Reproducibility of Scientific Results

(Central Banks)

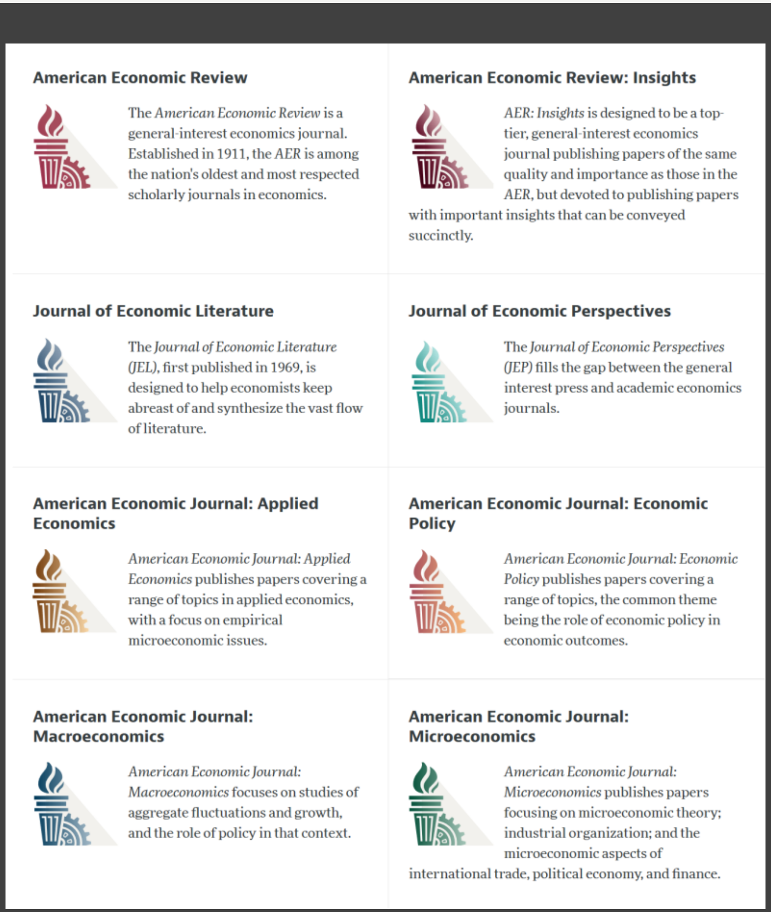
Lars Vilhuber

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.

Context

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.**

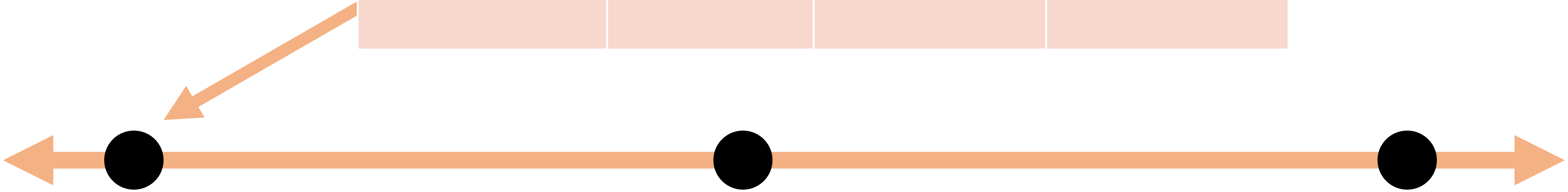
Current efforts at the AEA

- **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

Replication continuum

https://doi.org/10.17226/25303

Same data	Same code	Same methods	Same context



Reproducibility

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

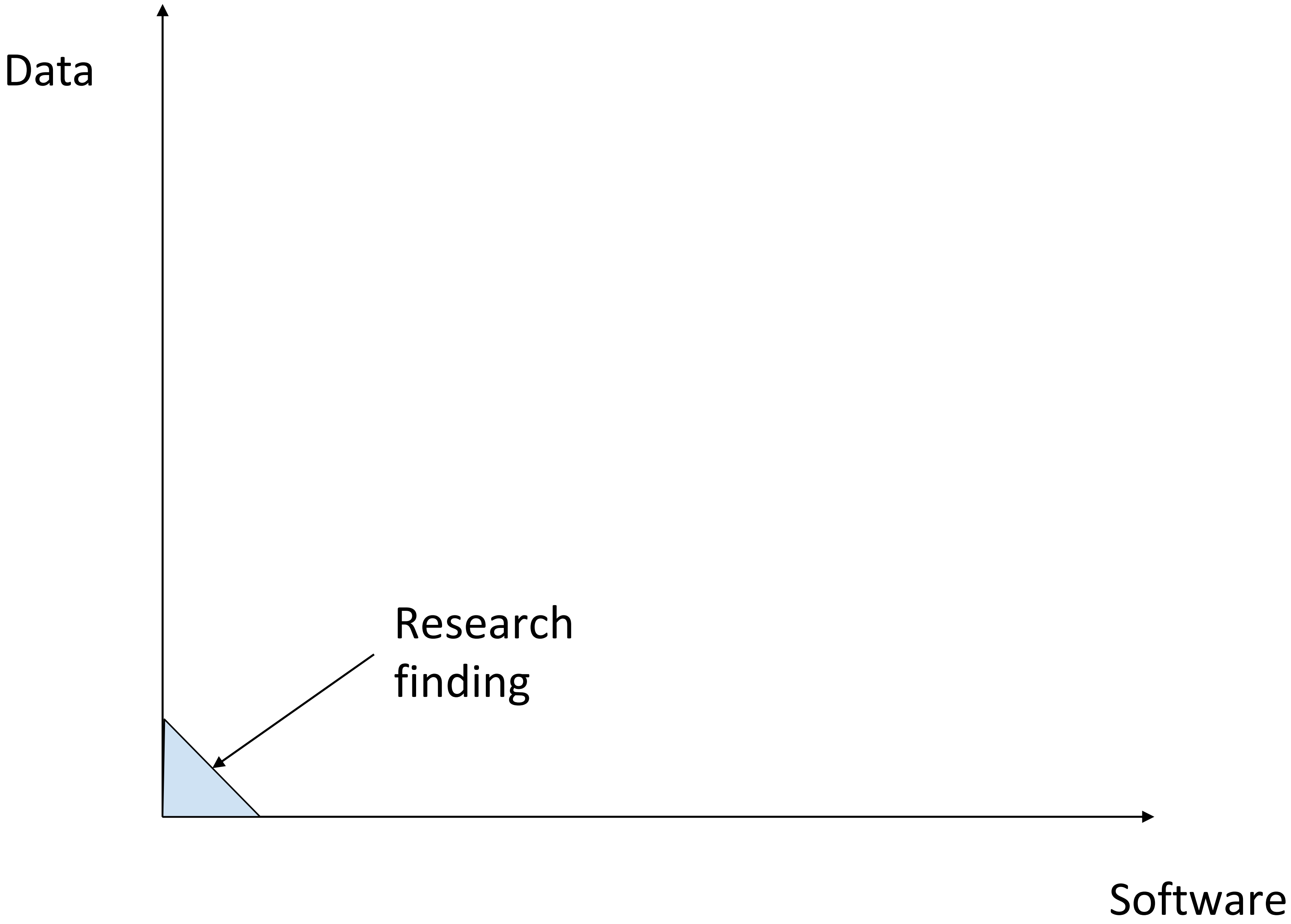
Replicability

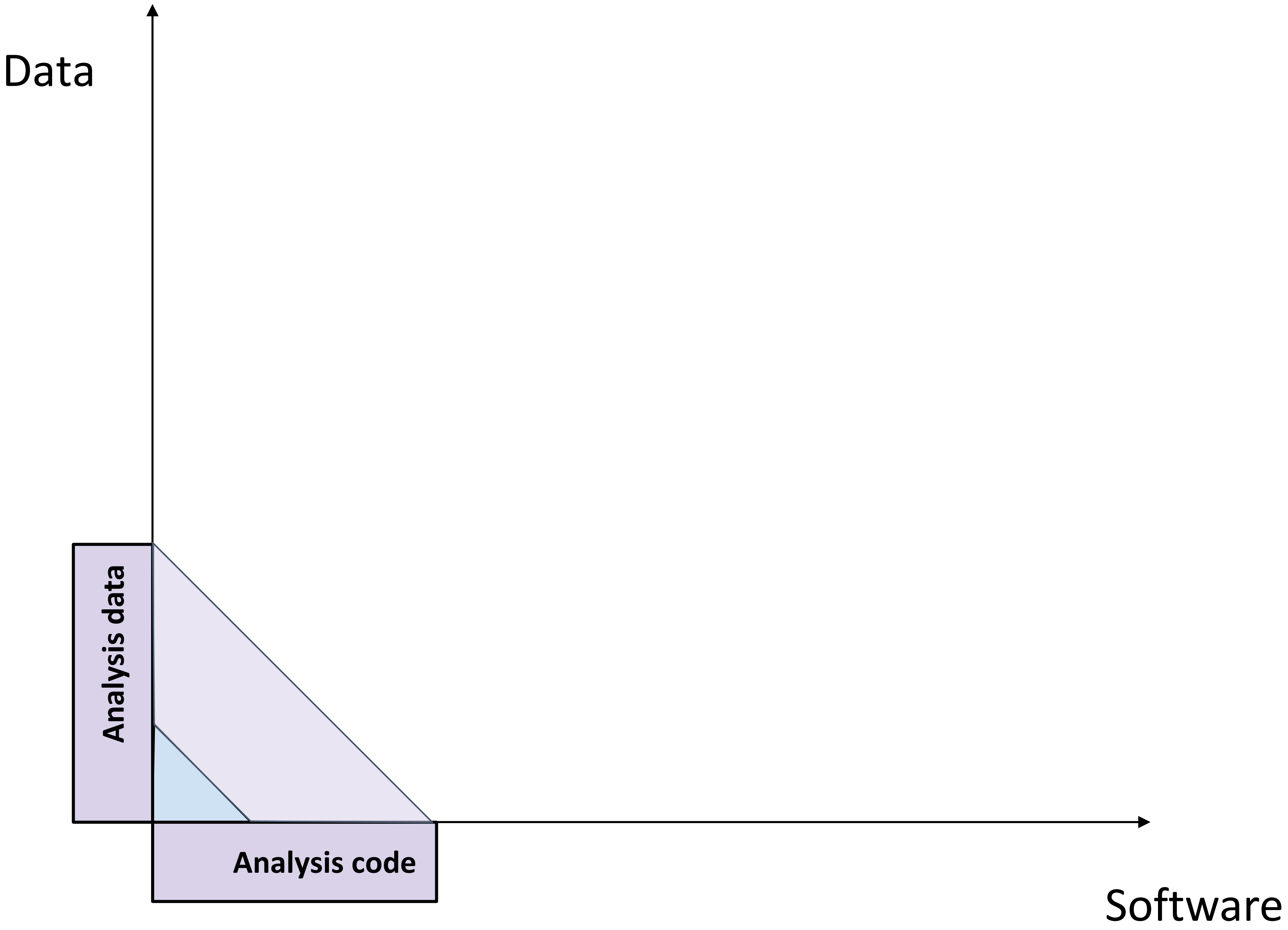
- Wide Replication (Pesaran 2003)
- Statistical Replication (Hamermesh 2007)
- Reproduction/Reanalysis (Clemens 2015)

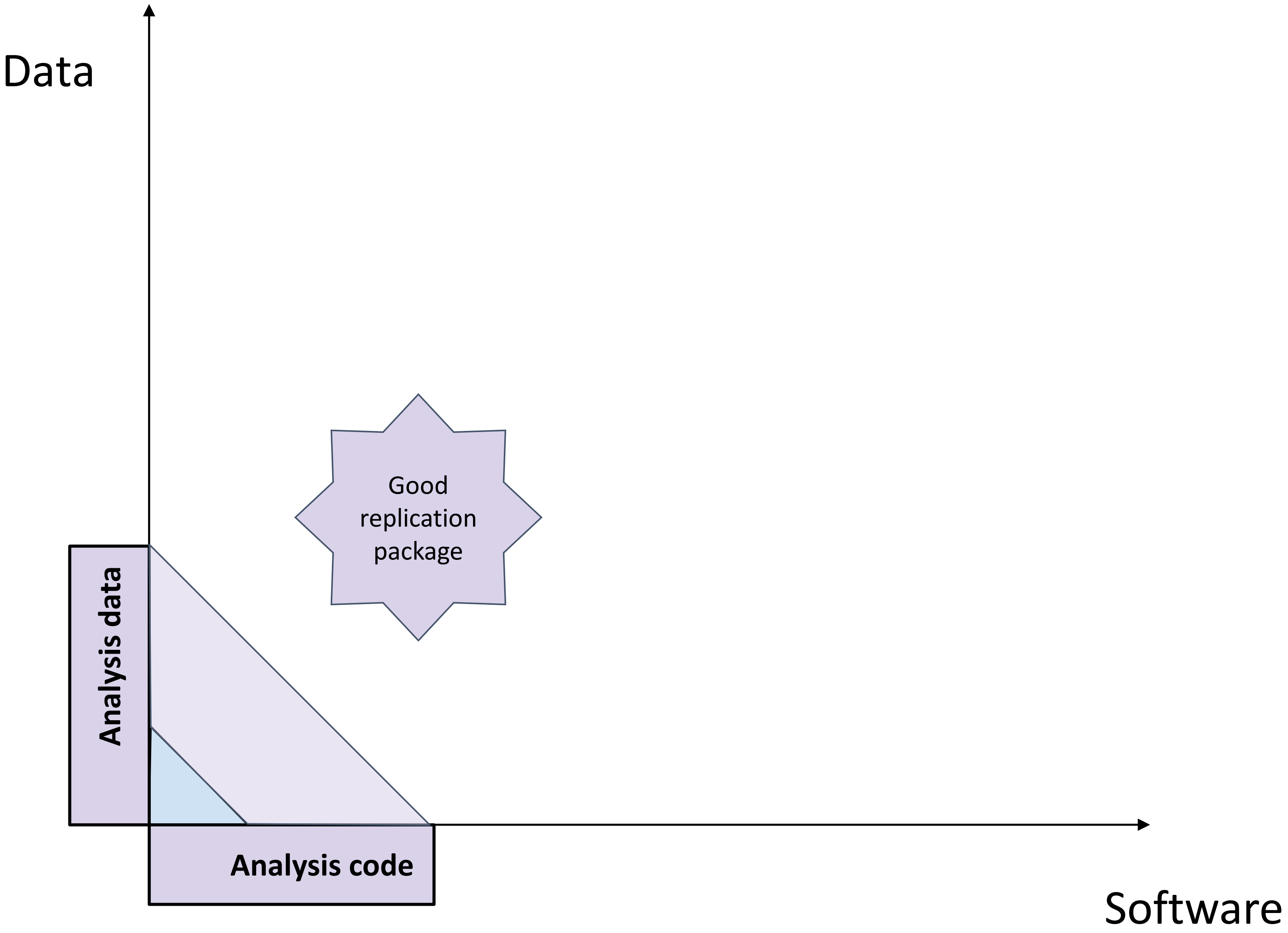
Generalizability

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

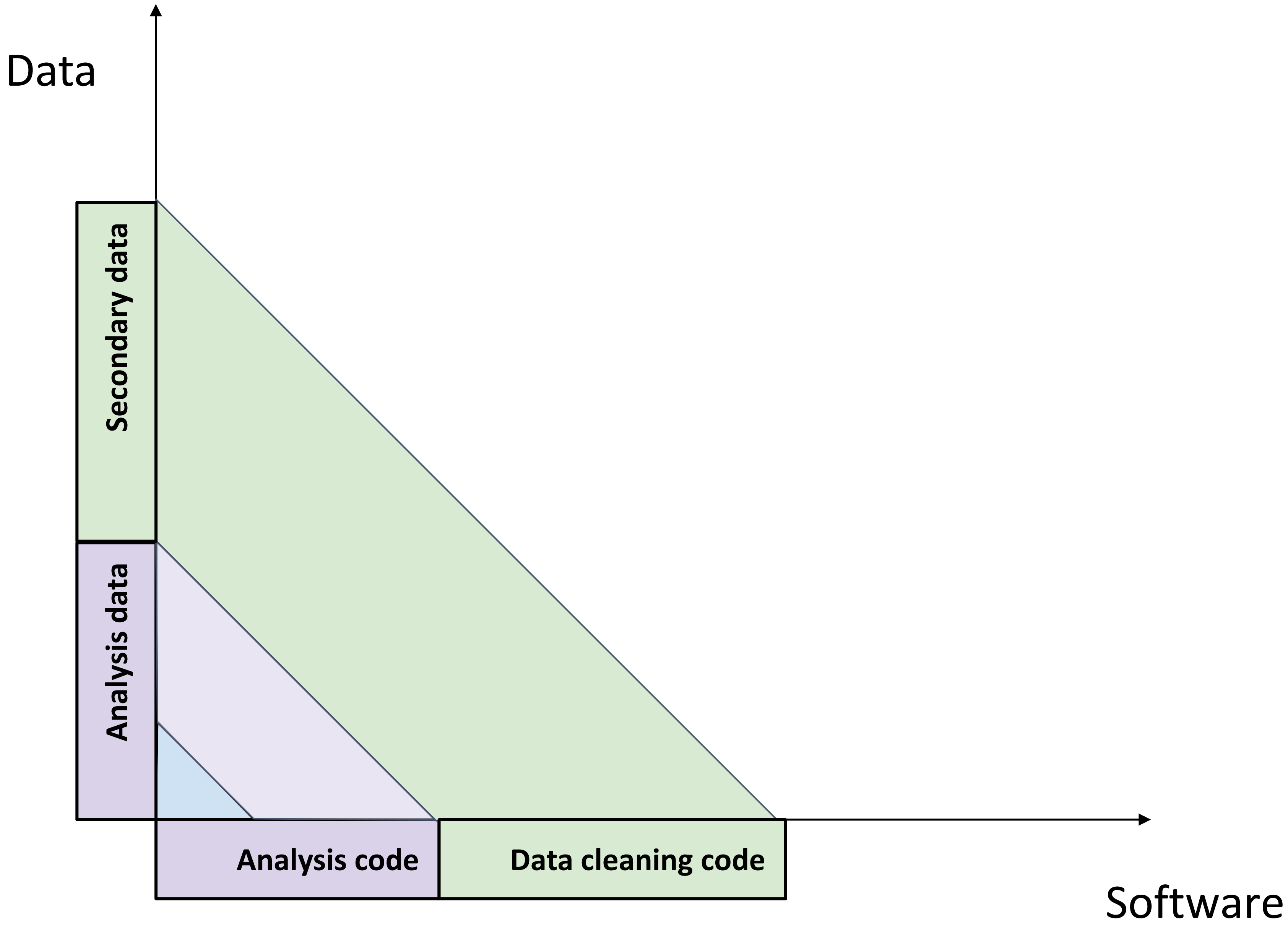
Status quo ante

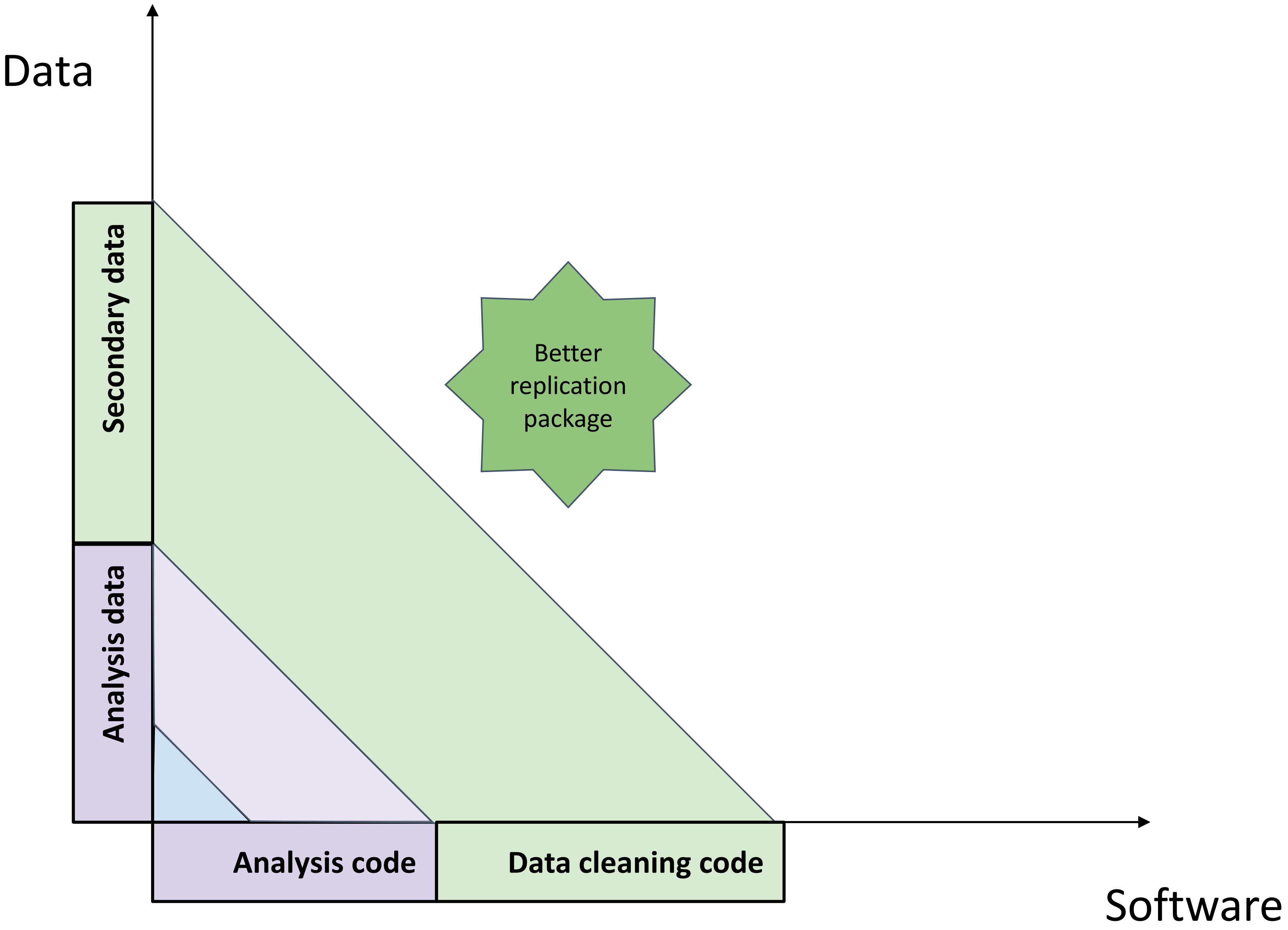

Data
Research finding
Software

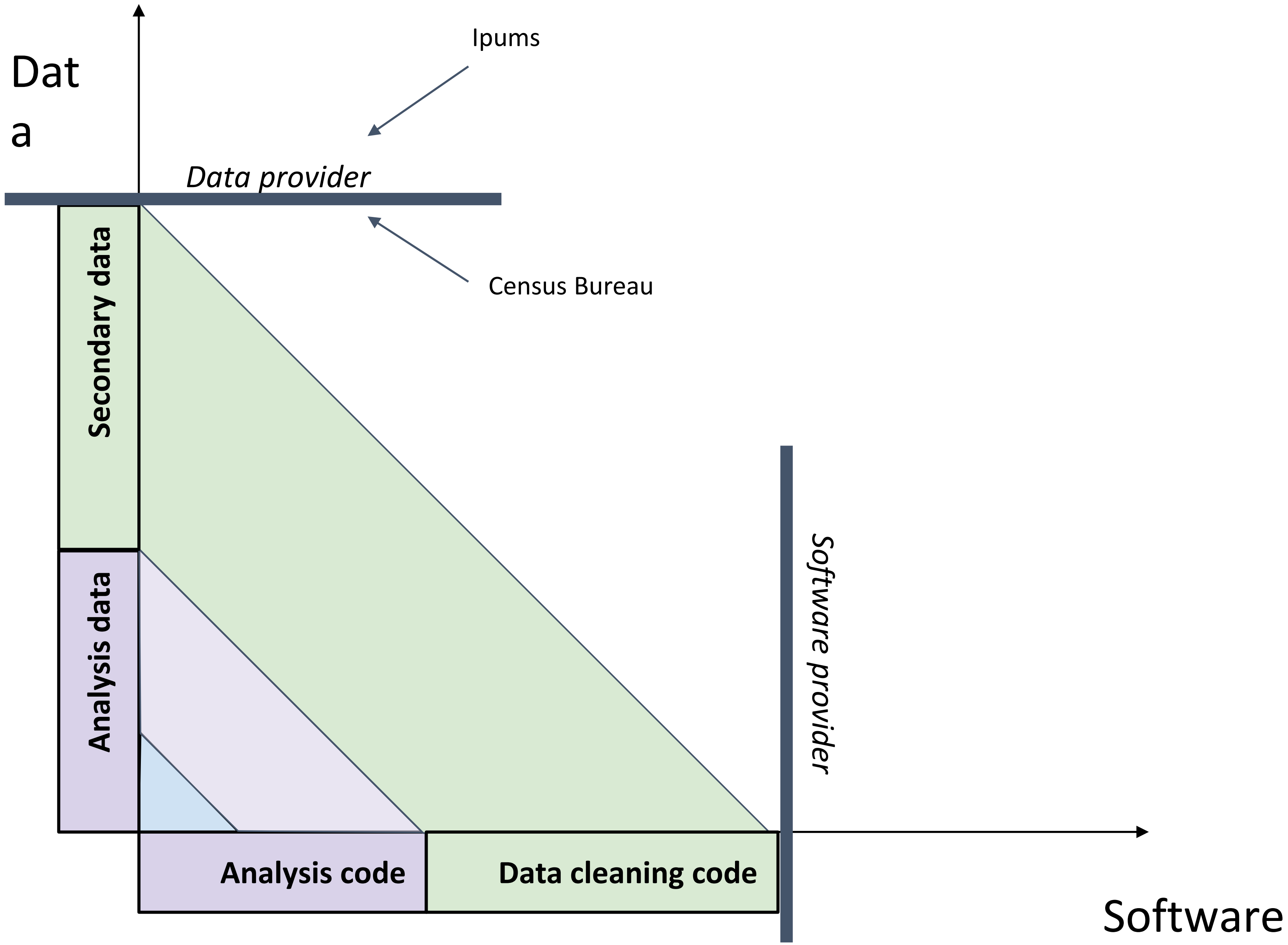

Data
Analysis data
Analysis code
Software


Data
Good replication package
Analysis data
Analysis code
Software

Status quo


Data
Secondary data
Analysis data
Analysis code
Data cleaning code
Software


Data
Secondary data
Analysis data
Better replication package
Analysis code
Data cleaning code
Software



Dat
a

Ipums

Data provider

Census Bureau

Secondary data

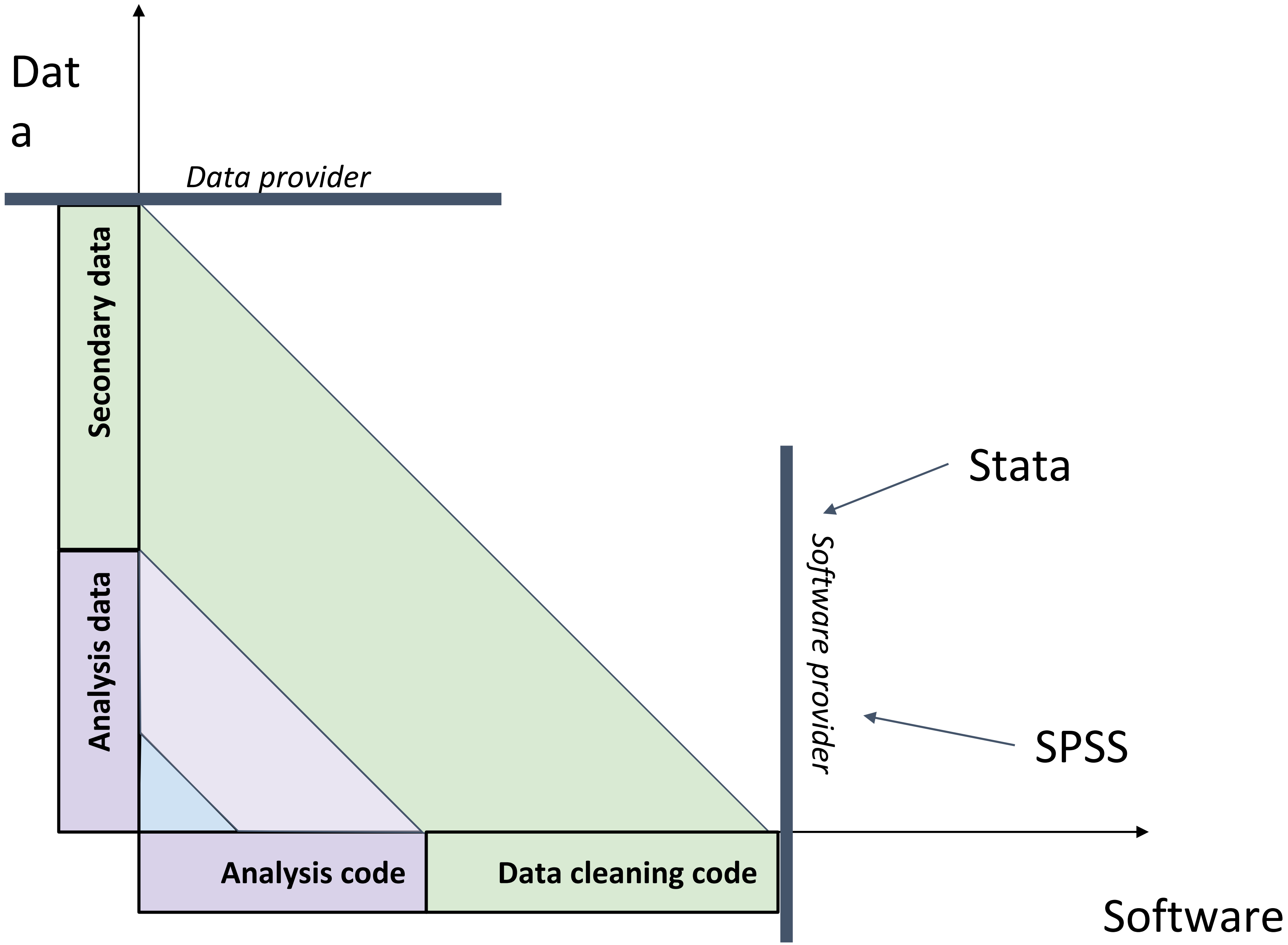
Analysis data

Software provider

Analysis code

Data cleaning code

Software


Dat
a
Data provider
Secondary data
Analysis data
Analysis code
Data cleaning code
Software provider
Stata
SPSS
Software

Topics

Topics

- Data Provenance
  - Typical Macro
  - Commercial data intermediaries
  - Commercial data sources

- Reproducibility in Secure Remote Access Systems
  - External pre-publication [cascad], as-publication [Marianne], post-publication [ReplicationWiki, SSRP]

Topics

- Is this an academic discussion only?
  - Policy briefs and reproducibility [Sylverie]
  - New types of publications (Dynamic / interactive documents) [Julia]
- Pushing the technological frontier
  - Docker for interactive use [Thibaud]
  - Continuous-integration in public and private spheres
  - Certificates of reproducibility

Data provenance

Poor citation practices

- **Macrodata:**

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

- **Microdata**:

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

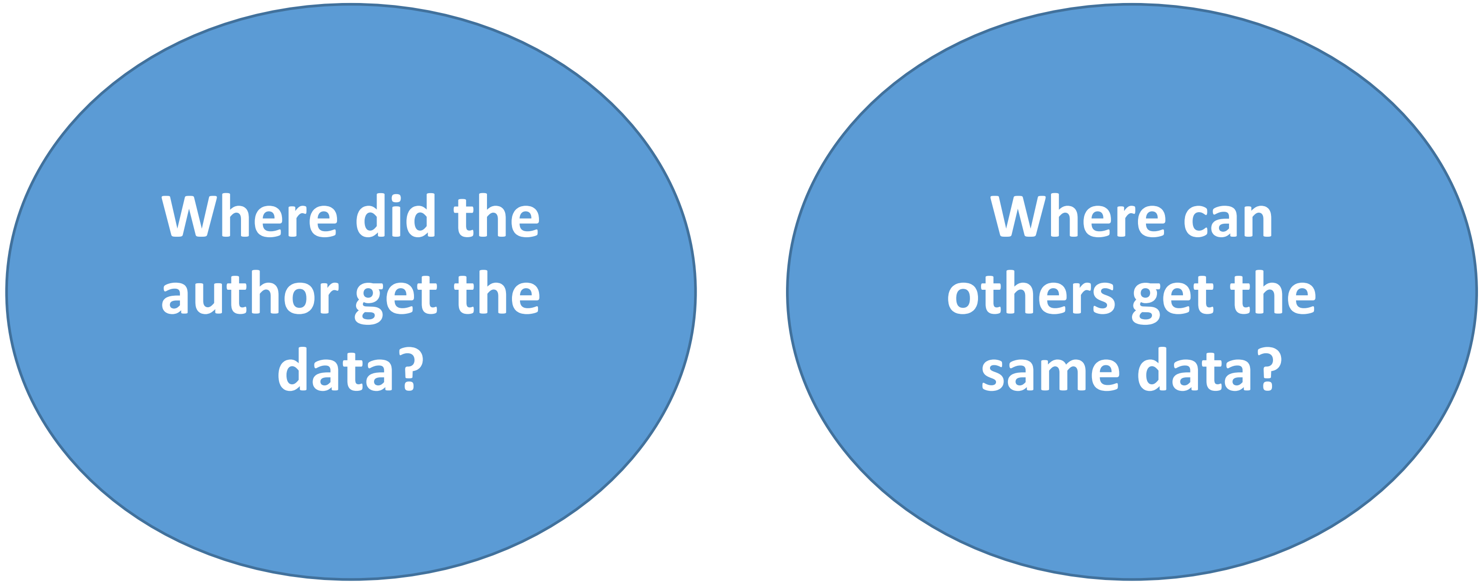
Three pieces

Where did the author get the data?

Where can others get the same data?

Where does the value-added data go?

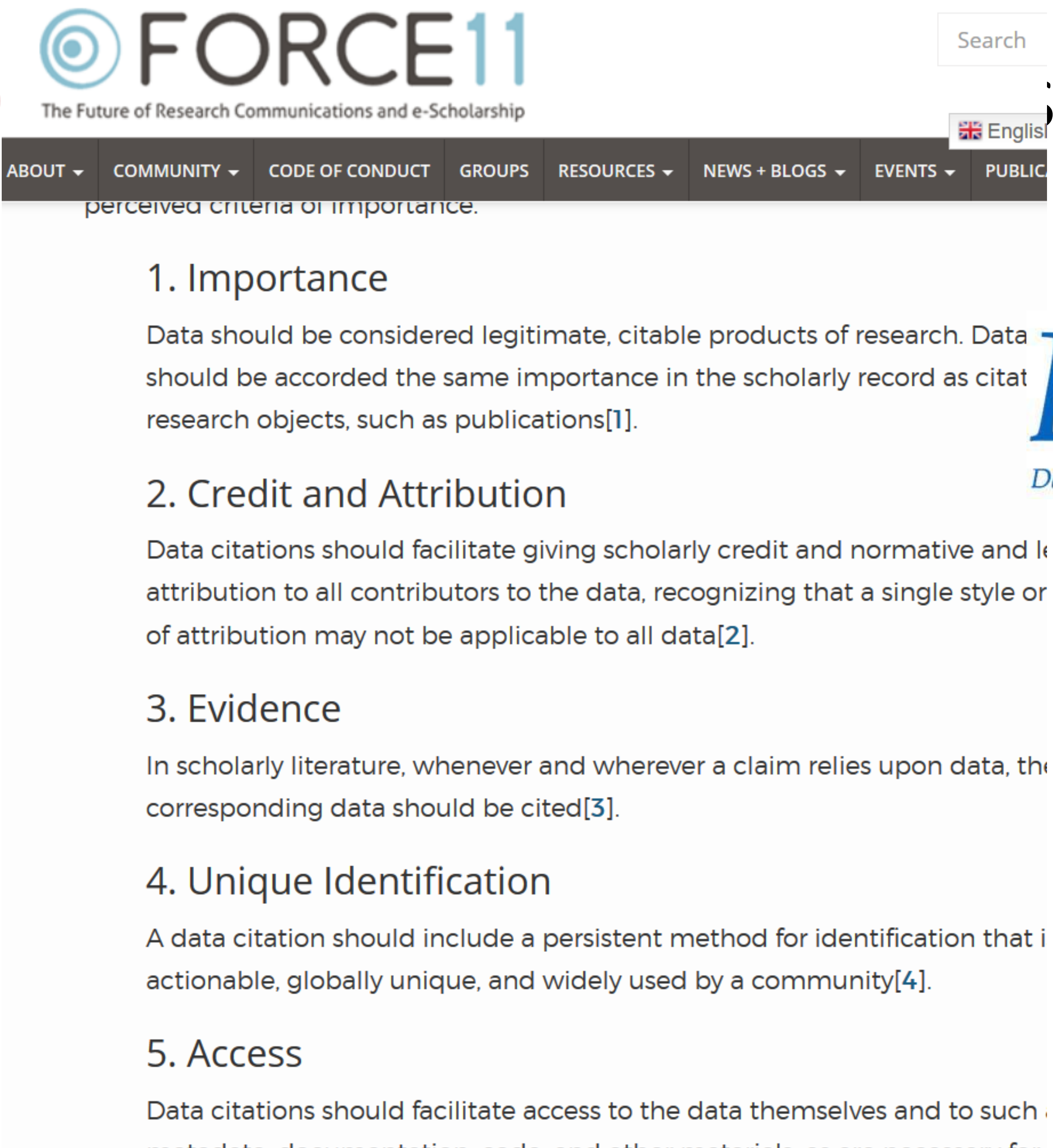
Three pieces



Where did the author get the data?

Where can others get the same data?

Challenge:
Authors are bad at documenting provenance



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

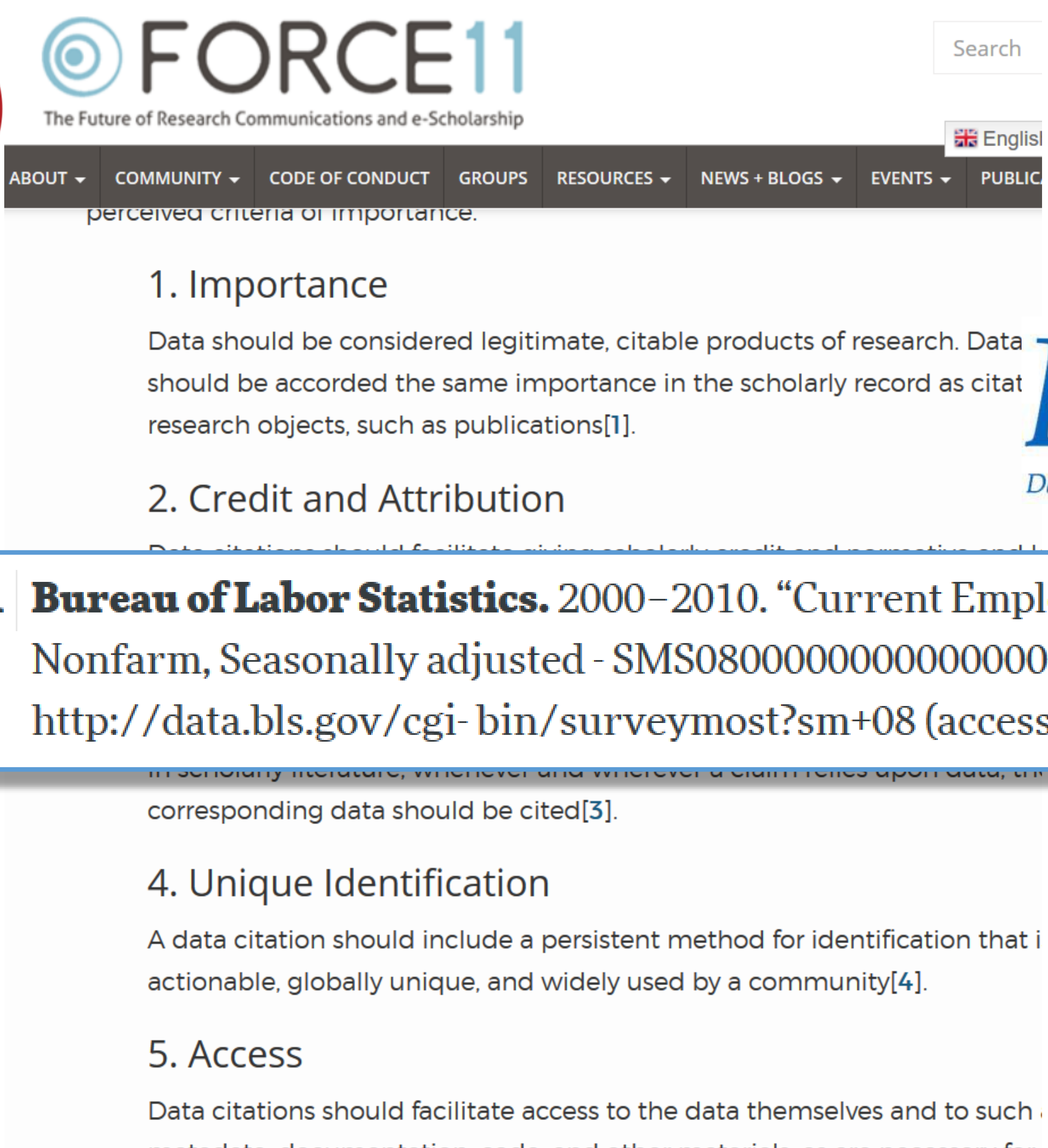
Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [https://www.force11.org/group/joint-declaration-data-citation-principles-final].

How do you document the provenance when you cannot use the data?

Wrong question!

How do you document data provenance?

- What do you need to request?
  - Name, specification, DOI, etc.
- Where do you need to request it?
  - Website, an archive, a Freedom of Information Act officer, etc.
- Details, details:
  - Copy of your request form?
  - Copy of your request letter?
  - Etc.
- Don’t assume (too much) prior knowledge!



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).

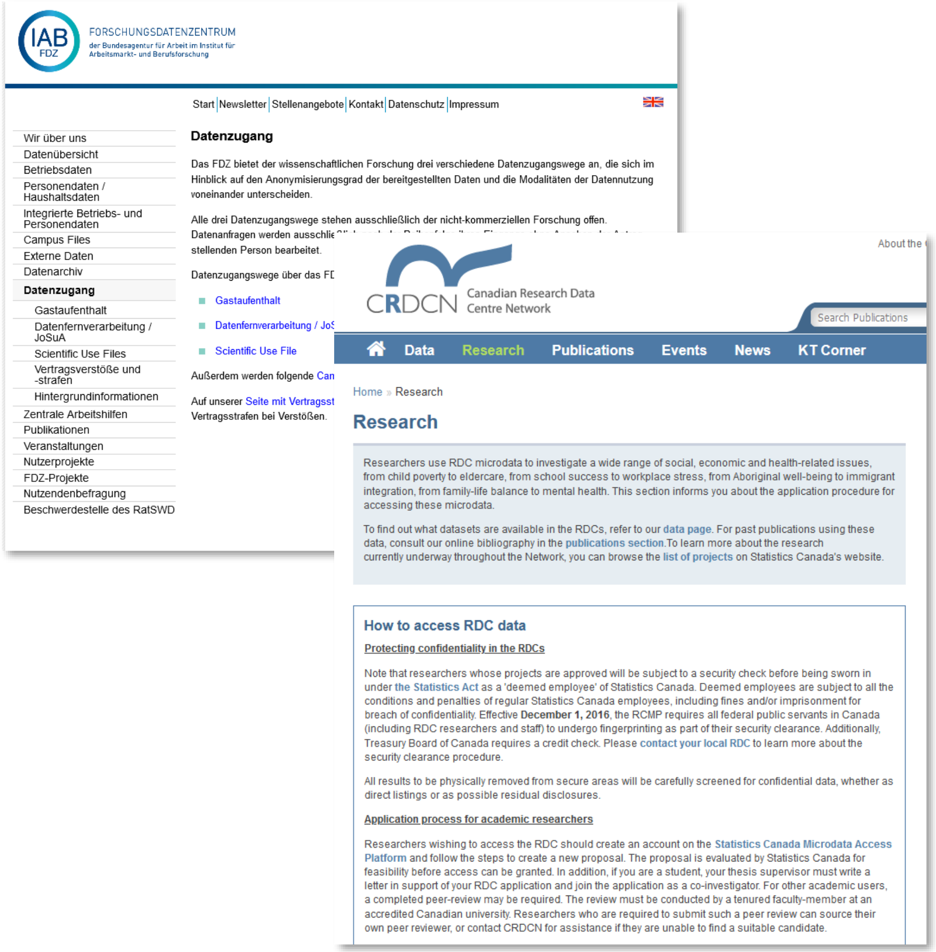
DC1
Data Citation Principles

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].

How did you get the data in first place?

- You **applied** for the data **through a process**
- You **purchased** the data from a provider
- You signed an **Non-Disclosure Agreement (NDA)** with a company
- Your **university** has an **agreement** with a data provider

…



IAB FDZ
FORSCHUNGSDATENZENTRUM
der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung
Start | Newsletter | Stellenangebote | Kontakt | Datenschutz | Impressum
Wir über uns
Datenübersicht
Betriebsdaten
Personendaten / Haushaltsdaten
Integrierte Betriebs- und Personendaten
Campus Files
Externe Daten
Datenarchiv
Datenzugang
Gastaufenthalt
Datenfernverarbeitung / JoSuA
Scientific Use Files
Vertragsverstöße und -strafen
Hintergrundinformationen
Zentrale Arbeitshilfen
Publikationen
Veranstaltungen
Nutzerprojekte
FDZ-Projekte
Nutzendenbefragung
Beschwerdestelle des RatSWD

Datenzugang

Das FDZ bietet der wissenschaftlichen Forschung drei verschiedene Datenzugangswege an, die sich im Hinblick auf den Anonymisierungsgrad der bereitgestellten Daten und die Modalitäten der Datennutzung voneinander unterscheiden.

Alle drei Datenzugangswege stehen ausschließlich der nicht-kommerziellen Forschung offen. Datenanfragen werden ausschlie stellenden Person bearbeitet.

Datenzugangswege über das FD

- Gastaufenthalt
- Datenfernverarbeitung / JoS
- Scientific Use File

Außerdem werden folgende Can

Auf unserer Seite mit Vertragsst Vertragsstrafen bei Verstößen.

Research

Researchers use RDC microdata to investigate a wide range of social, economic and health-related issues, from child poverty to eldercare, from school success to workplace stress, from Aboriginal well-being to immigrant integration, from family-life balance to mental health. This section informs you about the application procedure for accessing these microdata.

To find out what datasets are available in the RDCs, refer to our data page. For past publications using these data, consult our online bibliography in the publications section.To learn more about the research currently underway throughout the Network, you can browse the list of projects on Statistics Canada's website.

How to access RDC data

Protecting confidentiality in the RDCs

Note that researchers whose projects are approved will be subject to a security check before being sworn in under the Statistics Act as a 'deemed employee' of Statistics Canada. Deemed employees are subject to all the conditions and penalties of regular Statistics Canada employees, including fines and/or imprisonment for breach of confidentiality. Effective **December 1, 2016**, the RCMP requires all federal public servants in Canada (including RDC researchers and staff) to undergo fingerprinting as part of their security clearance. Additionally, Treasury Board of Canada requires a credit check. Please contact your local RDC to learn more about the security clearance procedure.

All results to be physically removed from secure areas will be carefully screened for confidential data, whether as direct listings or as possible residual disclosures.

Application process for academic researchers

Researchers wishing to access the RDC should create an account on the Statistics Canada Microdata Access Platform and follow the steps to create a new proposal. The proposal is evaluated by Statistics Canada for feasibility before access can be granted. In addition, if you are a student, your thesis supervisor must write a letter in support of your RDC application and join the application as a co-investigator. For other academic users, a completed peer-review may be required. The review must be conducted by a tenured faculty-member at an accredited Canadian university. Researchers who are required to submit such a peer review can source their own peer reviewer, or contact CRDCN for assistance if they are unable to find a suitable candidate.

How did you get the data in first place?

- You **applied** for the data **through a process**
- You **purchased** the data from a provider
- You signed an **Non-Disclosure Agreement (NDA)** with a company
- Your **university** has an **agreement** with a data provider

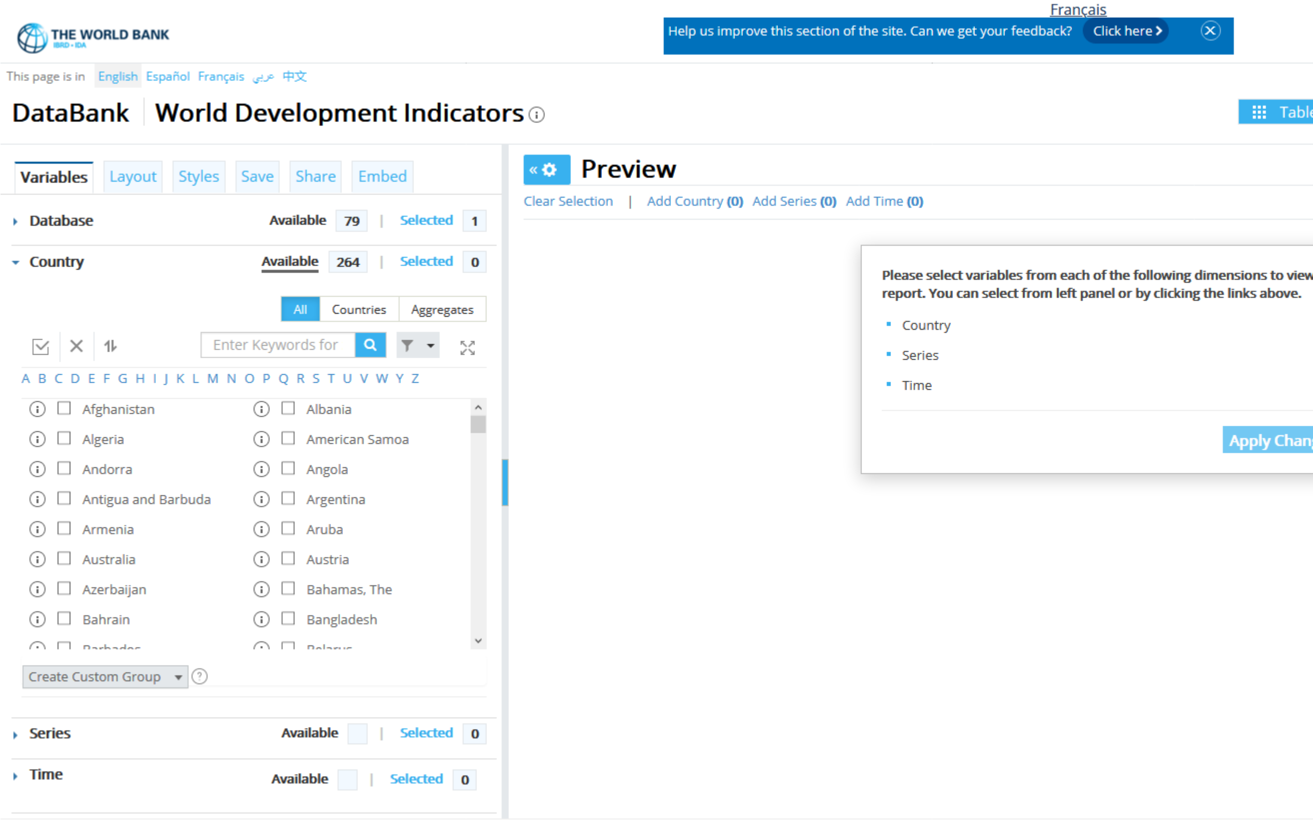
…

- Be thorough
  - Do not assume that the reader knows the details, or the conditions
- Be cognizant of what your university might have contributed
  - Maybe they set up a local access point
  - Maybe a safe room

You must have described the data

- You must have **named** the dataset you wanted
- You downloaded the data from from an **online query system**
- You **specified the extract** from a company database (in words, in SQL, etc.)

…



THE WORLD BANK
IBRD • IDA

Français

Help us improve this section of the site. Can we get your feedback? Click here ›

This page is in English Español Français عربي 中文

DataBank | World Development Indicators

Table

Variables | Layout | Styles | Save | Share | Embed

▸ Database — Available 79 | Selected 1

▾ Country — Available 264 | Selected 0

All | Countries | Aggregates

Enter Keywords for

A B C D E F G H I J K L M N O P Q R S T U V W Y Z

□ Afghanistan	□ Albania
□ Algeria	□ American Samoa
□ Andorra	□ Angola
□ Antigua and Barbuda	□ Argentina
□ Armenia	□ Aruba
□ Australia	□ Austria
□ Azerbaijan	□ Bahamas, The
□ Bahrain	□ Bangladesh
□ Barbados	□ Belarus

Create Custom Group

▸ Series — Available | Selected 0

▸ Time — Available | Selected 0

Preview

Clear Selection | Add Country **(0)** Add Series **(0)** Add Time **(0)**

Please select variables from each of the following dimensions to view report. You can select from left panel or by clicking the links above.

- Country
- Series
- Time

Apply Chan

You must have described the data

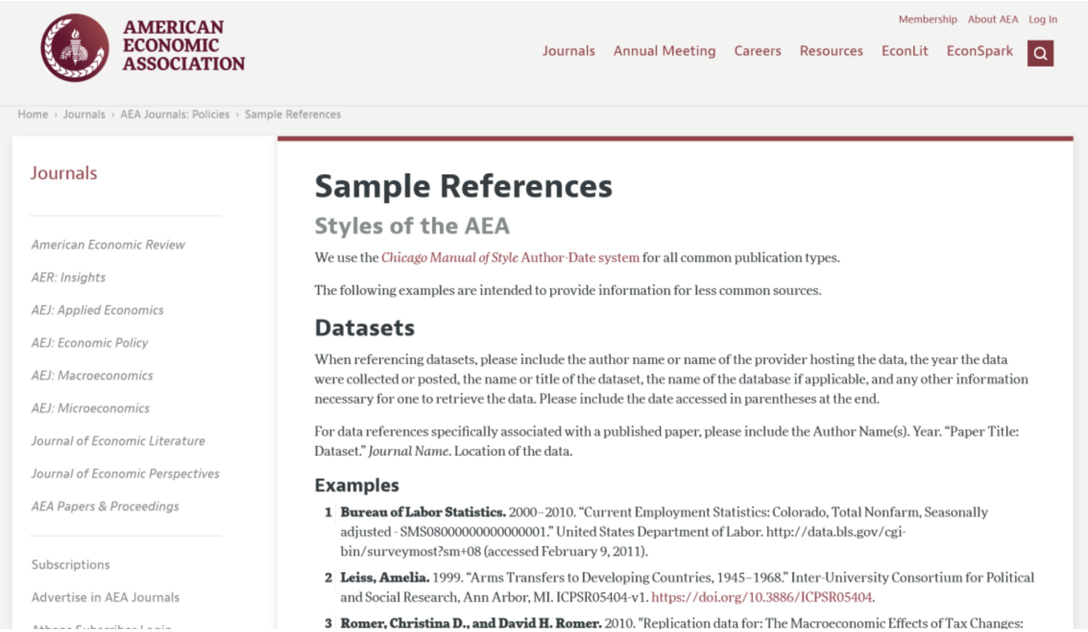
- You must have **named** the dataset you wanted
- You downloaded the data from from an **online query system**
- You **specified the extract** from a company database (in words, in SQL, etc.)

…

- Be thorough and precise
  - Is there a unique identifier?
  - Does your provider have a unique way of accessing it?
  - Is each access a custom dataset?
  - Does the provider keep older versions?

Guidance

Direct guidance



AMERICAN ECONOMIC ASSOCIATION
Membership About AEA Log In
Journals Annual Meeting Careers Resources EconLit EconSpark
Home › Journals › AEA Journals: Policies › Sample References
Journals
American Economic Review
AER: Insights
AEJ: Applied Economics
AEJ: Economic Policy
AEJ: Macroeconomics
AEJ: Microeconomics
Journal of Economic Literature
Journal of Economic Perspectives
AEA Papers & Proceedings
Subscriptions
Advertise in AEA Journals

Sample References

Styles of the AEA

We use the *Chicago Manual of Style* Author-Date system for all common publication types.

The following examples are intended to provide information for less common sources.

Datasets

When referencing datasets, please include the author name or name of the provider hosting the data, the year the data were collected or posted, the name or title of the dataset, the name of the database if applicable, and any other information necessary for one to retrieve the data. Please include the date accessed in parentheses at the end.

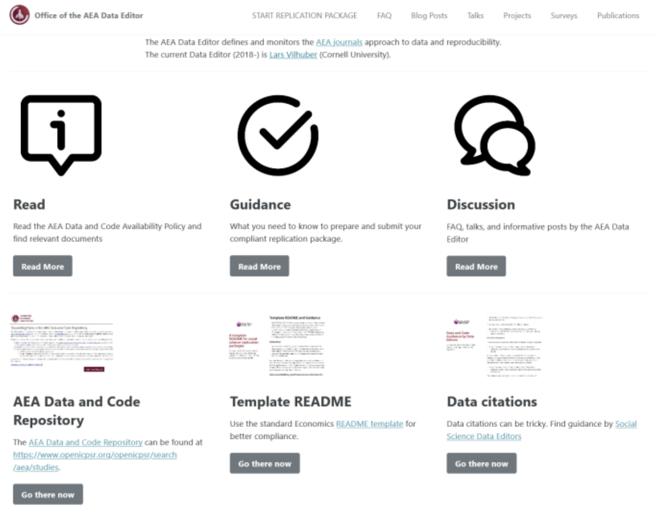
For data references specifically associated with a published paper, please include the Author Name(s). Year. "Paper Title: Dataset." *Journal Name*. Location of the data.

Examples

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).

2 **Leiss, Amelia.** 1999. "Arms Transfers to Developing Countries, 1945–1968." Inter-University Consortium for Political and Social Research, Ann Arbor, MI. ICPSR05404-v1. https://doi.org/10.3886/ICPSR05404.

3 **Romer, Christina D., and David H. Romer.** 2010. "Replication data for: The Macroeconomic Effects of Tax Changes:



The AEA Data Editor defines and monitors the AEA journals approach to data and reproducibility.
The current Data Editor (2018-) is Lars Vilhuber (Cornell University).

Read

Read the AEA Data and Code Availability Policy and find relevant documents

Guidance

What you need to know to prepare and submit your compliant replication package.

Discussion

FAQ, talks, and informative posts by the AEA Data Editor

AEA Data and Code Repository

The AEA Data and Code Repository can be found at https://www.openicpsr.org/openicpsr/search/aea/studies.

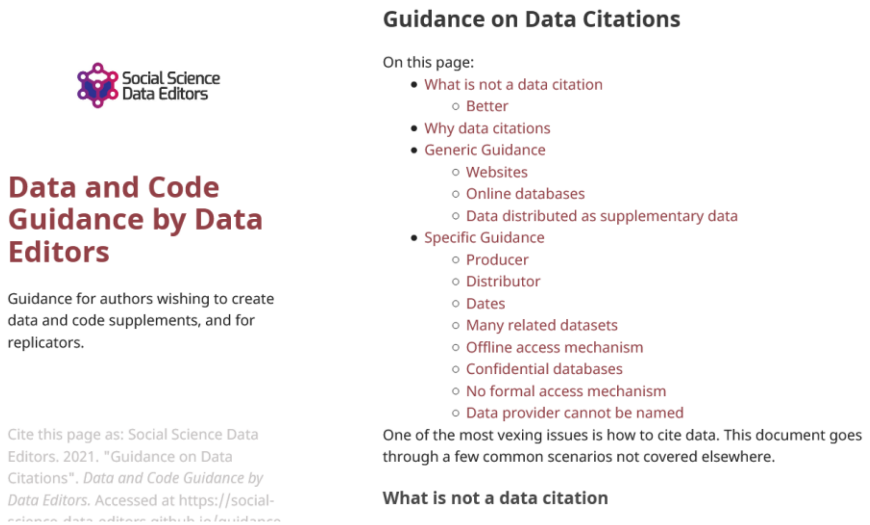
Template README

Use the standard Economics README template for better compliance.

Data citations

Data citations can be tricky. Find guidance by Social Science Data Editors

Enhanced guidance



Social Science Data Editors

Data and Code Guidance by Data Editors

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

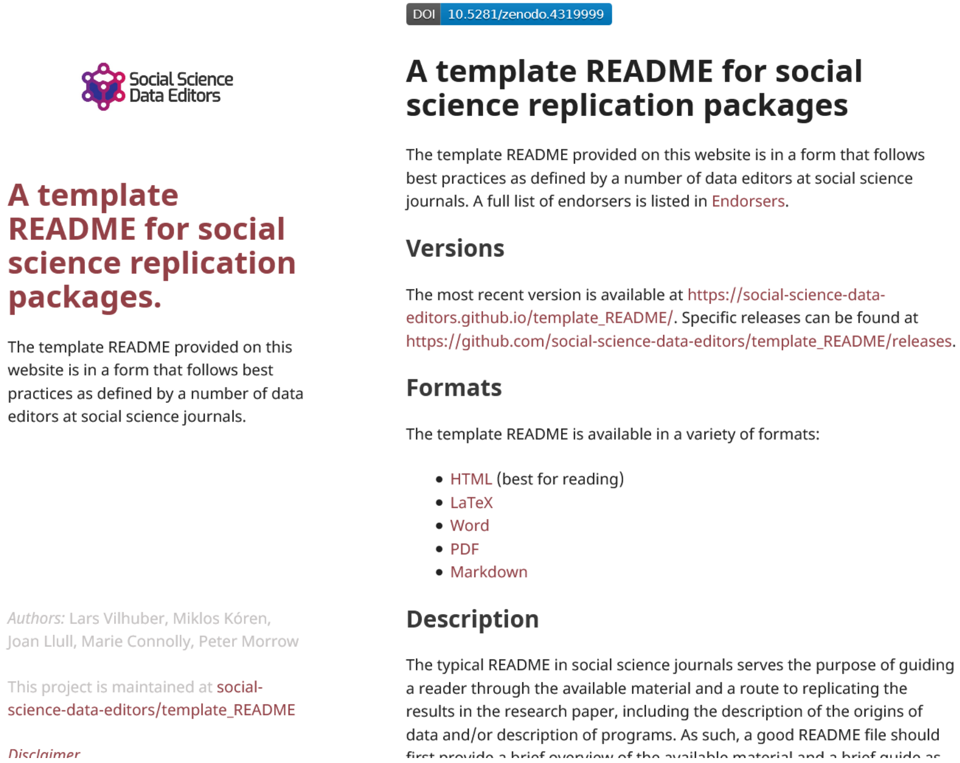
Cite this page as: Social Science Data Editors. 2021. "Guidance on Data Citations". *Data and Code Guidance by Data Editors*. Accessed at https://social-science-data-editors.github.io/guidance

Guidance on Data Citations

On this page:

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



Social Science Data Editors

A template README for social science replication packages.

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

Authors: Lars Vilhuber, Miklos Kóren, Joan Llull, Marie Connolly, Peter Morrow

This project is maintained at social-science-data-editors/template_README

Disclaimer

DOI 10.5281/zenodo.4319999

A template README for social science replication packages

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

Versions

The most recent version is available at https://social-science-data-editors.github.io/template_README/. Specific releases can be found at https://github.com/social-science-data-editors/template_README/releases.

Formats

The template README is available in a variety of formats:

- HTML (best for reading)
- LaTeX
- Word
- PDF
- Markdown

Description

The typical README in social science journals serves the purpose of guiding a reader through the available material and a route to replicating the results in the research paper, including the description of the origins of data and/or description of programs. As such, a good README file should first provide a brief overview of the available material and a brief guide as

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

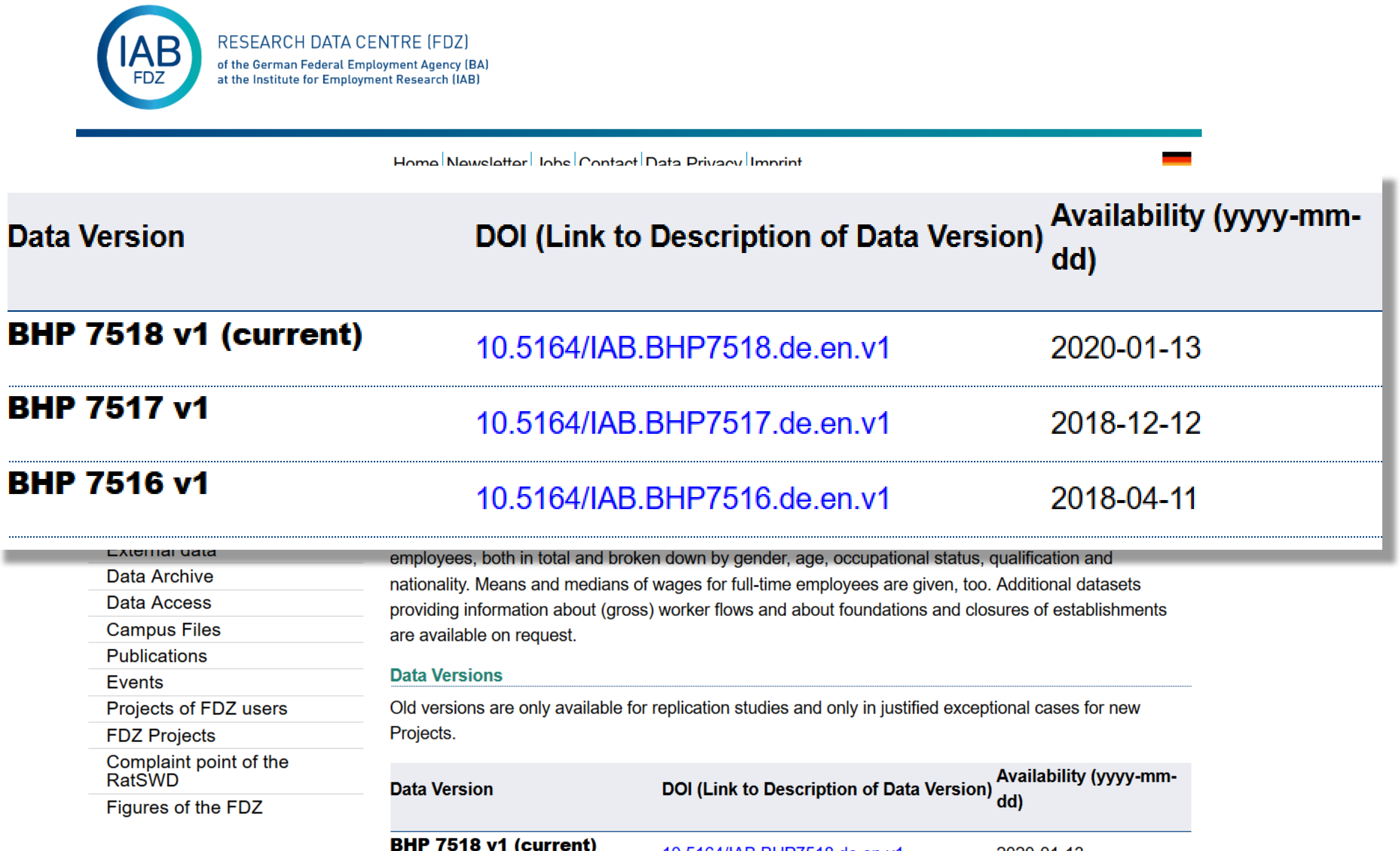
ICPSR notes that a citation should include the following items:

- 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.

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

Element of a (data) citation

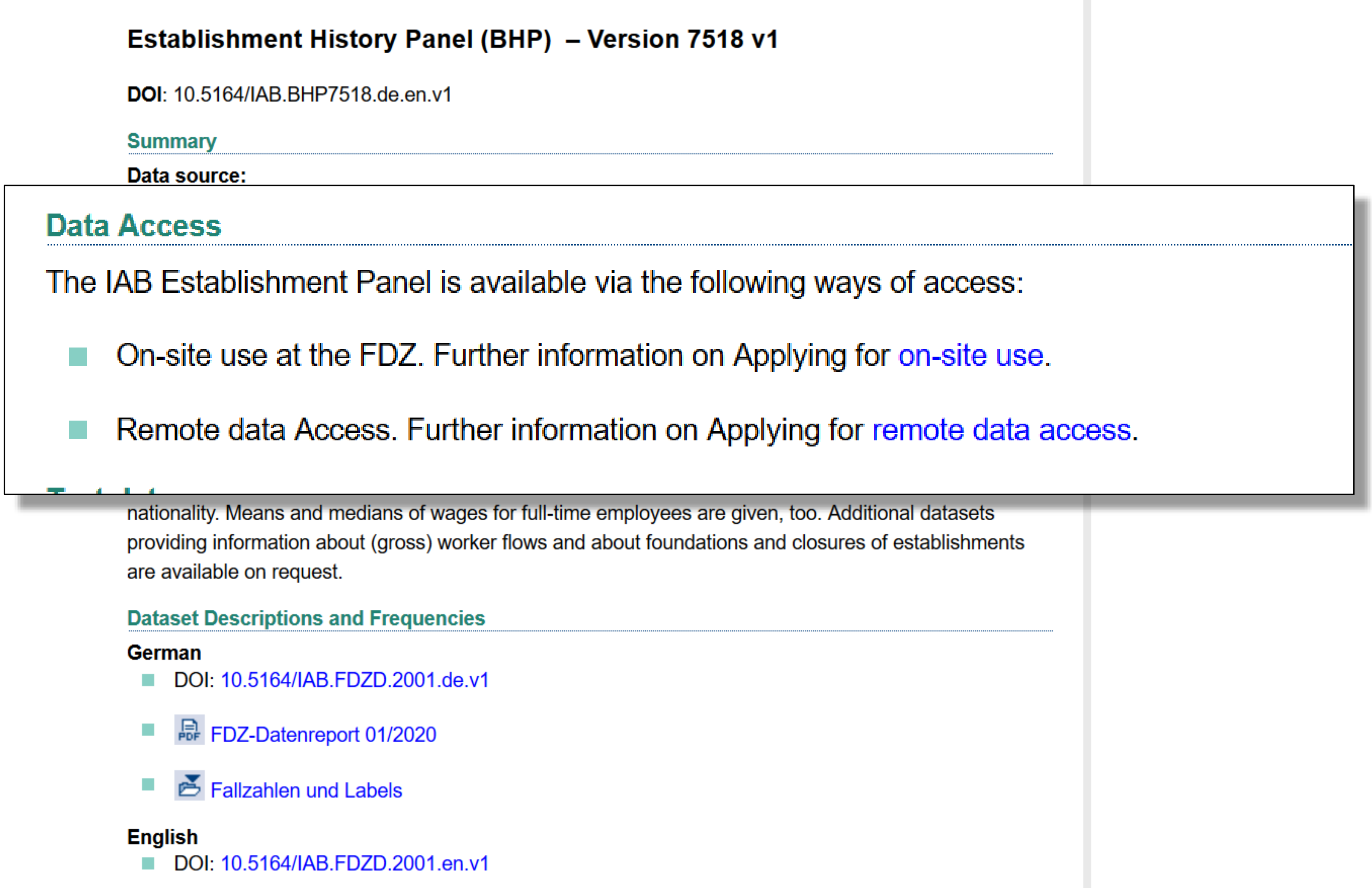
ICPSR notes that a citation should include the following items:

- 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.

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

Crafting data citations

In some cases, governments have list of their (named) registers. For instance, **Statistics Denmark** provides the full list of registers at

http://www.dst.dk/extranet/forskningvariabellister/Oversigt%20over%20registre.html.

These can be used to craft data citations:

https://social-science-data-editors.github.io/guidance/addtl-data-citation-guidance.html#confidential-databases

Statistics Denmark. 2020. "Døde i Danmark (DOD, Deaths in Denmark), 1970-2019 [database]", Danmarks Statistiks Forskningsservice, accessed (xxx).

- README can point to the codebook for each register, e.g., https://www.dst.dk/extranet/ForskningVariabellister/DOD%20-%20D%C3%B8de%20i%20Danmark.html for the aforementioned "DOD" register.

Element of a (data) citation

ICPSR notes that a citation should include the following items:

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

Constructed Citation:

Statistics Denmark, “Døde i Danmark (DOD, Deaths in Denmark), 1970-2019” [database]. Accessed via the Danmarks Statistiks Forskningsservice http://www.dst.dk/extranet/forskningvariabellister/Oversigt%20over%20registre.html June 26, 2020.

Crafting data citations

In some cases, governments have list of their (named) registers. For instance, **Statistics Denmark** provides the full list of registers at

http://www.dst.dk/extranet/forskningvariabellister/Oversigt%20over%20registre.html.

These can be used to craft data citations:

Statistics Denmark, "Døde i Danmark (DOD, Deaths in Denmark), 1970-2019" [database]. Accessed via the Danmarks Statistiks Forskningsservice http://www.dst.dk/extranet/forskningvariabellister/Oversigt%20over%20registre.html June 26, 2020.

- README can point to the codebook for each register, e.g., https://www.dst.dk/extranet/ForskningVariabellister/DOD%20-%20D%C3%B8de%20i%20Danmark.html for the aforementioned "DOD" register.

https://social-science-data-editors.github.io/guidance/addtl-data-citation-guidance.html#confidential-databases

Element of a (data) citation

ICPSR notes that a citation should include the following items:

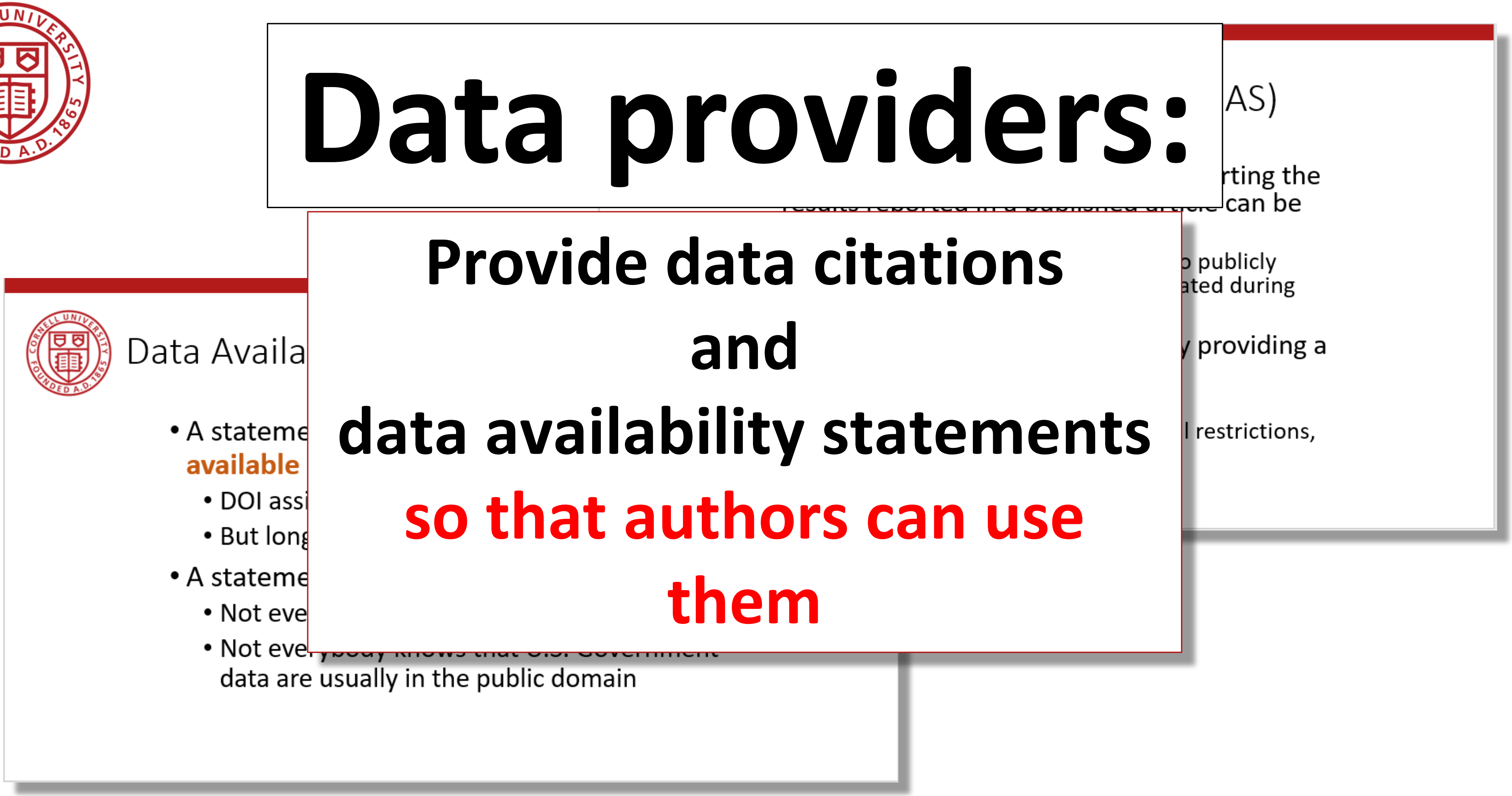
- Author
- Title
- Distributor
- Date
- Version
- Persistent identifier

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 providers:

Provide data citations
and
data availability statements
so that authors can use them



AS)

rting the

results reported in a published article can be

o publicly
ated during

y providing a

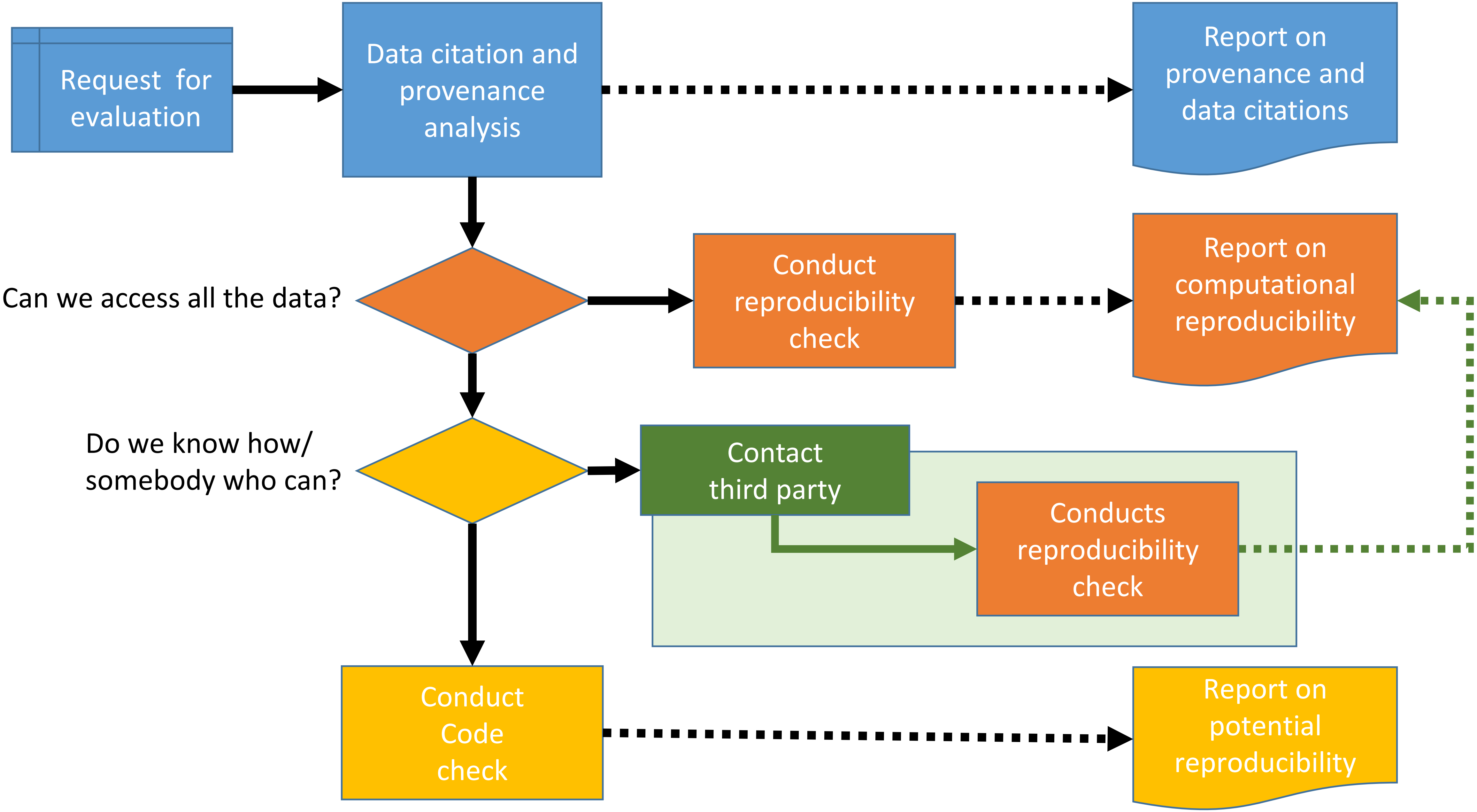
l restrictions,

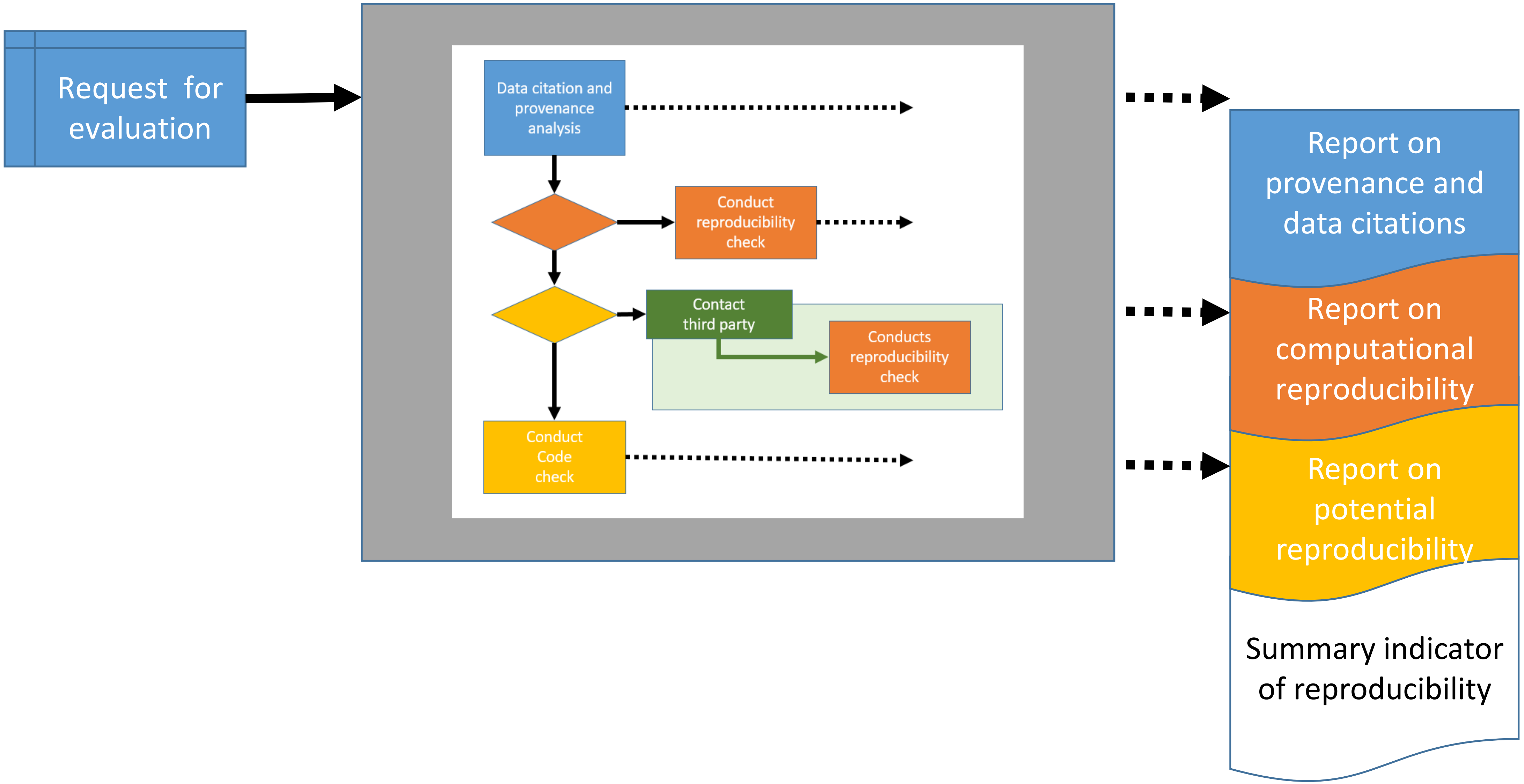
Data Availa

- A stateme
available
  - DOI assi
  - But long
- A stateme
  - Not eve
  - Not everybody knows that U.S. Government data are usually in the public domain

Reproducibility in RDCs

Challenge:
How to verify reproducibility when data is restricted?


Request for evaluation
Data citation and provenance analysis
Report on provenance and data citations
Can we access all the data?
Conduct reproducibility check
Report on computational reproducibility
Do we know how/ somebody who can?
Contact third party
Conducts reproducibility check
Conduct Code check
Report on potential reproducibility



Request for evaluation

Data citation and provenance analysis

Conduct reproducibility check

Contact third party

Conducts reproducibility check

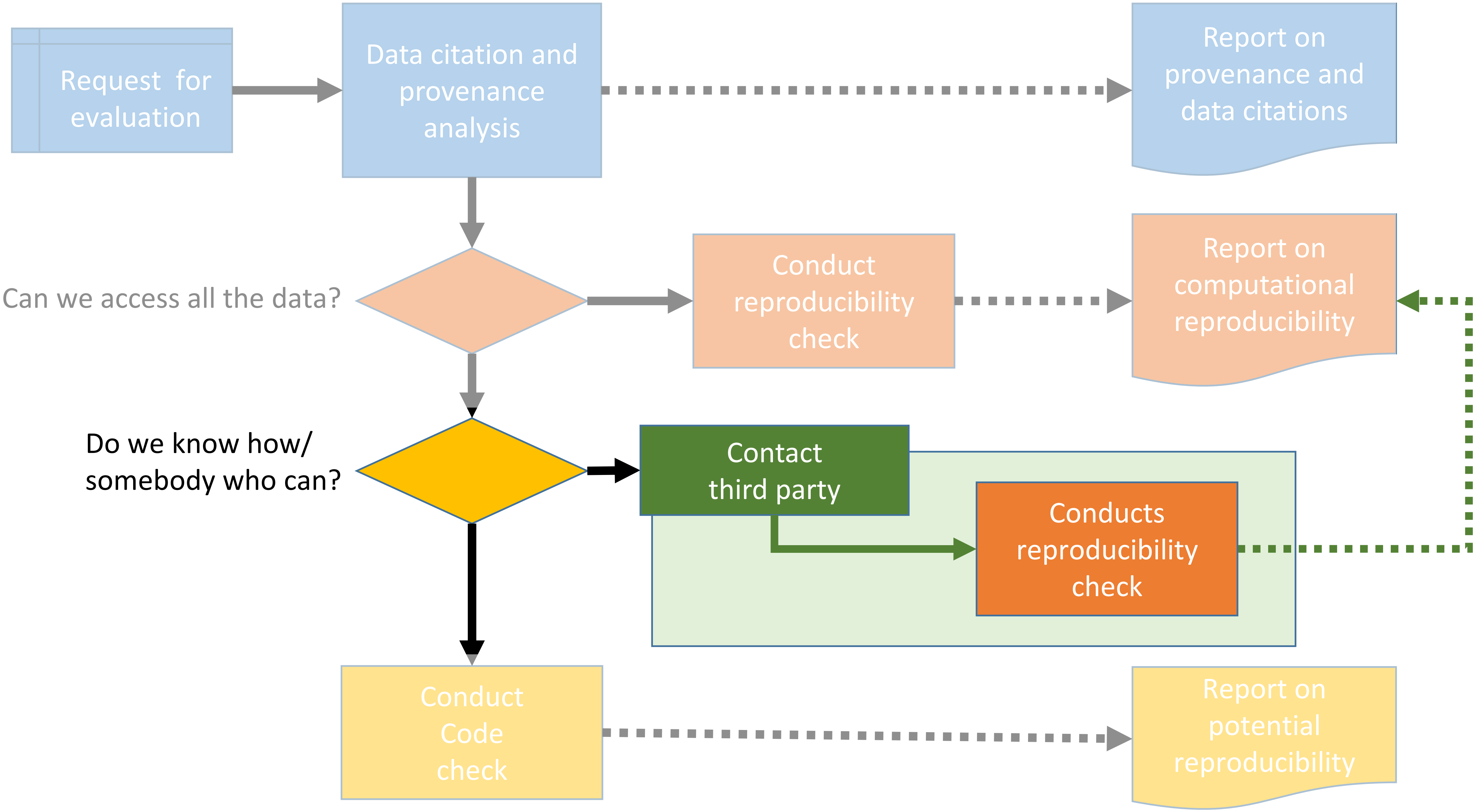
Conduct Code check

Report on provenance and data citations

Report on computational reproducibility

Report on potential reproducibility

Summary indicator of reproducibility



Request for evaluation

Data citation and provenance analysis

Report on provenance and data citations

Can we access all the data?

Conduct reproducibility check

Report on computational reproducibility

Do we know how/ somebody who can?

Contact third party

Conducts reproducibility check

Conduct Code check

Report on potential reproducibility

1

Challenge:
How to verify reproducibility when data is restricted?

Alternative methods

- **Request access to the data ourselves**
  - We requested access to SOEP data (Germany) ✗
  - We requested access to IAB, BLS ✓, others
  - We have access to Brazilian ✓ data
  - Sign DUA with eBay ✓, Kilts ✓, Zillow ⌛, etc.

2

Challenge:
How to verify reproducibility when data is restricted?

Verification services

cascad

the first certification agency for scientific code & data

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

Secure Data Hub

Travail, Emploi / 189

Société, Justice, Éducation / 113

Économie, Entreprises, Finance / 267

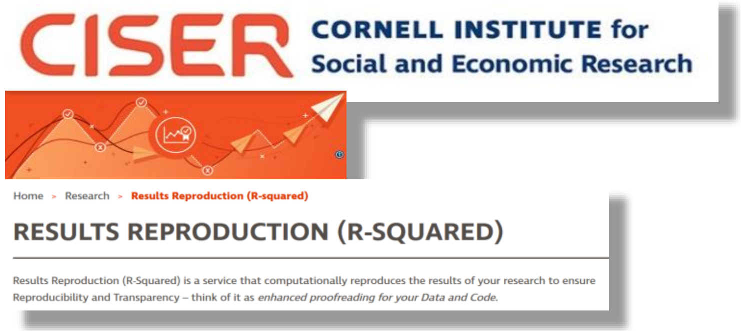
Environnement, Agriculture / 187

Santé / 244

Alternative methods

- **Request access to the data ourselves**
  - We requested access to SOEP data (Germany) ✗
  - We requested access to IAB, BLS ✓, others
  - We have access to Brazilian ✓ data
  - Sign DUA with eBay ✓, Kilts ✓, Zillow ⌛, etc.

- **Ask others (staff/ students) to run code for us**
  - **cascad** for Swedish ✓, French ✓ confidential data



CISER CORNELL INSTITUTE for Social and Economic Research

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.*

3

Challenge:
How to verify reproducibility when data is restricted?

Alternative methods

- **Request access to the data ourselves**
  - We requested access to SOEP data (Germany) ✗
  - We requested access to IAB, BLS ✓, others
  - We have access to Brazilian ✓ data
  - Sign DUA with eBay ✓, Kilts ✓, Zillow ⌛, etc.

- **Ask others (staff/ students) to run code for us**
  - **cascad** for Swedish ✓, French ✓ confidential data
  - Asked masters student at IFAU on different paper w/ Swedish ✓
  - Asked staff at IAB ✓, BLS ✓, Census Bureau ✗, **Banco de Portugal (BPLIM) ?**, etc.
  - Ask graduate students of the same research group (honors system) ✓ ✓ ✓ ✓ ✓

Alternative
3rd party?

Alternative

3rd party?

Look left/right

Consider the following “game”

- Think of your latest **term paper/ thesis proposal/ anything**
- Package it up
  - Has to be complete
- Put it on Dropbox/ floppy/ Github

Then:

Consider the following "game"

- Think of your latest **term paper/ thesis proposal/ anything**
- Package it up
  - Has to be complete
- Put it on Dropbox/ floppy/ Github

- Then:
  - Ask your neighbor to run your
    - get your data,
    - Run your code
  - WITHOUT EVER SPEAKING WITH YOU
    - Or emailing
    - Or texting
    - Or tik-toking
    - Or slacking
    - ….

Consider the following “game”

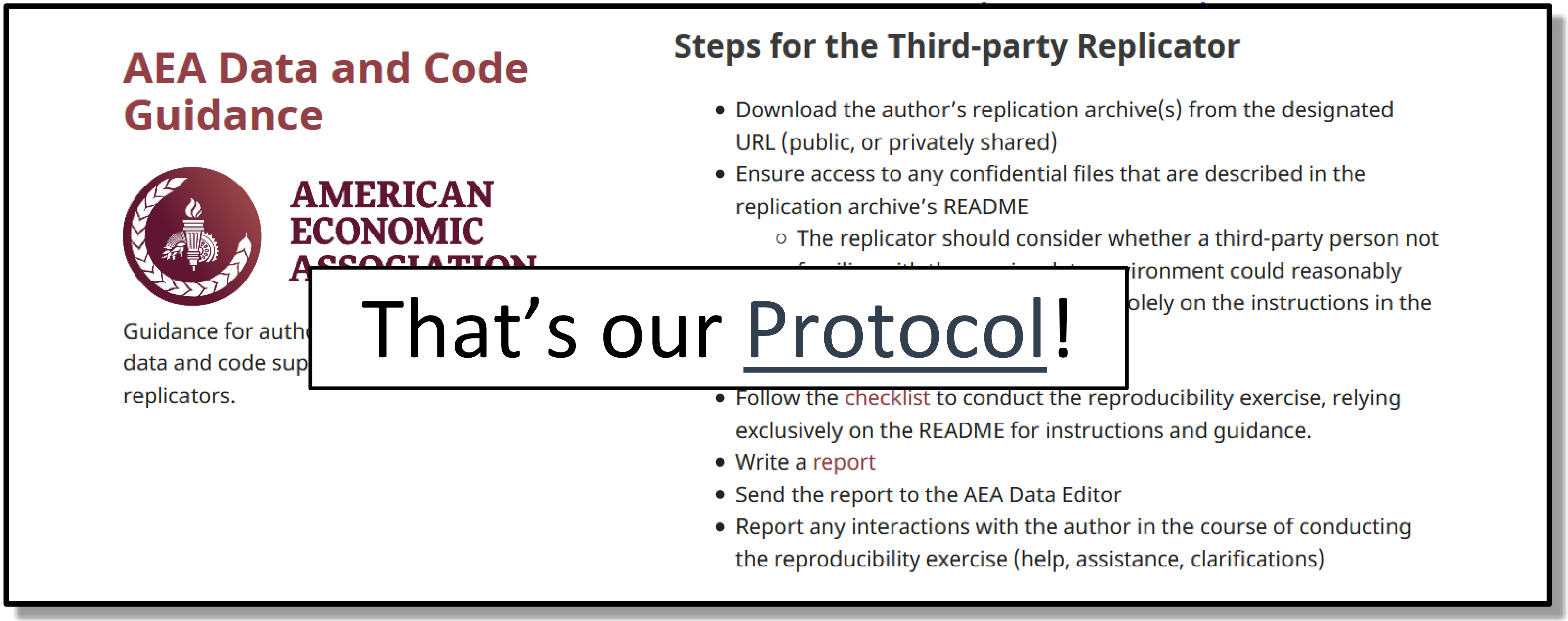
- Think of your latest **term paper/ thesis proposal/ anything**
- Packa
  - Ha
- Put it
  Githu

- Then:
  - Ask your neighbor to run your
  - AKING WITH
    - or slacking
    - ….

Do you think that would work?

How to prepare the replication package

- README
- Now ask an RA/ colleague/



AEA Data and Code Guidance

AMERICAN ECONOMIC ASSOCIATION

Guidance for auth
data and code sup
replicators.

Steps for the Third-party Replicator

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

That's our Protocol!

Academia only?

Paradigm shifts

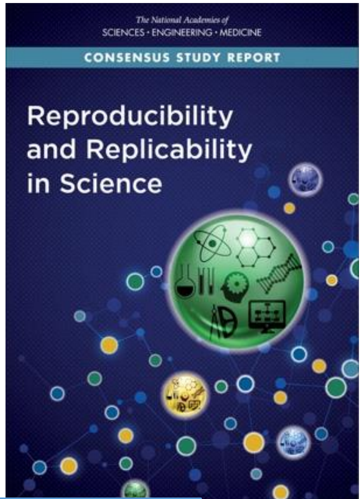
- How **academic research** is conducted & shared (open science, reproducibility & rigor)
  - Modernizing long-standing practices that support scrutiny, debate, self-correction, new discoveries
  - Beyond publication to sharing data, metadata, methods, software, and other outputs
  - Findable, Accessible, Interoperable, Reusable (FAIR) principles for data sharing
  - Greater emphasis on research integrity, reproducibility and replicability in scholarship

The National Academies of SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

OPEN SCIENCE BY DESIGN

Realizing a Vision for 21st Century Research



The National Academies of SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

Reproducibility and Replicability in Science

eLife

Reproducibility in Cancer Biology: Challenges for assessing replicability in preclinical cancer biology

Timothy M Errington, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, Brian A Nosek

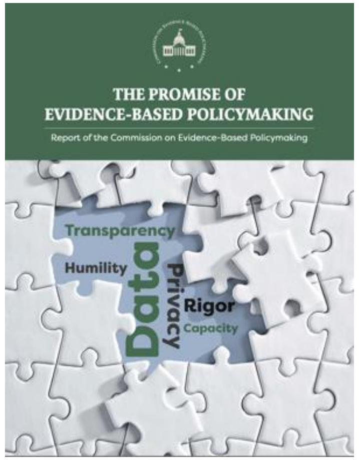
Center for Open Science, United States; Science Exchange, United States; University of Virginia, United States

Feature Article · Dec 7, 2021

Source: Sarah Nusser

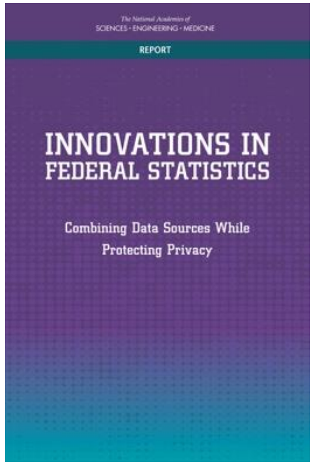
Paradigm shifts

- **Government policy** for data sharing and integration
  - Evidence-based Policy Making Act
  - Federal Data Strategy
- **Continued innovation in approaches, standards & tools for official statistics**
  - Expanding use of non-survey data sources in creating statistical products
  - International official statistics community practice and standards development



THE PROMISE OF EVIDENCE-BASED POLICYMAKING

Report of the Commission on Evidence-Based Policymaking

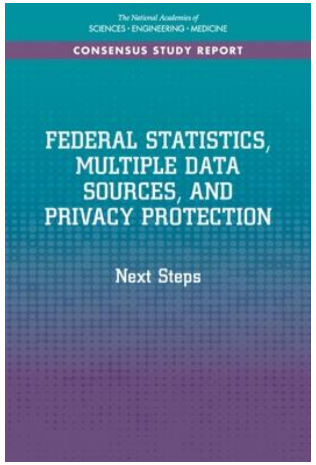


The National Academies of SCIENCES · ENGINEERING · MEDICINE

REPORT

INNOVATIONS IN FEDERAL STATISTICS

Combining Data Sources While Protecting Privacy



The National Academies of SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

FEDERAL STATISTICS, MULTIPLE DATA SOURCES, AND PRIVACY PROTECTION

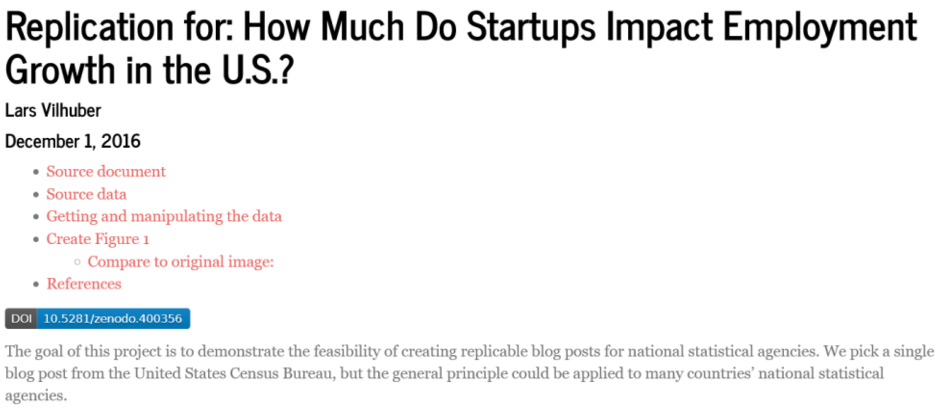
Next Steps

Source: Sarah Nusser

Reproducibility of Policy Briefs [Sylverie]

- Seems obvious: policy depends on it!
- Also infographics!

Simple example: https://larsvilhuber.github.io/jobcreationblog/README.html



Replication for: How Much Do Startups Impact Employment Growth in the U.S.?

Lars Vilhuber

December 1, 2016

- Source document
- Source data
- Getting and manipulating the data
- Create Figure 1
  - Compare to original image:
- References

DOI 10.5281/zenodo.400356

The goal of this project is to demonstrate the feasibility of creating replicable blog posts for national statistical agencies. We pick a single blog post from the United States Census Bureau, but the general principle could be applied to many countries' national statistical agencies.

- How to do it quickly?
  - Templates, competing teams, technology **[see next part]**
- Public or private?
  - Repositories can be internal, but should expect to be made public
- Risks
  - May be misconstrued
  - Will definitely be analyzed by graduate students all over the world!

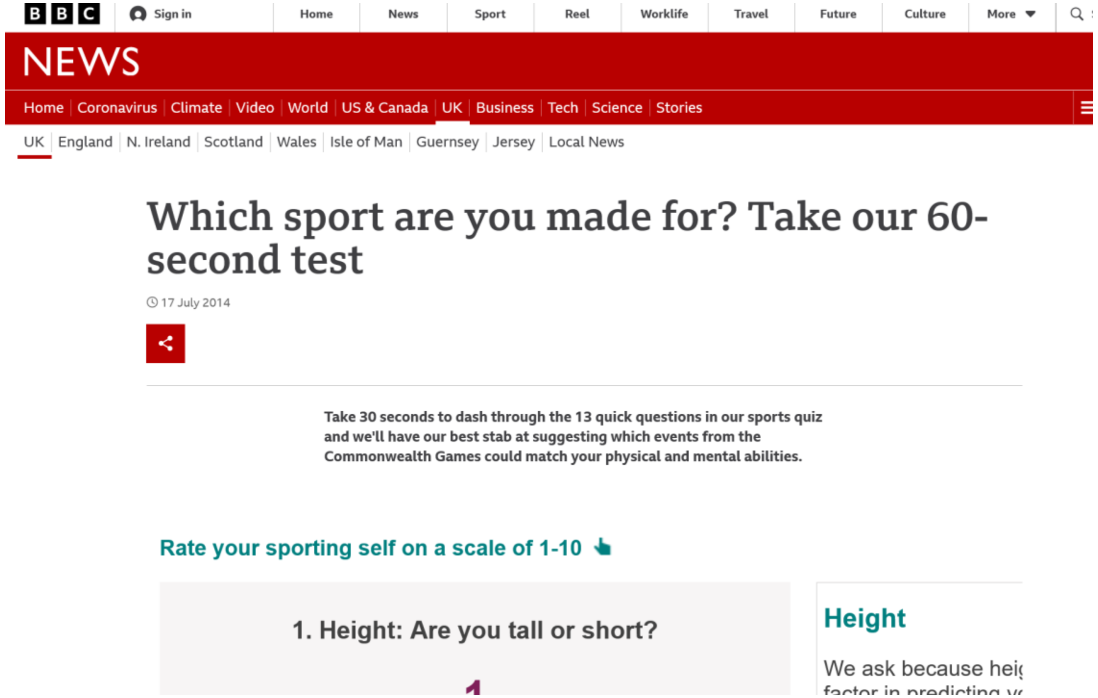
Dynamic Policy Briefs [Julia]

- Making simple infographics or policy briefs (partially) dynamics
  - Play with policy assumptions
  - Audience may be public, or decision makers

- How to do it quickly?
  - Templates, competing teams, technology **[see next part]**
- Public or private?
  - Repositories can be internal, but should expect to be made public
- Risks
  - May be misconstrued
  - Will definitely be analyzed by graduate students all over the world!

Dynamic Policy Briefs

- Making simple infographics or policy briefs (partially) dynamics
  - Play with policy assumptions
  - Audience may be public, or decision makers
- Regularly done by "data journalists"
  - [BBC sports]


BBC
Sign in
Home
News
Sport
Reel
Worklife
Travel
Future
Culture
More
NEWS
Home | Coronavirus | Climate | Video | World | US & Canada | UK | Business | Tech | Science | Stories
UK | England | N. Ireland | Scotland | Wales | Isle of Man | Guernsey | Jersey | Local News

Which sport are you made for? Take our 60-second test

17 July 2014

Take 30 seconds to dash through the 13 quick questions in our sports quiz and we'll have our best stab at suggesting which events from the Commonwealth Games could match your physical and mental abilities.

Rate your sporting self on a scale of 1-10

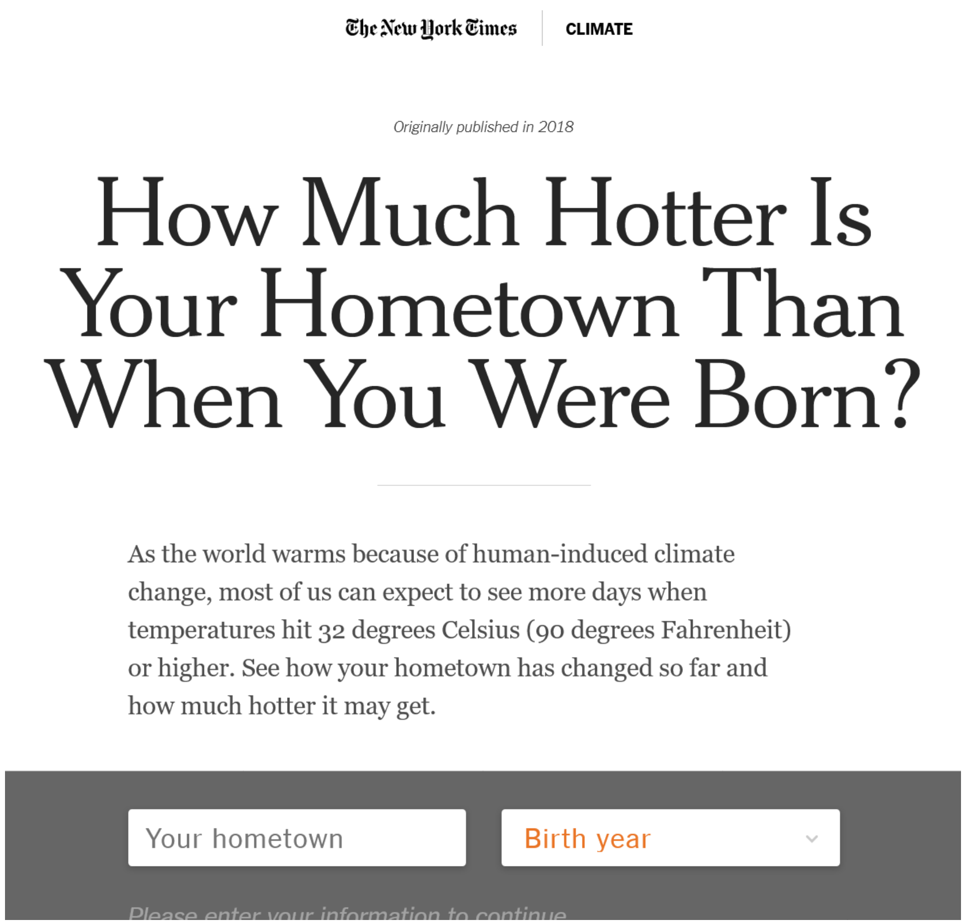
1. Height: Are you tall or short?

Height

We ask because hei
factor in predicting v

Dynamic Policy Briefs

- Making simple infographics or policy briefs (partially) dynamics
  - Play with policy assumptions
  - Audience may be public, or decision makers
- Regularly done by “data journalists”
  - BBC sports
  - NY Times



The New York Times | CLIMATE

Originally published in 2018

How Much Hotter Is Your Hometown Than When You Were Born?

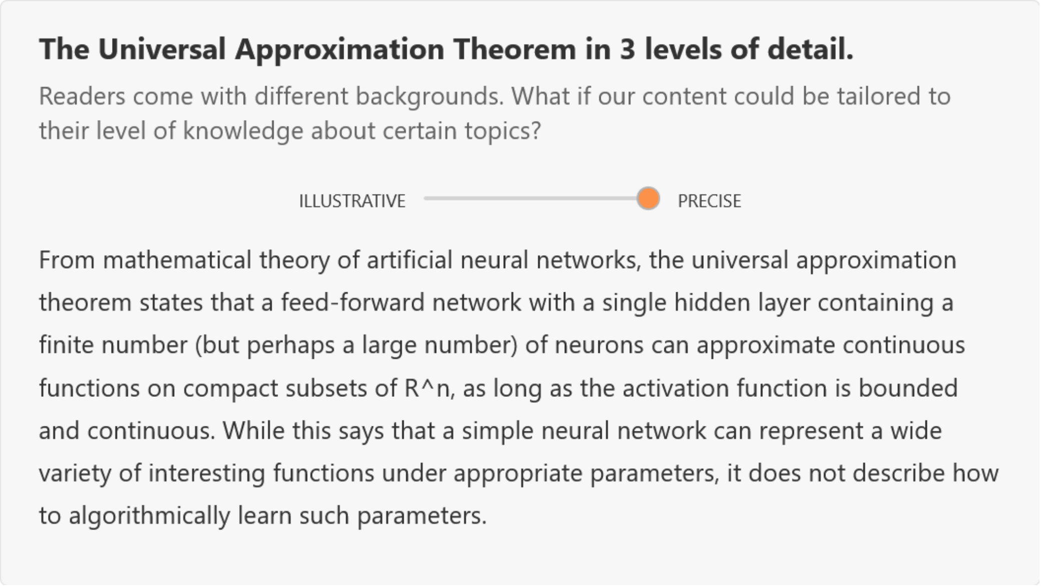
As the world warms because of human-induced climate change, most of us can expect to see more days when temperatures hit 32 degrees Celsius (90 degrees Fahrenheit) or higher. See how your hometown has changed so far and how much hotter it may get.

Please enter your information to continue

Dynamic Policy Briefs [Julia]

- Making simple infographics or policy briefs (partially) dynamics
  - Play with policy assumptions
  - Audience may be public, or decision makers
- Regularly done by "data journalists"
- Some new article types
  - Distill (Hohman, Fred, Matthew Conlen, Jeffrey Heer, and Duen Horng (Polo) Chau. "Communicating with Interactive Articles." *Distill* 5, no. 9 (September 11, 2020): e28. https://doi.org/10.23915/distill.00028.)

- Reducing cognitive load



The Universal Approximation Theorem in 3 levels of detail.

Readers come with different backgrounds. What if our content could be tailored to their level of knowledge about certain topics?

ILLUSTRATIVE ● PRECISE

From mathematical theory of artificial neural networks, the universal approximation theorem states that a feed-forward network with a single hidden layer containing a finite number (but perhaps a large number) of neurons can approximate continuous functions on compact subsets of R^n, as long as the activation function is bounded and continuous. While this says that a simple neural network can represent a wide variety of interesting functions under appropriate parameters, it does not describe how to algorithmically learn such parameters.

Dynamic Policy Briefs

- Making simple infographics or policy briefs (partially) dynamics
  - Play with policy assumptions
  - Audience may be public, or decision makers
- Regularly done by “data journalists”
- Some new article types
  - Distill (Hohman, Fred, Matthew Conlen, Jeffrey Heer, and Duen Horng (Polo) Chau. “Communicating with Interactive Articles.” *Distill* 5, no. 9 (September 11, 2020): e28. https://doi.org/10.23915/distill.00028.)
  - Stencila/ eLife

- Reducing cognitive load
- Exposing assumptions

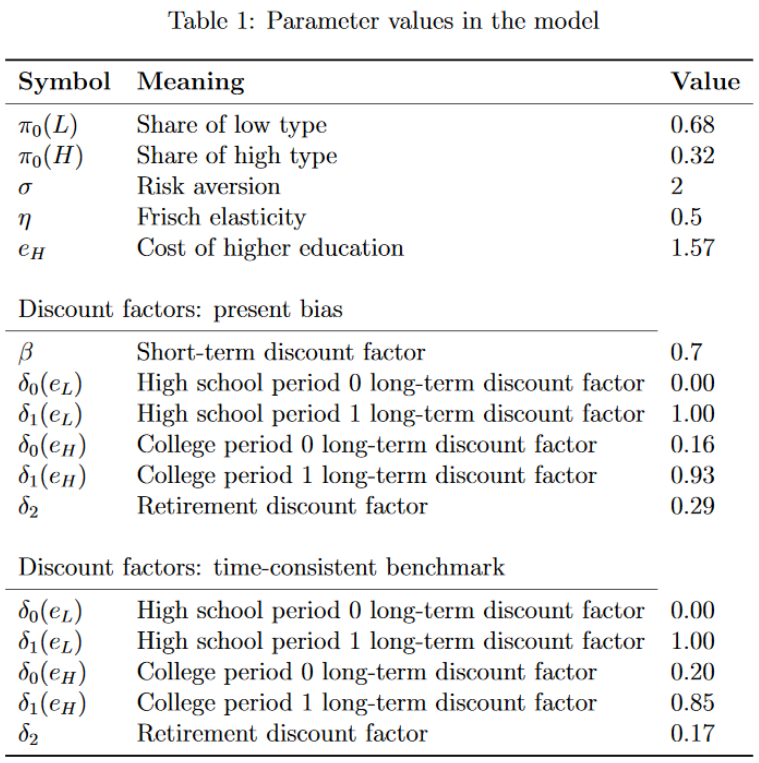


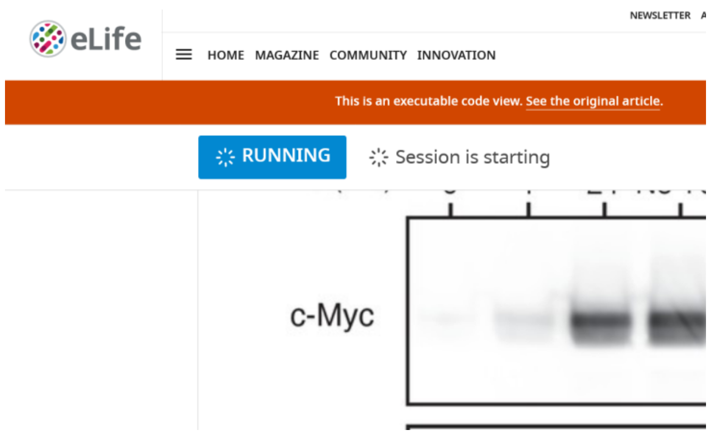
Table 1: Parameter values in the model

Symbol	Meaning	Value
$\pi_0(L)$	Share of low type	0.68
$\pi_0(H)$	Share of high type	0.32
σ	Risk aversion	2
η	Frisch elasticity	0.5
e_H	Cost of higher education	1.57
Discount factors: present bias		
β	Short-term discount factor	0.7
$\delta_0(e_L)$	High school period 0 long-term discount factor	0.00
$\delta_1(e_L)$	High school period 1 long-term discount factor	1.00
$\delta_0(e_H)$	College period 0 long-term discount factor	0.16
$\delta_1(e_H)$	College period 1 long-term discount factor	0.93
δ_2	Retirement discount factor	0.29
Discount factors: time-consistent benchmark		
$\delta_0(e_L)$	High school period 0 long-term discount factor	0.00
$\delta_1(e_L)$	High school period 1 long-term discount factor	1.00
$\delta_0(e_H)$	College period 0 long-term discount factor	0.20
$\delta_1(e_H)$	College period 1 long-term discount factor	0.85
δ_2	Retirement discount factor	0.17

Dynamic Policy Briefs

- Making simple infographics or policy briefs (partially) dynamics
  - Play with policy assumptions
  - Audience may be public, or decision makers
- Regularly done by "data journalists"
- Some new article types
  - Distill (Hohman, Fred, Matthew Conlen, Jeffrey Heer, and Duen Horng (Polo) Chau. "Communicating with Interactive Articles." *Distill* 5, no. 9 (September 11, 2020): e28. https://doi.org/10.23915/distill.00028.)
  - Stencila/ eLife (uses R)

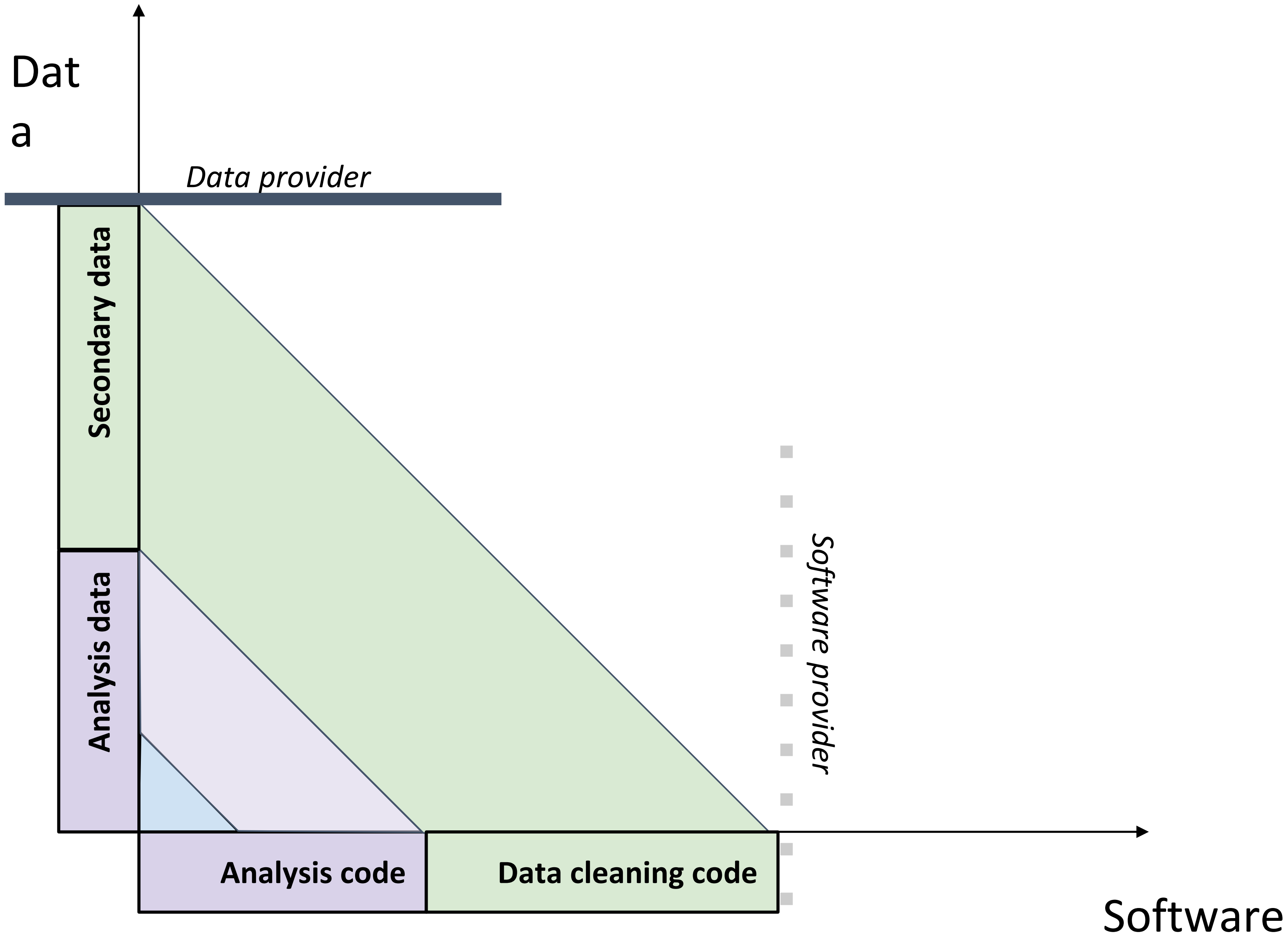
- Reducing cognitive load
- Exposing assumptions

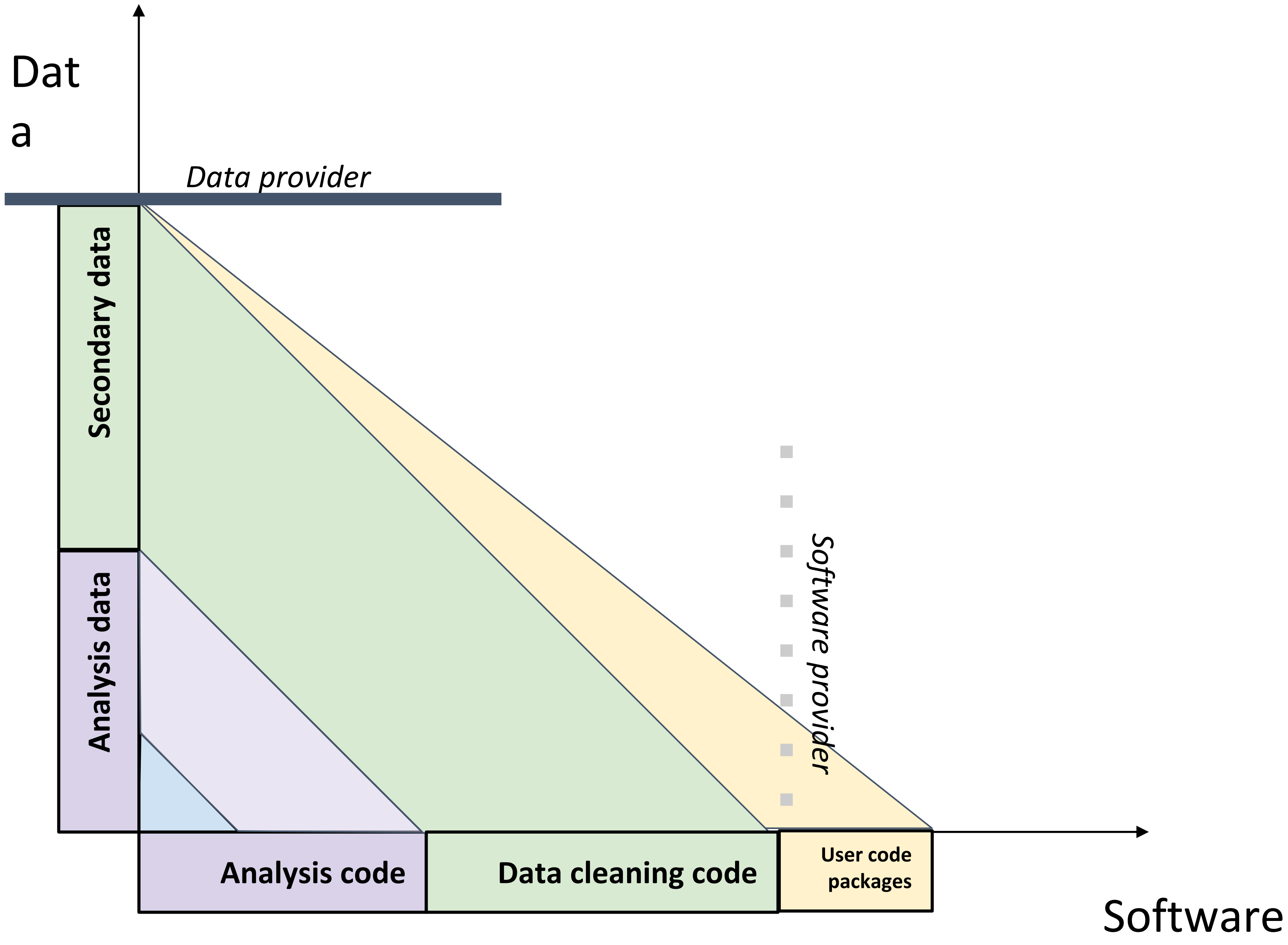


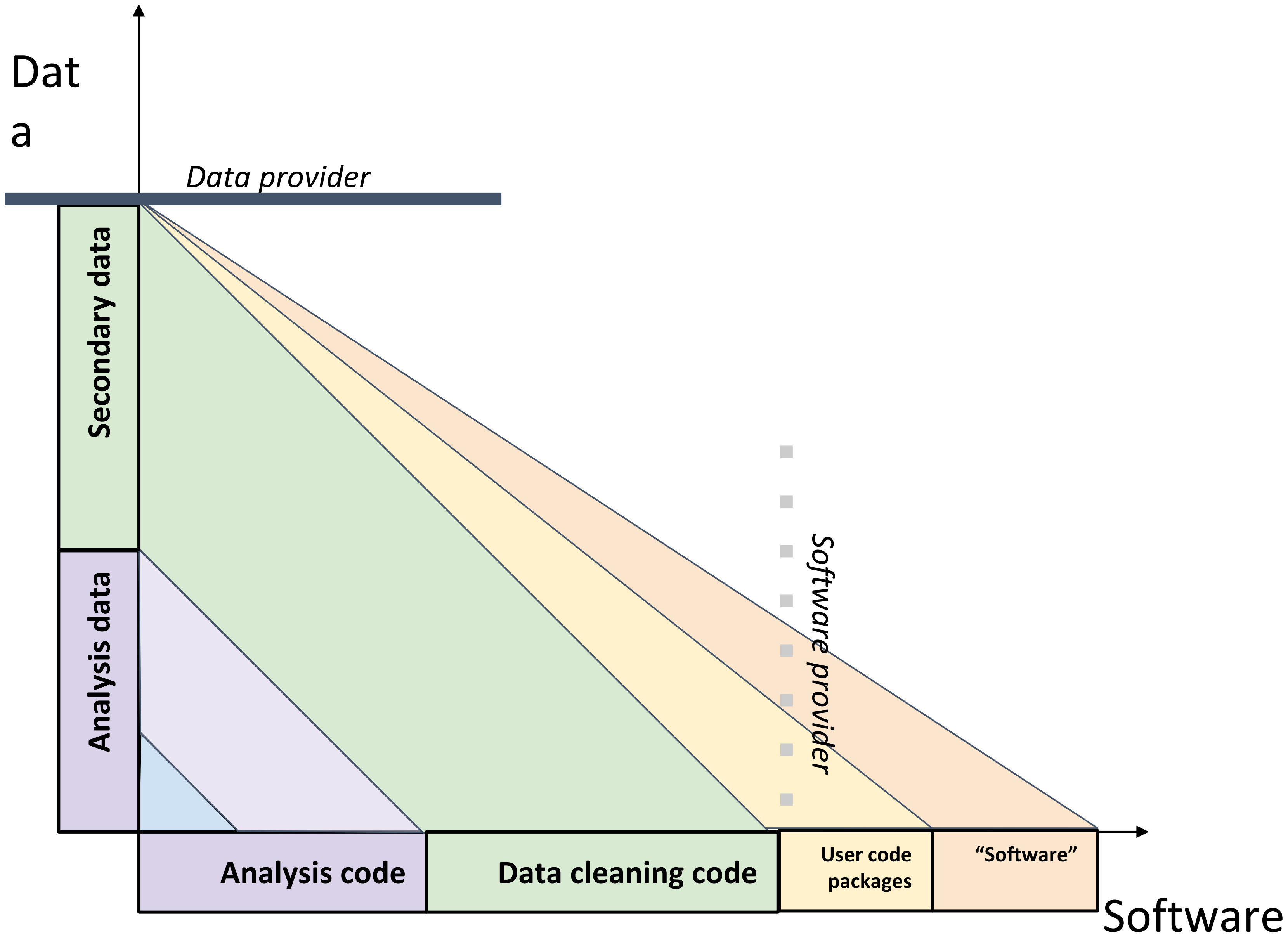
This is an executable code view. See the original article.

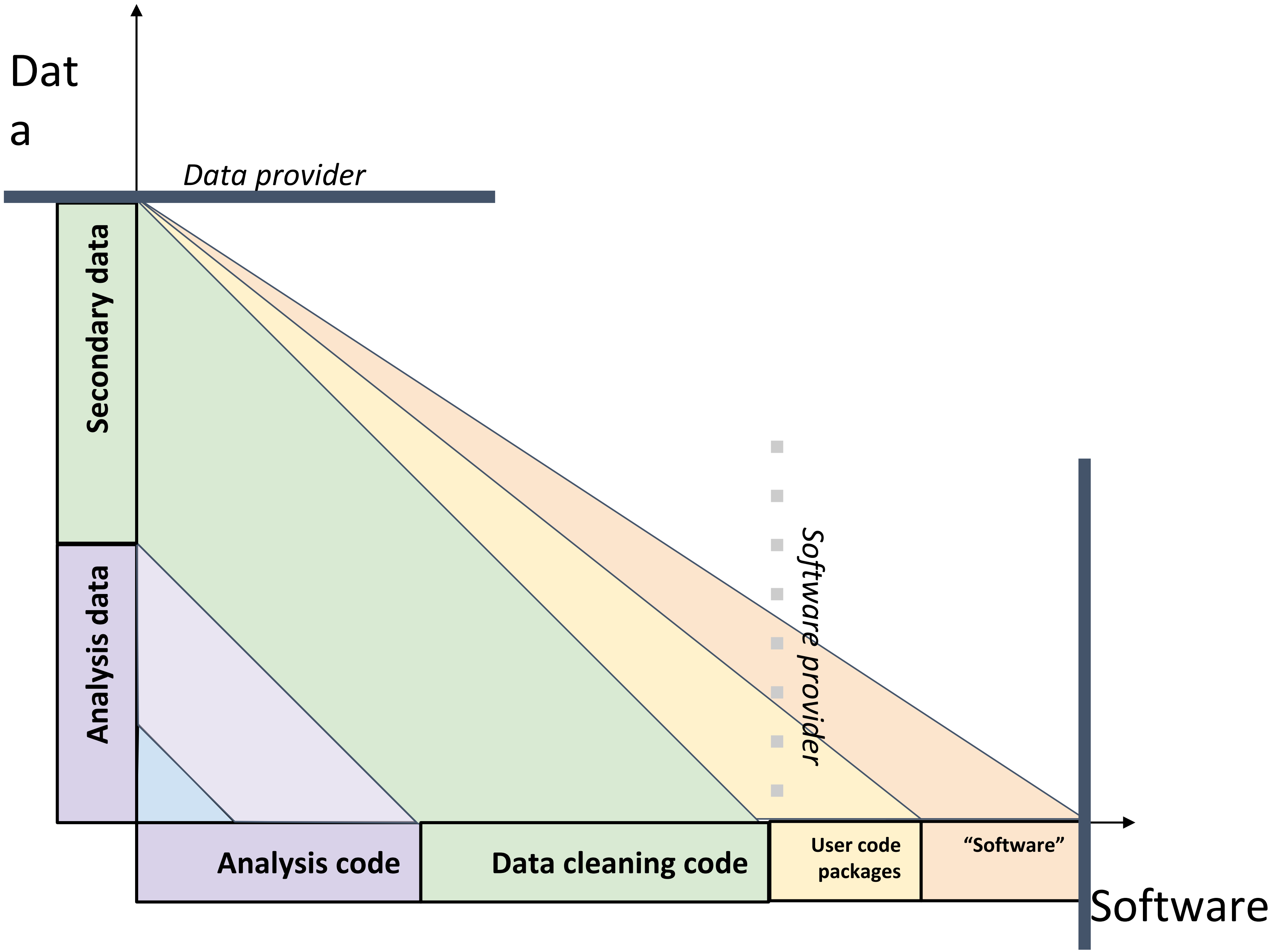
Session is starting

Technological frontier

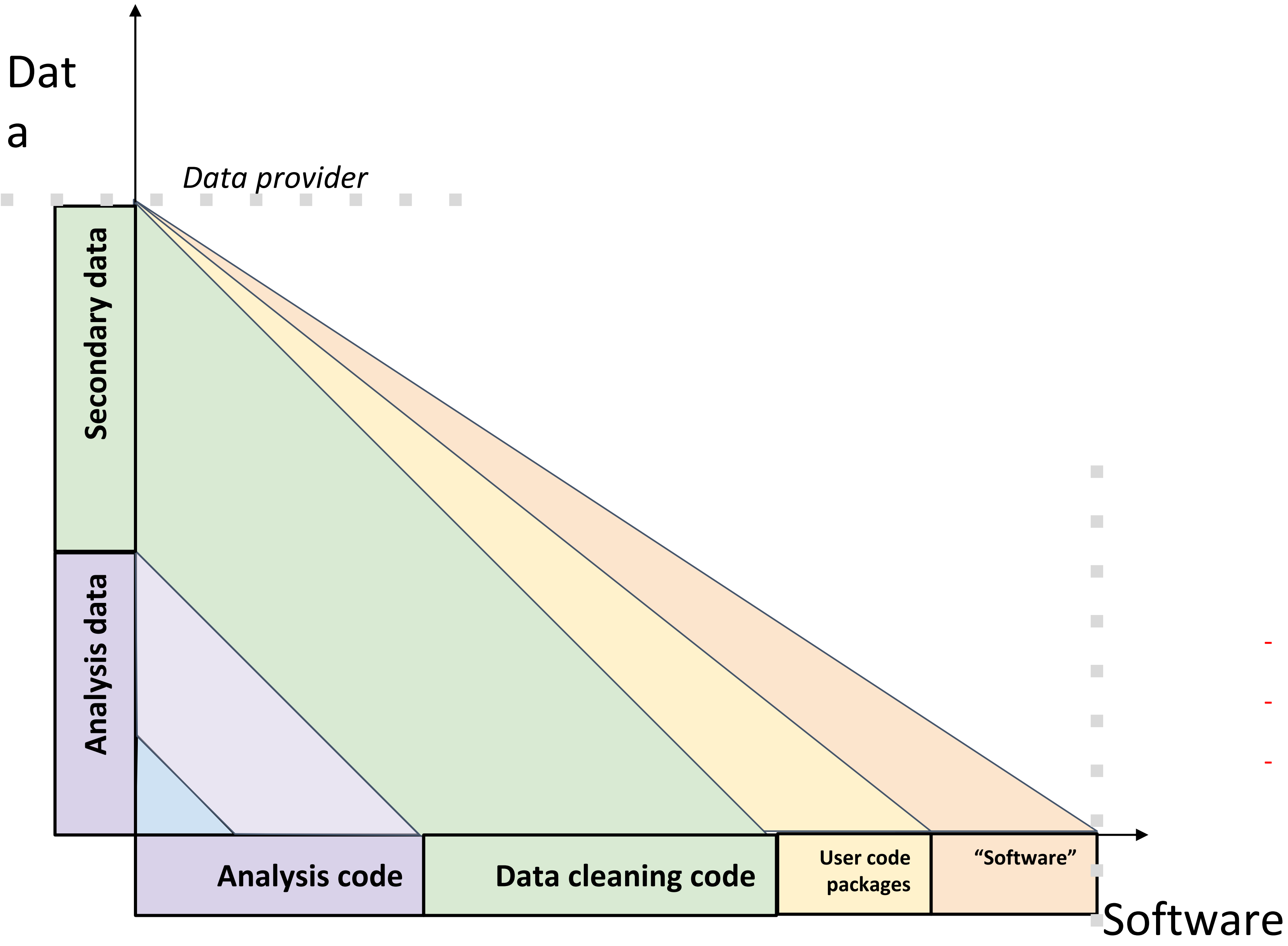

Dat
a
Data provider
Secondary data
Analysis data
Software provider
Analysis code
Data cleaning code
Software


Dat
a
Data provider
Secondary data
Analysis data
Software provider
Analysis code
Data cleaning code
User code packages
Software


Dat
a
Data provider
Secondary data
Analysis data
Software provider
Analysis code
Data cleaning code
User code packages
"Software"
Software

Dat
a
Data provider
Secondary data
Analysis data
Software provider
Analysis code
Data cleaning code
User code packages
"Software"
Software

What if *software* were part of the replication package



Dat
a

Data provider

Secondary data

Analysis data

Analysis code

Data cleaning code

User code packages

“Software”

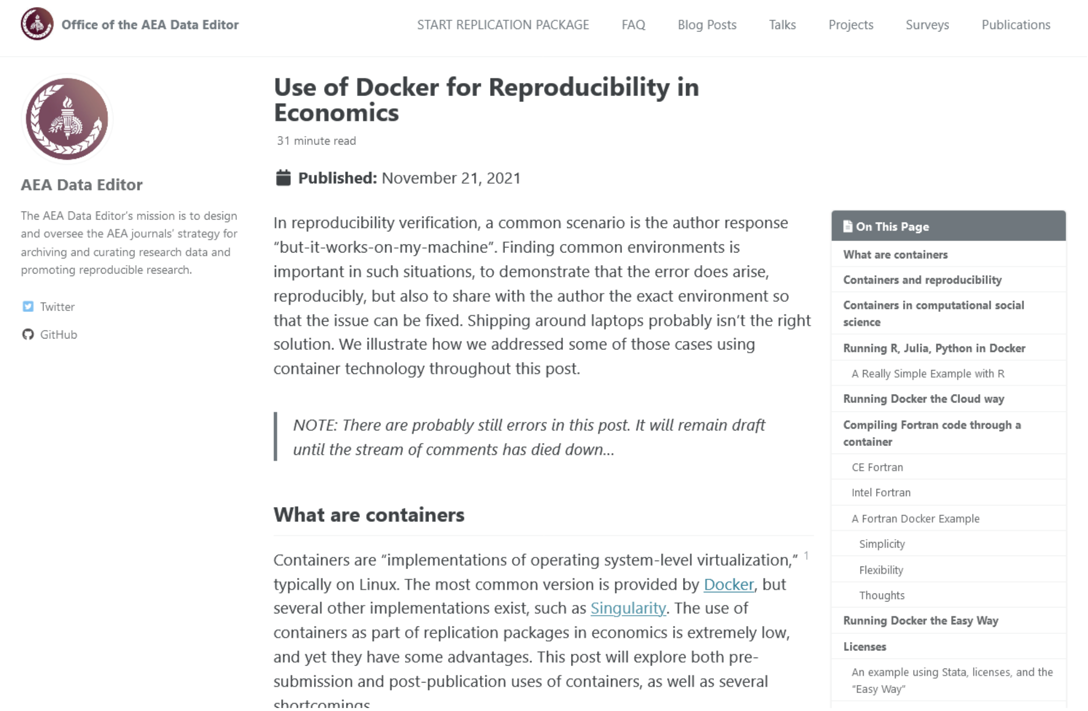
Software

- **Compilers**
- **Libraries**
- **Source code**

What if *the computer* were part of the replication package

Use of virtual environments (Docker, VM)

- https://aeadataeditor.github.io/posts/2021-11-16-docker



AEA Data Editor

The AEA Data Editor's mission is to design and oversee the AEA journals' strategy for archiving and curating research data and promoting reproducible research.

Use of Docker for Reproducibility in Economics

31 minute read

Published: November 21, 2021

In reproducibility verification, a common scenario is the author response "but-it-works-on-my-machine". Finding common environments is important in such situations, to demonstrate that the error does arise, reproducibly, but also to share with the author the exact environment so that the issue can be fixed. Shipping around laptops probably isn't the right solution. We illustrate how we addressed some of those cases using container technology throughout this post.

> *NOTE: There are probably still errors in this post. It will remain draft until the stream of comments has died down…*

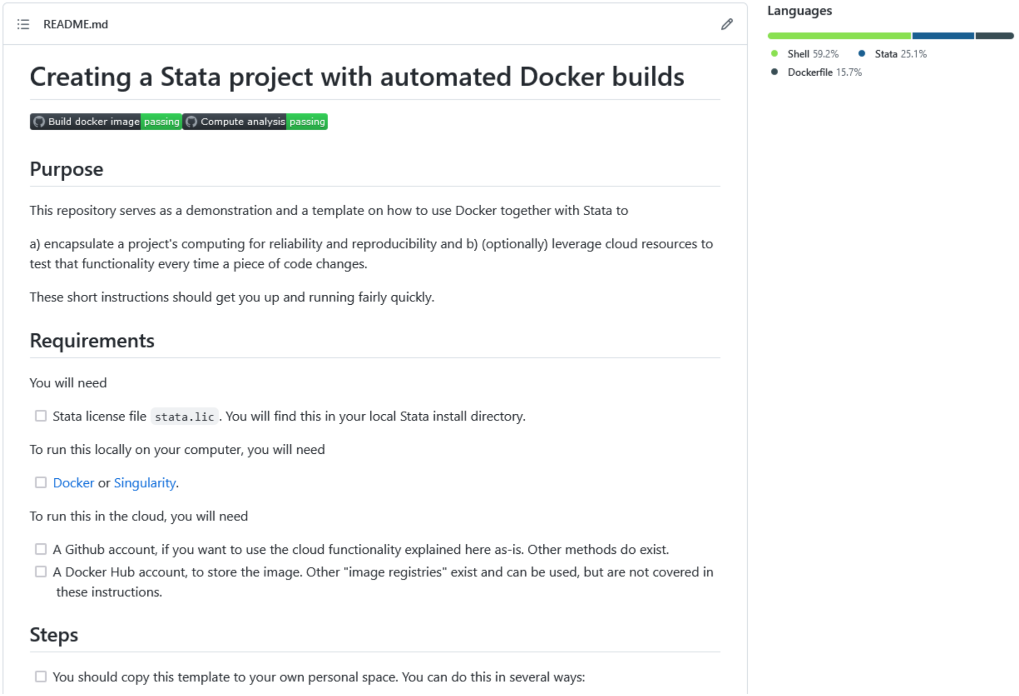
What are containers

Containers are "implementations of operating system-level virtualization,"[1] typically on Linux. The most common version is provided by Docker, but several other implementations exist, such as Singularity. The use of containers as part of replication packages in economics is extremely low, and yet they have some advantages. This post will explore both pre-submission and post-publication uses of containers, as well as several shortcomings

On This Page

Use of virtual environments (Docker, VM)

- https://aeadataeditor.github.io/posts/2021-11-16-docker
- https://github.com/AEADataEditor/stata-project-with-docker



Creating a Stata project with automated Docker builds

Purpose

This repository serves as a demonstration and a template on how to use Docker together with Stata to

a) encapsulate a project's computing for reliability and reproducibility and b) (optionally) leverage cloud resources to test that functionality every time a piece of code changes.

These short instructions should get you up and running fairly quickly.

Requirements

You will need

- □ Stata license file `stata.lic`. You will find this in your local Stata install directory.

To run this locally on your computer, you will need

- □ Docker or Singularity.

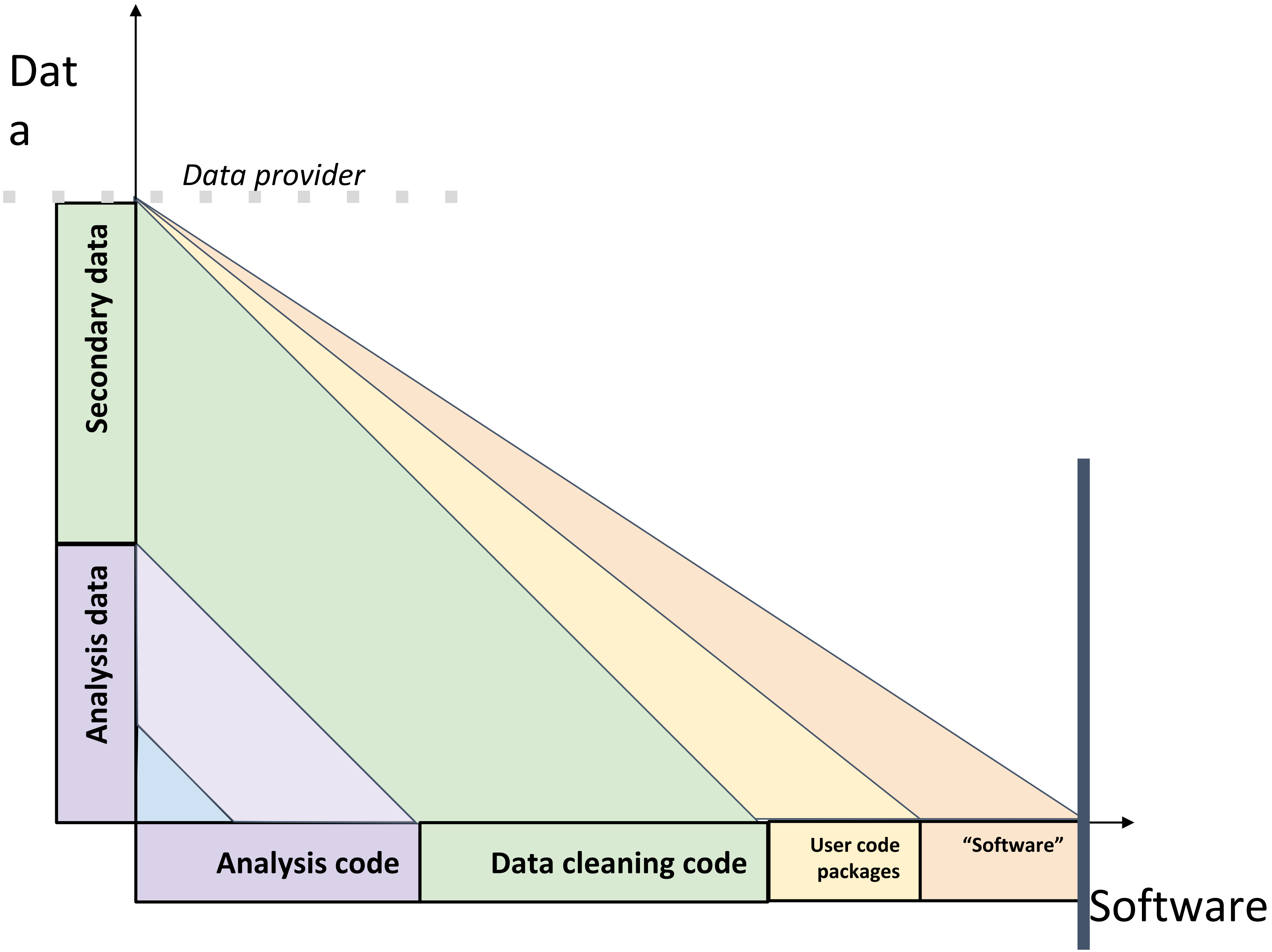
To run this in the cloud, you will need

- □ A Github account, if you want to use the cloud functionality explained here as-is. Other methods do exist.
- □ A Docker Hub account, to store the image. Other "image registries" exist and can be used, but are not covered in these instructions.

Steps

- □ You should copy this template to your own personal space. You can do this in several ways:

What if *data providers* were more transparent?



Dat
a

Data provider

Secondary data

Analysis data

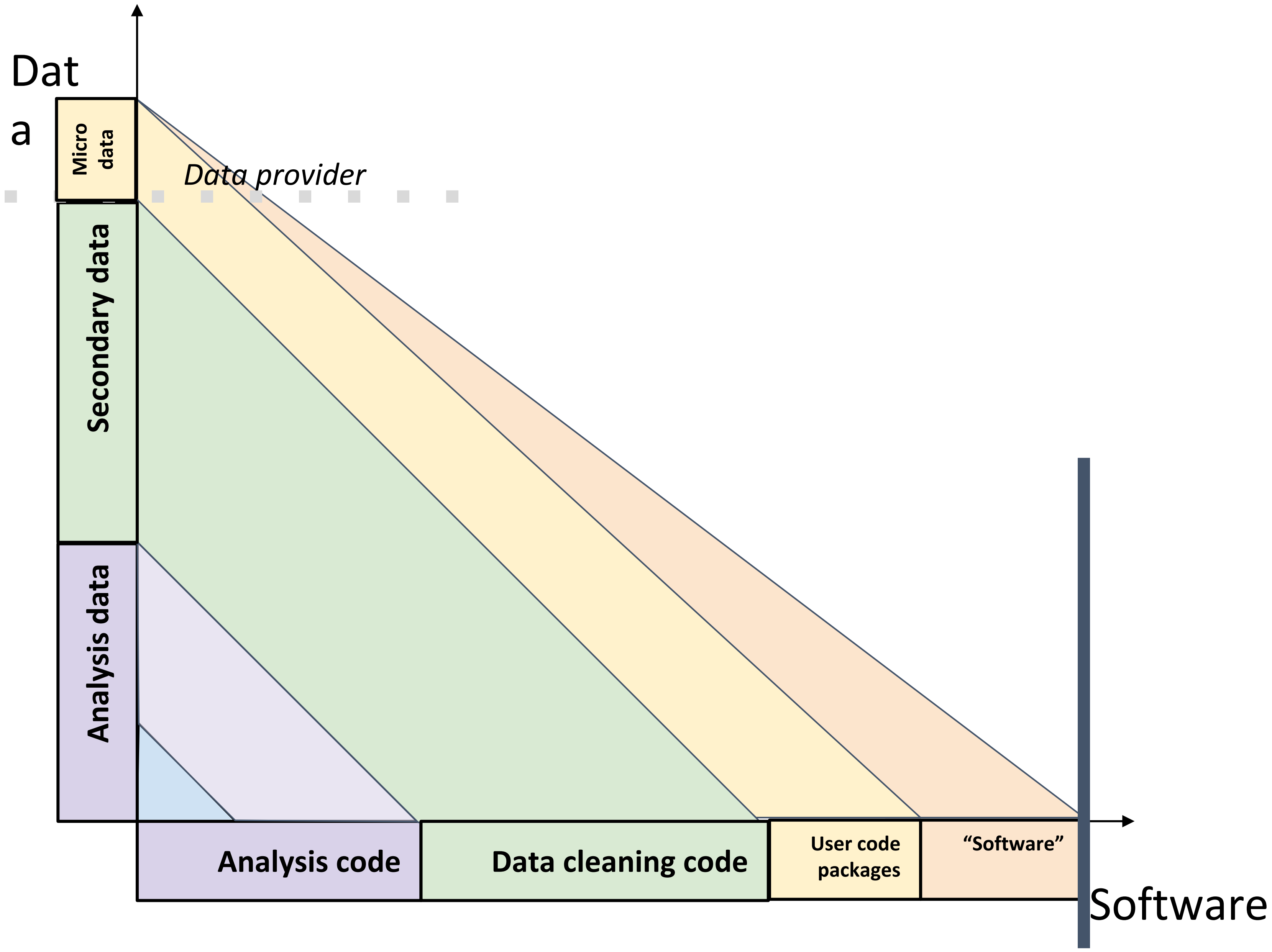
Analysis code

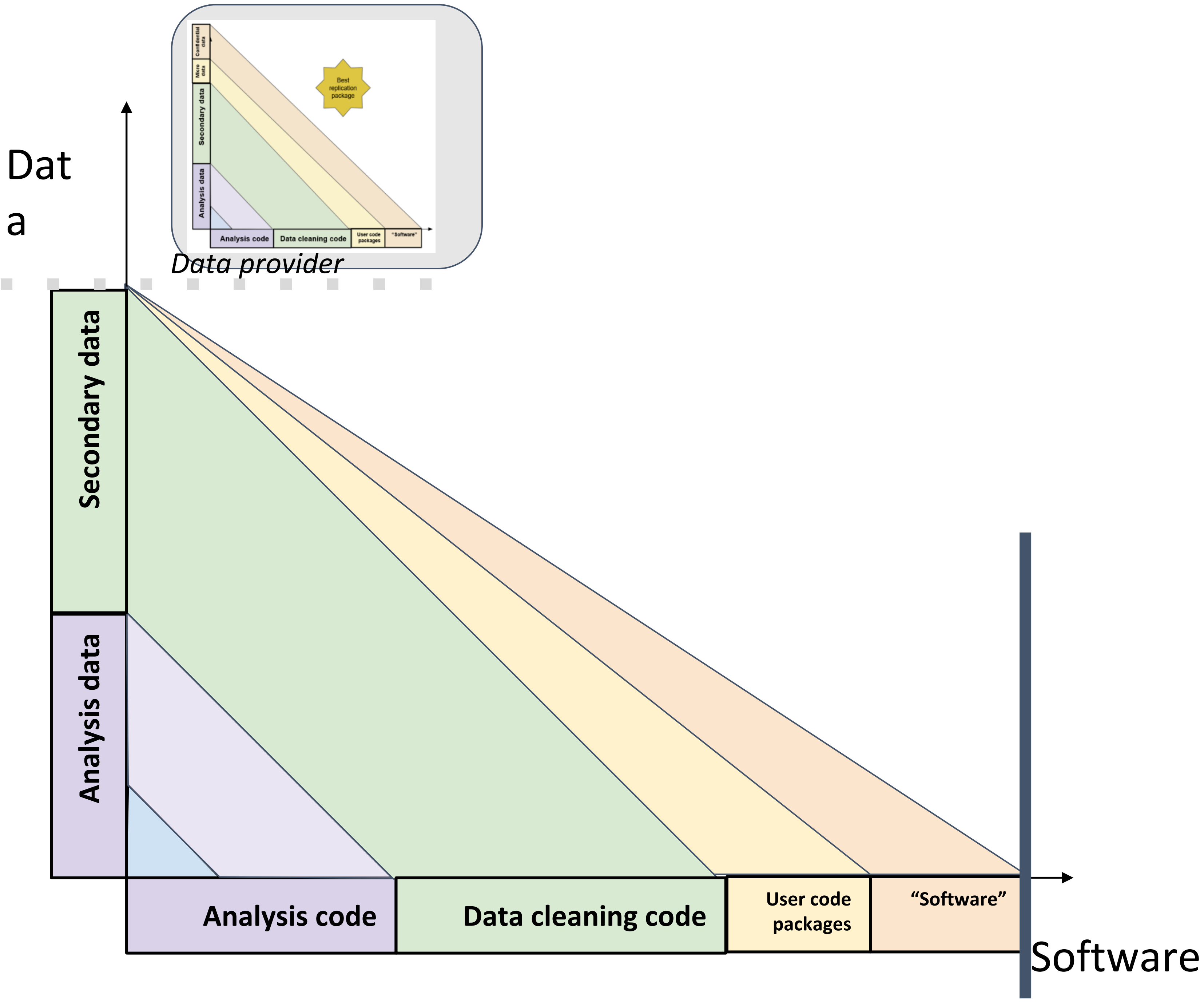
Data cleaning code

User code packages

“Software”

Software


Dat
a
Micro data
Data provider
Secondary data
Analysis data
Analysis code
Data cleaning code
User code packages
"Software"
Software


Confidential data
Micro data
Secondary data
Analysis data
Best replication package
Analysis code
Data cleaning code
User code packages
"Software"
Data
Data provider
Secondary data
Analysis data
Analysis code
Data cleaning code
User code packages
“Software”
Software

What if the paper were
constantly recomputed?

What if the paper were *constantly recomputed?*

- Recompute Jupyter notebook, Rmarkdown upon every change to the code?

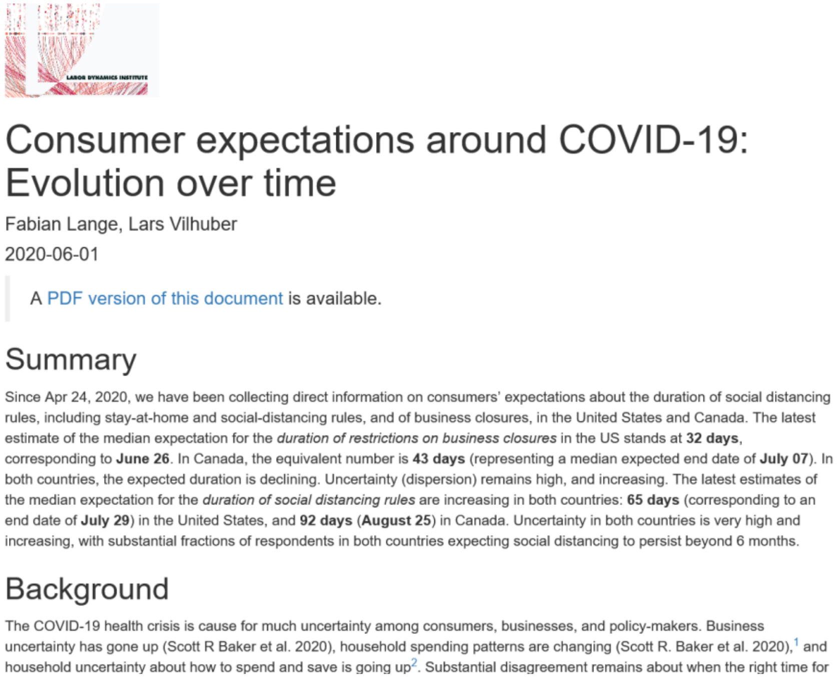
- It’s called “continuous integration” in software development
  - Easy to do if you have the infrastructure
  - Usually combines two services (Travis CI was quite popular)
  - More and more integrated (Github Workflows, Bitbucket Pipelines, Gitlab … something)

What if the paper were *constantly recomputed?*

- Recompute Jupyter notebook, Rmarkdown upon every change to the code?

- It's called "continuous integration" in software development

Summary
Background
Expectations for duration of restrictions in the United States
Expectations for duration of restrictions in Canada
References



Consumer expectations around COVID-19: Evolution over time

Fabian Lange, Lars Vilhuber

2020-06-01

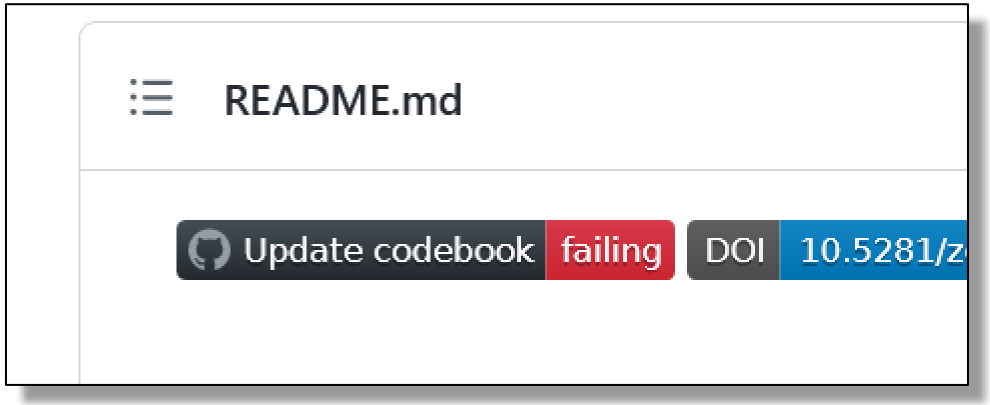
A PDF version of this document is available.

Summary

Since Apr 24, 2020, we have been collecting direct information on consumers' expectations about the duration of social distancing rules, including stay-at-home and social-distancing rules, and of business closures, in the United States and Canada. The latest estimate of the median expectation for the *duration of restrictions on business closures* in the US stands at **32 days**, corresponding to **June 26**. In Canada, the equivalent number is **43 days** (representing a median expected end date of **July 07**). In both countries, the expected duration is declining. Uncertainty (dispersion) remains high, and increasing. The latest estimates of the median expectation for the *duration of social distancing rules* are increasing in both countries: **65 days** (corresponding to an end date of **July 29**) in the United States, and **92 days** (**August 25**) in Canada. Uncertainty in both countries is very high and increasing, with substantial fractions of respondents in both countries expecting social distancing to persist beyond 6 months.

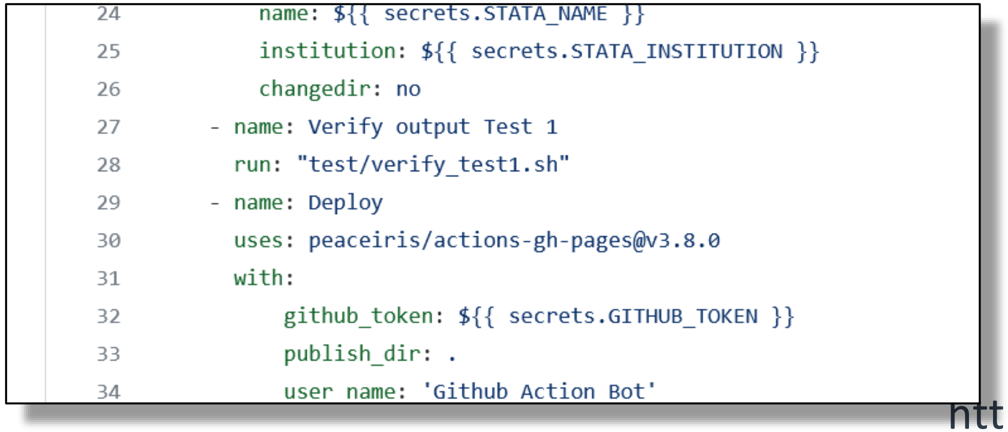
Background

The COVID-19 health crisis is cause for much uncertainty among consumers, businesses, and policy-makers. Business uncertainty has gone up (Scott R Baker et al. 2020), household spending patterns are changing (Scott R. Baker et al. 2020),[1] and household uncertainty about how to spend and save is going up[2]. Substantial disagreement remains about when the right time for



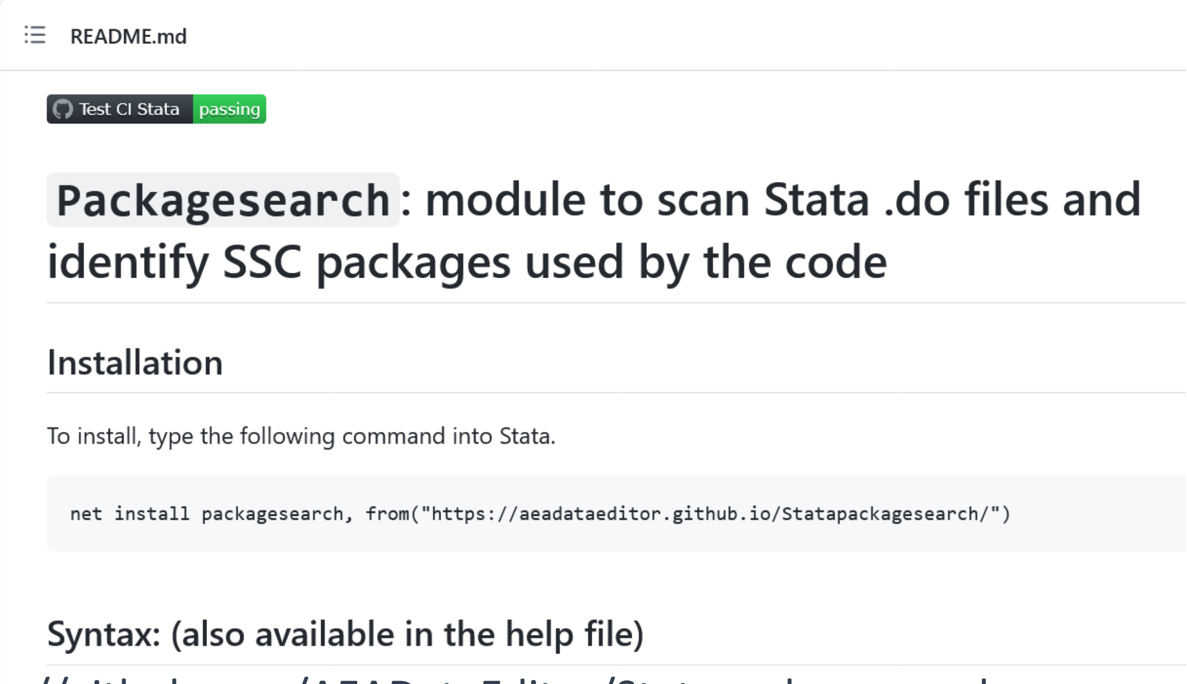
What if the paper were *constantly recomputed?*

- Recompute Jupyter notebook, Rmarkdown upon every change to the code?
- Recompute **Stata** upon every change to the code???



```
        name: ${{ secrets.STATA_NAME }}
        institution: ${{ secrets.STATA_INSTITUTION }}
        changedir: no
    - name: Verify output Test 1
      run: "test/verify_test1.sh"
    - name: Deploy
      uses: peaceiris/actions-gh-pages@v3.8.0
      with:
          github_token: ${{ secrets.GITHUB_TOKEN }}
          publish_dir: .
          user_name: 'Github Action Bot'
```

- It's called "continuous integration" in software development



README.md

Test CI Stata passing

`Packagesearch`: module to scan Stata .do files and identify SSC packages used by the code

Installation

To install, type the following command into Stata.

```
net install packagesearch, from("https://aeadataeditor.github.io/Statapackagesearch/")
```

Syntax: (also available in the help file)

https://github.com/AEADataEditor/Statapackagesearch

The role for journals

Goal: Transportability

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

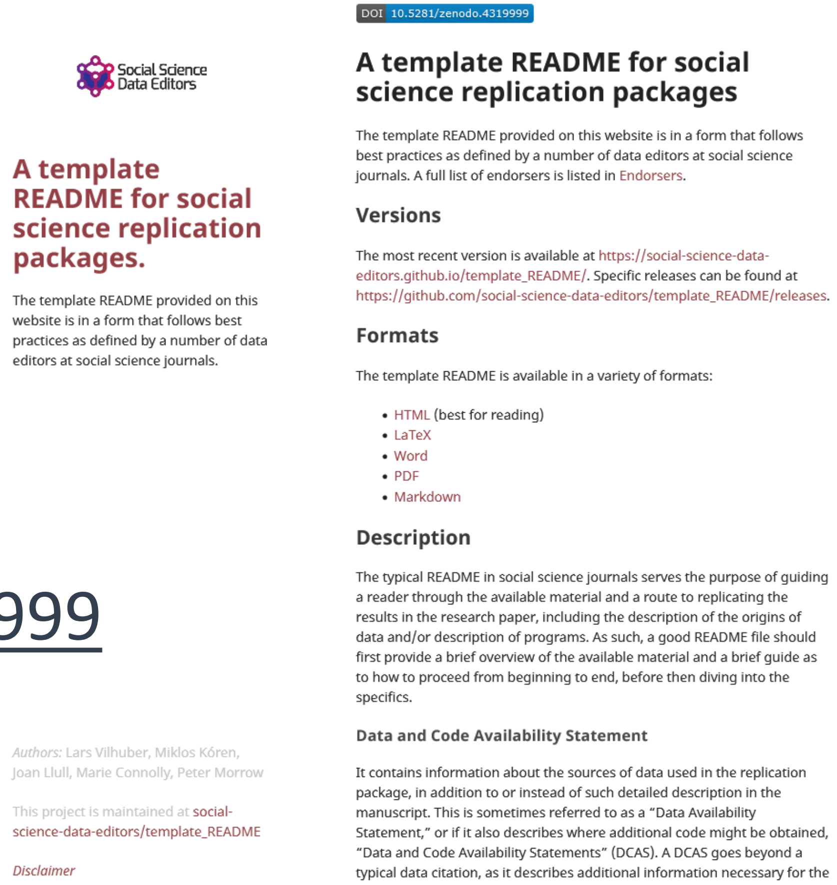
Social science “guild”

Social Science Data Editors

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

Template for README

- https://social-science-data-editors.github.io/template_README/
- https://doi.org/10.5281/zenodo.4319999



Social Science Data Editors

A template README for social science replication packages.

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

Authors: Lars Vilhuber, Miklos Kóren, Joan Llull, Marie Connolly, Peter Morrow

This project is maintained at social-science-data-editors/template_README

Disclaimer

DOI 10.5281/zenodo.4319999

A template README for social science replication packages

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

Versions

The most recent version is available at https://social-science-data-editors.github.io/template_README/. Specific releases can be found at https://github.com/social-science-data-editors/template_README/releases.

Formats

The template README is available in a variety of formats:

- HTML (best for reading)
- LaTeX
- Word
- PDF
- Markdown

Description

The typical README in social science journals serves the purpose of guiding a reader through the available material and a route to replicating the results in the research paper, including the description of the origins of data and/or description of programs. As such, a good README file should first provide a brief overview of the available material and a brief guide as to how to proceed from beginning to end, before then diving into the specifics.

Data and Code Availability Statement

It contains information about the sources of data used in the replication package, in addition to or instead of such detailed description in the manuscript. This is sometimes referred to as a "Data Availability Statement," or if it also describes where additional code might be obtained, "Data and Code Availability Statements" (DCAS). A DCAS goes beyond a typical data citation, as it describes additional information necessary for the

Thank you!

Recommendation for data providers

Recommendations (data providers)

Clear re-usable provenance statements

- Provide pre-written statements
  - Clear access rules
  - Clear timelines
  - Clear restrictions

Clear citation

- Use pre-written/ customizable citations
  - Various styles (APA, Chicago,…)
  - Various bibliographic managers
- Make landing page citation-friendly
  - DC Terms!

Corollary

Stable access

- Provide stable mechanisms
  - For static packages (URL)
  - For dynamic queries (cart!)
- Ideally PIDs (DOIs) prominently displayed
- Clear versioning (even if offline)
  - But provide an access mechanism that actually works!

Extension

Support for researcher-generated files (data and code)

- Provide a repository for both **distribution-restricted public data** AND **restricted (confidential) data**
- Link to code examples in the literature (replication pckgs!)