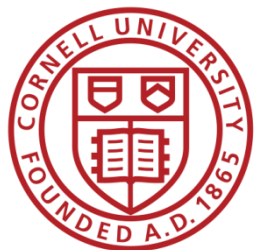


# Implementing Increased Transparency and Reproducibility in Economics: Lessons learned

Lars Vilhuber  
Cornell University

Original version: <http://doi.org/10.5281/zenodo.3911311>

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.



Would you buy a car from this guy?





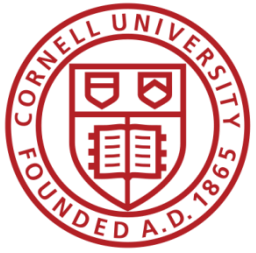
# Provenance!

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

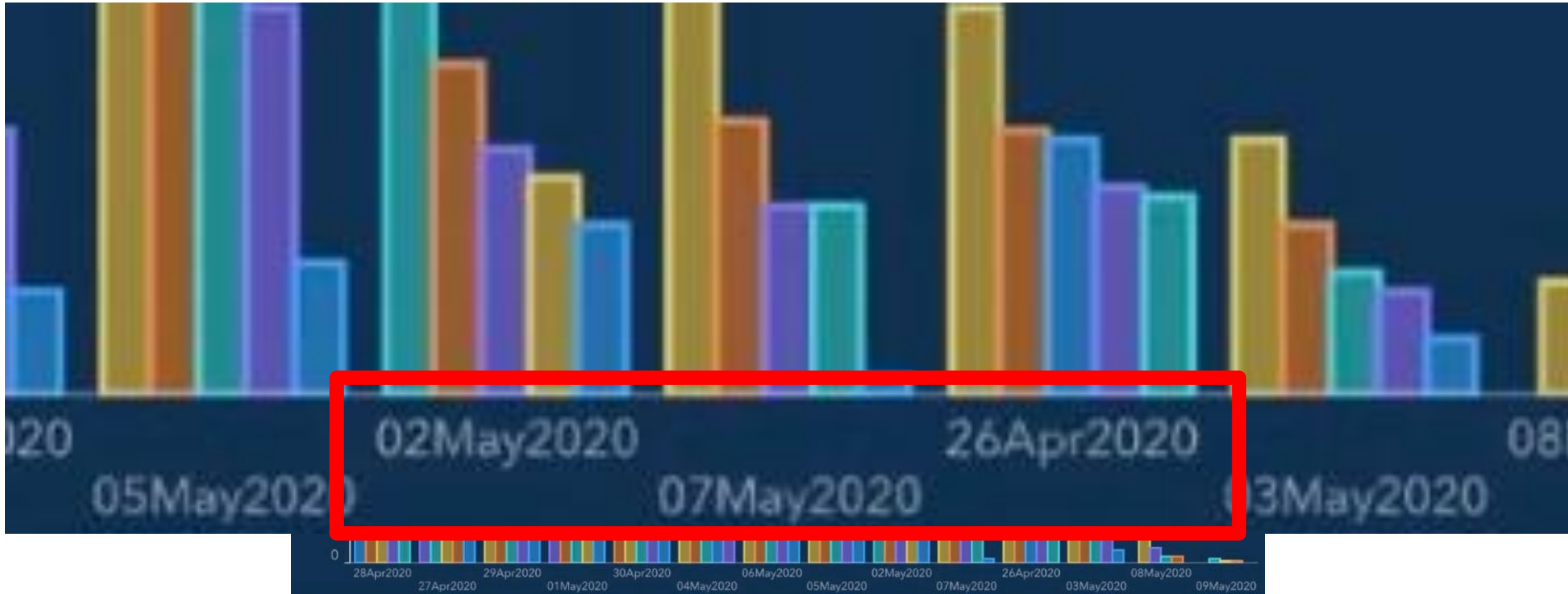


# Would you use this data?

```
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!

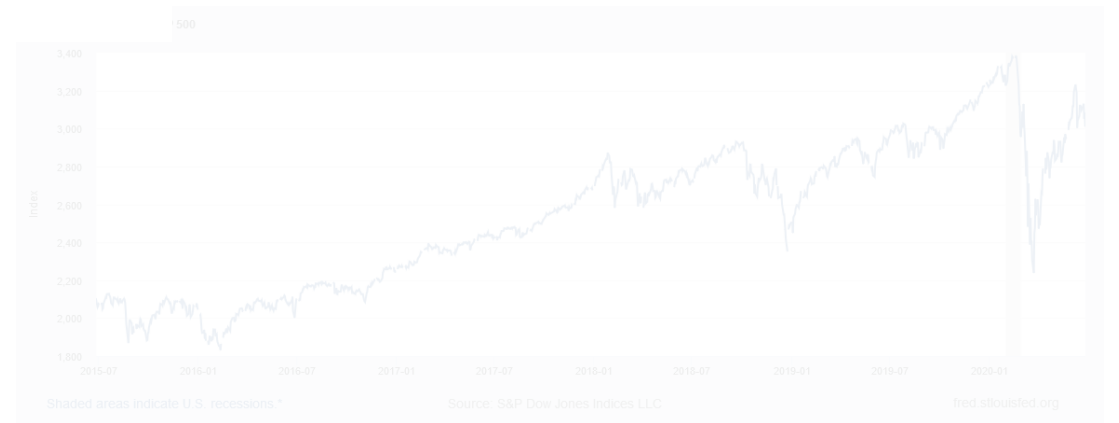
- Does the provider have a good record?
- Where do the data come from?
- What do we know about the data?

# Metadata!

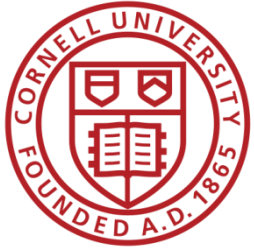


“It’s a file called stockmarket.xlsx”

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

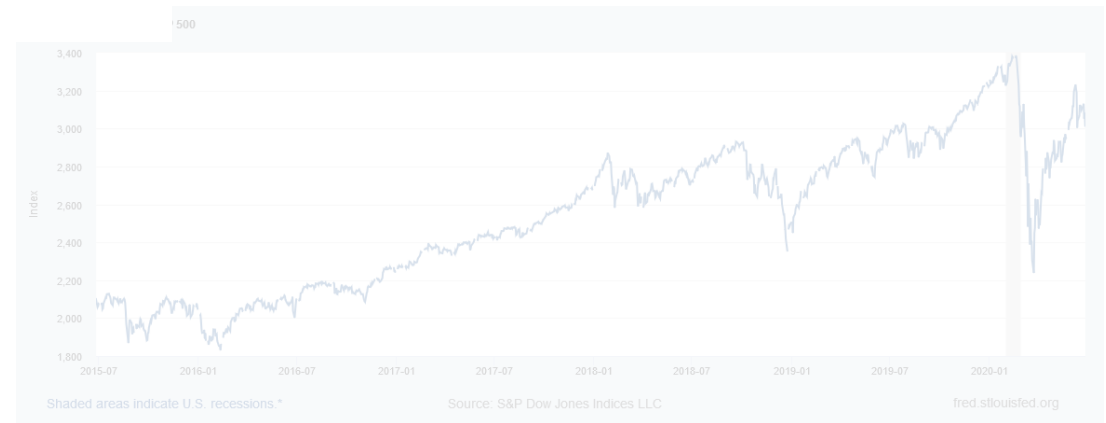
S&P 500, Index, Daily,  
Not Seasonally Adjusted

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

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.



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



# 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	2075.12
2015-07-14	2076.78
2015-07-15	2076.78
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.





# 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?





# 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



**ECONOMIC POLICY UNCERTAINTY**

[Home](#)

[Methodology](#)

[Media](#)

[Research & Applications](#)

[About Us](#)

## EPU Indices

[All Country-Level Data](#)

<a href="#">Global</a>	<a href="#">USA</a>
<a href="#">Australia</a>	<a href="#">Brazil</a>
<a href="#">Canada</a>	<a href="#">Chile</a>
<a href="#">China</a>	<a href="#">Colombia</a>
<a href="#">Croatia</a> <b>New</b>	<a href="#">France</a>
<a href="#">Germany</a>	<a href="#">Greece</a>
<a href="#">Hong Kong</a>	<a href="#">India</a>
<a href="#">Ireland</a>	<a href="#">Italy</a>
<a href="#">Japan</a>	<a href="#">South Korea</a>

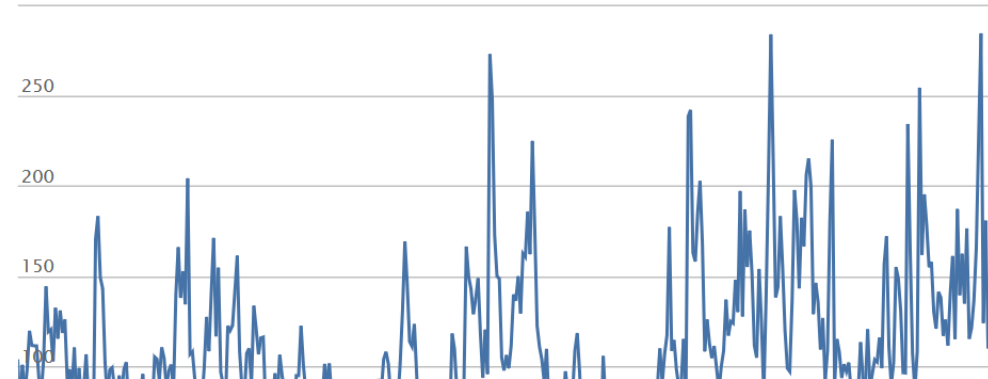
## Economic Policy Uncertainty Index

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

INTERNET ARCHIVE  
WayBack Machine

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

Go

SEP DEC JAN

[103 captures](#)

18 Aug 2012 - 14 Dec 2019

14  
2018 2019 2020



ECONOMIC POLICY UNCERTAINTY

[Home](#)

[Methodology](#)

[Media](#)

[Research & Applications](#)

[About Us](#)

## EPU Indices

[All Country-Level Data](#)

[Global](#)

[Australia](#)

[Canada](#)

[China](#)

[Croatia](#)

[Germany](#)

[Hong Kong](#)

[Ireland](#)

[Japan](#)

[France](#)

[Greece](#)

[India](#)

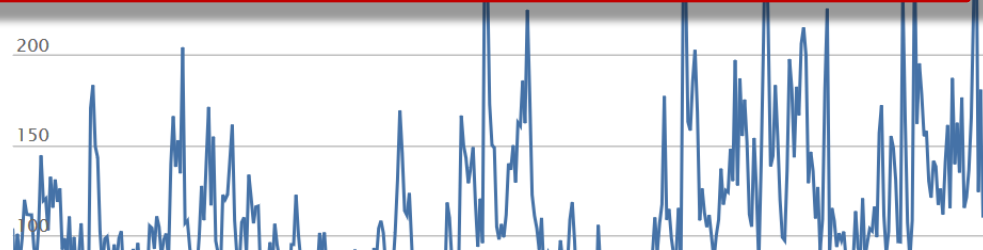
[Italy](#)

[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!





## 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.

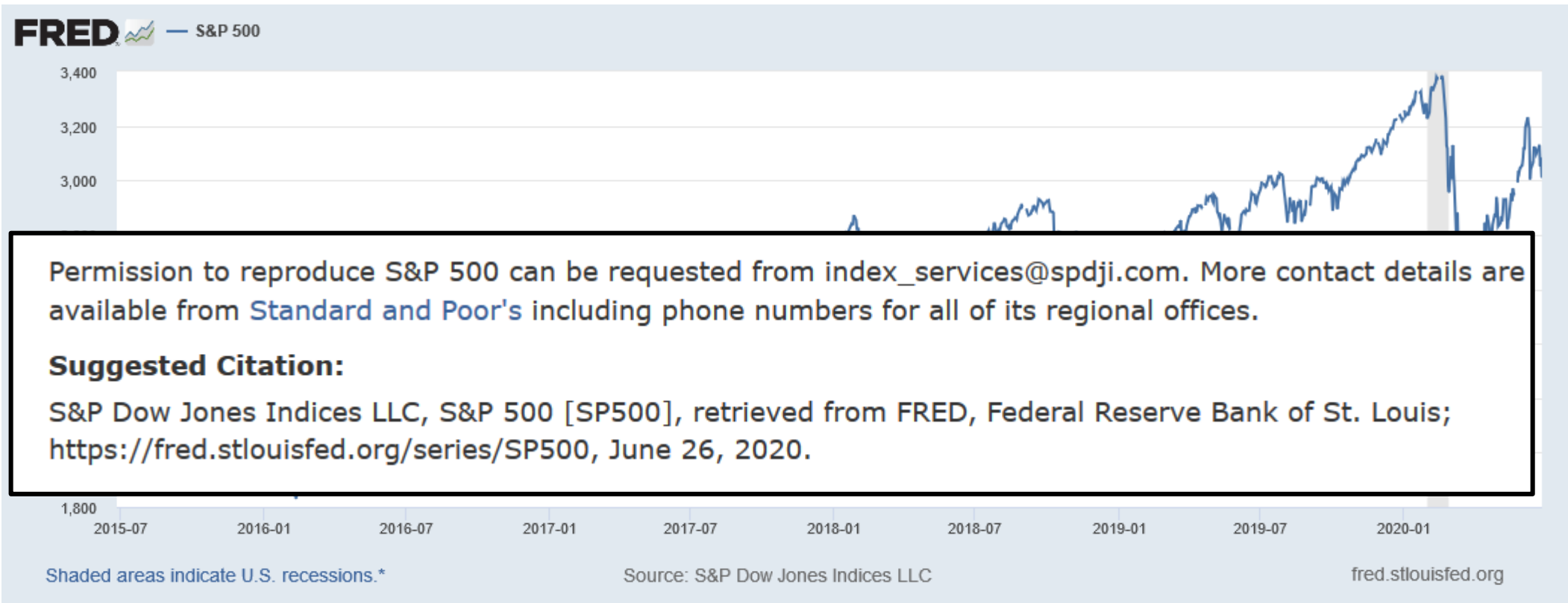


## Response

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.



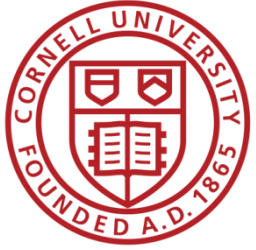
# Example 3: FRED (St. Louis Fed) and SP500





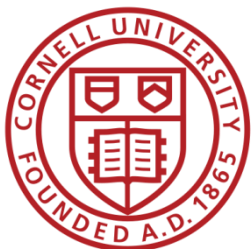
# Restricted-access data

- 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)
<b>BHP 7518 v1 (current)</b>	<a href="#">10.5164/IAB.BHP7518.de.en.v1</a>	2020-01-13
<b>BHP 7517 v1</b>	<a href="#">10.5164/IAB.BHP7517.de.en.v1</a>	2018-12-12
<b>BHP 7516 v1</b>	<a href="#">10.5164/IAB.BHP7516.de.en.v1</a>	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

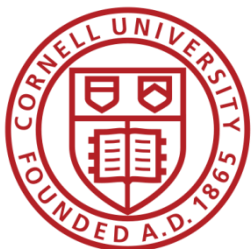
Old versions are only available for replication studies and only in justified exceptional cases for new Projects.

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)
<b>BHP 7518 v1 (current)</b>	<a href="#">10.5164/IAB.BHP7518.de.en.v1</a>	2020-01-13
<b>BHP 7517 v1</b>	<a href="#">10.5164/IAB.BHP7517.de.en.v1</a>	2018-12-12
<b>BHP 7516 v1</b>	<a href="#">10.5164/IAB.BHP7516.de.en.v1</a>	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”

# Thank you!

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