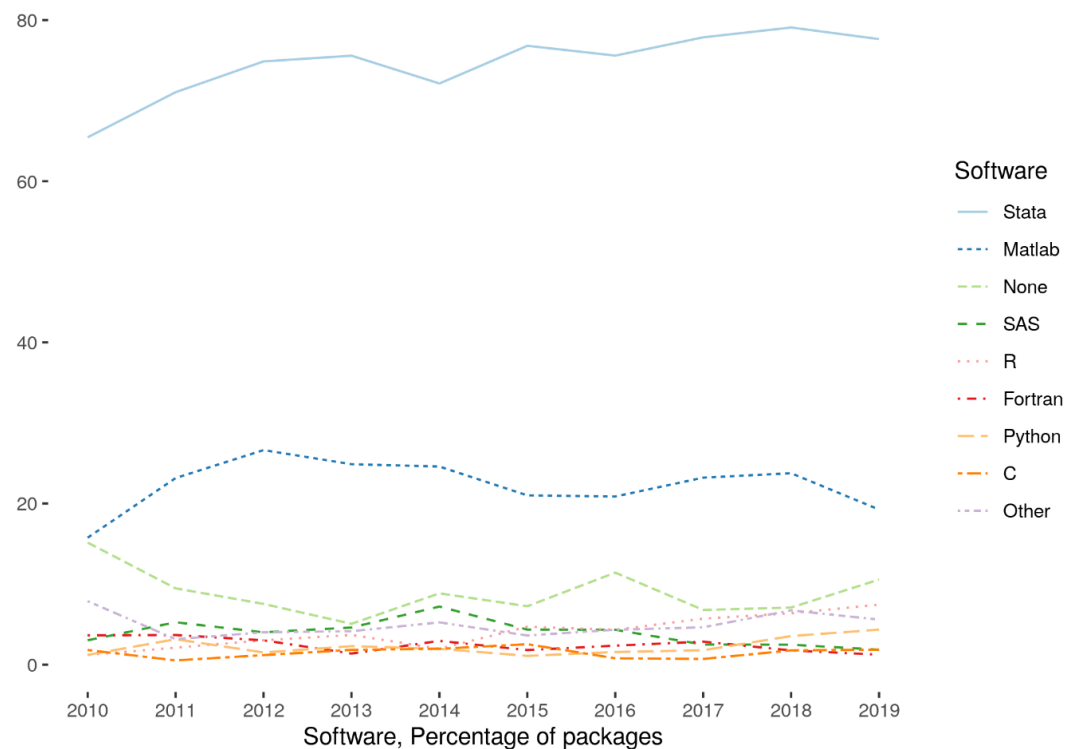
Who is doing that?

- Earlier reproducibility work: **Flavio Stanchi** (now at AirBnb), **Sylvérie 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



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%**)



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

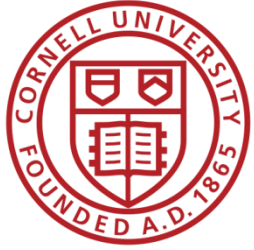
[MC number] [Manuscript Title] Validation and Replication results

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 July 16, 2020 (1 year), the AEA Data Editor team conducted

- **866 assessments**
- Of which **277 manuscripts** have been “accepted”



Also work with 3rd parties

- cascada – Using confidential data!
- CISER (R-Squared)
- Various contributors with access to confidential data
 - Authors
 - Their institutions
 - Network of known groups with access (less frequent)



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



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

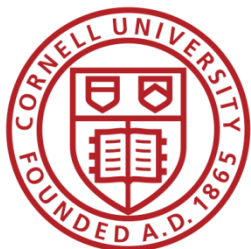
RESULTS REPRODUCTION (R-SQUARED)

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

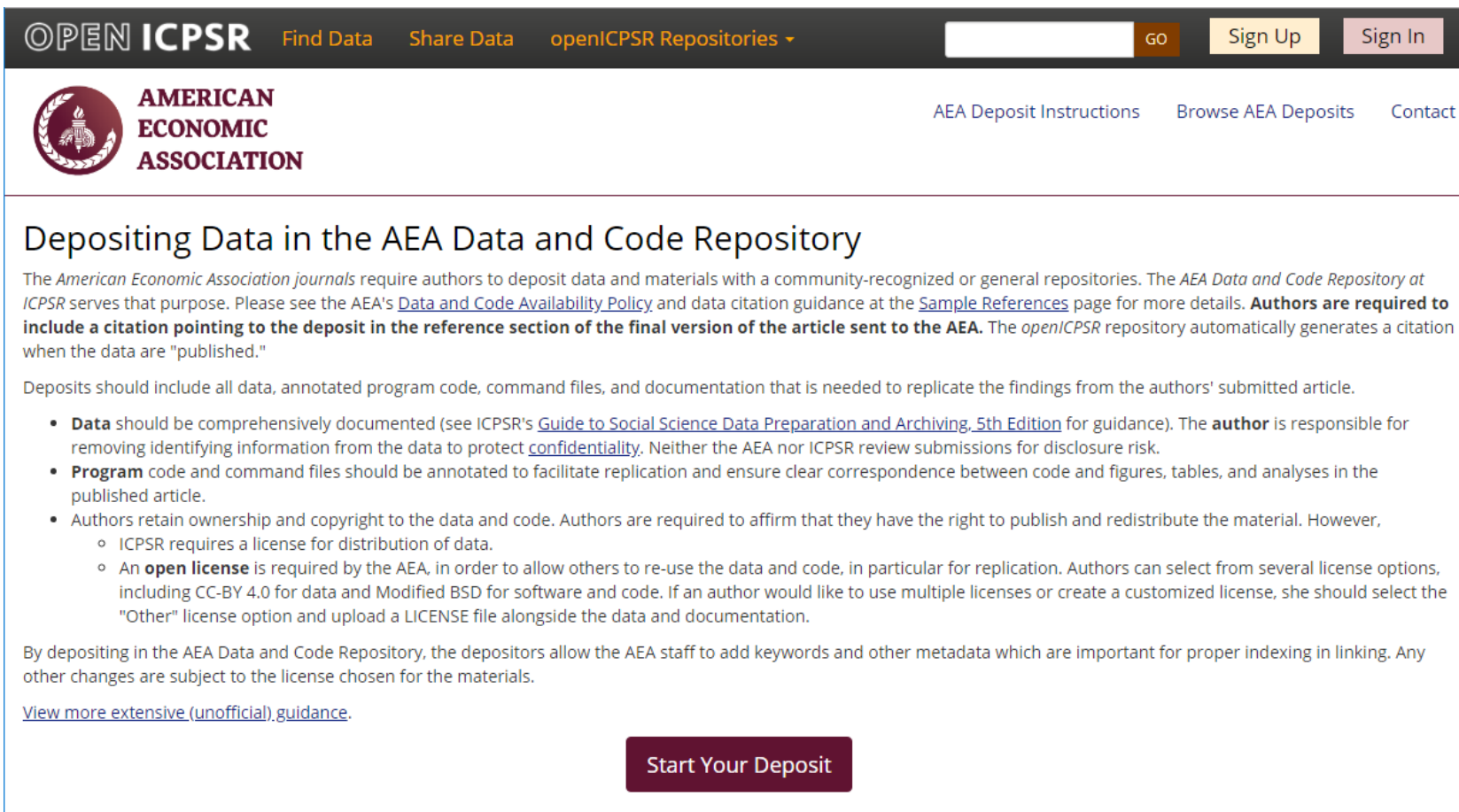


Current efforts at the AEA

- **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



The screenshot shows the top navigation bar of the openICPSR website. It includes the "OPEN ICPSR" logo, navigation links for "Find Data", "Share Data", and "openICPSR Repositories", a search bar with a "GO" button, and "Sign Up" and "Sign In" buttons. Below this is the American Economic Association (AEA) logo and text, along with links for "AEA Deposit Instructions", "Browse AEA Deposits", and "Contact". The main content area is titled "Depositing Data in the AEA Data and Code Repository" and contains a paragraph of introductory text, a list of requirements for deposits, and a "Start Your Deposit" button.

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."

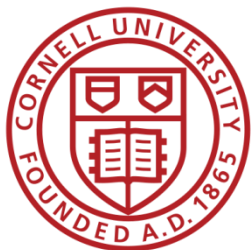
Deposits should include all data, annotated program code, command files, and documentation that is needed to replicate the findings from the authors' submitted article.

- **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](#)



Action: Data citations and metadata

What is **FAIR**?

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

A screenshot of the FORCE11 website. The logo "FORCE11" is at the top, with the tagline "The Future of Research Communications and e-Scholarship" below it. A navigation bar contains "ABOUT", "COMMUNITY", and "CODE OF CON". The breadcrumb trail reads "FORCE11 » Groups » The FAIR Data Principles". The main heading is "THE FAIR DATA PRINCIPLES". Below it is a link "JOIN IN THE DISCUSSION - LEA FAIR Data Principles" and a section titled "Preamble" with the text "One of the grand challenges of data-intensiv".

FORCE11
The Future of Research Communications and e-Scholarship

ABOUT ▾ COMMUNITY ▾ CODE OF CON

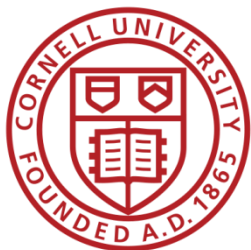
FORCE11 » Groups » The FAIR Data Principles

THE FAIR DATA PRINCIPLES

[JOIN IN THE DISCUSSION - LEA FAIR Data Principles](#)

Preamble

One of the grand challenges of data-intensiv



FAIR data principles rely on metadata

— Scope of Project

Subject Terms ?

Do not copy/paste multiple terms into this field. Terms must be entered individually.

Russia Industry Factories Russian Empire Corporations

JEL Classification ?

L20 General N63 Europe: Pre-1913 O43 Institutions and Growth

Manuscript Number ?

AER-2015-1656.R3 [edit](#) [remove](#)

Geographic Coverage ? [+ add value](#)

European Russia (Russian Empire) [edit](#) [remove](#)

Time Period(s) ? [+ add value](#)

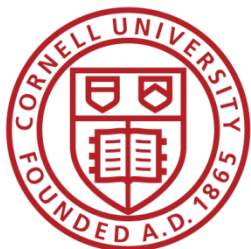
1894 – 1908 (Three years: 1894, 1900, and 1908) [edit](#) [remove](#)

Collection Date(s) ? [+ add value](#)

Universe ?

Manufacturing establishments in the European part of the Russian Empire. [edit](#) [remove](#)

Data Type(s) ?



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:

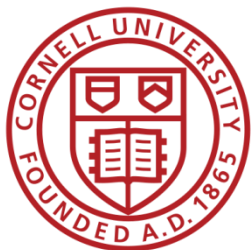
Gregg, Amanda. Imperial Russian Factory Database, 1894-1908. Nashville, TN: American Economic Association [publisher], 2020. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020-01-29. <https://doi.org/10.3886/E110681V1>

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>.



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



```

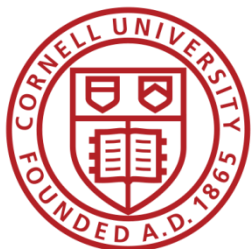
<meta name="DC.identifier" content="10.3886/E110681V1" />
<meta name="DC.title" content="Imperial Russian Factory Database, 1894-1908" />

  <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" />

```

	1908MicroData.xlsx	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
	AG_Corp_Prod_Database.dta	application/x-stata	11.9 MB	10/08/2014



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



<script type="application/ld+json">

```
{
  "name": "Imperial Russian Factory Database, 1894-1908",
  "identifier": "http://doi.org/10.3886/E110681V1",
  "description": "This database dig...
  manufacturing censuses. For each factory, the database includes industry, province, enterprise form, total workers, total revenue, and identifiers that...
  .1908 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.&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": ["Middlebur...
  "name": "openICPSR Self-Deposit Archive", "url": "http://www.openicpsr.org/", "@type": "DataCatalog"},
  "funder": [{"name": "Economic History Association", "@ty...
  "directorate for Social, Behavioral and Economic Sciences", "@type": "Organization"},
  {"name": "Yale Economic Growth Center", "@type": "Organization"},
  {"name...
  "fund", "@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/V1...
  stata", "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.org/openicpsr/project/110681/version/V1/download/terms?path=/openicpsr/110681/fcr:versions/V1/AG_Corp_Prod_Database.dta&type=application/x-stata", "encodingFormat": "application/zip"},
  {"file...
  'openicpsr/project/110681/version/V1/download/terms?path=/openicpsr/110681/fcr:versions/V1/AG_Corp_Prod_Database_withAksiz.dta&type=application/x-sta...
  "fileFormat": "stata", "contentURL": "https://www.openicpsr.org/openicpsr/project/110681/version/V1/download/terms?path=/openicpsr/110681/fcr:versions/V1...
  stata", "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
		application/x-stata	11.9 KB	10/08/2014



... and findability relies on metadata

Google

imperial russian factory



1 dataset found



Imperial Russian Factory
Database, 1894-1908

www.openicpsr.org
search.datacite.org
+1 more

stata

Updated Jan 29, 2020



Not seeing a result you expected?
[Learn](#) how you can add new
datasets to our index.



AMERICAN
ECONOMIC
ASSOCIATION

Imperial Russian Factory Database, 1894-1908

[Explore at openICPSR](#)

[Explore at search.datacite.org](#)

[Explore at www.da-ra.de](#)

2 scholarly articles cite this dataset ([View in Google Scholar](#))

stata

Unique identifier

<https://doi.org/10.3886/E110681V1>

Dataset updated Jan 29, 2020

Dataset provided by

[American Economic Association](#)

Authors

Amanda Gregg

License

[Attribution 4.0 \(CC BY 4.0\)](#)

License information was derived automatically

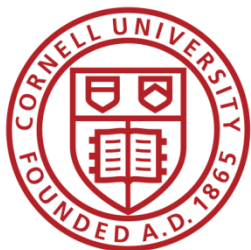
Area covered

European Russia (Russian Empire)



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



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 citable research objects, such as publications[1].

DC¹

Data Citation Principles

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].

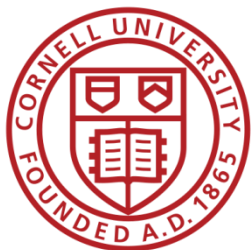
4. Unique Identification

A data citation should include a persistent method for identification that is 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>].



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 citable research objects, such as publications[1].



Data Citation Principles

2. Credit and Attribution

Data citations should facilitate citation of data and attribution of credit.

1 Bureau of Labor Statistics. 2000–2010. “Current Employment Statistics: Colorado, Total Nonfarm, Seasonally adjusted - SMS080000000000000001.” United States Department of Labor. <http://data.bls.gov/cgi-bin/surveymost?sm+08> (accessed February 9, 2011).

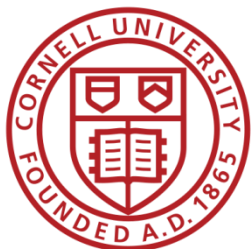
In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited[3].

4. Unique Identification

A data citation should include a persistent method for identification that is 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 citations

- Creating specific guidance in the absence of strong discipline-specific guidance



Data and Code Guidance by Data Editors

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

Guidance on Data Citations

On this page:

- Better
- Websites
- Online databases
- Data distributed as supplementary data
- Producer
- Distributor
- Dates
- Offline access mechanism
- Confidential databases
- No formal access mechanism

One of the most vexing issues is how to cite data. This document goes through a few common scenarios not covered elsewhere.

What is not a data citation

Many authors initially neglect to add data citations, or do not know how to add a data citation. Often, we see authors cite papers with supplementary data, but not databases or other data:

<https://social-science-data-editors.github.io/guidance/addtl-data-citation-guidance.html>

Some practical tips
(based on 800+
assessments)



Provenance!

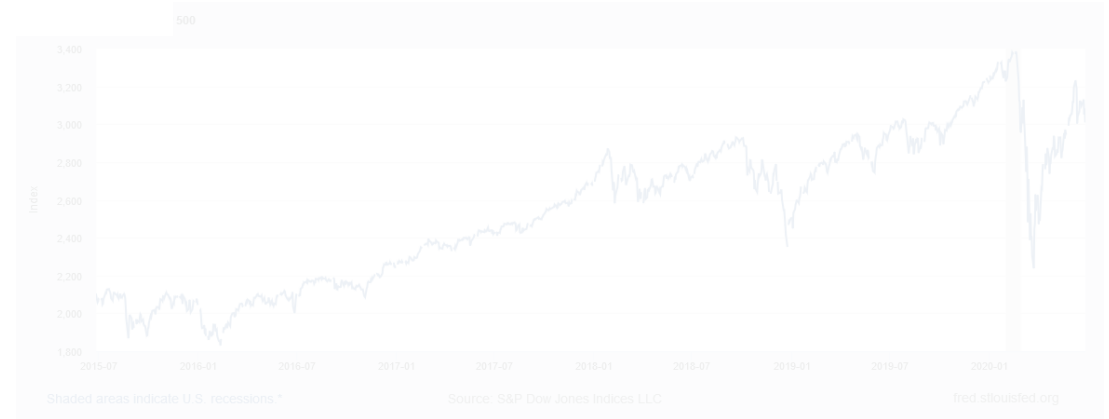
- Does the provider have a good record?
- Where do the data come from?
- What do we know about the data?

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



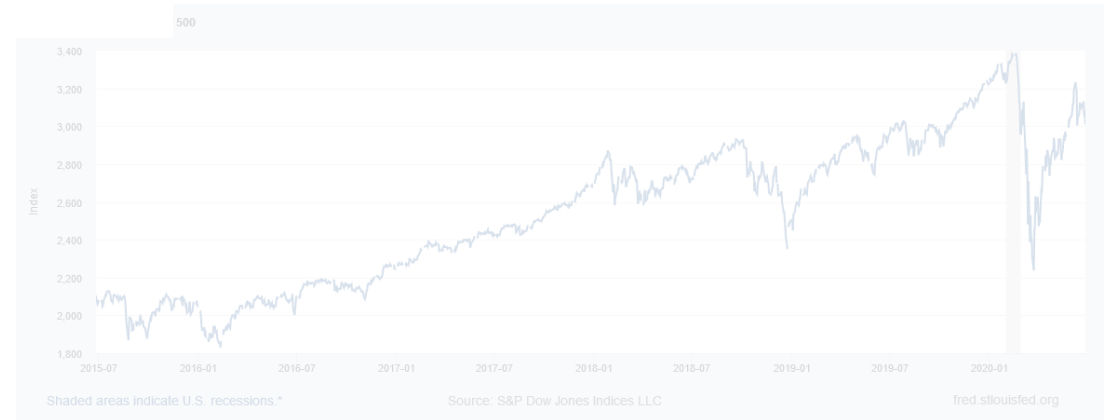


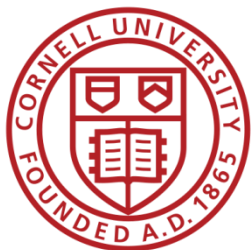
“It’s a file called SP500.xlsx”

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



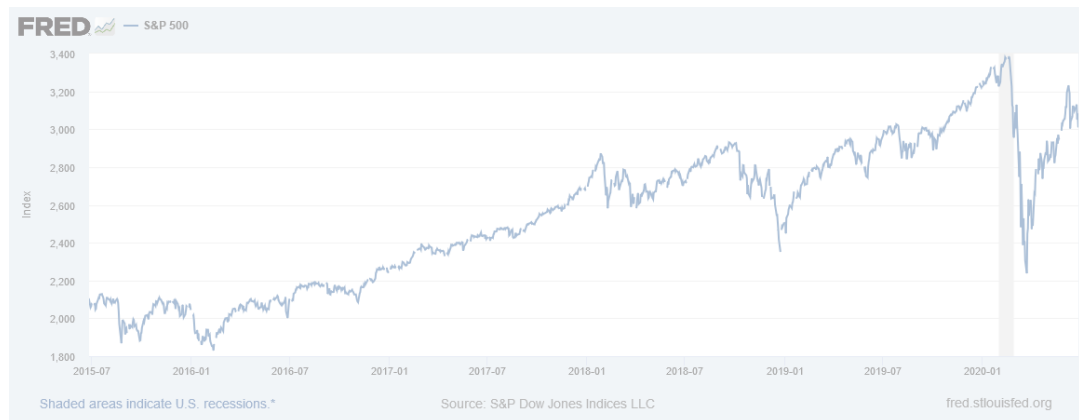


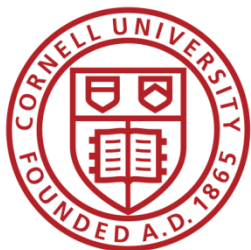
“It’s a file called SP500.xlsx, downloaded from FRED.”

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.



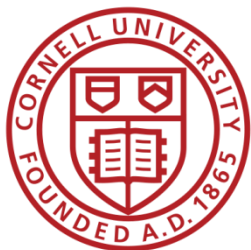


“SP500.xlsx, from S&P (2020). Not provided as part of replication package because © S&P.”

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.”



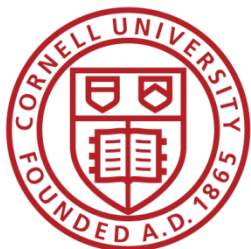
S&P 500, Index, Daily,
Not Seasonally Adjusted

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

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

2015-07-08	2101.49
2015-07-09	2057.64
2015-07-10	2063.11
2015-07-13	2077.12
2015-07-14	2076.78
2015-07-15	2077.00
2015-07-16	2068.76
2015-07-17	2081.34
2015-07-20	2046.68
	2051.31
	2076.62
	2099.60
	2108.95
	2107.40
	2124.29
	2126.64
	2128.28

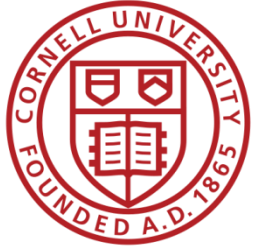
S&P Dow Jones Indices LLC. 2020. “S&P 500 [SP500] [dataset]”, retrieved from FRED, Federal Reserve Bank of St. Louis; <https://fred.stlouisfed.org/series/SP500>, June 26, 2020.





Citing restricted-access data

“Well, I can’t download the data, so I can’t cite it.”



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?
- Is there additional information access?
 - → URL leads to exact data?
 - → URL leads to application procedure?
 - → other access procedure is described?





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



[Home](#)

[Methodology](#)

[Media](#)

[Research & Applications](#)

[About Us](#)

EPU Indices

[All Country-Level Data](#)

Global	USA
Australia	Brazil
Canada	Chile
China	Colombia
Croatia New	France
Germany	Greece
Hong Kong	India
Ireland	Italy
Japan	South Korea

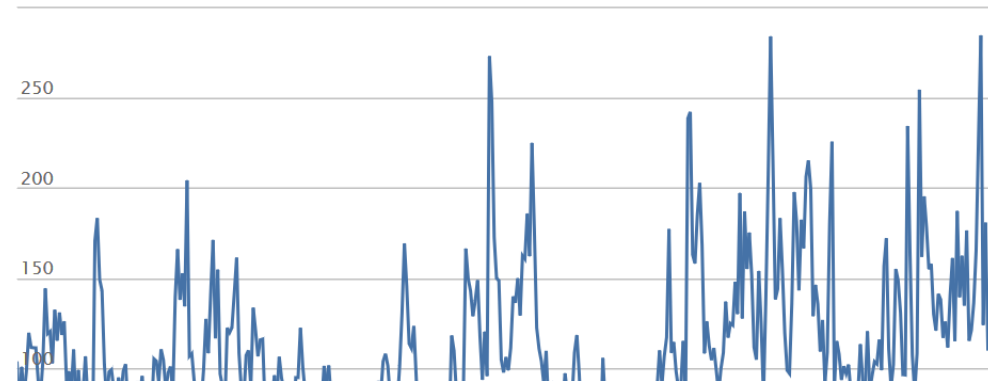
Economic Policy Uncertainty Index

We develop indices of economic policy uncertainty for countries around the world.

Monthly US Economic Policy Uncertainty Index



Zoom [1m](#) [3m](#) [6m](#) [1y](#) [7y](#) [All](#)





Example 2: Academic data publisher



https://www.policyuncertainty.com/index.html

Go

SEP DEC JAN

103 captures

18 Aug 2012 - 14 Dec 2019



2018 2019 2020



Home

Methodology

Media

Research & Applications

About Us

EPU Indices

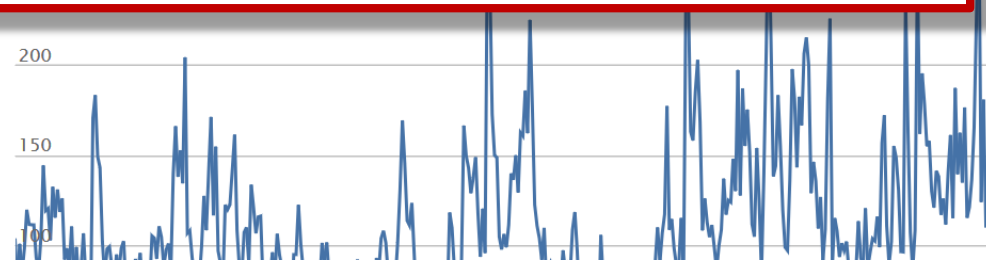
All Country-Level Data

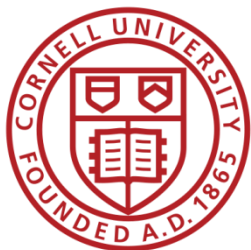
- [Global](#)
- [Australia](#)
- [Canada](#)
- [China](#)
- [Croatia](#)
- [France](#)
- [Germany](#)
- [Greece](#)
- [Hong Kong](#)
- [India](#)
- [Ireland](#)
- [Italy](#)
- [Japan](#)
- [South Korea](#)

Economic Policy Uncertainty Index


We develop indices of economic policy uncertainty for countries around the world.

© 2012-2018 by Economic Policy Uncertainty





Example 2: Academic data publisher-new!

 **ECONOMIC POLICY UNCERTAINTY**

[Home](#) [Methodology](#) [Media](#) [Research & Applications](#) [About Us](#)

EPU Indices


[All Country-Level Data](#)

[Global](#) [USA](#)

Economic Policy Uncertainty Index

We develop indices of economic policy uncertainty for countries around the world.

[Monthly US Economic Policy Uncertainty Index](#)



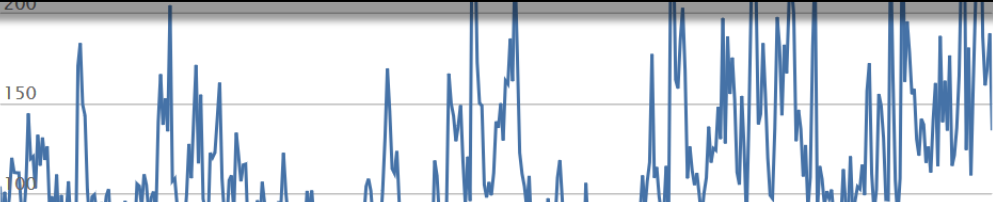
This work is licensed under a [Creative Commons Attribution 4.0 International License](#)

[Germany](#) [Greece](#)

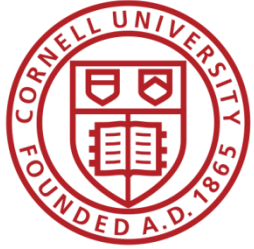
[Hong Kong](#) [India](#)

[Ireland](#) [Italy](#)

[Japan](#) [South Korea](#)



The chart displays the Economic Policy Uncertainty Index over time, with a y-axis ranging from 100 to 200. The index shows significant volatility, with major peaks around 2008 and 2014-2015, and a notable dip around 2011. The data points are connected by a blue line.



Rights to use data

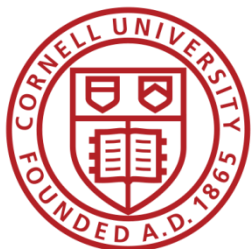
- You browsed a website
- You purchased the data
- You signed a data use agreement
- You created the data (lab experiment)
- You had survey respondents consent to use (IRB approval!)



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.**



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](#)

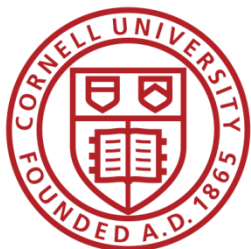
[Figures of the FDZ](#)

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](https://doi.org/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](https://doi.org/10.5164/IAB.FDZD.2001.de.v1)
-  [FDZ-Datenreport 01/2020](#)
-  [Fallzahlen und Labels](#)

English

- DOI: [10.5164/IAB.FDZD.2001.en.v1](https://doi.org/10.5164/IAB.FDZD.2001.en.v1)



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.



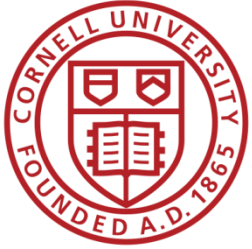
Element of a (data) citation

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.



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?

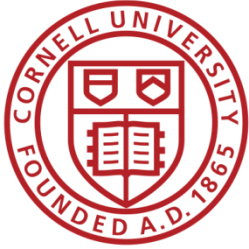
Please use the following citation when referring to this file in the different versions:

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 or 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?

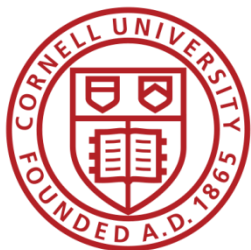
Please use the following citation when referring to this file in the different versions:

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?



Data Availability Statements (DAS)

- A statement about **where data** supporting the results reported in a published article can be

to publicly
ated during
y providing a
restrictions,

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



Data Availa

- A statement **available**
 - DOI ass
 - But long
- A statement about **usage rights**
 - Not every dataset is in the public domain
 - Not everybody knows that U.S. Government data are usually in the public domain

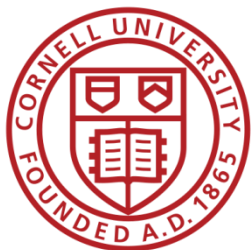
Take-away




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



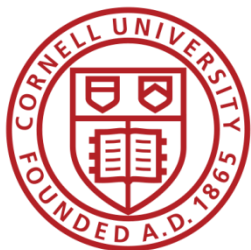
Streamlining replication packages

- 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!



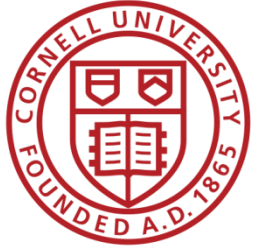
Extreme examples

- 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"
-



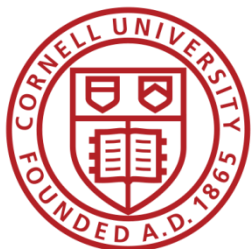
Extreme examples

- 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



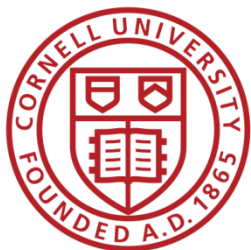
AMERICAN
ECONOMIC
ASSOCIATION

Guidance for authors
data and code sup
replicators.


That's our Protocol!

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 familiar with the replication environment could reasonably rely solely 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)




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



It's only 3/4 full!

Hey, it's not empty!

The image shows a glass of water. A red arrow points to the water level with the text "It's only 3/4 full!". A blue arrow points to the empty space above the water with the text "Hey, it's not empty!".



Verification services

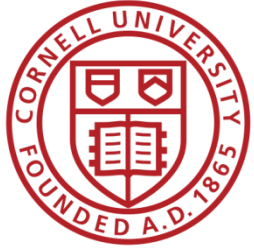
 **cascad**
*the first certification
agency for scientific
code & data*

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



The screenshot shows the CASD website interface. At the top, there is a dark navigation bar with a hamburger menu icon on the left, the 'CASD' logo in the center, and the words 'PROJETS' and 'DONNÉES' on the right. Below the navigation bar, the main content area has a dark purple background with a network-like pattern of white lines and dots. The text 'Secure Data Hub' is prominently displayed in white, with a small network icon to its right. Below this, there are several lines of text representing different categories and their associated counts:

- Travail, Emploi / 189
- Société, Justice, Éducation / 113
- Économie, Entreprises, Finance / 267
- Environnement, Agriculture / 187
- Santé / 244



Assume we can access the data

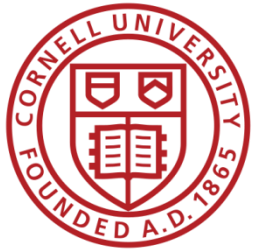
Sometimes we cannot

- 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 with
 - French, Brazilian, and US confidential admin data
 - Purchased commercial data (Twitter, Indian GDP)
 - Proprietary data under NDA/DUA (Ebay)
 - Data with application procedure (Chinese Panel, Demographic and Health Survey, European establishment data)
 - Remotely or locally

The role for journals



Goal: Transportability

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



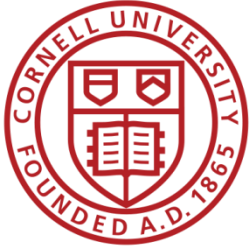
Social science “guild”



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

Thank you!

<https://doi.org/10.5281/zenodo.3981458>



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](https://doi.org/10.5164/IAB.BHP7516.de.en.v1)
- French verification service “cascad” within French RDC CASD
 - <https://www.casd.eu/en/le-centre-dacces-secure-aux-donnees-casd/certification-de-resultats-cascad-casd/>

