ILR LDI

Labor Dynamics Institute

Computational Reproducibility, Transparency, and Credibility of Official Statistics

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 National Academies, the American Economic Association, or any of the funding agencies.

United Nations: Fundamental Principles of Official Statistics

Principle 3: *Accountability and Transparency*

To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics.

Source

Principles and Practices for a Federal Statistical Agency

Principle 2: *Credibility among Data Users*

A federal statistical agency must have credibility with those who use its data and information.

Source

Audiences of Credible Official Statistics

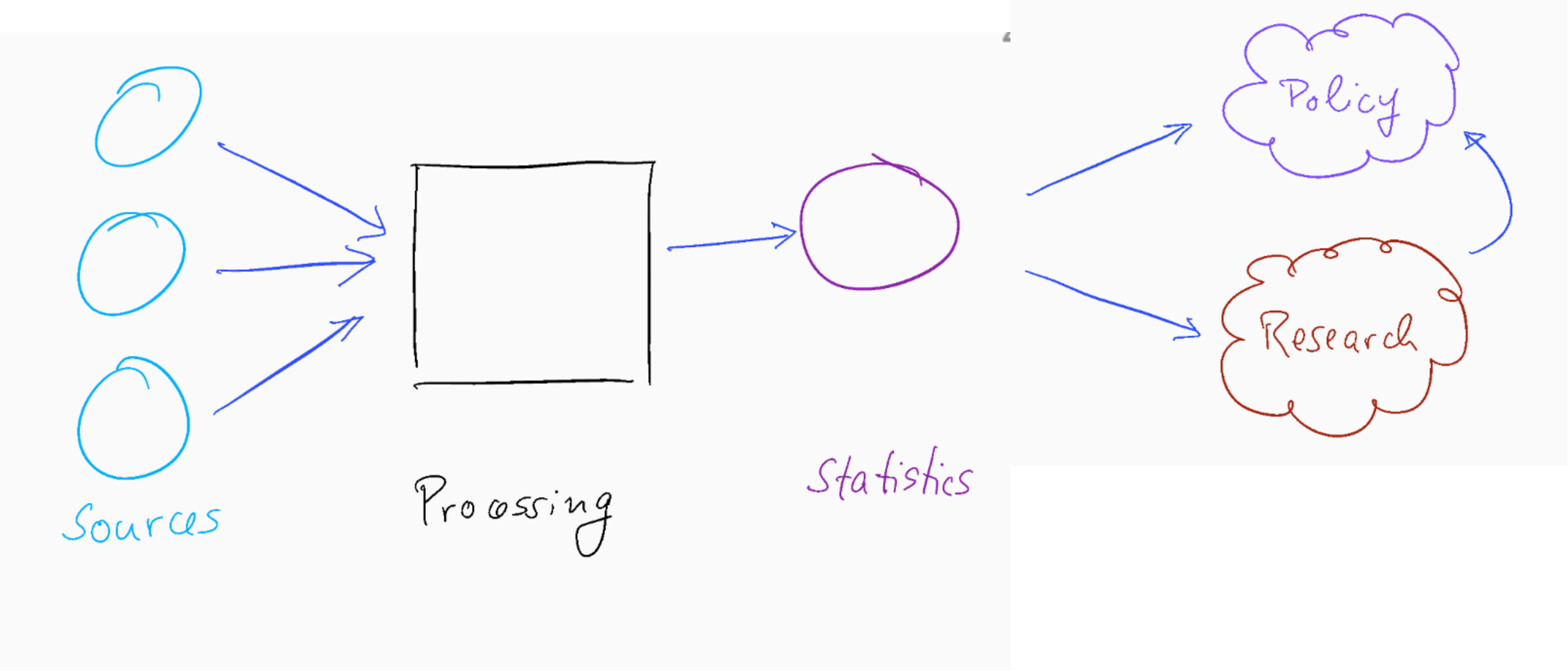
Audience 1:

- General public
- Policymakers
  - Federal
  - Sub-national

Audience 2

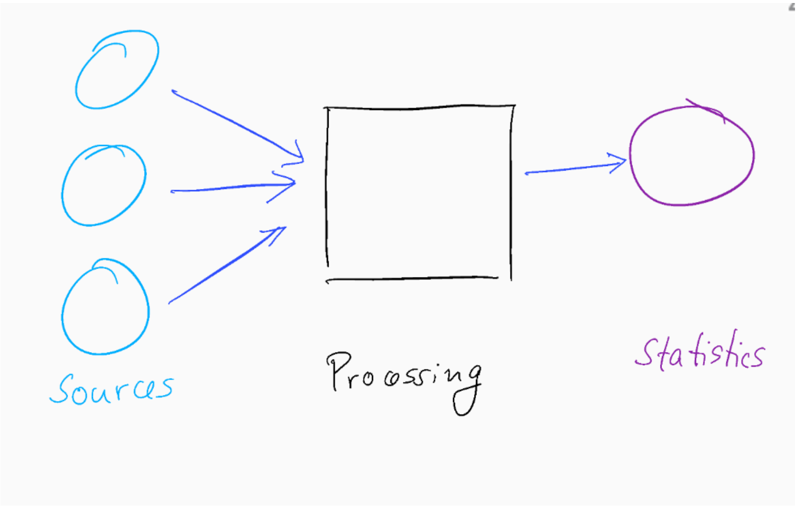
- Researchers
  - Not just academic

Basic setup



Computational Reproducibility and Official Statistics

- Detailed information on sources
  - Instructions on information is collected
    - Surveys
    - Administrative data
- Availability of “computing instructions”
  - Code
  - Including for disclosure avoidance
- Availability of reliable, trusted data archives
  - Of released data – for audience 1 & 2 – ability to reproduce downstream uses
  - Of source data – for audience 2 – ability to reproduce released data

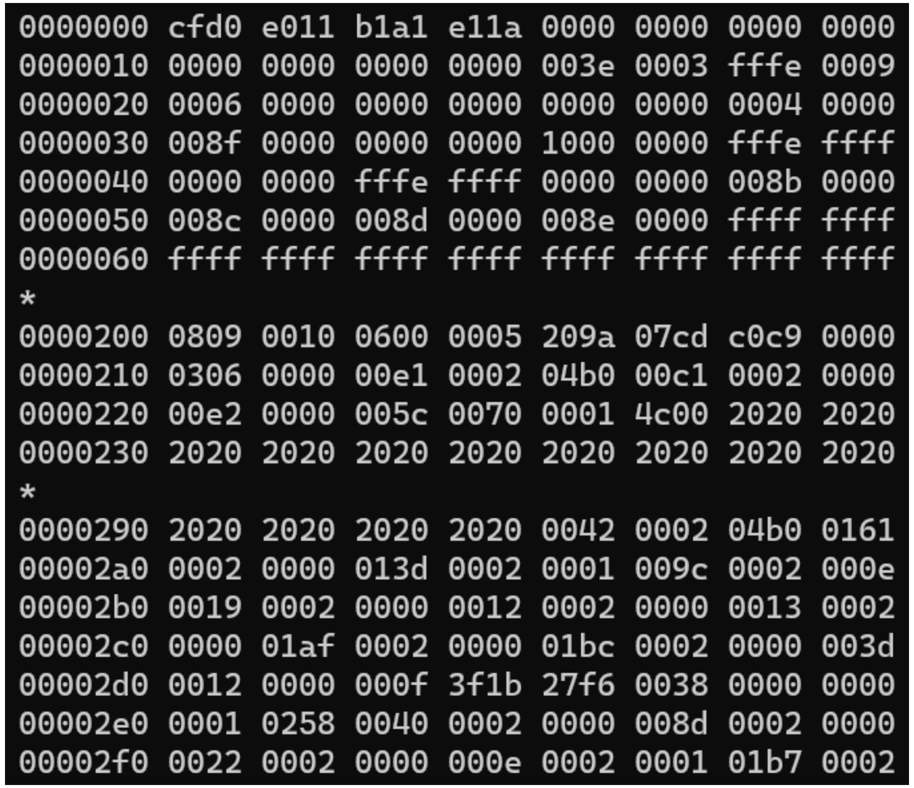


Would you buy a car from this guy?

Provenance!

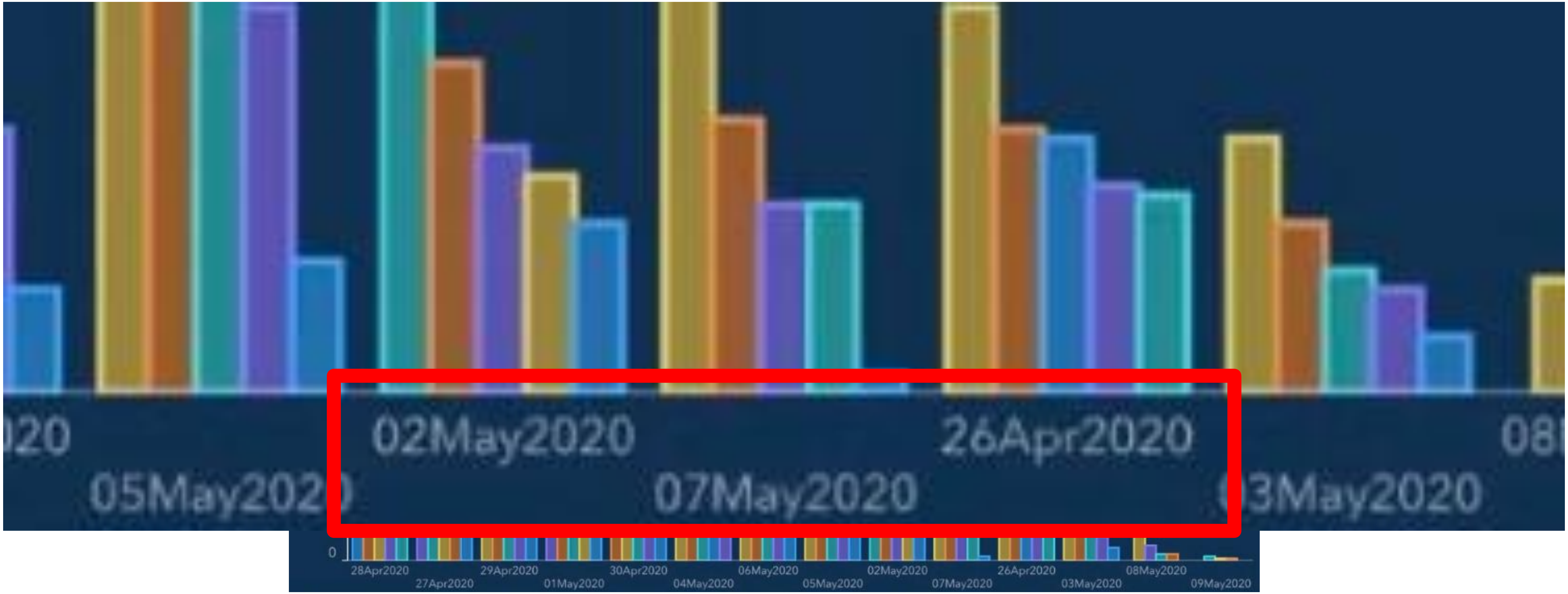
- Does the sales person have a good record?
- Where does the car come from?
- What do we know about the car?

Would you use this data?

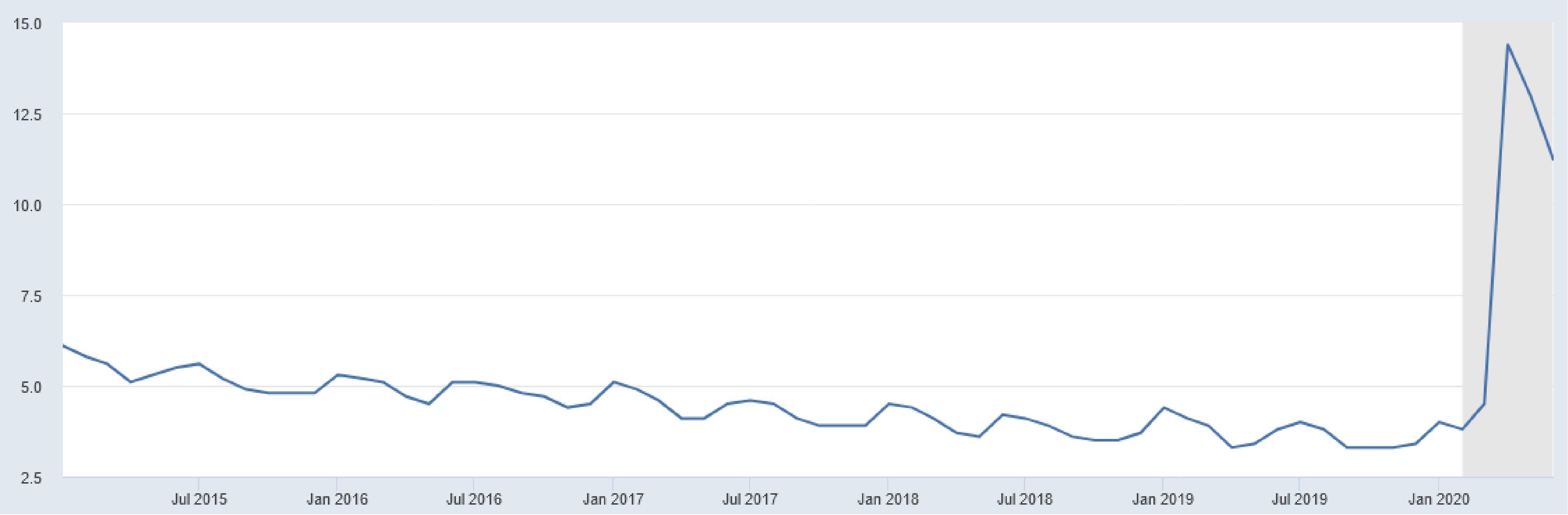


```
0000000 cfd0 e011 b1a1 e11a 0000 0000 0000 0000
0000010 0000 0000 0000 0000 003e 0003 fffe 0009
0000020 0006 0000 0000 0000 0000 0000 0004 0000
0000030 008f 0000 0000 0000 1000 0000 fffe ffff
0000040 0000 0000 fffe ffff 0000 0000 008b 0000
0000050 008c 0000 008d 0000 008e 0000 ffff ffff
0000060 ffff 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?



Or would you trust this data?



U.S. Bureau of Labor Statistics, Unemployment Rate [UNRATENSA], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/UNRATENSA, August 3, 2020.

Provenance!

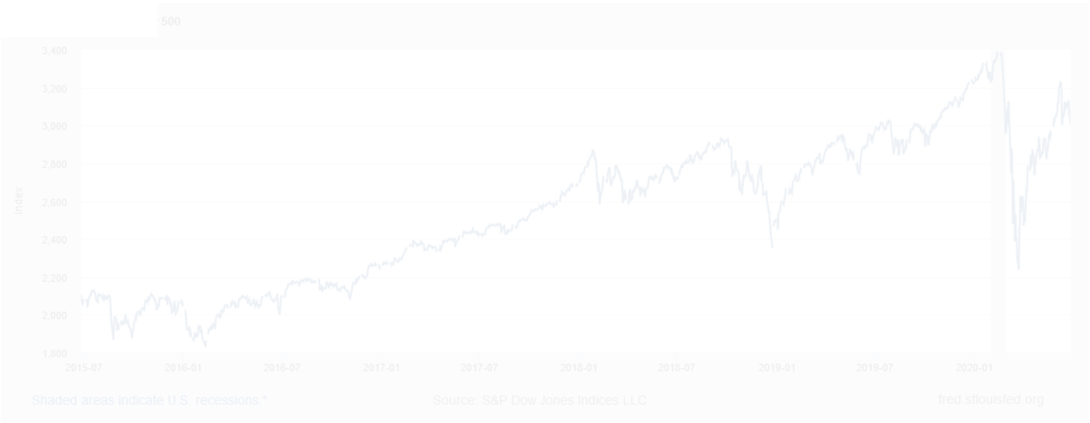
- Does the provider have a good record?
- Where do the data come from?
- What do we know about the data?

Metadata!

Context:
from data to trusted data provenance

“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



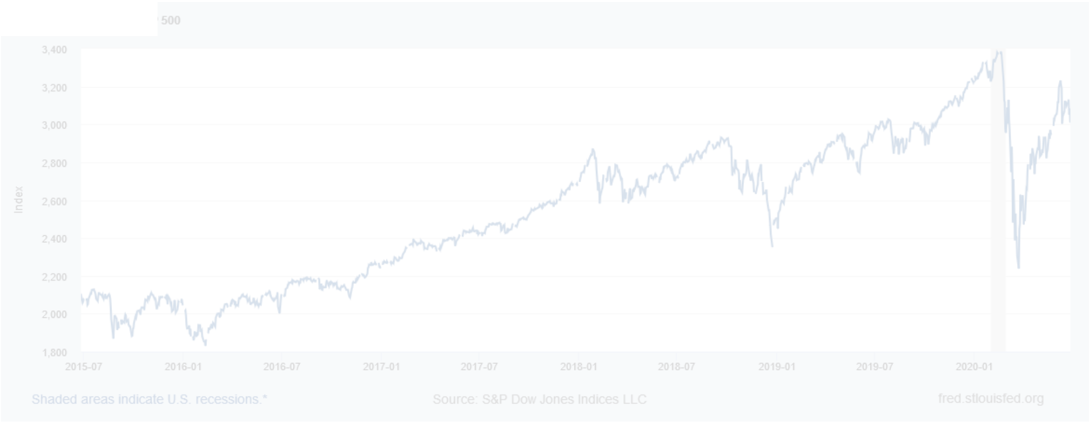
Shaded areas indicate U.S. recessions.*

Source: S&P Dow Jones Indices LLC

fred.stlouisfed.org

“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



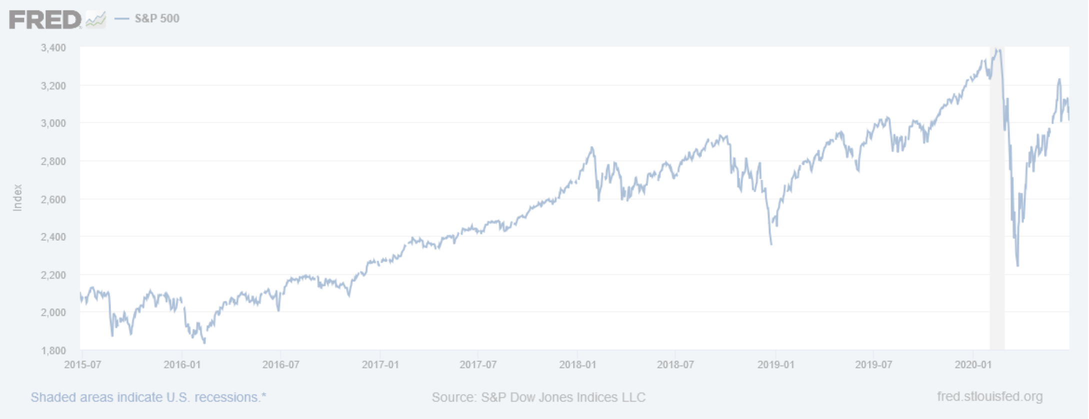
Shaded areas indicate U.S. recessions.*

Source: S&P Dow Jones Indices LLC

fred.stlouisfed.org

“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



Shaded areas indicate U.S. recessions.*

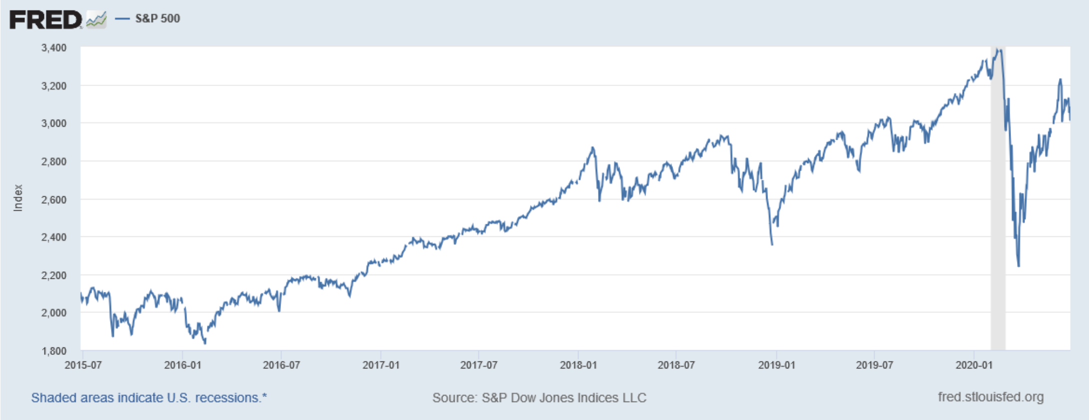
Source: S&P Dow Jones Indices LLC

fred.stlouisfed.org

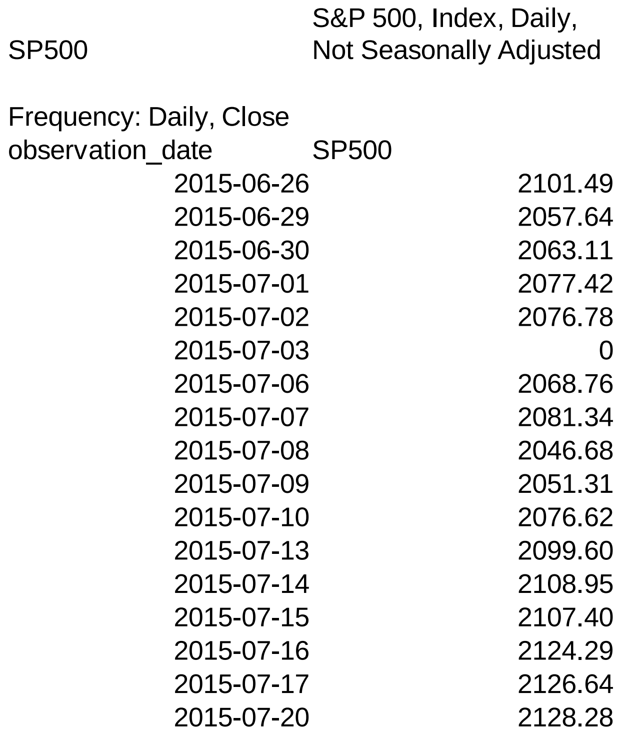
“It’s a file called SP500.xlsx, downloaded from FRED.”

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.

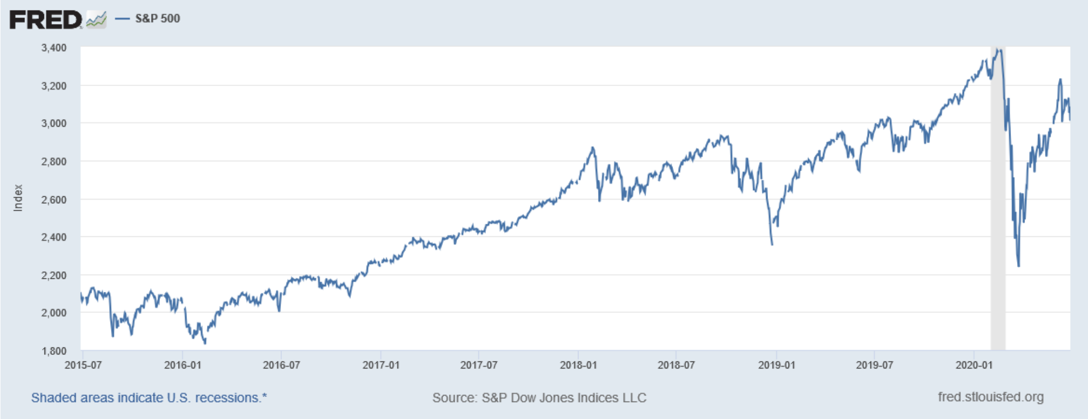


“SP500.xlsx, from S&P (2020). Not provided as part of replication package because © S&P. ”

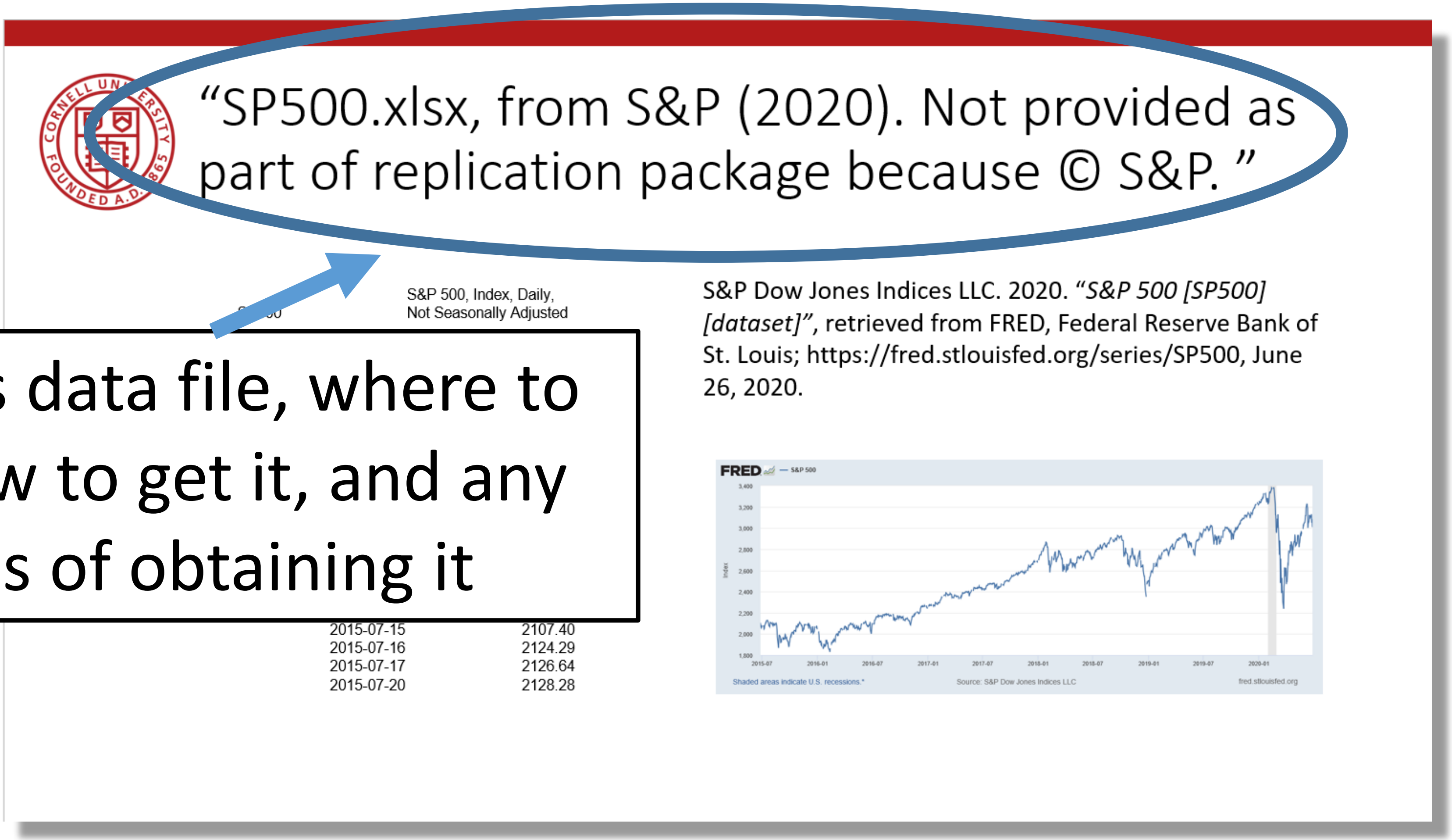


SP500	S&P 500, Index, Daily, Not Seasonally Adjusted
Frequency: Daily, Close	
observation_date	SP500
2015-06-26	2101.49
2015-06-29	2057.64
2015-06-30	2063.11
2015-07-01	2077.42
2015-07-02	2076.78
2015-07-03	0
2015-07-06	2068.76
2015-07-07	2081.34
2015-07-08	2046.68
2015-07-09	2051.31
2015-07-10	2076.62
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
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.

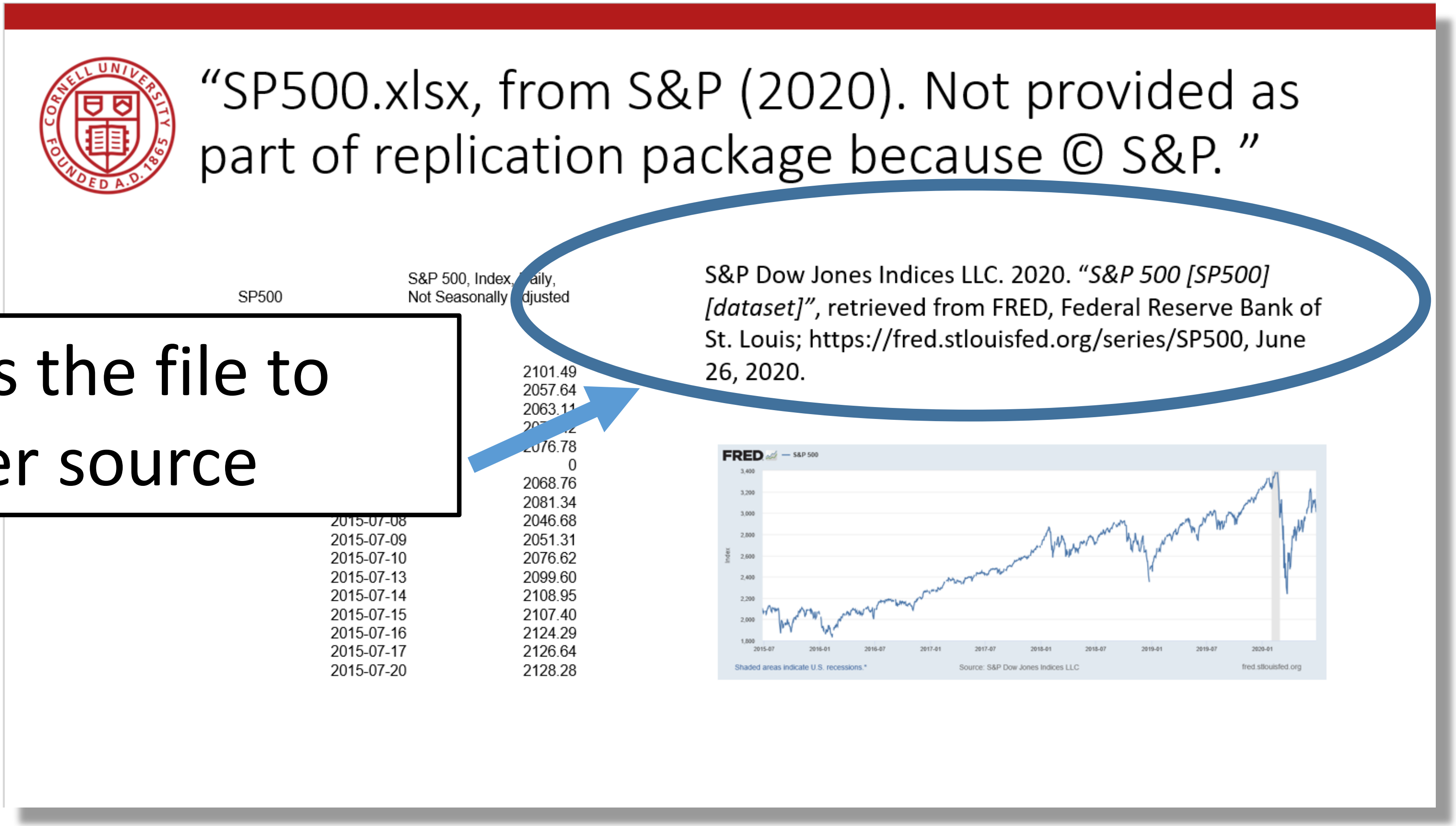
Shaded areas indicate U.S. recessions.*

Source: S&P Dow Jones Indices LLC

fred.stlouisfed.org

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

Data Citation



“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
	2101.49
	2057.64
	2063.11
	[illegible]
	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

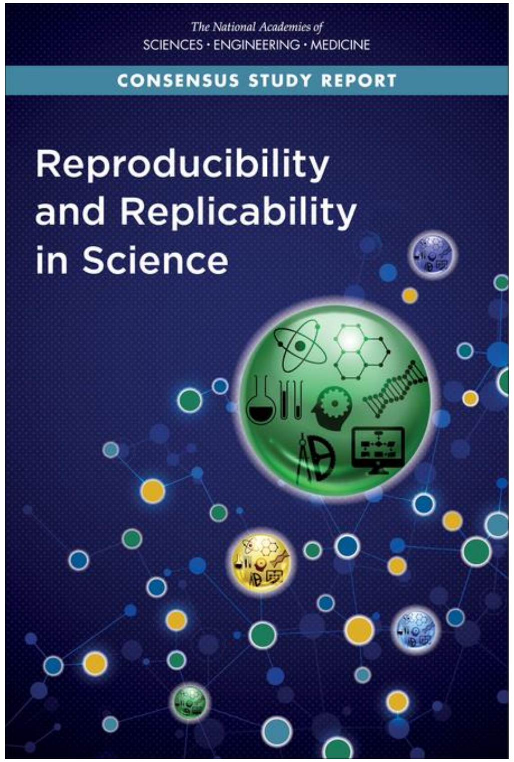
S&P Dow Jones Indices LLC. 2020. “*S&P 500 [SP500] [dataset]*”, retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SP500, June 26, 2020.

Attributes the file to the proper source

Background
What is reproducibility and replicability?

Replication continuum

https://doi.org/10.17226/25303



The National Academies of
SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

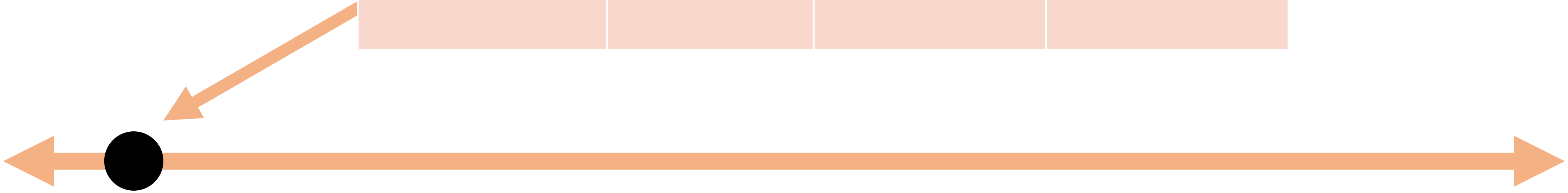
Reproducibility and Replicability in Science

Reproducibility

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

Replication continuum

Same data	Same code	Same methods	Same context

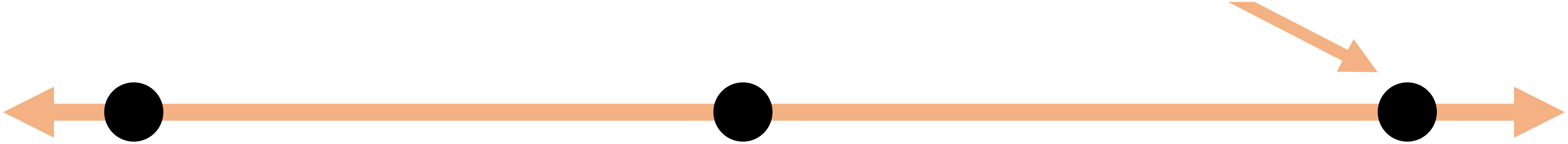


Reproducibility

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

Replication continuum

Different data	Different code or software	Different methods	Different 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

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

ICPSR
Sharing data to advance science

CASD

CRDCN Canadian Research Data Centre Network

FSRDC

IPUMS USA

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

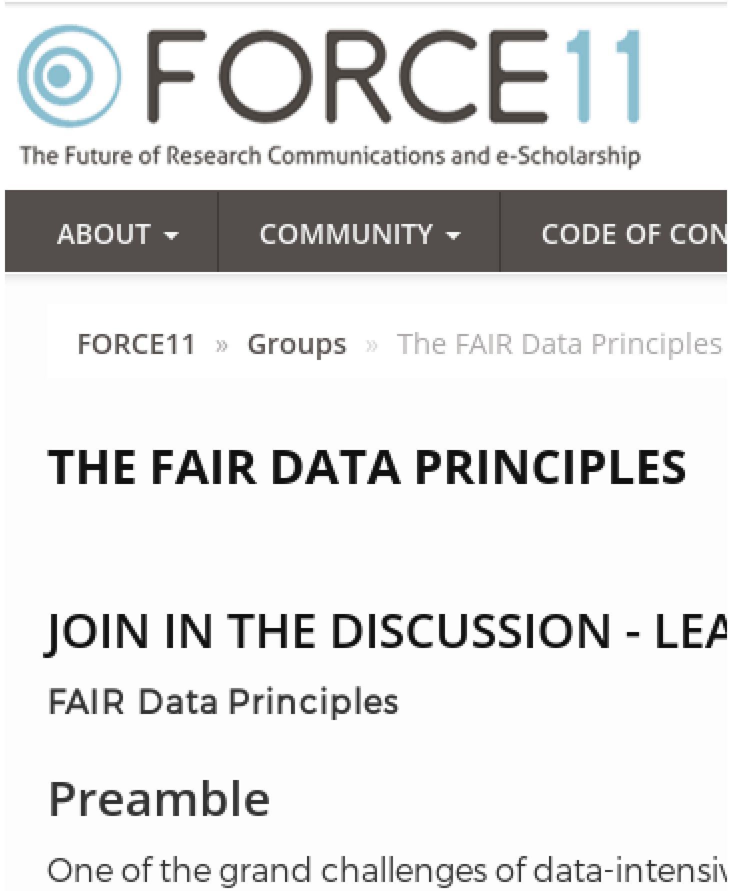
IAB INSTITUT FÜR ARBEITSMARKT- UND BERUFSFORSCHUNG
Die Forschungseinrichtung der Bundesagentur für Arbeit

CIQSS | QICSS
CENTRE INTERUNIVERSITAIRE QUÉBÉCOIS DE STATISTIQUES SOCIALES
QUEBEC INTER-UNIVERSITY CENTER FOR SOCIAL STATISTICS

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

FORCE11 » Groups » The FAIR Data Principles

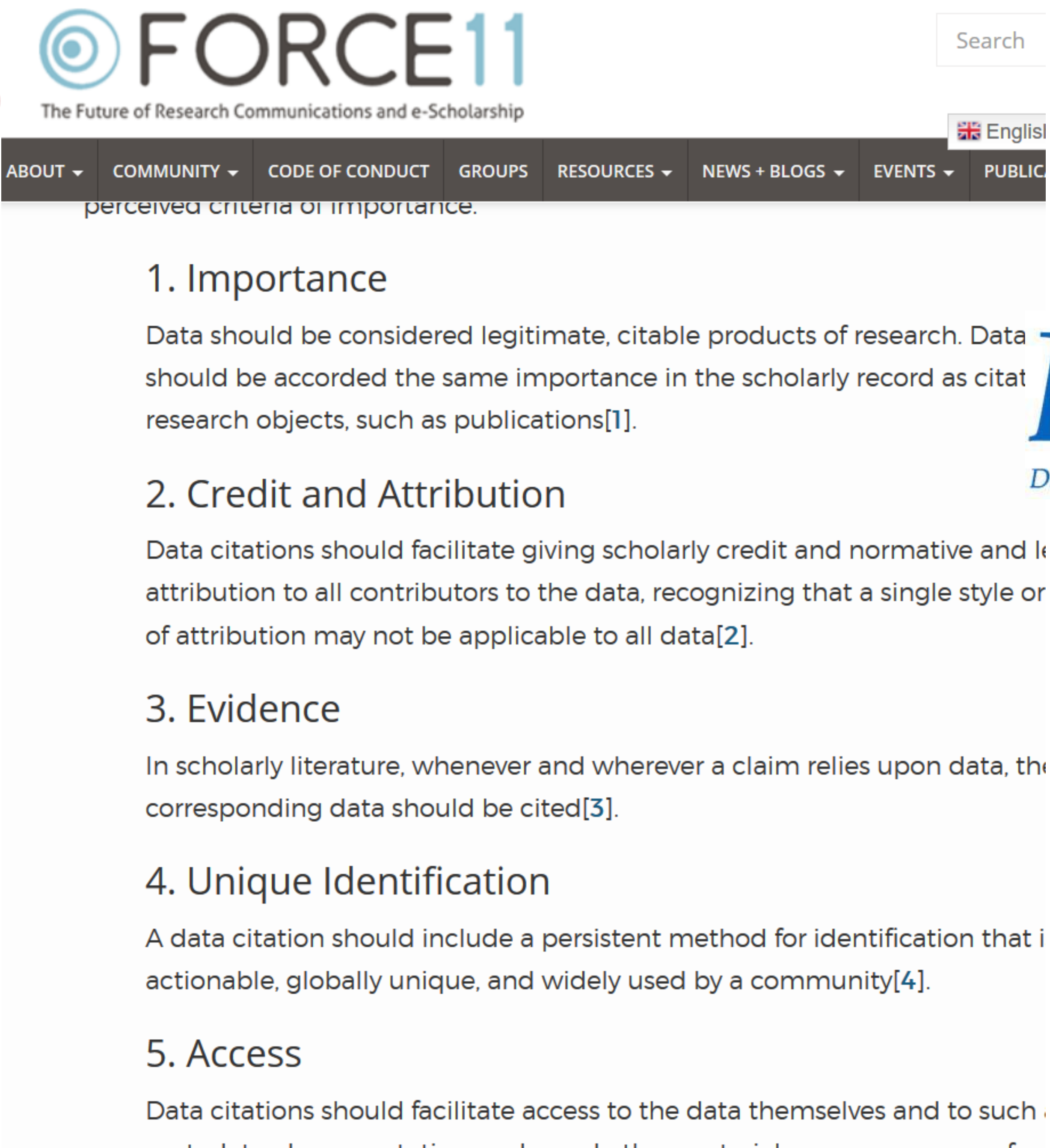
THE FAIR DATA PRINCIPLES

JOIN IN THE DISCUSSION - LEA

FAIR Data Principles

Preamble

One of the grand challenges of data-intensiv



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 l attribution to all contributors to the data, recognizing that a single style or of attribution may not be applicable to all data[2].

3. Evidence

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

4. Unique Identification

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

5. Access

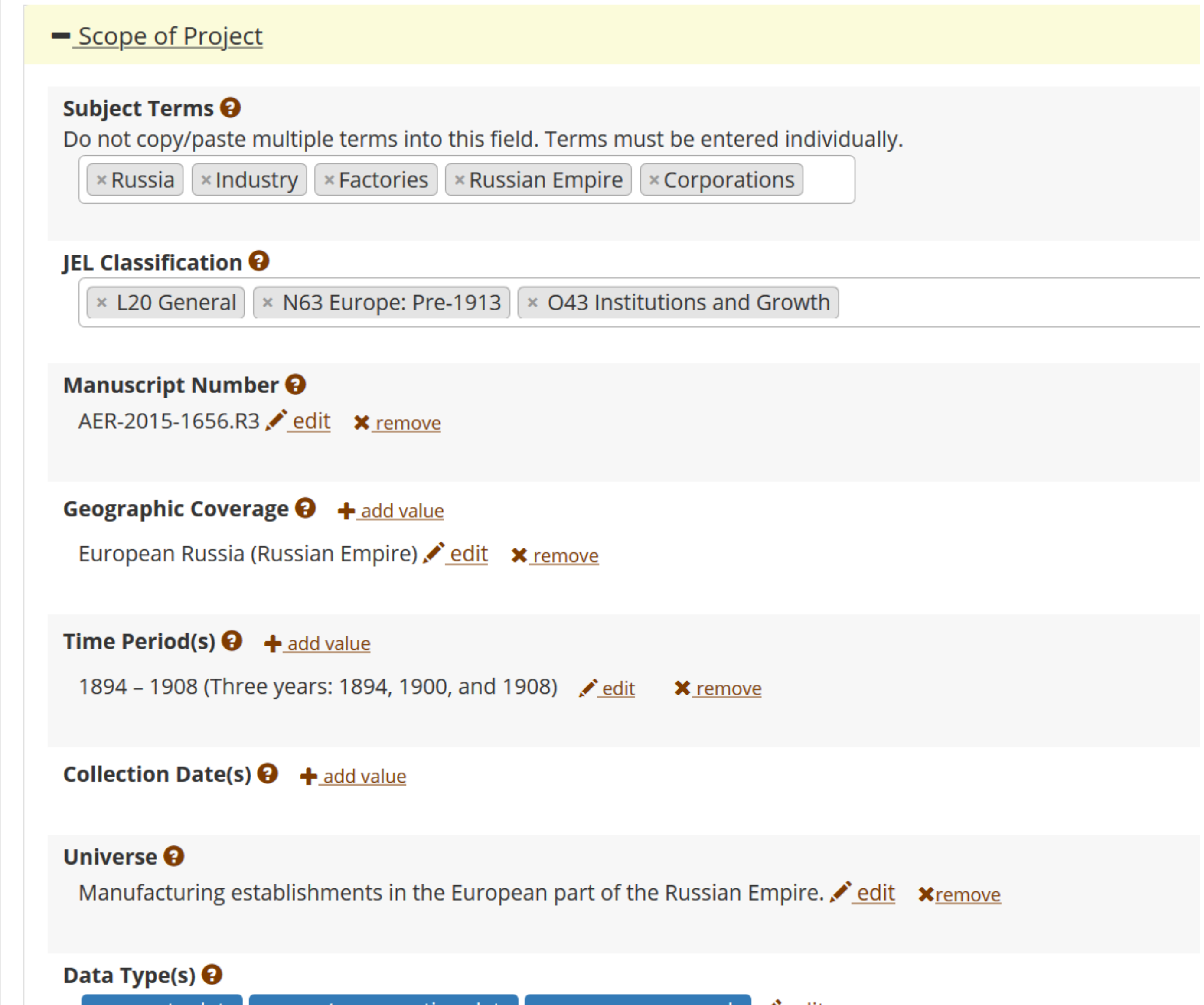
Data citations should facilitate access to the data themselves and to such

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].

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

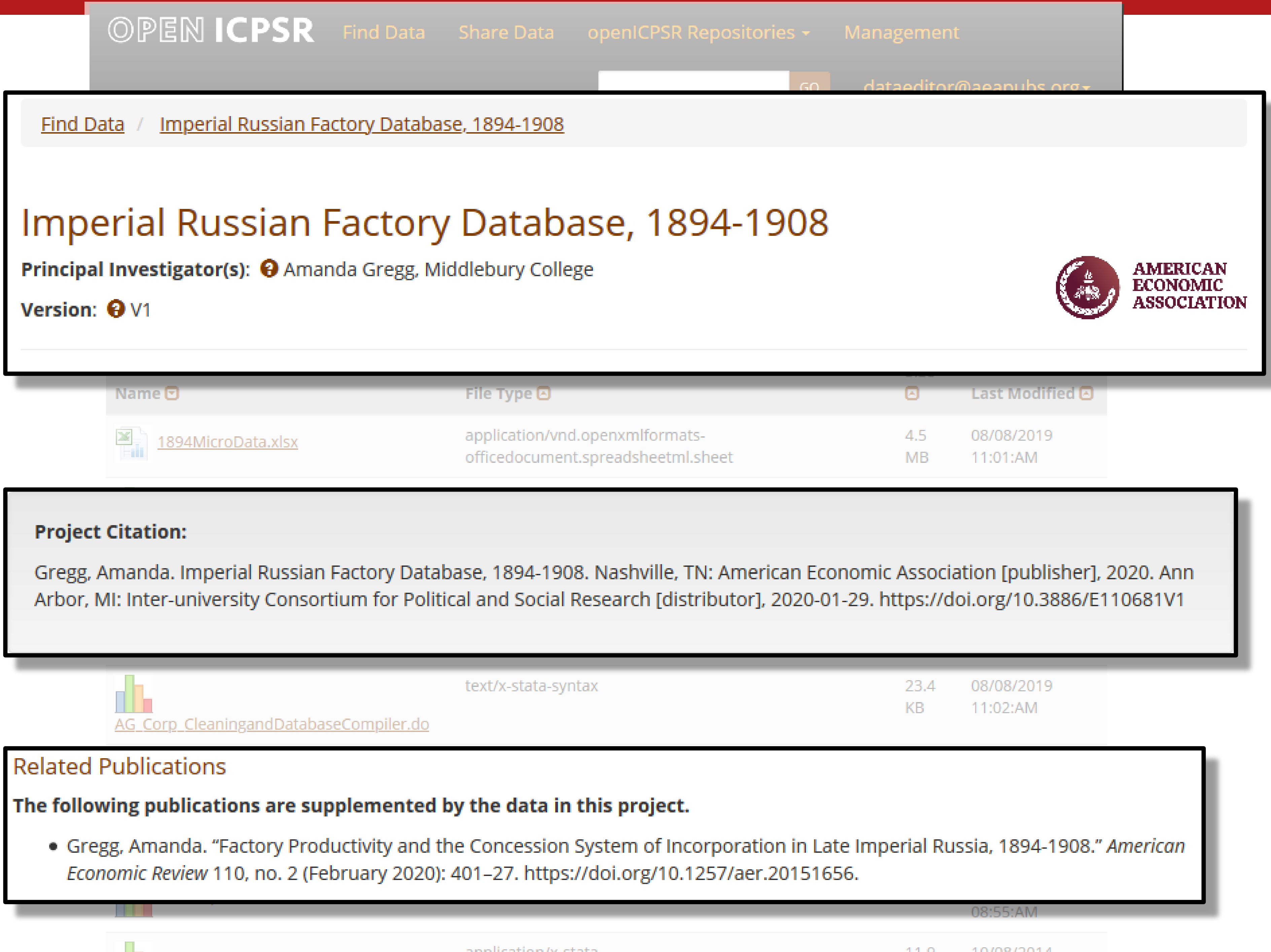
1894 – 1908 (Three years: 1894, 1900, and 1908) edit remove

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	Size	Last Modified
1894MicroData.xlsx	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet	4.5 MB	08/08/2019 11:01:AM
AG_Corp_CleaningandDatabaseCompiler.do	text/x-stata-syntax	23.4 KB	08/08/2019 11:02:AM
			08:55:AM
	application/x-stata	11.9	10/08/2014

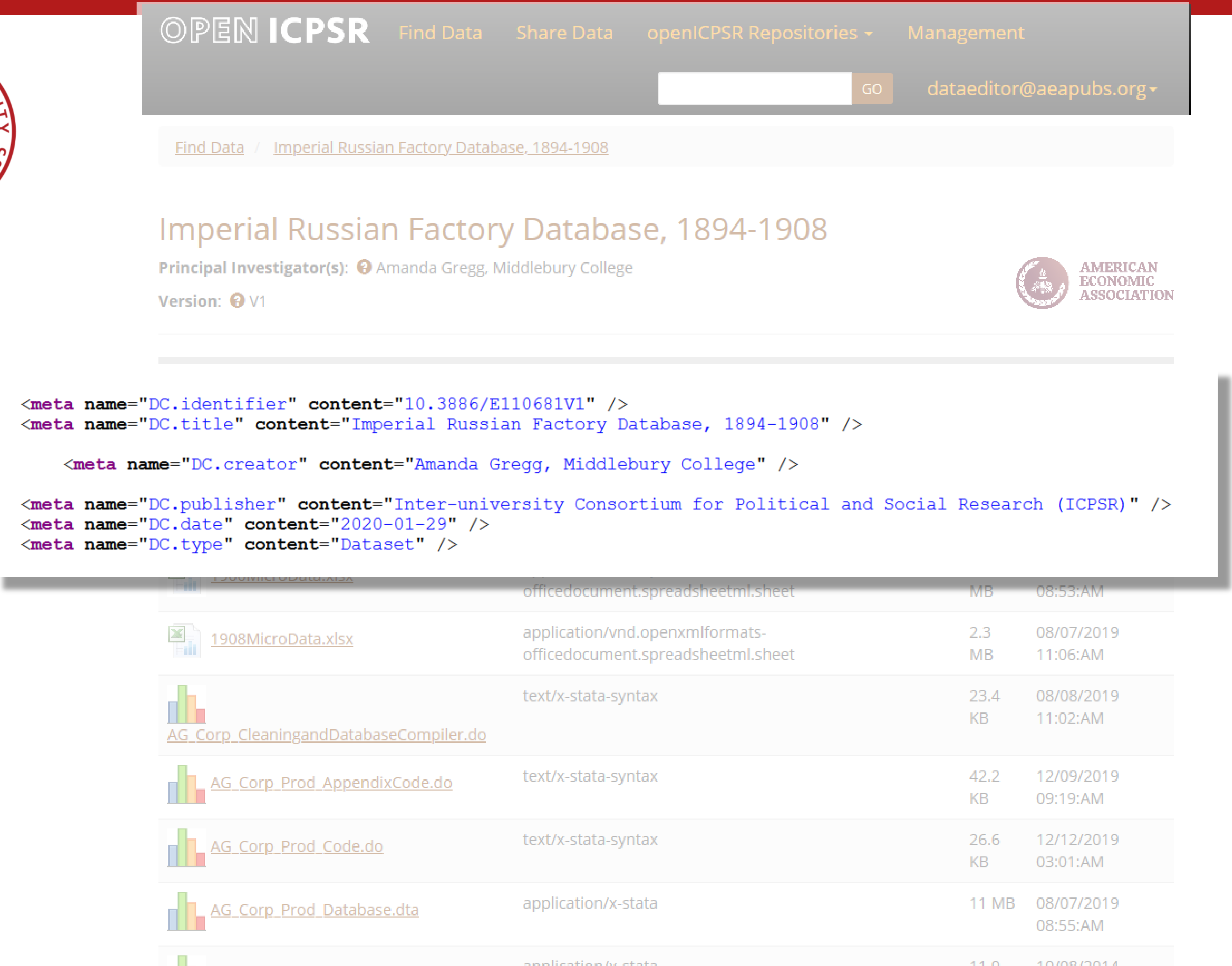
Project Citation:

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

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.



Find Data / Imperial Russian Factory Database, 1894-1908

Imperial Russian Factory Database, 1894-1908

Principal Investigator(s): Amanda Gregg, Middlebury College

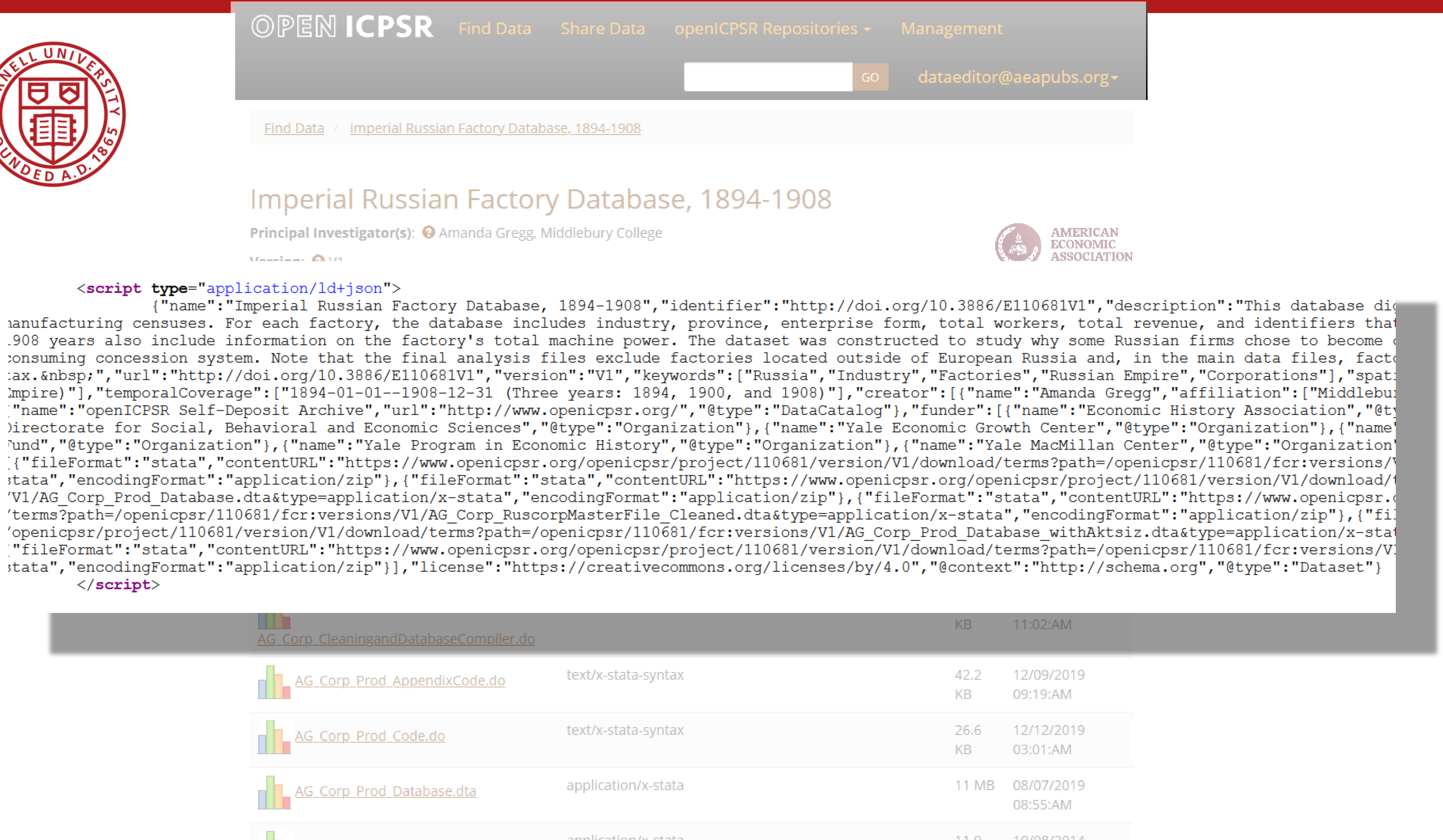
Version: V1

```
<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" />
```

	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		



Find Data / Imperial Russian Factory Database, 1894-1908

Imperial Russian Factory Database, 1894-1908

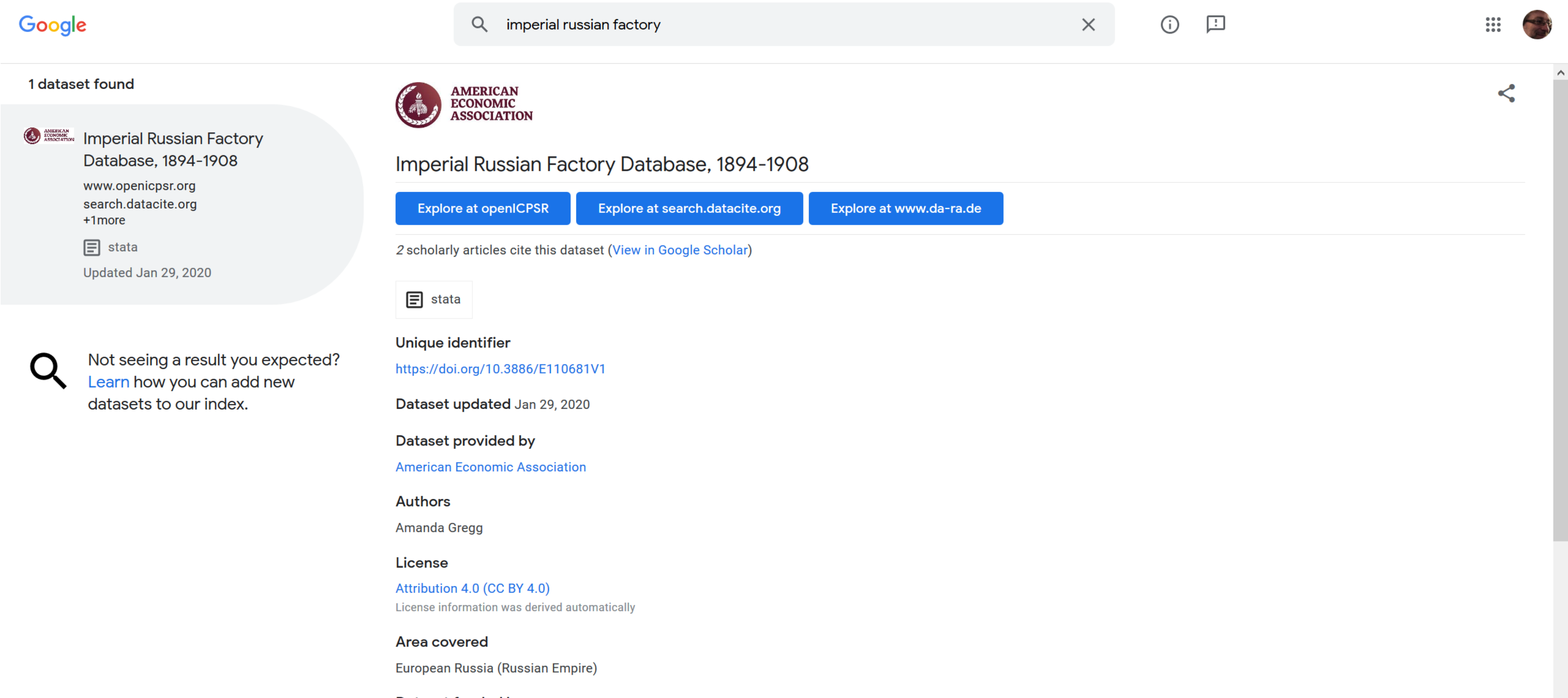
Principal Investigator(s): Amanda Gregg, Middlebury College

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
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 
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"],"spat
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"},{"fi
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/V
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
	application/x-stata	11.9	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
+1more

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)

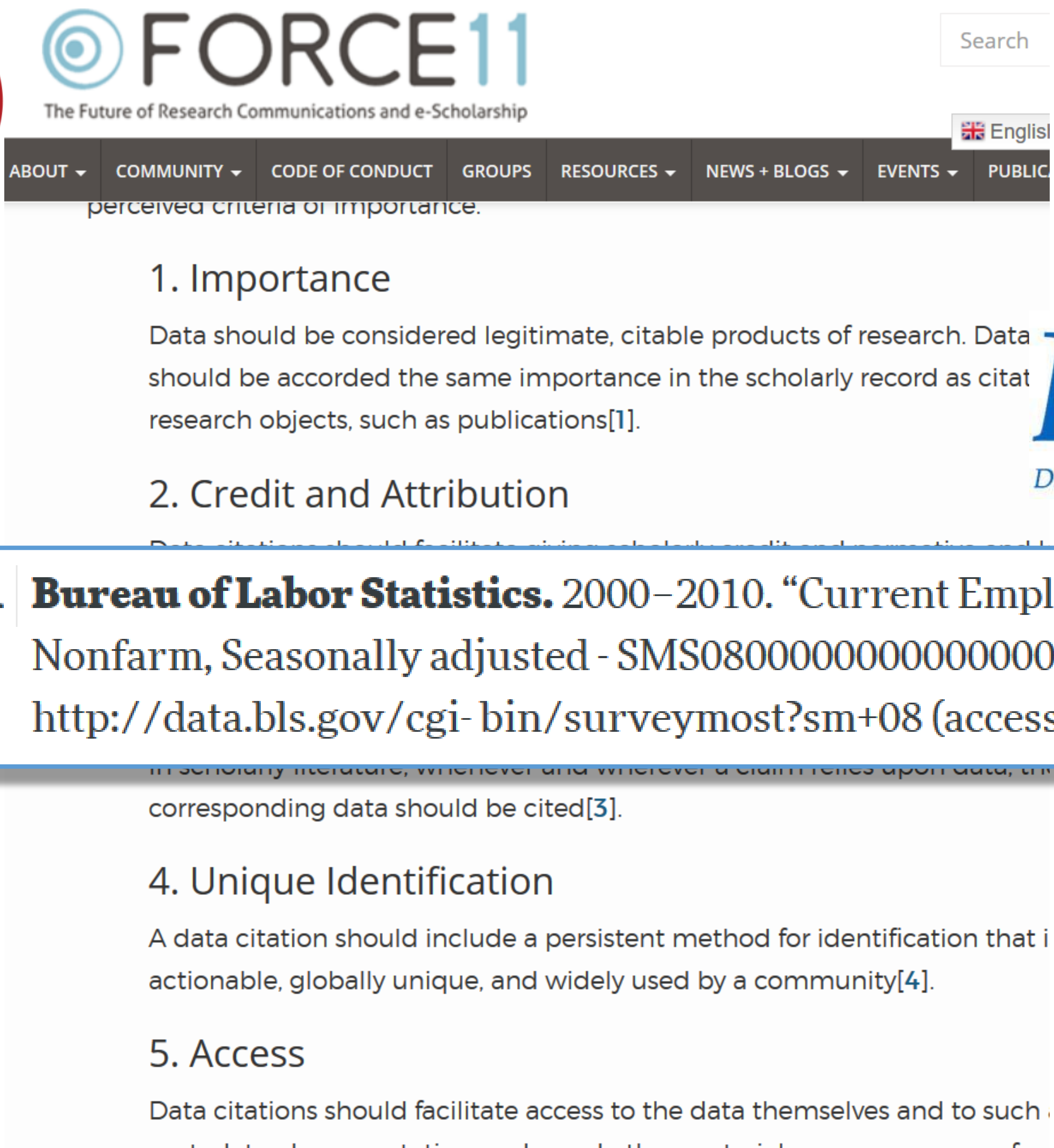
License information was derived automatically

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



FORCE11

The Future of Research Communications and e-Scholarship

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 l

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

4. Unique Identification

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

5. Access

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

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).

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].

Examples in Official Statistics

Examples in Official Statistics

- Trusted archives: none
- Reliable versioning of output data:
  - Good example: BEA
  - Counter example: BLS (great accessibility, poor versioning)
- Availability of computing instructions
  - Good example: BLS (CPS, unemployment, inflation rate)
  - Counter example: much else
- Transparency of disclosure avoidance
  - Good example: Census 2020, but also CPS
  - Bad example: almost everything else

Trusted archives

What is a “trusted” archive?

“A **reliable digital repository** is one whose mission is to provide **long-term access** to managed digital resources; that accepts responsibility for the long-term maintenance of digital resources **on behalf** of its depositors and **for the benefit of current and *future* users**; [...] that establishes methodologies for system evaluation that meet **community expectations of trustworthiness**; that can be **depended upon** to carry out its long-term responsibilities to depositors and users openly and explicitly; and whose policies, practices, and performance can be audited and measured.”

source

What is a “trusted” archive?

Various definitions, certifications, criteria:

- CoreTrustSeal (http://doi.org/10.5281/zenodo.3638211)
- Trusted Repositories Audit & Certification (TRAC), (https://www.crl.edu/PDF/trac.pdf)
    - includes many research libraries as well as US National Archives (NARA)
- See also DataScience@NIH

What is a “trusted” archive?

Transparency, long-term preservation, access,…

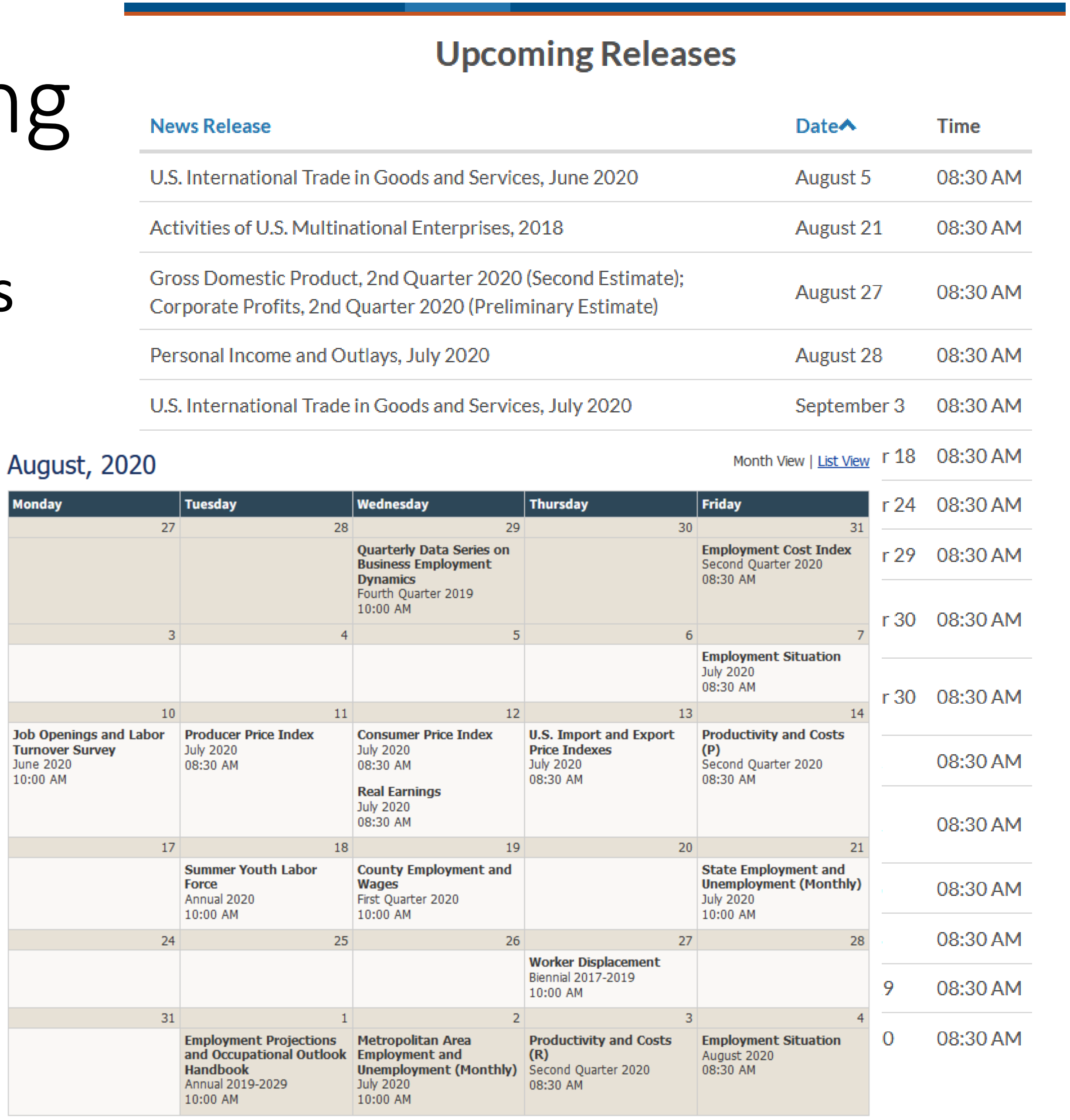
- Includes **documentation** of these processes
- Includes **version control** strategy
- Includes **maintenance of persistent identifiers**

These are only starting to emerge amongst statistical agencies in the US!

Reliable versioning of output data

Reliable versioning

- Ability to access data as it was when it was released
  - Might be a single file
  - Might be an indicator of any revisions of a data item
- Example:
  - BEA release ***schedule***
  - BLS release ***schedule***



Upcoming Releases

News Release	Date	Time
U.S. International Trade in Goods and Services, June 2020	August 5	08:30 AM
Activities of U.S. Multinational Enterprises, 2018	August 21	08:30 AM
Gross Domestic Product, 2nd Quarter 2020 (Second Estimate); Corporate Profits, 2nd Quarter 2020 (Preliminary Estimate)	August 27	08:30 AM
Personal Income and Outlays, July 2020	August 28	08:30 AM
U.S. International Trade in Goods and Services, July 2020	September 3	08:30 AM
	r 18	08:30 AM
	r 24	08:30 AM
	r 29	08:30 AM
	r 30	08:30 AM
	r 30	08:30 AM
		08:30 AM
		08:30 AM
		08:30 AM
		08:30 AM
	9	08:30 AM
	0	08:30 AM

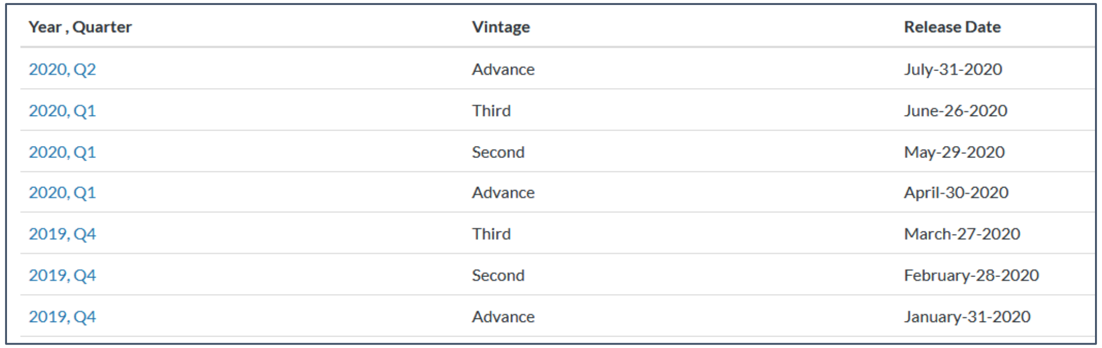
August, 2020

Month View | List View

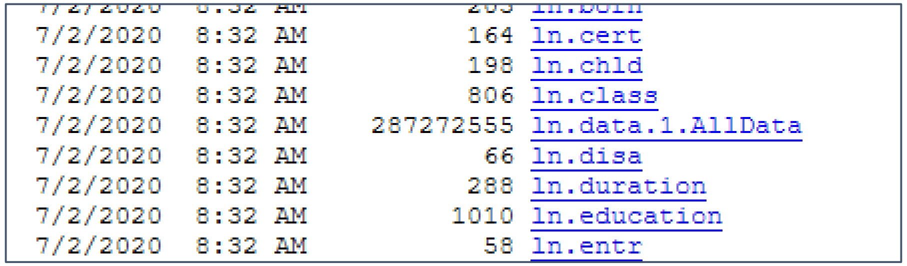
Monday	Tuesday	Wednesday	Thursday	Friday
27	28	29 **Quarterly Data Series on Business Employment Dynamics** Fourth Quarter 2019 10:00 AM	30	31 **Employment Cost Index** Second Quarter 2020 08:30 AM
3	4	5	6	7 **Employment Situation** July 2020 08:30 AM
10 **Job Openings and Labor Turnover Survey** June 2020 10:00 AM	11 **Producer Price Index** July 2020 08:30 AM	12 **Consumer Price Index** July 2020 08:30 AM **Real Earnings** July 2020 08:30 AM	13 **U.S. Import and Export Price Indexes** July 2020 08:30 AM	14 **Productivity and Costs (P)** Second Quarter 2020 08:30 AM
17	18 **Summer Youth Labor Force** Annual 2020 10:00 AM	19 **County Employment and Wages** First Quarter 2020 10:00 AM	20	21 **State Employment and Unemployment (Monthly)** July 2020 10:00 AM
24	25	26	27 **Worker Displacement** Biennial 2017-2019 10:00 AM	28
31	1 **Employment Projections and Occupational Outlook Handbook** Annual 2019-2029 10:00 AM	2 **Metropolitan Area Employment and Unemployment (Monthly)** July 2020 10:00 AM	3 **Productivity and Costs (R)** Second Quarter 2020 08:30 AM	4 **Employment Situation** August 2020 08:30 AM

Reliable versioning

- Ability to access data as it was when it was released
  - Might be a single file
  - Might be an indicator of any revisions of a data item
- Example:
  - BEA release ***files***: one per release
  - BLS release ***files***: one continually updated, no (systematic) versioning



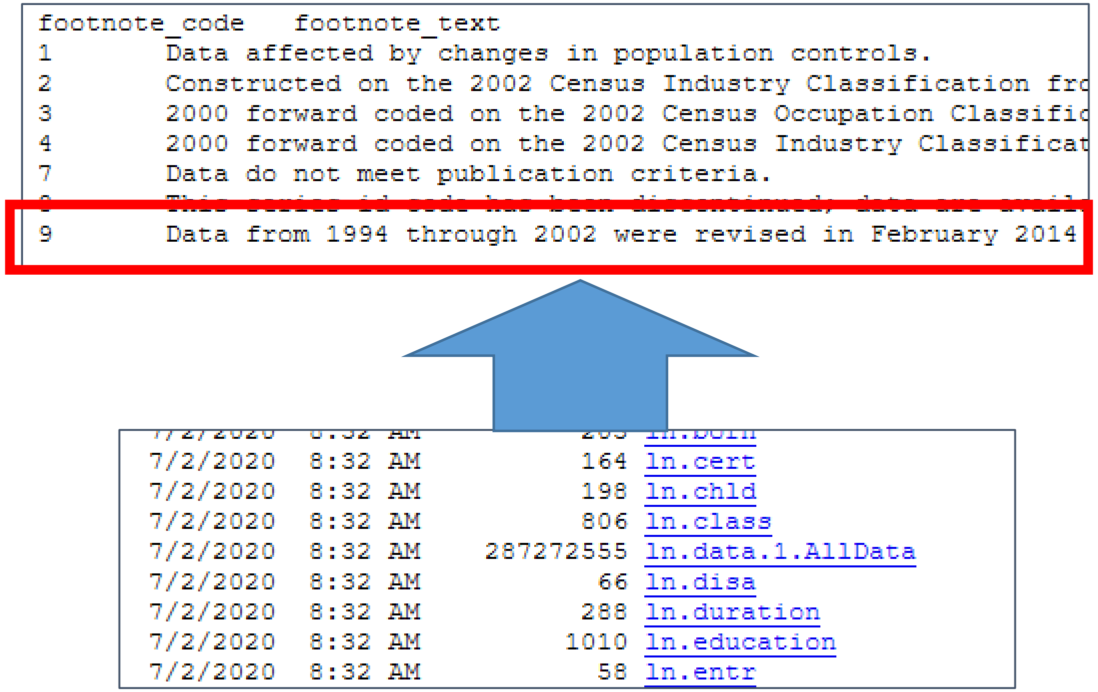
Year , Quarter	Vintage	Release Date
2020, Q2	Advance	July-31-2020
2020, Q1	Third	June-26-2020
2020, Q1	Second	May-29-2020
2020, Q1	Advance	April-30-2020
2019, Q4	Third	March-27-2020
2019, Q4	Second	February-28-2020
2019, Q4	Advance	January-31-2020



```
7/2/2020   8:32 AM          203 ln.born
7/2/2020   8:32 AM          164 ln.cert
7/2/2020   8:32 AM          198 ln.chld
7/2/2020   8:32 AM          806 ln.class
7/2/2020   8:32 AM    287272555 ln.data.1.AllData
7/2/2020   8:32 AM           66 ln.disa
7/2/2020   8:32 AM          288 ln.duration
7/2/2020   8:32 AM         1010 ln.education
7/2/2020   8:32 AM           58 ln.entr
```

Reliable versioning

- Ability to access data as it was when it was released
  - Might be a single file
  - Might be an indicator of any revisions of a data item
- Example:
  - BEA release ***files***: one per release
  - BLS release ***files***: one continually updated, no (systematic) versioning



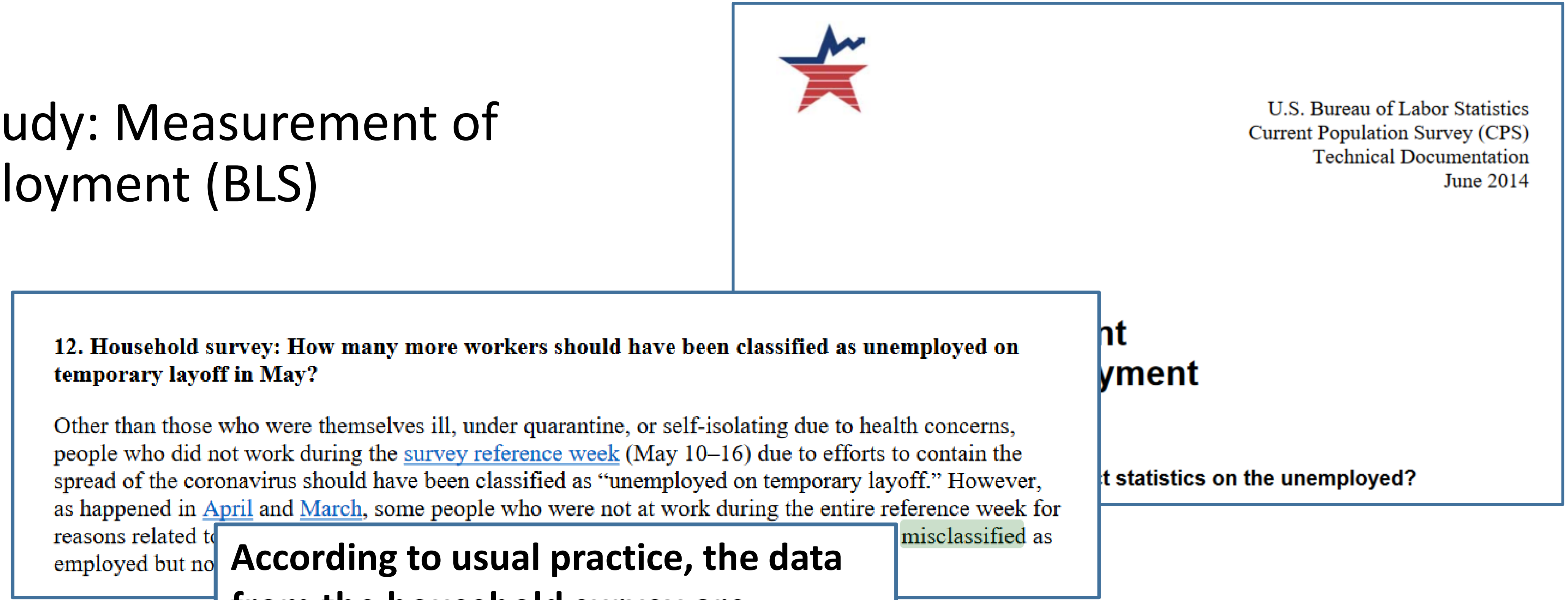
```
footnote_code   footnote_text
1       Data affected by changes in population controls.
2       Constructed on the 2002 Census Industry Classification fro
3       2000 forward coded on the 2002 Census Occupation Classific
4       2000 forward coded on the 2002 Census Industry Classificat
7       Data do not meet publication criteria.
8       This series id code has been discontinued; data are avail
9       Data from 1994 through 2002 were revised in February 2014
```

```
7/2/2020   8:32 AM         203 ln.born
7/2/2020   8:32 AM         164 ln.cert
7/2/2020   8:32 AM         198 ln.chld
7/2/2020   8:32 AM         806 ln.class
7/2/2020   8:32 AM   287272555 ln.data.1.AllData
7/2/2020   8:32 AM          66 ln.disa
7/2/2020   8:32 AM         288 ln.duration
7/2/2020   8:32 AM        1010 ln.education
7/2/2020   8:32 AM          58 ln.entr
```

Computing instructions

Computing instructions

- Case study: Measurement of unemployment (BLS)



U.S. Bureau of Labor Statistics
Current Population Survey (CPS)
Technical Documentation
June 2014

nt
yment

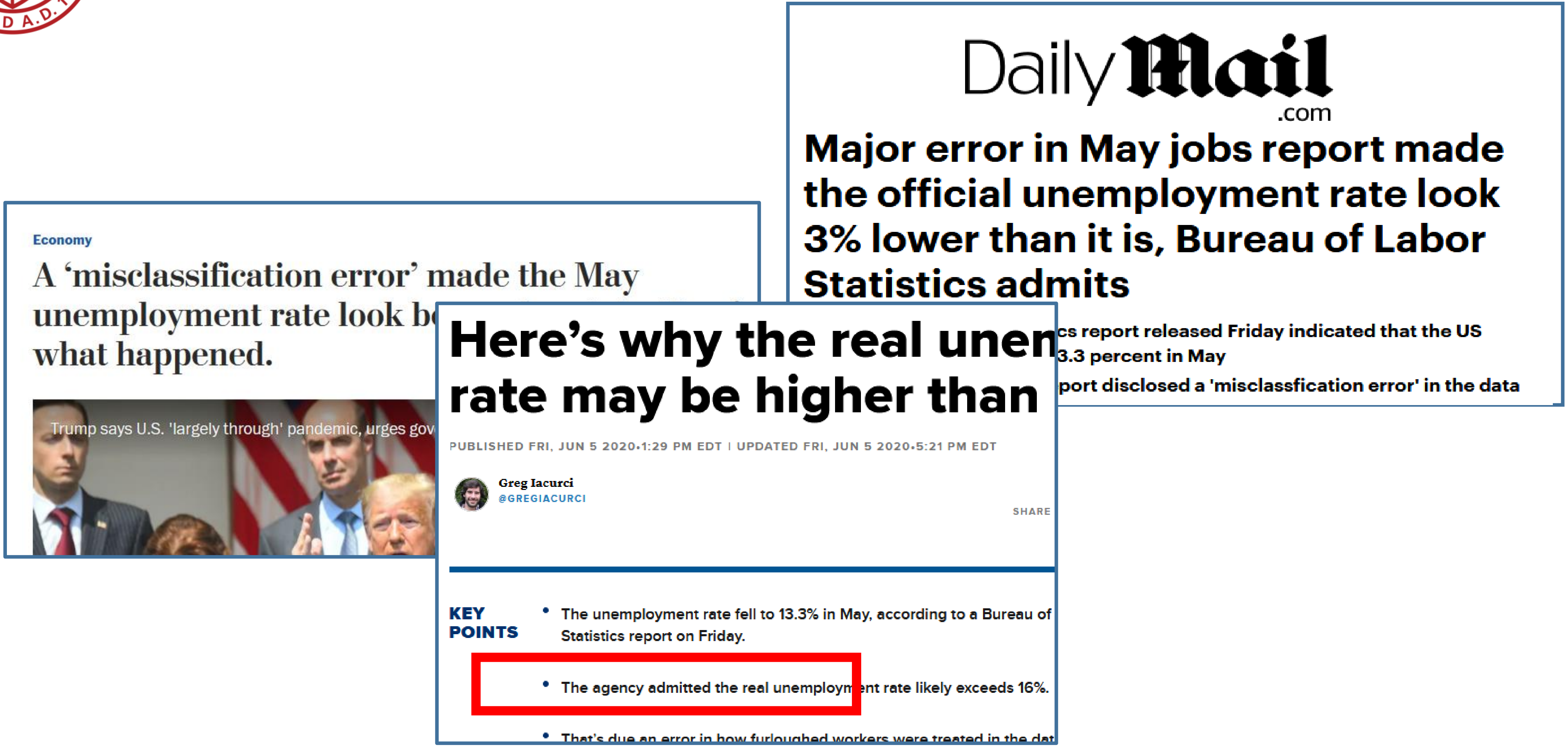
t statistics on the unemployed?

12. Household survey: How many more workers should have been classified as unemployed on temporary layoff in May?

Other than those who were themselves ill, under quarantine, or self-isolating due to health concerns, people who did not work during the survey reference week (May 10–16) due to efforts to contain the spread of the coronavirus should have been classified as “unemployed on temporary layoff.” However, as happened in April and March, some people who were not at work during the entire reference week for reasons related t misclassified as employed but no

According to usual practice, the data from the household survey are *accepted as recorded*. To maintain data integrity, no ad hoc actions are taken to reassign survey responses.

Transparency can be hard…



Economy

A ‘misclassification error’ made the May unemployment rate look be

what happened.

Trump says U.S. 'largely through' pandemic, urges gov

Daily Mail .com

Major error in May jobs report made the official unemployment rate look 3% lower than it is, Bureau of Labor Statistics admits

cs report released Friday indicated that the US

3.3 percent in May

port disclosed a 'misclassfication error' in the data

Here’s why the real unen rate may be higher than

PUBLISHED FRI, JUN 5 2020•1:29 PM EDT | UPDATED FRI, JUN 5 2020•5:21 PM EDT

Greg Iacurci
@GREGIACURCI

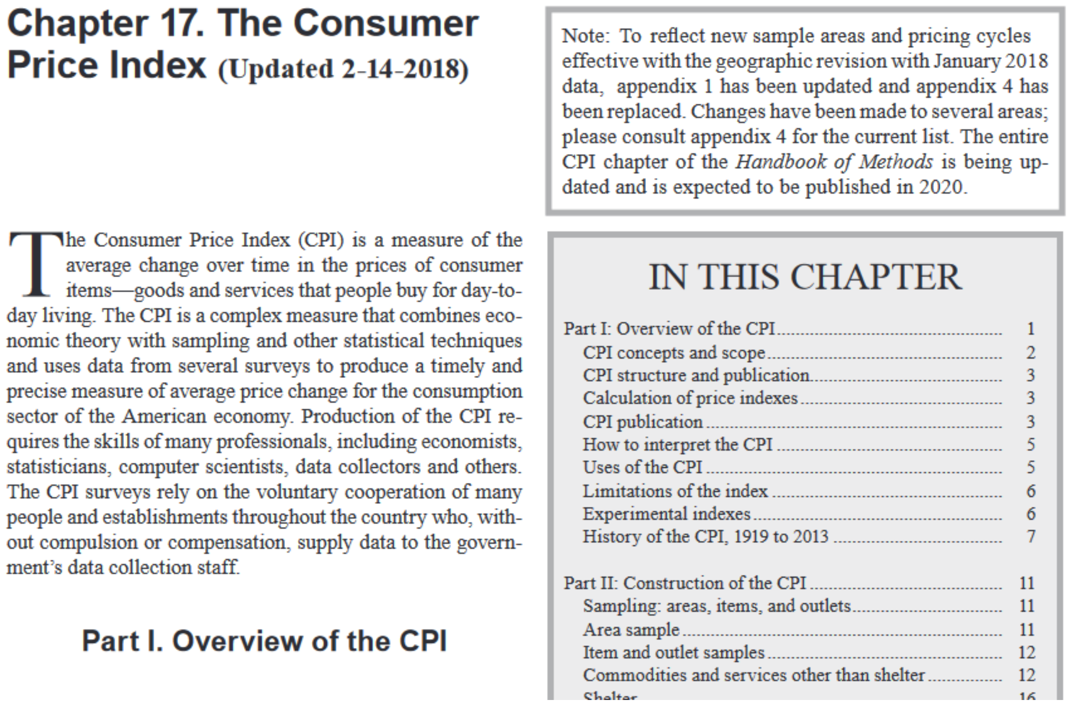
SHARE

KEY POINTS

- The unemployment rate fell to 13.3% in May, according to a Bureau of Statistics report on Friday.
- The agency admitted the real unemployment rate likely exceeds 16%.
- That’s due an error in how furloughed workers were treated in the dat

Computing instructions

Case study: Measurement of consumer price index (CPI)



Chapter 17. The Consumer Price Index (Updated 2-14-2018)

Note: To reflect new sample areas and pricing cycles effective with the geographic revision with January 2018 data, appendix 1 has been updated and appendix 4 has been replaced. Changes have been made to several areas; please consult appendix 4 for the current list. The entire CPI chapter of the *Handbook of Methods* is being updated and is expected to be published in 2020.

The Consumer Price Index (CPI) is a measure of the average change over time in the prices of consumer items—goods and services that people buy for day-to-day living. The CPI is a complex measure that combines economic theory with sampling and other statistical techniques and uses data from several surveys to produce a timely and precise measure of average price change for the consumption sector of the American economy. Production of the CPI requires the skills of many professionals, including economists, statisticians, computer scientists, data collectors and others. The CPI surveys rely on the voluntary cooperation of many people and establishments throughout the country who, without compulsion or compensation, supply data to the government's data collection staff.

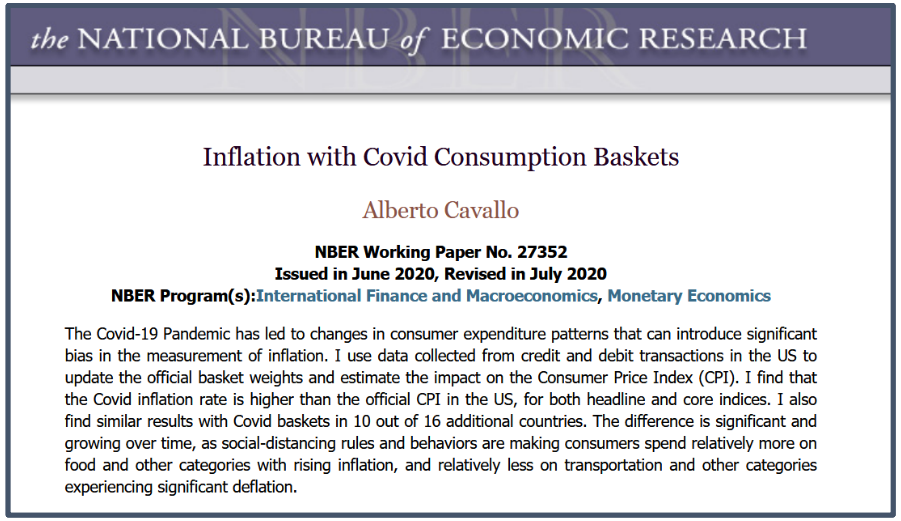
Part I. Overview of the CPI

IN THIS CHAPTER

Computing instructions

Case study: Measurement of consumer price index (CPI)

- Allows for researchers to replicate and investigate



the NATIONAL BUREAU of ECONOMIC RESEARCH

Inflation with Covid Consumption Baskets

Alberto Cavallo

NBER Working Paper No. 27352
Issued in June 2020, Revised in July 2020
NBER Program(s): International Finance and Macroeconomics, Monetary Economics

The Covid-19 Pandemic has led to changes in consumer expenditure patterns that can introduce significant bias in the measurement of inflation. I use data collected from credit and debit transactions in the US to update the official basket weights and estimate the impact on the Consumer Price Index (CPI). I find that the Covid inflation rate is higher than the official CPI in the US, for both headline and core indices. I also find similar results with Covid baskets in 10 out of 16 additional countries. The difference is significant and growing over time, as social-distancing rules and behaviors are making consumers spend relatively more on food and other categories with rising inflation, and relatively less on transportation and other categories experiencing significant deflation.

Chapter 17. The Consumer Price Index (Updated 2-14-2018)

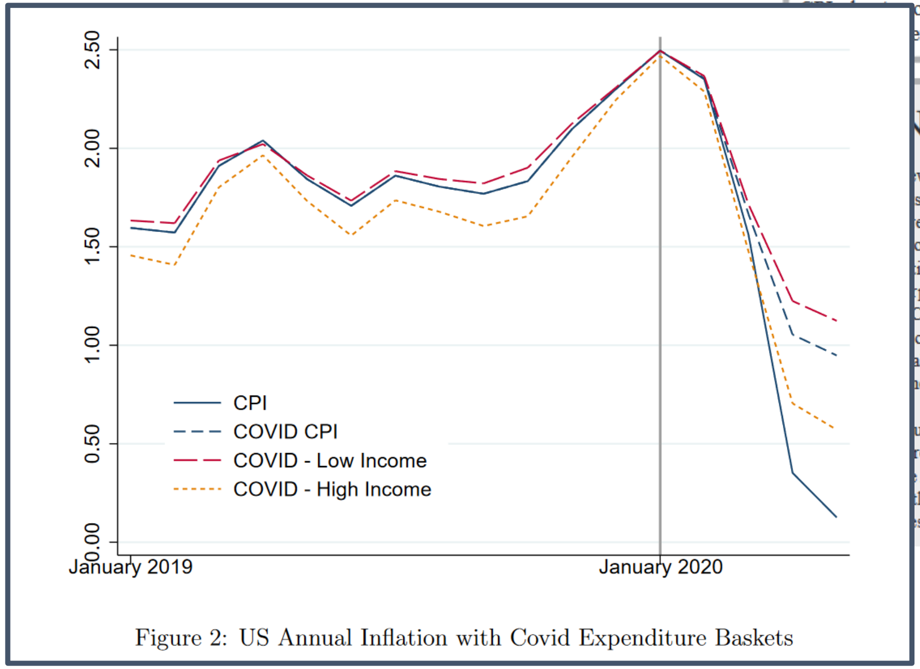
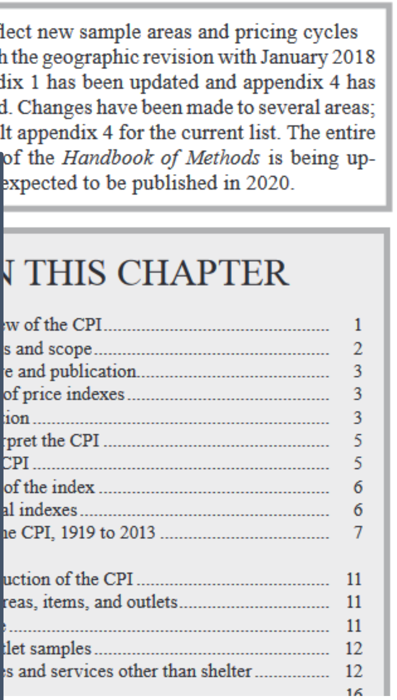


Figure 2: US Annual Inflation with Covid Expenditure Baskets



Note: To reflect new sample areas and pricing cycles effective with the geographic revision with January 2018 data, appendix 1 has been updated and appendix 4 has been replaced. Changes have been made to several areas; please consult appendix 4 for the current list. The entire of the *Handbook of Methods* is being up- expected to be published in 2020.

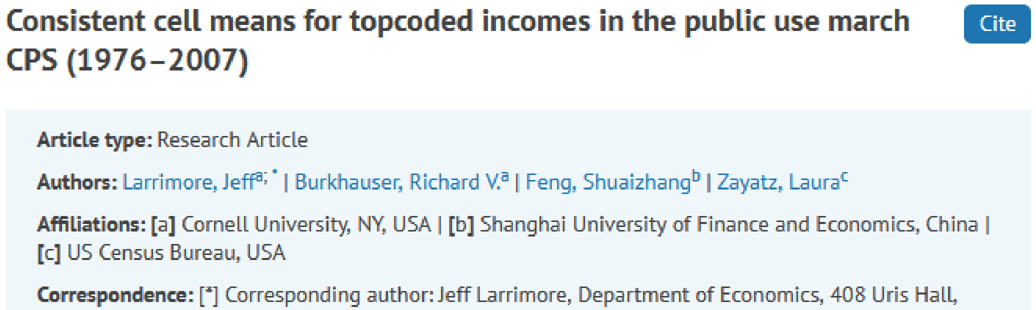
N THIS CHAPTER

Computing instructions

Case study: Topcoding in CPS

(Larrimore, Burkhauser, Feng, Zayatz, 2008)

- Topcoding affects trends in income inequality
- Ability to diagnose the problem
- (FSRDC) Ability to investigate and fix the problem



Consistent cell means for topcoded incomes in the public use march CPS (1976–2007)

Cite

Article type: Research Article

Authors: Larrimore, Jeff[a;*] | Burkhauser, Richard V.[a] | Feng, Shuaizhang[b] | Zayatz, Laura[c]

Affiliations: [a] Cornell University, NY, USA | [b] Shanghai University of Finance and Economics, China | [c] US Census Bureau, USA

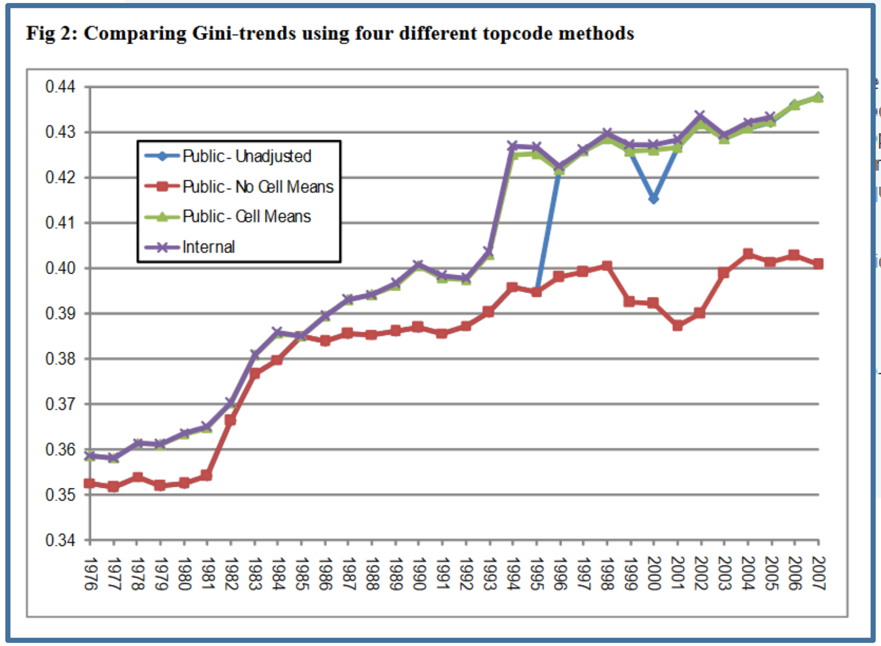
Correspondence: [*] Corresponding author: Jeff Larrimore, Department of Economics, 408 Uris Hall, 255 2818; E-mail:

to the larger research ove the topcode for any pcoded since 1976. We r with the public use ualities of the US

icient, topcoding, and cell

-128, 2008

Fig 2: Comparing Gini-trends using four different topcode methods



Computing instructions

Case study: Topcoding in CPS
(Larrimore, Burkhauser, Feng, Zayatz, 2008)

- Topcoding affects trends in income inequality
- Ability to diagnose the problem
- (FSRDC) Ability to investigate and fix the problem

Even here there are problems (versioning!)

- "… even the internal March CPS data, which are not subject to top coding, have been censored to various degrees over time…"
- "… the U.S. Census Bureau does not maintain any versions of the internal March CPS data that are not subject to this form of censoring."

Transparency in disclosure avoidance

Computing instructions

Case study: Topcoding in CPS

(Larrimore, Burkhauser, Feng, Zayatz, 2008)

- Topcoding
income in
- Ability to
- (FSRDC) A
and fix the

Even here there are problems (versioning!)

rch CPS
ect to
ensored
time…”
au does
ns of the
that are
of

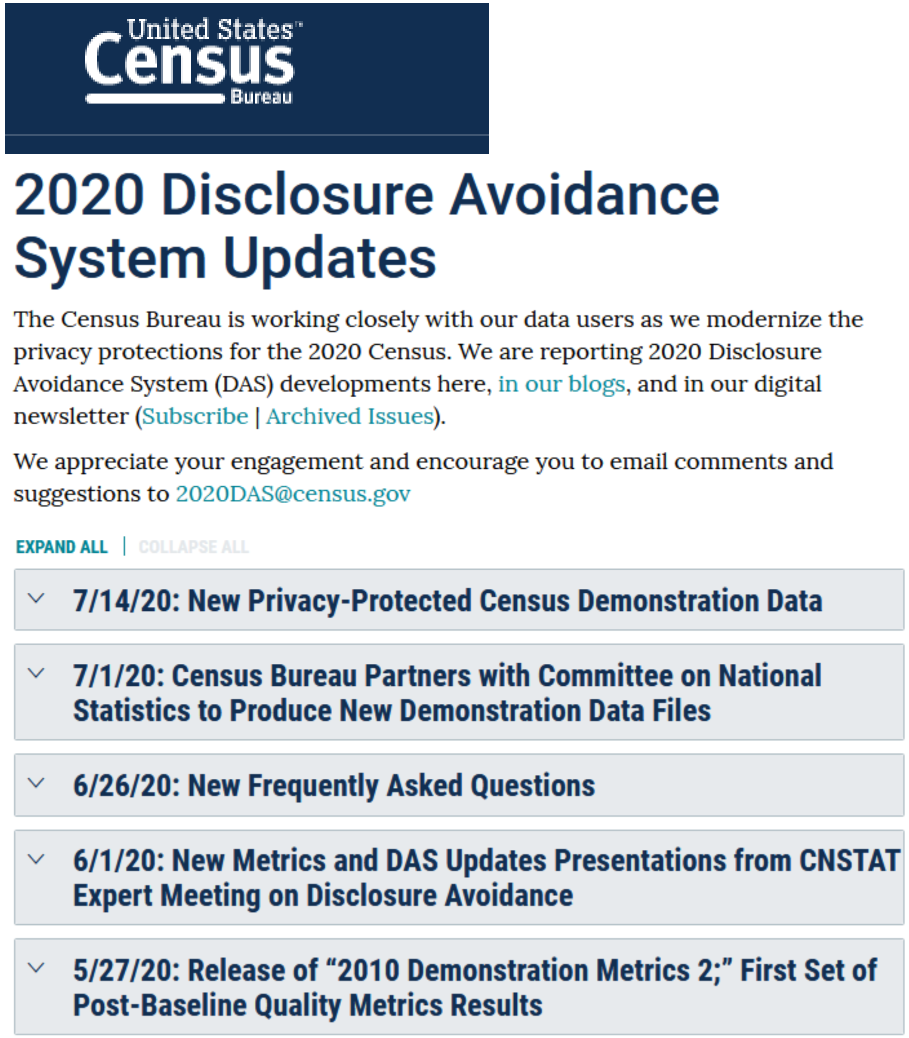
Most of the U.S. Census Bureau procedures for creating cell means for topcoded values in the public use March CPS data can be found in the 2007 Current Population Survey Annual Demographic File Technical Documentation [10], but in some cases we learned about them via conversations with various U.S. Census Bureau employees charged with creating the cell means

Consistent Cell Means for Topcoded Incomes in the Public Use March CPS (1976-2007)

US Census Bureau Center for Economic Studies Paper No. CES-WP-08-06

Disclosure avoidance

- New disclosure avoidance in 2020 Decennial Census of Population (source)



United States Census Bureau

2020 Disclosure Avoidance System Updates

The Census Bureau is working closely with our data users as we modernize the privacy protections for the 2020 Census. We are reporting 2020 Disclosure Avoidance System (DAS) developments here, in our blogs, and in our digital newsletter (Subscribe | Archived Issues).

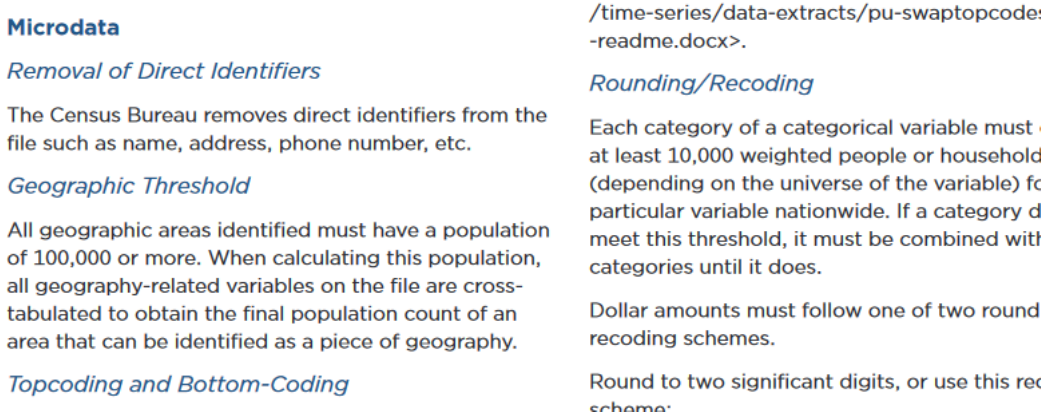
We appreciate your engagement and encourage you to email comments and suggestions to 2020DAS@census.gov

- 7/14/20: New Privacy-Protected Census Demonstration Data
- 7/1/20: Census Bureau Partners with Committee on National Statistics to Produce New Demonstration Data Files
- 6/26/20: New Frequently Asked Questions
- 6/1/20: New Metrics and DAS Updates Presentations from CNSTAT Expert Meeting on Disclosure Avoidance
- 5/27/20: Release of “2010 Demonstration Metrics 2;” First Set of Post-Baseline Quality Metrics Results

Disclosure avoidance

- New disclosure avoidance in 2020 Decennial Census of Population
- Recent summary of CPS disclosure avoidance measures (source)

Research and Methodology Directorate
A History of the Current Population Survey and Disclosure Avoidance
Census



Microdata

Removal of Direct Identifiers

The Census Bureau removes direct identifiers from the file such as name, address, phone number, etc.

Geographic Threshold

All geographic areas identified must have a population of 100,000 or more. When calculating this population, all geography-related variables on the file are cross-tabulated to obtain the final population count of an area that can be identified as a piece of geography.

Topcoding and Bottom-Coding

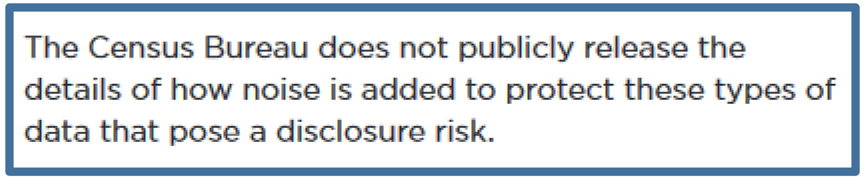
/time-series/data-extracts/pu-swaptopcode -readme.docx>.

Rounding/Recoding

Each category of a categorical variable must at least 10,000 weighted people or household (depending on the universe of the variable) fc particular variable nationwide. If a category d meet this threshold, it must be combined witl categories until it does.

Dollar amounts must follow one of two round recoding schemes.

Round to two significant digits, or use this rec scheme:



The Census Bureau does not publicly release the details of how noise is added to protect these types of data that pose a disclosure risk.

Examples in Official Statistics

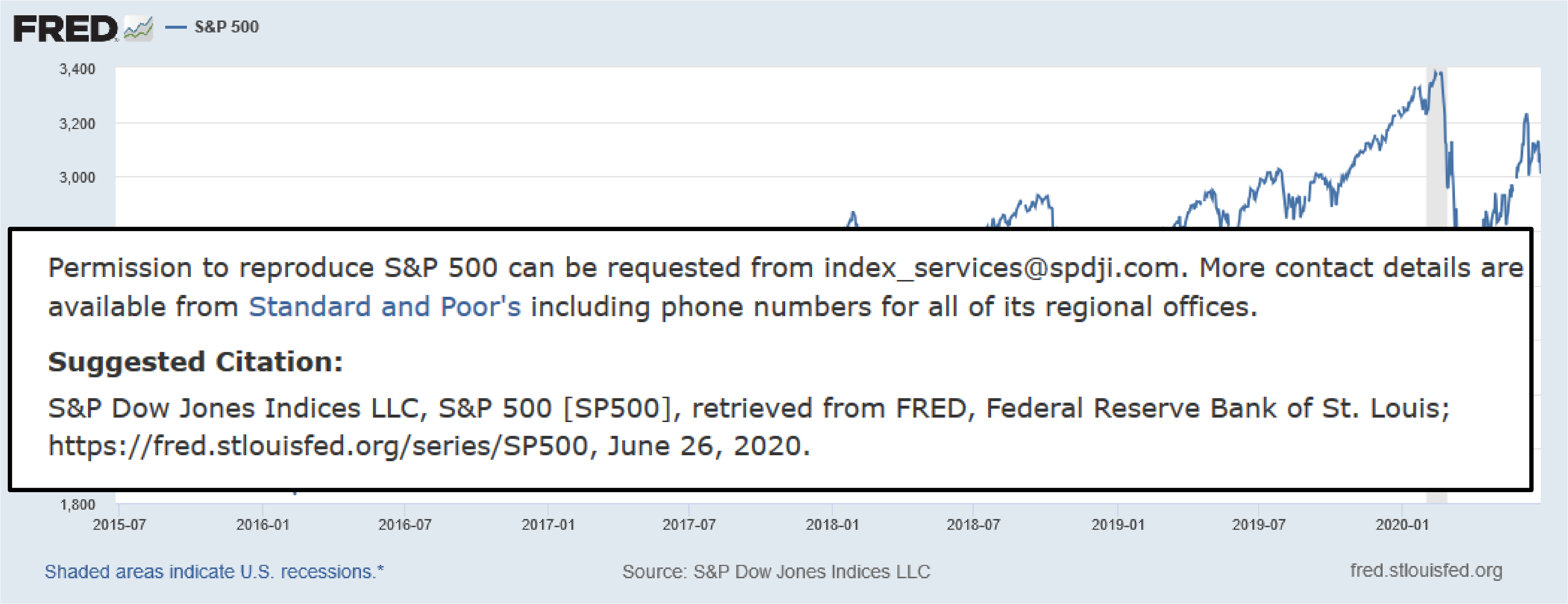
- Trusted archives: **none**
- **Reliable versioning** of output data:
    - Good example: BEA
    - Counter example: BLS (great accessibility, poor versioning)
    - **No persistent identifiers**!
- Availability of **computing instructions**
    - Good example: BLS (CPS, unemployment, inflation rate)
    - Counter example: much else
    - **Almost never code!**
- **Transparency of disclosure avoidance**
    - Good example: Census 2020, but also CPS
    - Bad example: much else

Mechanisms available in 2020

- Documentation of methods
  - Standard for surveys (but often not in a standard way)
  - Not standard for most non-survey data
  - Standards available (DDI, SDMX, DCAT-US, etc.)
- Code releases
  - Open source analogy
    - Consider encryption library SSL: widely used, open source.
    - Errors are detected occasionally, not always by the authors
- Access to data
  - Open Data is good, but also needs reliably versioned data
  - Tiered Access: Previous presentation

Some easy patches

Provide data citations, permissions



Permission to reproduce S&P 500 can be requested from index_services@spdji.com. More contact details are available from Standard and Poor's including phone numbers for all of its regional offices.

Suggested Citation:

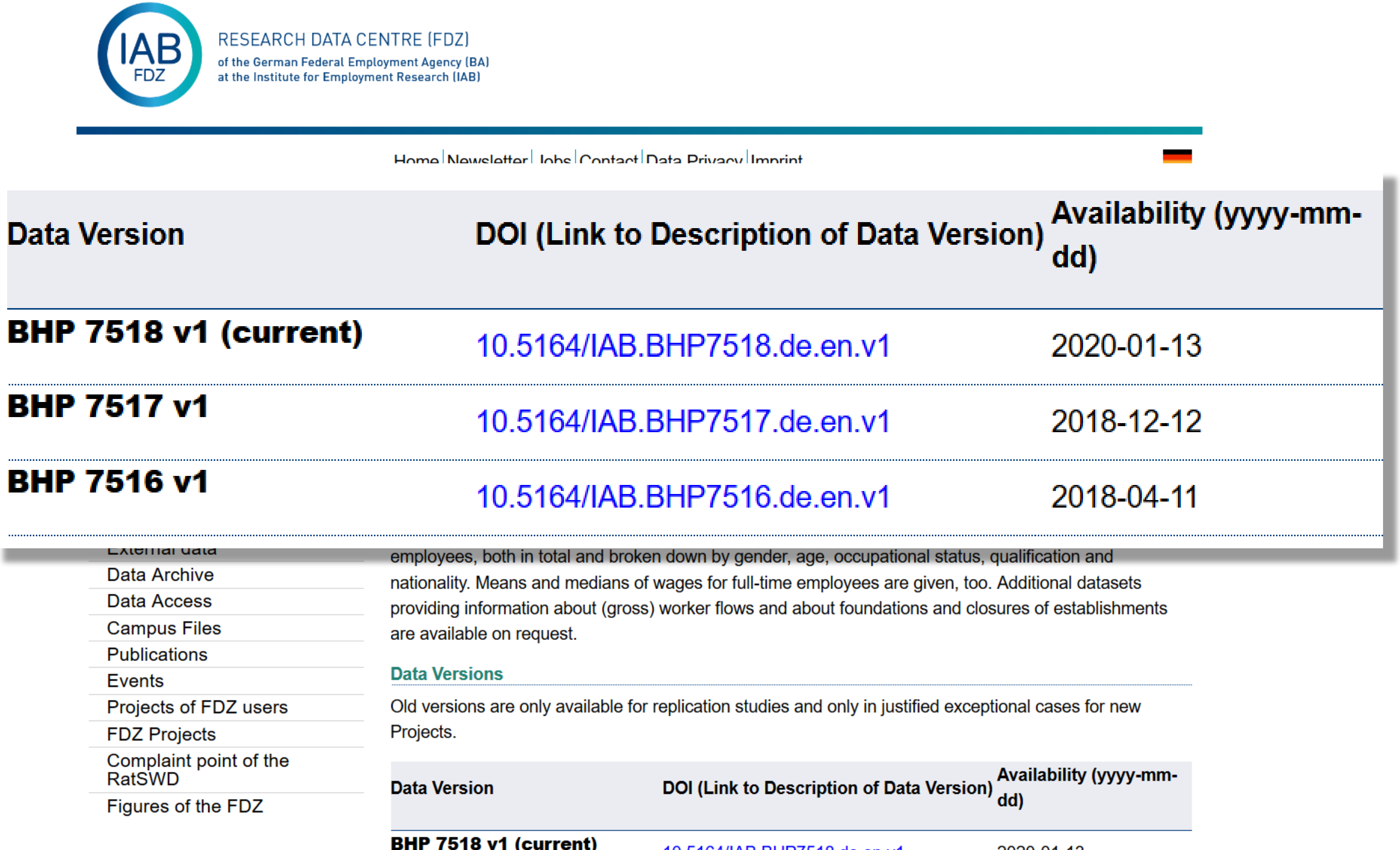
S&P Dow Jones Indices LLC, S&P 500 [SP500], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SP500, June 26, 2020.

Shaded areas indicate U.S. recessions.*

Source: S&P Dow Jones Indices LLC

fred.stlouisfed.org

Assign PID to data assets – even confidential



RESEARCH DATA CENTRE (FDZ)
of the German Federal Employment Agency (BA)
at the Institute for Employment Research (IAB)

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

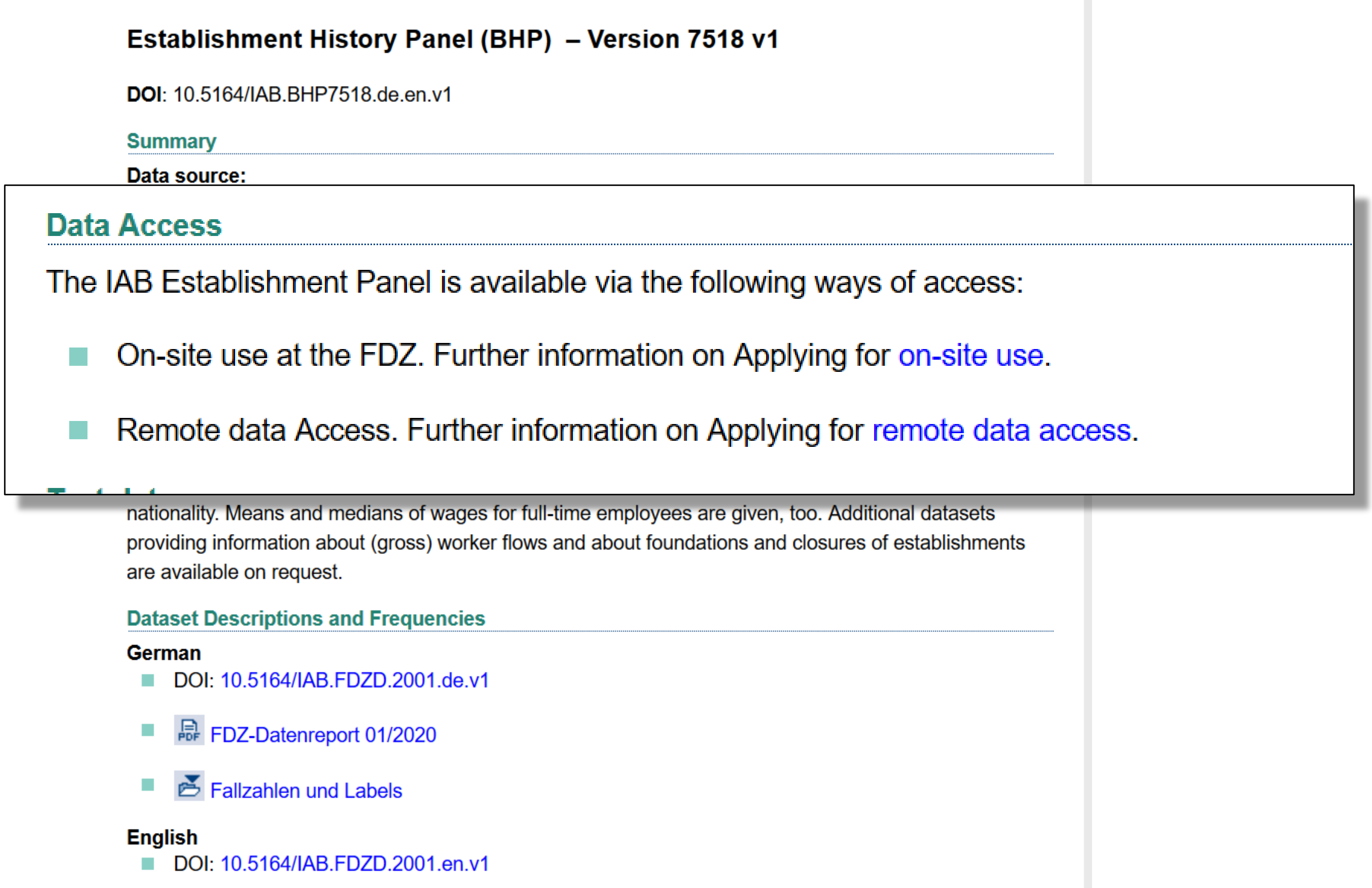
employees, both in total and broken down by gender, age, occupational status, qualification and nationality. Means and medians of wages for full-time employees are given, too. Additional datasets providing information about (gross) worker flows and about foundations and closures of establishments are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

Conclusion

- Transparency (of sources, of processing) is a key requirement of official statistics
- While some good examples (and benefits) exist, no consistency across US official statistics
- Transparency can carry reputational risks – need robust institutions
- Some low-hanging fruit could increase transparency and assist computational reproducibility

Thank you!

https://doi.org/10.5281/zenodo.3974666