

# Continuous Transparency and Reproducibility in Academic Publications

Lars Vilhuber  
Cornell University

*2023-04-27*

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.



Transparency and reproducibility of scientific publications does not stop at publication

# 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

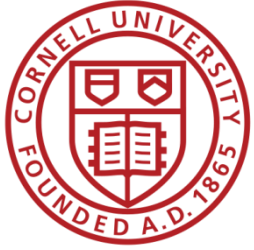


*American Economic Journal: Microeconomics* publishes papers focusing on microeconomic theory; industrial organization; and the microeconomic aspects of international trade, political economy, and finance.



# 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.**



# Reproducibility is assessed

Since July 16, 2019, the AEA Data Editor team has conducted reproducibility assessments

- for 1663 **manuscripts** (1369 full papers, 293 P&P) as of this morning



**AEA Data Editor** @AeaData · 1h

Normal 0%

At the start of summer of 2022, we have prepared about 1900 reports on about 1300 manuscripts (about 1050 if excluding the P&P). To infinity and beyond!



5



[Show this thread](#)



# Defining “reproducible research”

“Reproducibility” refers to the ability of a researcher to duplicate the results of a prior study using the **same materials** and **procedures** as were used by the original investigator.

Bollen et al. 2015. “Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science.” National Science Foundation.  
[https://www.nsf.gov/sbe/AC\\_Materials/SBE\\_Robust\\_and\\_Reliable\\_Research\\_Report.pdf](https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf).



# What do we check?

- How can somebody access the data?
- How can somebody assemble the computational resources?
- How can somebody obtain the figures and tables in the paper?
- How long will that take?



## 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 Kören, Joan Llull, Marie Connolly, Peter Morrow

This project is maintained at [social-science-data-editors/template\\_README](https://github.com/social-science-data-editors/template_README)

*Disclaimer*

DOI [10.5281/zenodo.4319999](https://doi.org/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/](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](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.





# Guidelines? Template README

<https://doi.org/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.

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

This project is maintained at [social-science-data-editors/template\\_README](https://github.com/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/](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](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.

# Computational empathy



# Computational empathy

- Focal reader: your next RA in 4 years
- Interaction: you hand them your README, but don't have time to go through all the details...
- Budget constraint: It shouldn't take too many RA hours
- Time constraint: It shouldn't take more than 1 week to "get it"



## 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 Kören, Joan Llull, Marie Connolly, Peter Morrow

This project is maintained at [social-science-data-editors/template\\_README](https://social-science-data-editors/template_README)

*Disclaimer*

DOI [10.5281/zenodo.4319999](https://doi.org/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/](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](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.

# Research compendium



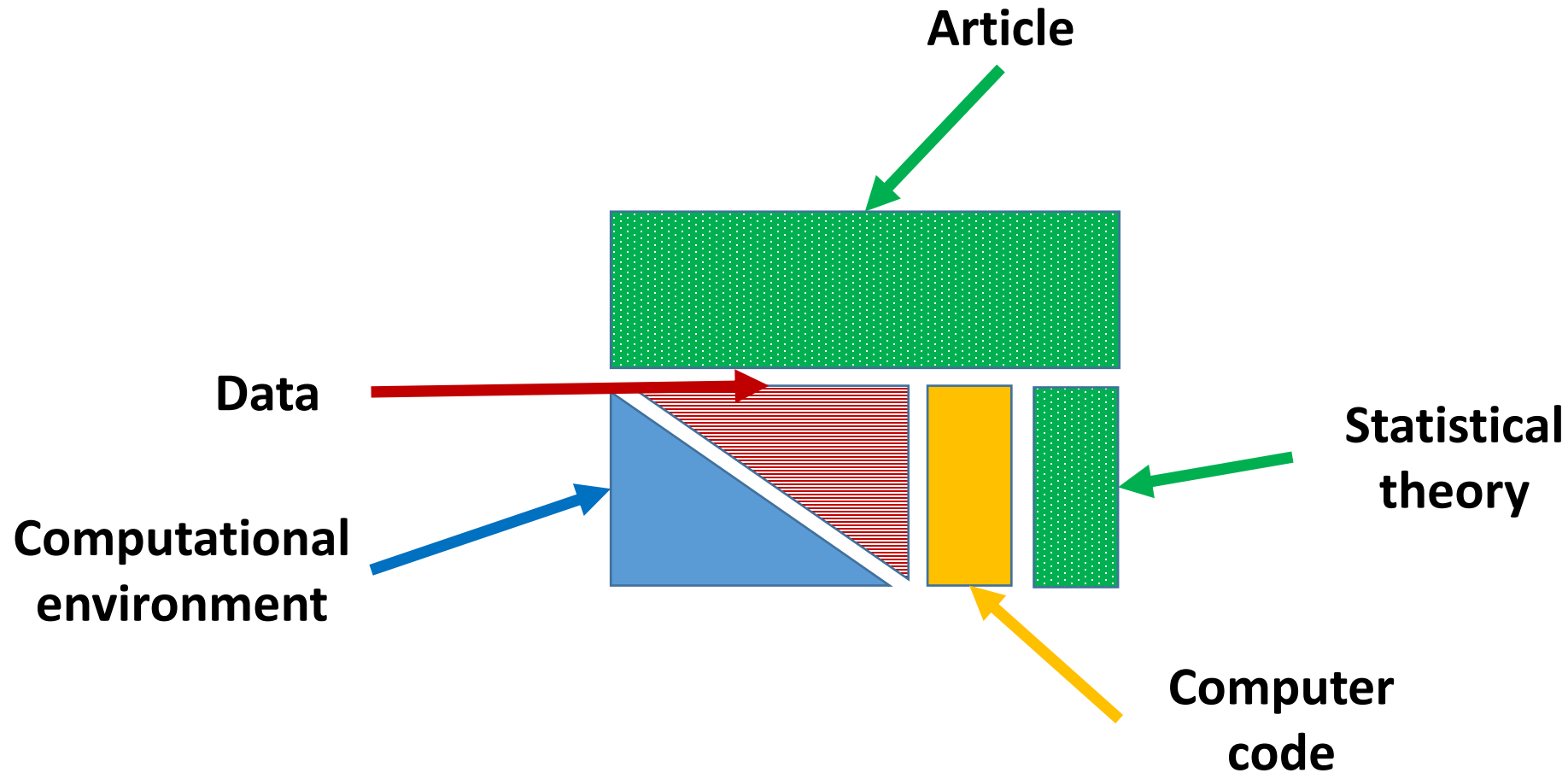
# Research compendium

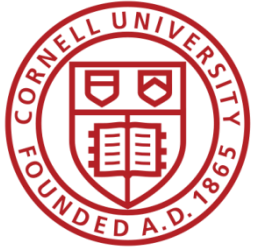
In the words of the slogan popularized by Buckheit and Donoho (1995),

***“a scientific publication is [...] merely advertising of the scholarship: [...] the complete software development environment and the complete set of instructions which generated the figures.”***

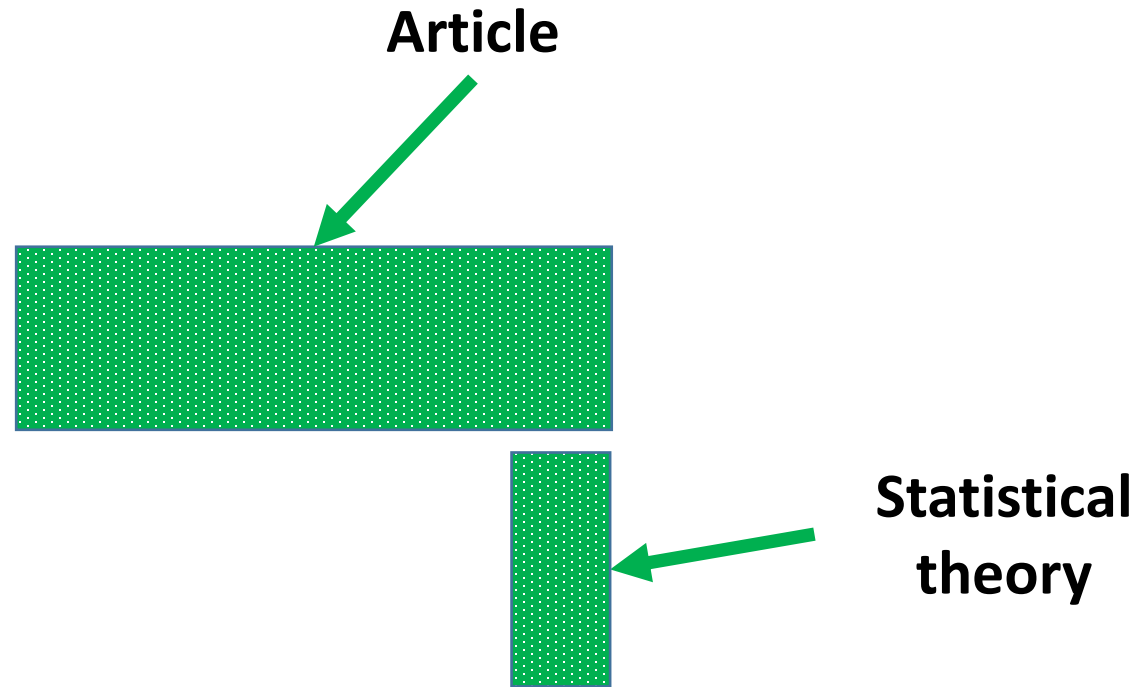


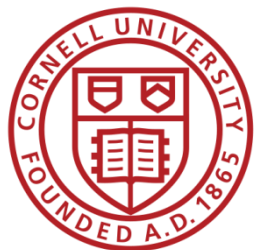
# Research compendium



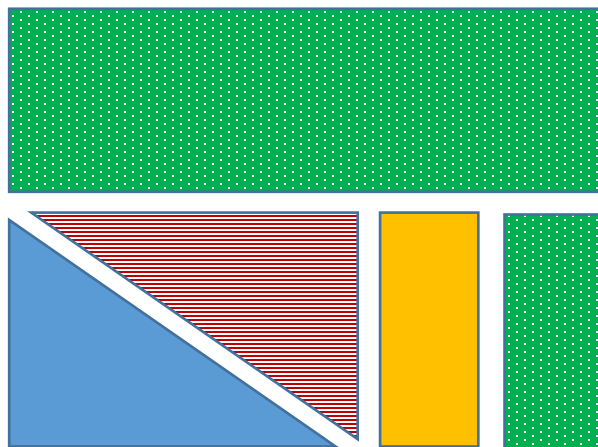


# Partial research compendium is not robust

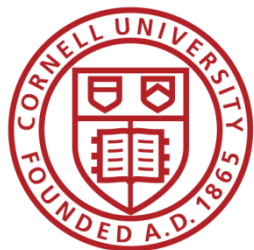




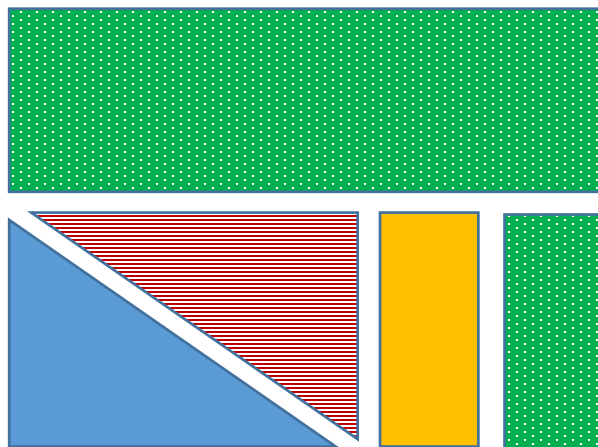
# Research compendium

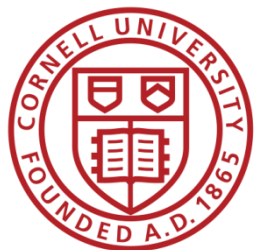




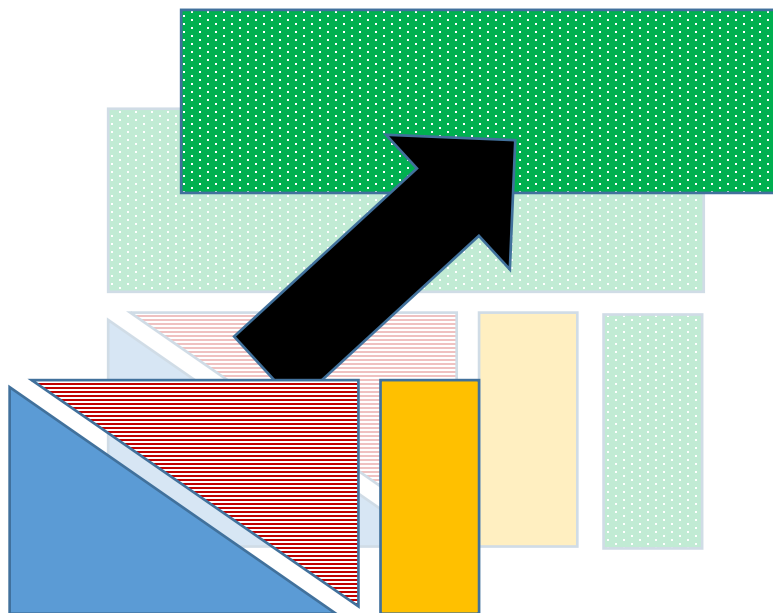


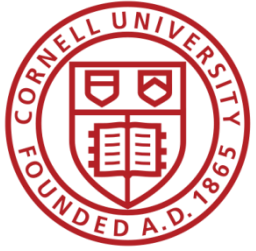
Do we have all the pieces?



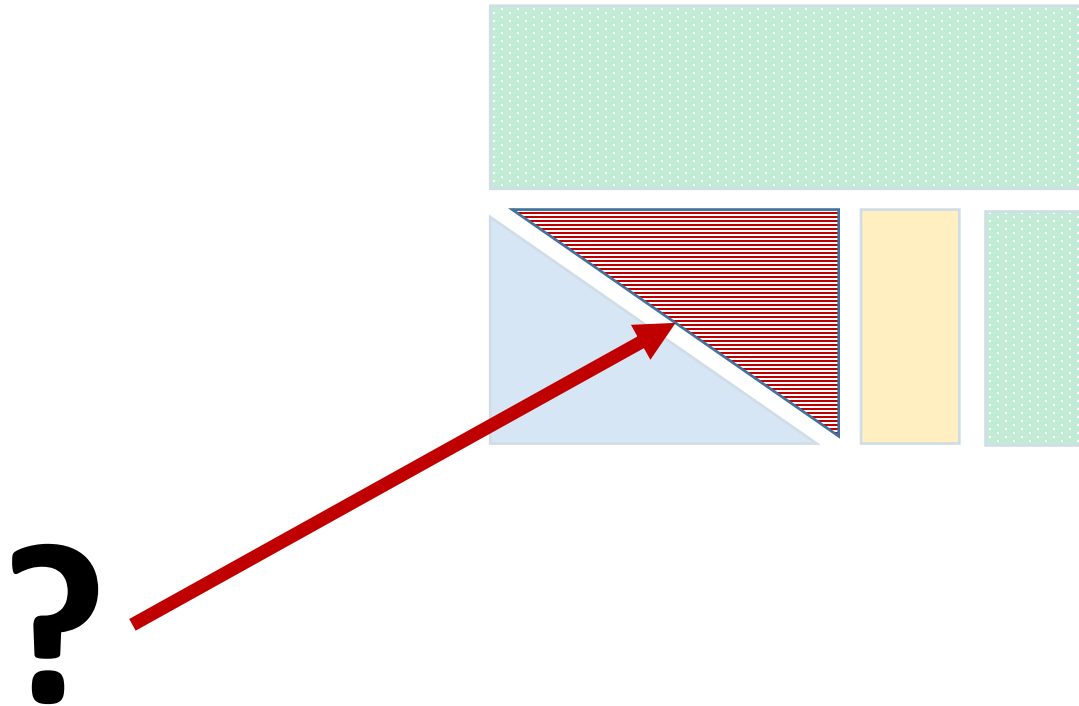


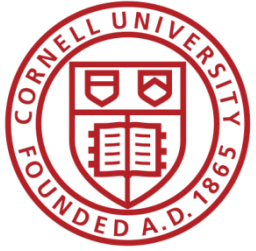
Does code + data + compute create output?



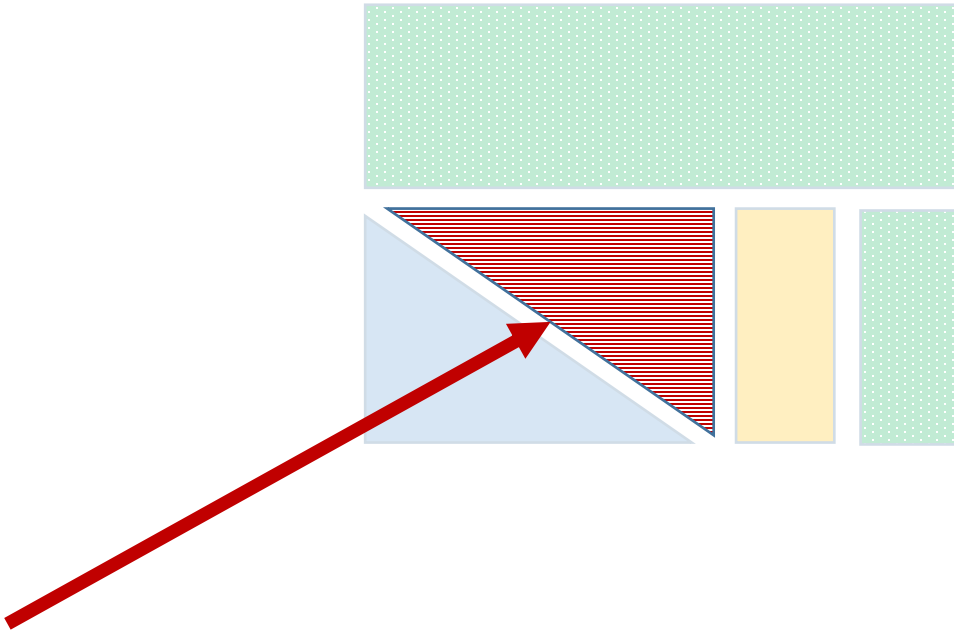
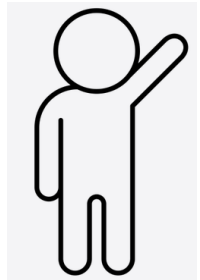


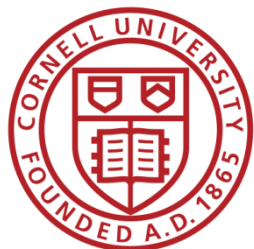
# Data provenance is important





# Data access is important





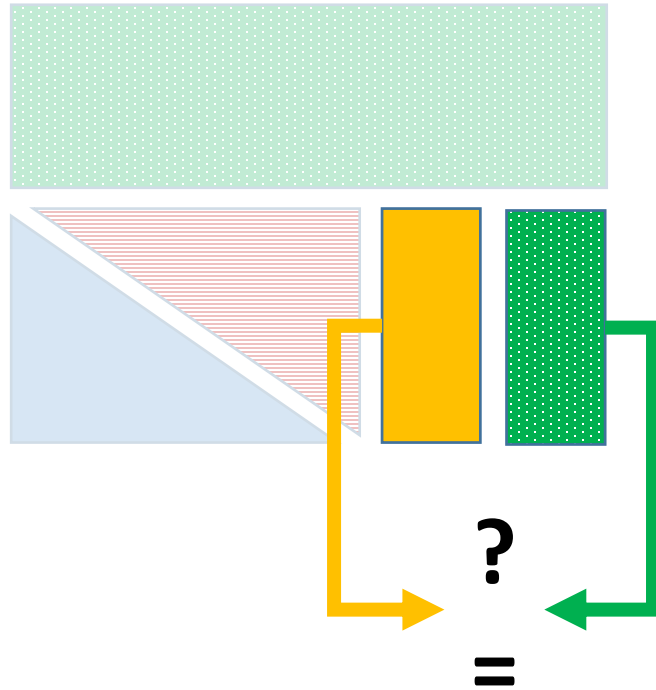
# Data access

- Public-use and shared confidential data





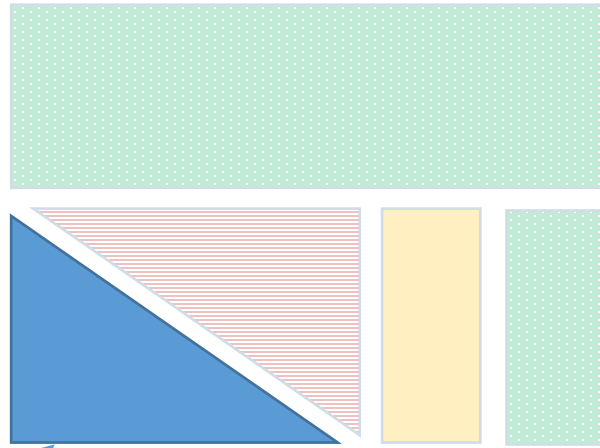
# Correctness of code is important





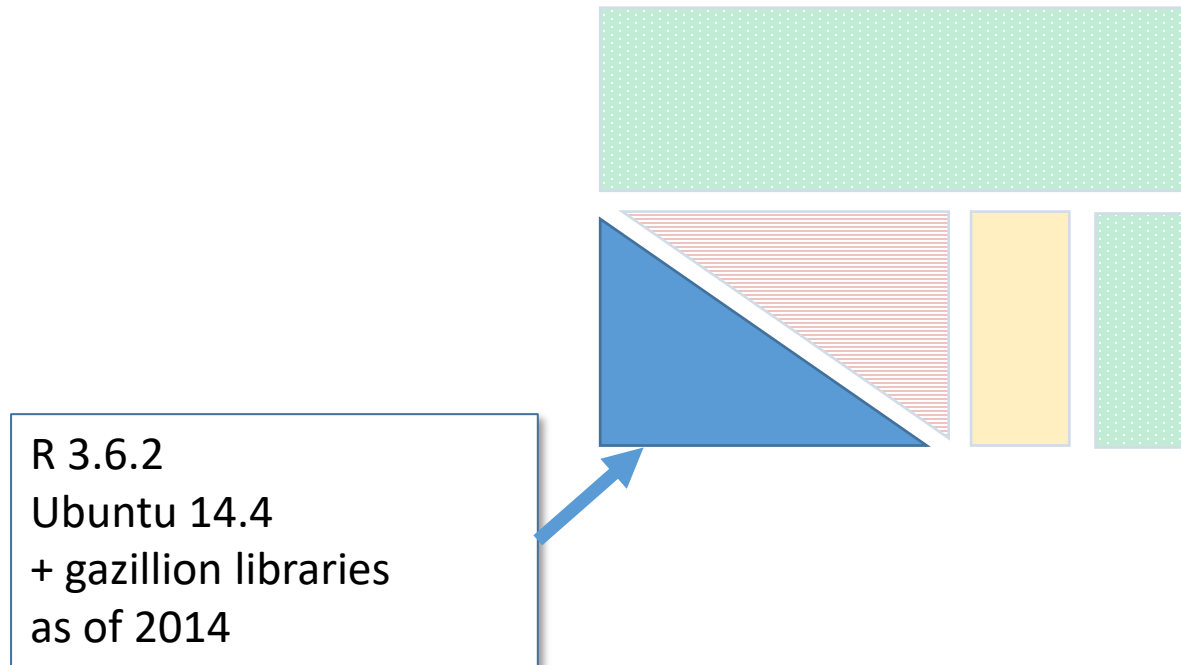
# Is the computational environment completely defined?

Stata 14  
MacOS 11.2  
+ 10 packages  
as of 2014





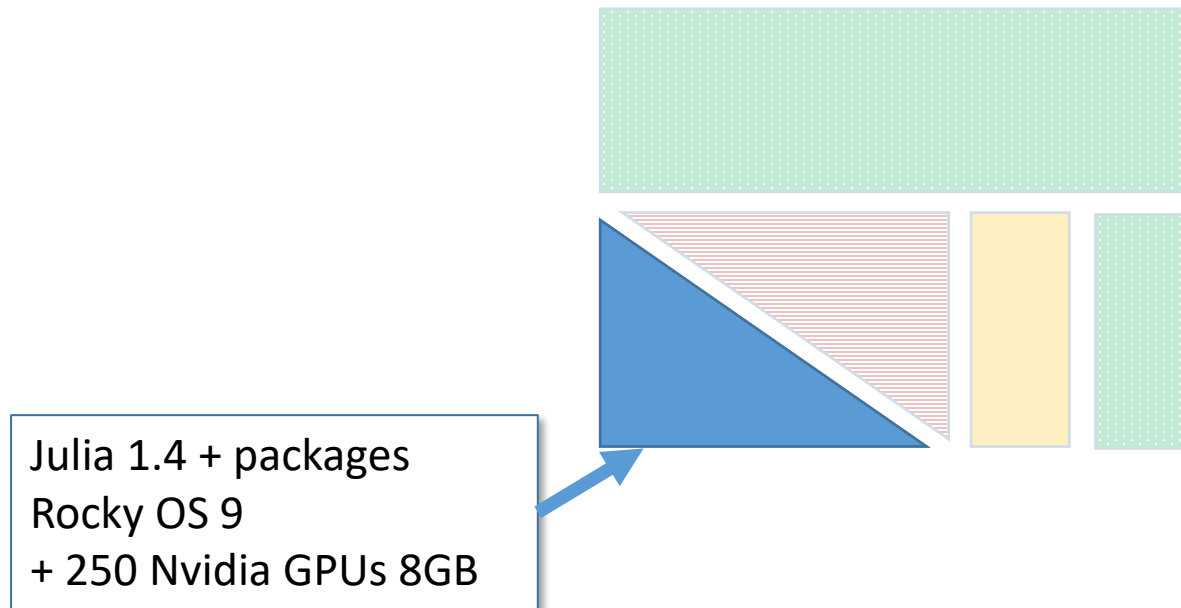
# Is the computational environment completely defined?







# Is the computational environment completely defined?



Computational  
empathy  
= “pity the poor  
replicator”



# Experience of irreproducibility as a risk factor for poor mental health in biomedical science doctoral students: A survey and interview-based study

AUTHORS

Nasser Lubega, Abigail Anderson, [Nicole Nelson](#)

Computational  
empathy  
= transparency

Monitoring  
transparency



# Journals check different things

- **AEA**: after acceptance, is code complete, is data cited, does code run (if data accessible)
- **AJPS**: after acceptance, is data documented, is code complete, does code run (if data accessible)?
- **CJE**: after acceptance, is package plausibly complete?
- **Astrophysics**: are data properly cited, are interactive graphics also accessible, is data standardized?

Monitoring  
transparency



# Journals are not the end...

- Reproducibility and replicability is checked in graduate class work (even in undergraduate!)
- Replication Games!
- Articles build on the foundations laid by others
  - Rely on previous findings
  - Question and recompute previous findings





# Painful when it fails

## THE REVIEW OF ECONOMIC STUDIES

Issues More content ▼ JEL ▼ Submit ▼ Purchase About ▼

The Review of Economic



Volume 90, Issue 2  
March 2023

< Previous

### JOURNAL ARTICLE

#### Retraction of: Growing up in a Recession

*The Review of Economic Studies*, Volume 90, Issue 2, March 2023, Page 1009,  
<https://doi.org/10.1093/restud/rdac085>

**Published:** 11 January 2023

This is a retraction to: *The Review of Economic Studies*, Volume 81, Issue 2, April 2014,  
Pages 787–817, <https://doi.org/10.1093/restud/rdt040>

 PDF  Split View  Cite  Permissions  Share ▼

**Issue Section:** [Retraction](#)

This is a retraction of: Paola Giuliano, Antonio Spilimbergo, Growing up in a  
Recession, *The Review of Economic Studies*, Volume 81, Issue 2, April 2014, Pages

## The Journal of FINANCE

The Journal of THE AMERICAN FINANCE ASSOCIATION

RETRACTION |  Free Access

### Retracted: Risk Management in Financial Institutions

 This article retracts the following: ▼

First published: 05 July 2021 | <https://doi.org/10.1111/jofi.13064>

 PDF  TOOLS 

[This article retracts the following:](#)

Adriano A. Rampini, S. Viswanathan, Guillaume Vuillemeys

<https://doi.org/10.1111/jofi.12868>

First published: 12 December 2019



# Correction of the academic record

## Growing up in a recession

P Giuliano, A Spilimbergo - *The Review of Economic Studies*, 2014 - JSTOR

Does the historical macroeconomic environment affect preferences for redistribution? We find that individuals who experienced a recession when young believe that success in life ...

☆ Save  Cite Cited by 840 Related articles



# Powerful when it works

## Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends

Jonathan Roth

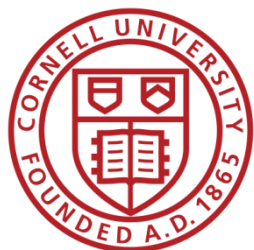
AMERICAN ECONOMIC REVIEW: INSIGHTS  
VOL. 4, NO. 3, SEPTEMBER 2022  
(pp. 305-22)

[Download Full Text PDF](#)

### Article Information

#### Abstract

This paper discusses two important limitations of the common practice of testing for preexisting differences in trends ("pre-trends") when using difference-in-differences and related methods. First, conventional pre-trends tests may have low power. Second, conditioning the analysis on the result of a pretest can distort estimation and inference, potentially exacerbating the bias of point estimates and under-coverage of confidence intervals. I analyze these issues both in theory and in simulations calibrated to a survey of recent papers in leading economics journals, which suggest that these limitations are important in practice. I conclude with practical recommendations for mitigating these issues.



# Availability key

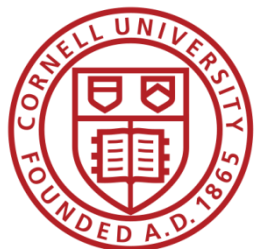
I searched on Google Scholar for occurrences of the phrase “event study” in papers published in the *American Economic Review*, *American Economic Journal: Applied Economics*, and *American Economic Journal: Economic Policy* between 2014 and June 2018.<sup>1</sup> I chose the phrase “event study” since researchers often evaluate pre-trends in an event-study plot.

The search returned 70 total papers that include a figure that the authors describe as an event-study plot. I exclude 43 papers for which data to replicate the main event-study plot were unavailable.<sup>2</sup> I further exclude 9 papers that do not report standard errors<sup>3</sup> and 3 that do not normalize their estimates relative to a pretreatment period.<sup>4</sup> Finally, I exclude 3 papers that do not attribute a causal interpretation to their estimates so that I can benchmark the magnitude of biases from differential trends relative to the estimated causal effects. This yields a final sample of 12 papers. For papers that present multiple event-study plots, I focus on the first plot meeting the criteria above, which I view as a reasonable proxy for the main specification.



# Roth (2022) re-computes papers

- Takes 12 papers with accessible papers
- Re-runs the analysis for each of them, adding his special twist (in this case, looking at the power of pre-trend tests)
- Not possible as meta-analysis: requires access to code and data



# Similarly: de Chaisemartin + Ramirez-Cuellar (forthcoming, AEJ:Applied)

## At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments?

Clément de Chaisemartin

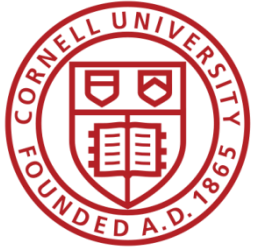
Jaime Ramirez-Cuellar

AMERICAN ECONOMIC JOURNAL: APPLIED ECONOMICS (FORTHCOMING)

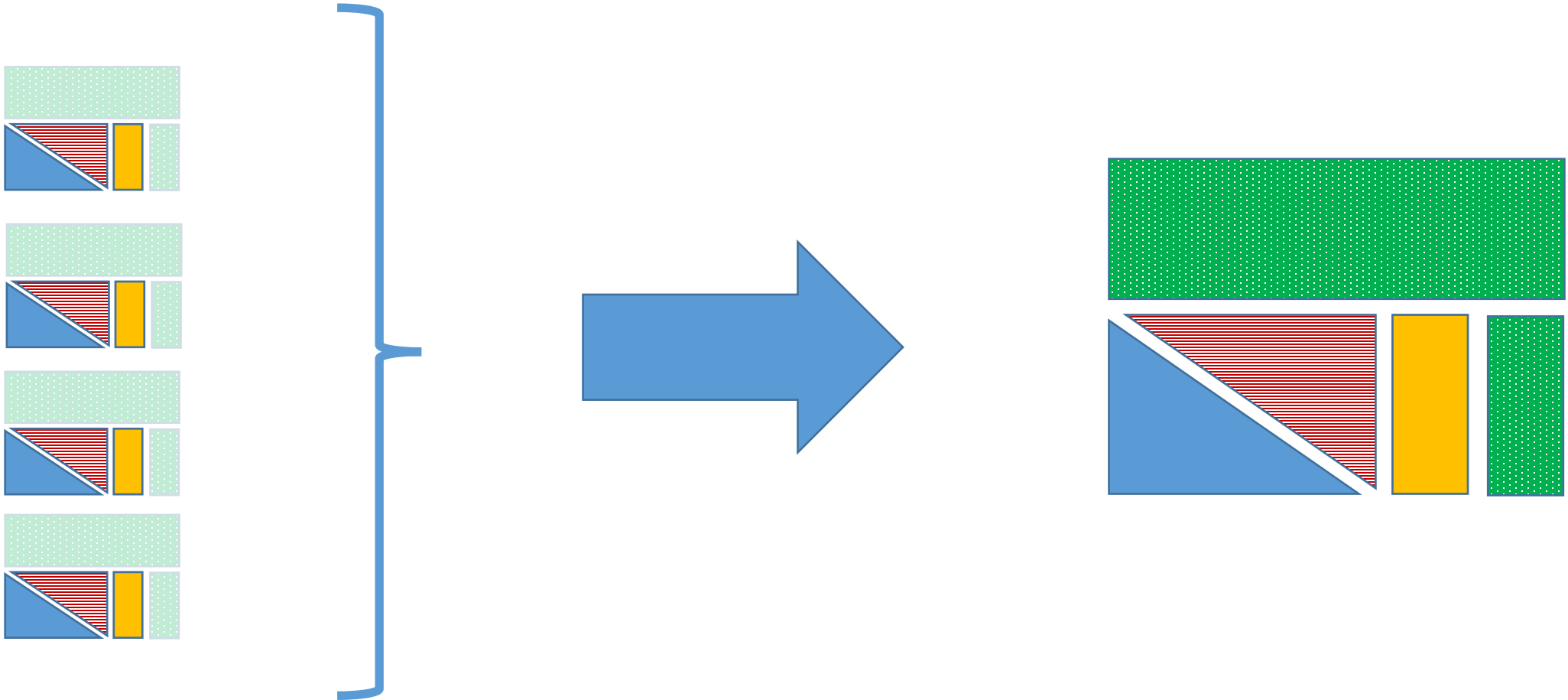
### Article Information

#### Abstract

In matched-pairs experiments in which one cluster per pair of clusters is assigned to treatment, to estimate treatment effects, researchers often regress their outcome on a treatment indicator and pair fixed effects, clustering standard errors at the unit-of-randomization level. We show that even if the treatment has no effect, a 5%-level t-test based on this regression will wrongly conclude that the treatment has an effect up to 16.5% of the time. To fix this problem, researchers should instead cluster standard errors at the pair level. Using simulations, we show that similar results apply to clustered experiments with small strata.



# Do we have all the pieces?



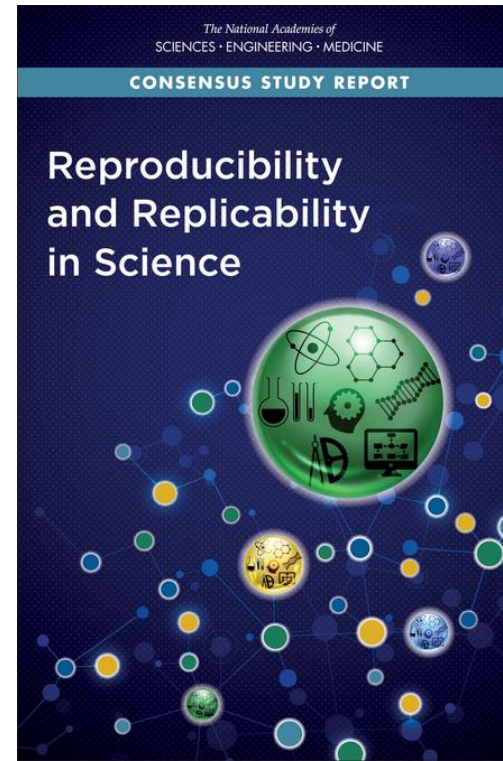
Why





# Replication continuum

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

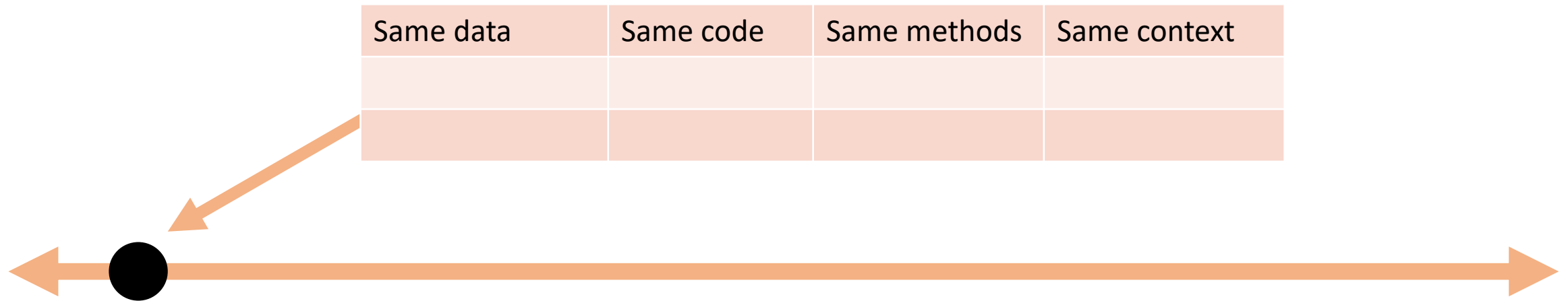


## Reproducibility

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



# Replication continuum



## Reproducibility

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



# Replication continuum



## Reproducibility

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

## Replicability

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



# Replication continuum

Same data	<b>Different code or software</b>	Same methods	Same context



**Reproducibility**

**Replicability**

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

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



# Replication continuum

New data	Same code	Same methods	Same context
collection			



**Reproducibility**

**Replicability**

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

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



# Replication continuum



## Reproducibility

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

## Replicability

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

## Generalizability

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



# Replication continuum

Different data	Different code	Different	Different
	or software	methods	context or
			country



**Reproducibility**

**Replicability**

**Generalizability**

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

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

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

# Tools





# Continuous updating of replication packages

AEA policy allows for revisions of packages after publication

- The initial package remains the “version of record”
- Additional versions are linked

The screenshot shows the American Economic Association (AEA) website. The header includes the AEA logo and navigation links: Membership, About AEA, Journals, Annual Meeting, Careers, Resources, and More +. A breadcrumb trail reads: Home > Journals > AEA Data and Code Policies and Guidance > Policy on Revisions of Data and Code Deposits in the AEA Data and Code Repository. The main content area is titled "Policy on Revisions of Data and Code Deposits in the AEA Data and Code Repository" and lists the following policy points:

- Policy
  - No Replacement
  - Infringement



# Continuous updating of replication packages

AEA policy allows for revisions of packages after publication

- The initial package remains the “version of record”
- Additional versions are linked

**OPENICPSR**Find Data

[Find Data](#) / [Process data for the AEA Pre-publication Verification Service](#)

A [newer version](#) of this project is available. See below for other available versions.

Process data for the AEA Pre-publication verification S

Principal Investigator(s): [Lars Vilhuber](#), Cornell University

Version: [V1](#)

**Published Versions**  
[V3 \[2022-05-24\]](#)  
[V2 \[2021-04-30\]](#)  
[V1 \[2020-02-24\]](#)



# Continuous updating of replication packages

- Data Editor monitors and responds to inquiries
- About 3% of yearly activity are revisions

## Industrial Espionage and Productivity

Albrecht Glitz  
Erik Meyersson

AMERICAN ECONOMIC REVIEW  
VOL. 110, NO. 4, APRIL 2020  
(pp. 1055-1103)

[Download Full Text PDF](#)

### Article Information

#### Abstract

In this paper, we investigate the economic returns to industrial espionage. We show that the flow of information provided by East German informants in the West over the period 1970–1989 led to a significant narrowing of sectoral TFP gaps between West and East Germany. These economic returns were primarily driven by relatively few high-quality pieces of information and particularly large in sectors closer to the West German technological

## OPENICPSR

[Find Data](#) / [Data and Code for: Industrial Espionage and Productivity](#)

A [newer version](#) of this project is available. See below for other available versions.

### Data and Code for: Industrial Espionage and Productivity

**Principal Investigator(s):** Albrecht Glitz, Universitat Pompeu Fabra, IPEG and Barcelona GSE; Erik Meyersson, Handelsbanken Capital Markets

**Version:** V1

### Published Versions

[V2 \[2022-10-06\]](#)

[V1 \[2020-03-23\]](#)



# Linking to private sharing sites

- Links to Github repositories
  - NOTE: Github.com is **NOT** an archive!
- Links to source websites

Added as part of a “related publication” taxonomy

**OPENICPSR**Find Data

[Find Data](#) / [Data and code for: Synthetic Difference in Differences](#)

Data and code for: Synthetic Difference in Differences

**Principal Investigator(s):** [Dmitry Arkhangelsky](#), CEMFI; Susan Athey, Stanford University; David A. Hirshberg, Emory University; Guido W. Imbens, Stanford University; Stefan Wager, Stanford University

**Version:** [V1](#)

Name	File Type	Size	Last Modified
<a href="#">synthdid-sdid-paper</a>			07/28/2021 07:09:PM

**Related Publications**

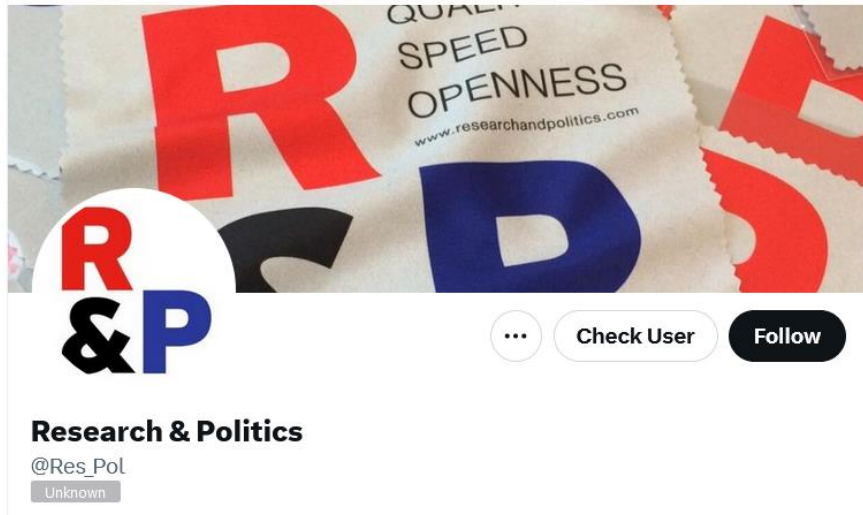
The following publications cite the data in this project.

- Arkhangelsky, Dmitry, Susan Athey, David Hirshberg, Imbens Guido, and Wager Stefan. “Synthdid: Synthetic Difference in Differences Estimation,” September 3, 2021. <https://github.com/synth-inference/synthdid>.



# Special journals

- Journal of Comments and Replications in Economics (JCRE)
- Journal of Applied Econometrics
- Research & Politics

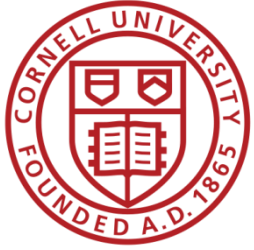


## EXTENSION OF THE REPLICATION SECTION'S COVERAGE

The *JAE* Replication Section, introduced in January 2003 under the editorship of Badi H. Baltagi, was initially devoted exclusively to the issue of replication of empirical results published in papers of the *Journal of Applied Econometrics*.

Given the encouraging response, we extended the coverage of the section to include replication of empirical results that have been published as regular papers in the following additional journals:



- *Econometrica*
- *American Economic Review* (except Papers & Proceedings)
- *Journal of Political Economy*
- *Quarterly Journal of Economics*
- *Review of Economics and Statistics*
- *Review of Economic Studies*
- *Journal of Econometrics*




# New types of journals?


BRIEF REPORT Check for updates

**REVISED** Studying the prominence effect amid the COVID-19 crisis: implications for public health policy decision-making. [version 2; peer review: 2 approved]

✉ Yossi Maaravi <sup>1</sup>, Ben Heller <sup>2</sup>

[Author details](#)


 This article is included in the **Emerging Diseases and Outbreaks** gateway.


 This article is included in the **Coronavirus** collection.


**ALL METRICS**


1522  
VIEWS


101  
DOWNLOADS


 Get PDF

 Get XML

 Cite

 Export

 Track

 Share

## Abstract

The novel coronavirus disease 2019 (COVID-19) has brought with it crucial policy- and decision-making situations, especially when making judgments between financial and health concerns. One particularly relevant decision-making phenomenon is the prominence effect, where decision-makers base their decisions on the most prominent attribute of the object at hand (e.g., health concerns) rather than weigh all the attributes together. This bias diminishes when the decision-making mode inhibits heuristic processes. In this study, we tested the prominence of health vs. financial concerns across two decision-making modes - choice (prone to heuristics) and matching (mitigates heuristics) - during the peak of the COVID-19 in the UK using Tversky *et al.*'s classic experimental paradigm. We added to the classic experimental design a priming condition. Participants were presented with two casualty-minimization programs, differing in lives saved and costs: program X would save 100 lives at the cost of 55-million-pound sterling

## Open Peer Review

Reviewer Status   

## Reviewer Reports

*Invited Reviewers*

1

2


**Version 2**  
(revision)  
07 May 21

  
**read**

**Version 1**  
20 Nov 20

↑  
  
**read**

  
**read**

1. **Roshen Fernando** , Australian National University, Canberra, Australia
2. **Peterson K Ozili**, University of Essex, Wivenhoe Park, UK





# Generating findings

UC San Diego



## SAN DIEGO'S REPLICATION GAMES.

📅 JULY 1ST, 2023

📍 UCSD, DEPARTMENT OF POLITICAL SCIENCE, CA

We are looking for professors, post-docs, and PhD students interested in a one-day replication challenge.

Participants will be **granted co-authorship** on a meta-paper combining the replications and will have the **opportunity to publish** their work. Participants will be matched based on field, and a study from a leading economics or political science journal will be assigned to each team based on interests.

The event will take place in person at the UC San Diego. Virtual participants are also welcome. The local organizer is Scott Desposato from the Department of Political Science at UCSD. The event is also co-sponsored by the Western Economic Association International (<https://weai.org/conferences/view/13/98th-Annual-Conference>).

Interested researchers or teams should send their field of study and preferred statistical software to:

**ABEL BRODEUR**  
abrodeur@uottawa.ca



## VIENNA'S REPLICATION GAMES.

📅 MAY 5TH, 2023

📍 WU (VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS), VIENNA, AUSTRIA

We are looking for professors, post-docs, and PhD students interested in a one-day replication challenge.

Participants will be **granted co-authorship** on a meta-paper combining the replications and will have the **opportunity to publish** their work. Participants will be matched based on field, and a study from a leading economics or political science journal will be assigned to each team based on interests.

The event will take place in person at WU (Vienna University of Economics and Business). Virtual participants are also welcome. The local organizer is Christoph Huber.

Interested researchers or teams should send their field of study and preferred statistical software to:

**ABEL BRODEUR**  
abrodeur@uottawa.ca



## LES JEUX DE RÉPLICATION DE MONTRÉAL.

📅 14 JUIN 2023

📍 Université du Québec à Montréal, CANADA

Nous sommes à la recherche de chercheurs.ses, post-doctorant.e.s et doctorant.e.s intéressé.e.s à prendre part à un défi de réplication d'une journée.

Les participant.e.s **deviendront co-auteurs** sur un méta-papier combinant les répliques et auront l'**opportunité de publier** leur travail.

Les participant.e.s seront apparié.e.s selon leur champ et une étude d'une revue en économique ou science politique sera attribuée à chaque équipe selon leurs intérêts.

L'événement aura lieu à l'Université du Québec à Montréal. Une participation virtuelle est également bienvenue.

Les chercheurs.ses et/ou équipes intéressées doivent envoyer leur champ d'études et leur logiciel statistique préféré à :

**ABEL BRODEUR**  
abrodeur@uottawa.ca



# Replication games, challenges, etc.

- <https://i4replication.org/description.html>
- “Hackathon” for data
- 70+ people in multiple teams, not just econ



[Home](#) [Replication Reports](#) [Replication Instructions](#) ▾ [People](#) [Teaching](#) [Publishing](#) [Repl](#)

## Replication Games

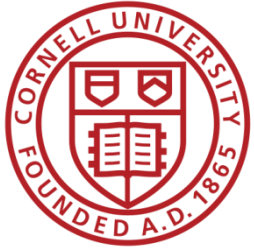
**Replication Games: A Collaborative Research Initiative:** The Replication Games is a one-day event that brings researchers together to collaborate on replicating papers published in high-ranking journals. Replication is a crucial aspect of scientific research, ensuring that results are reliable and reproducible. By participating in the Replication Games, you will not only contribute to the integrity of research in your field but also have the opportunity to network with fellow researchers and develop your coding skills.

For the Replication Games, we are looking for professors, post-docs, graduate students and more broadly researchers. See [here](#) for a list of upcoming Replication Games.

### What to Expect:

Researchers participating in the Replication Games will join a small team of 3-5 members with similar research interests. Teams may either conduct a robustness replication, which is the ability to duplicate the results of a prior study using the same data but different procedures as were used by the original investigator - or recode the study using the raw or intermediate data. Teams may also





# Classwork, but where to go?

Welcome to the  
**Social Science  
Reproduction Platform**

Accelerating computational  
reproducibility in the social sciences

GET STARTED!

Search reproductions by paper title, DOI, or author. Do a blank search to see all.

**RR  
REPLICATION  
WIKI**

Main page Discussion

Welcome to  
**Replication in the social sciences, particularly economics!**

The Wiki has undergone a technical relaunch.  
Users who log in will be asked to change their password  
from the old version.

17,168 articles

Search:

Find one of 4,072 studies to replicate	Browse in 275 journals, 38 working paper series & blogs and 29 books	Browse by methods, software used, JEL codes, keywords, authors, data type, geographical origin of data, and data source	Find one of 794 replications, 27 corrections & 16 retractions
---	--	--	--

Each study has its own page  
with bibliographic  
information

3 of 11

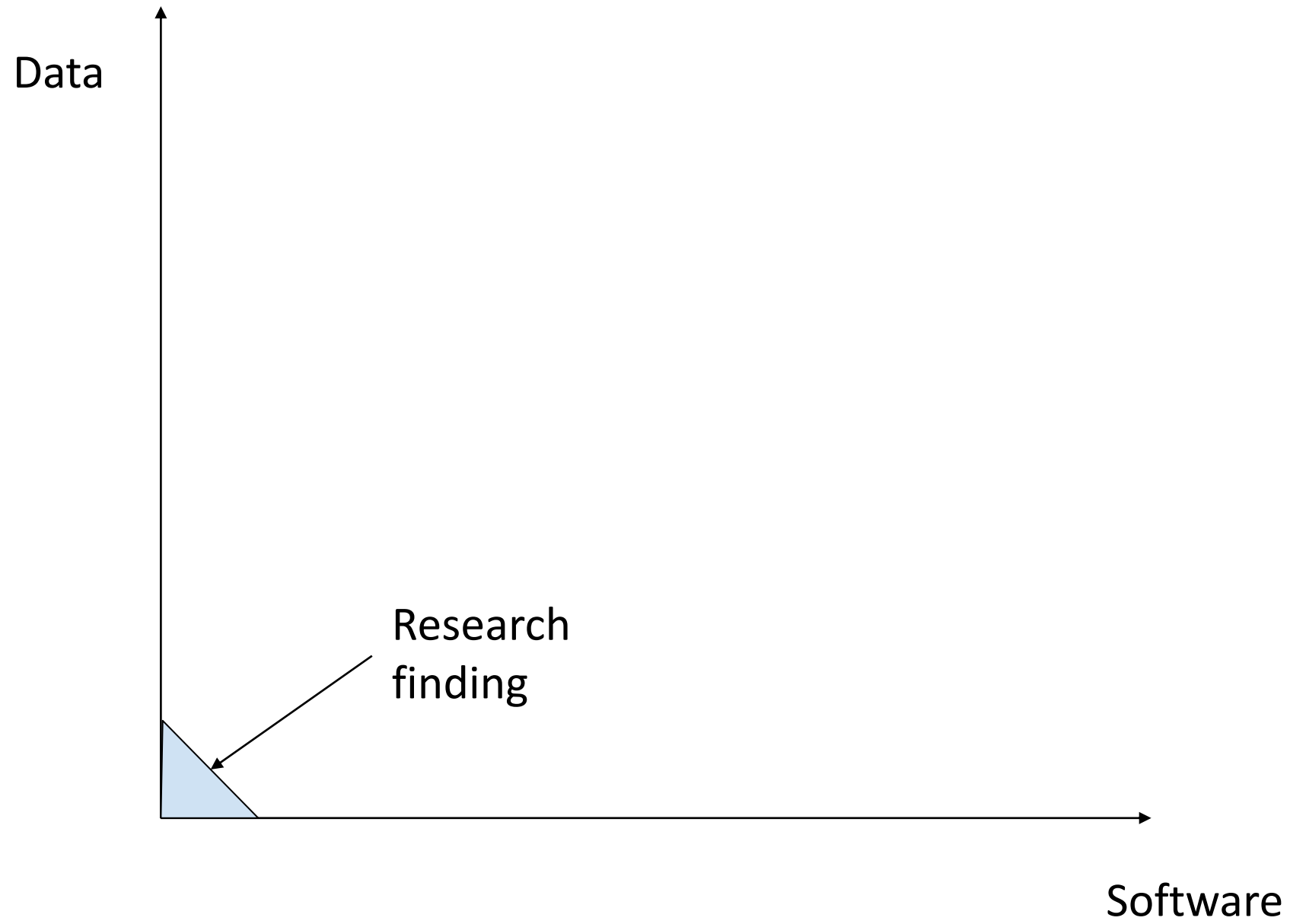
**Join the proje**  
You can join the proj  
find **most relevant** to  
announce news and

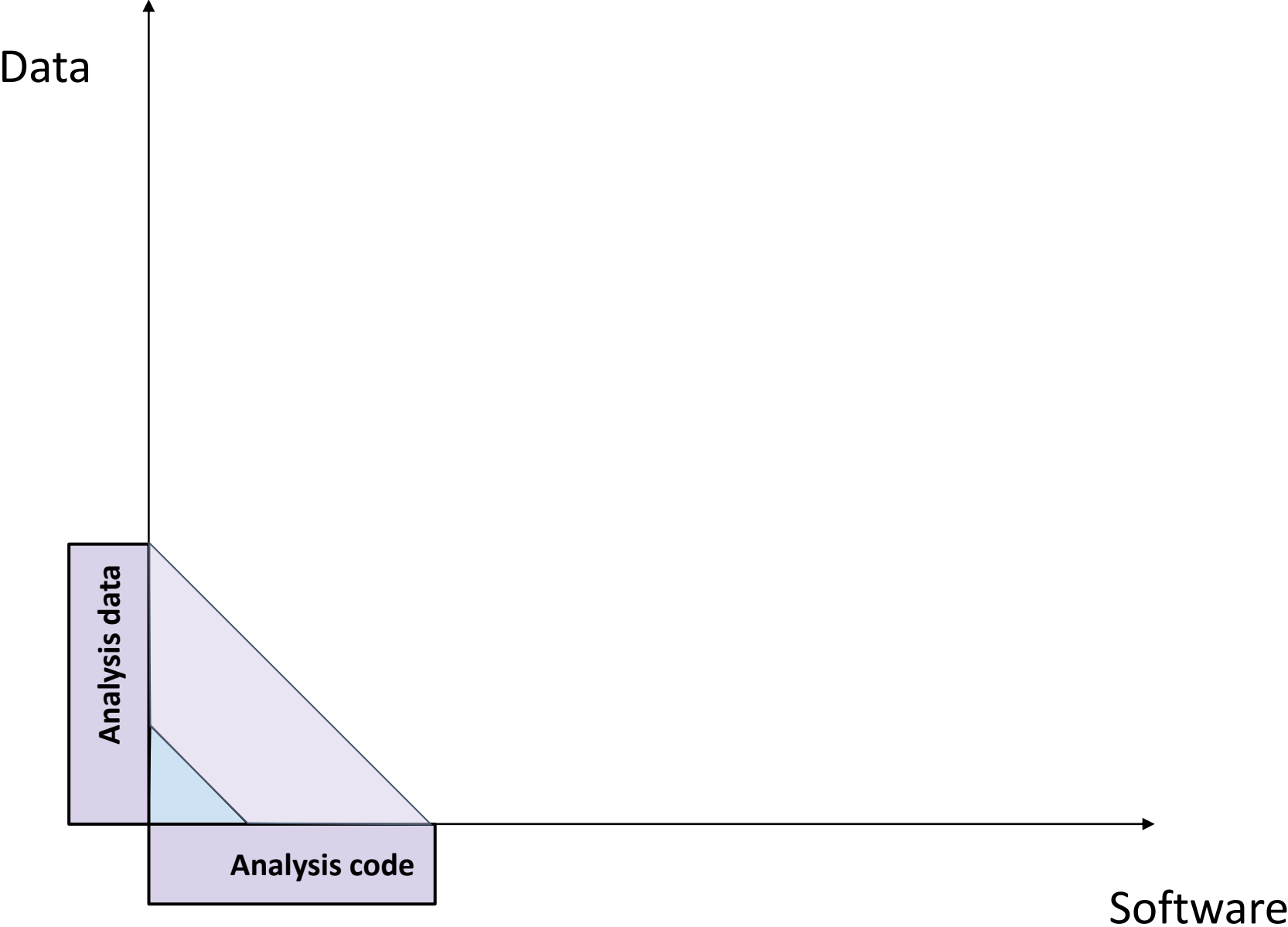
**Most recently**

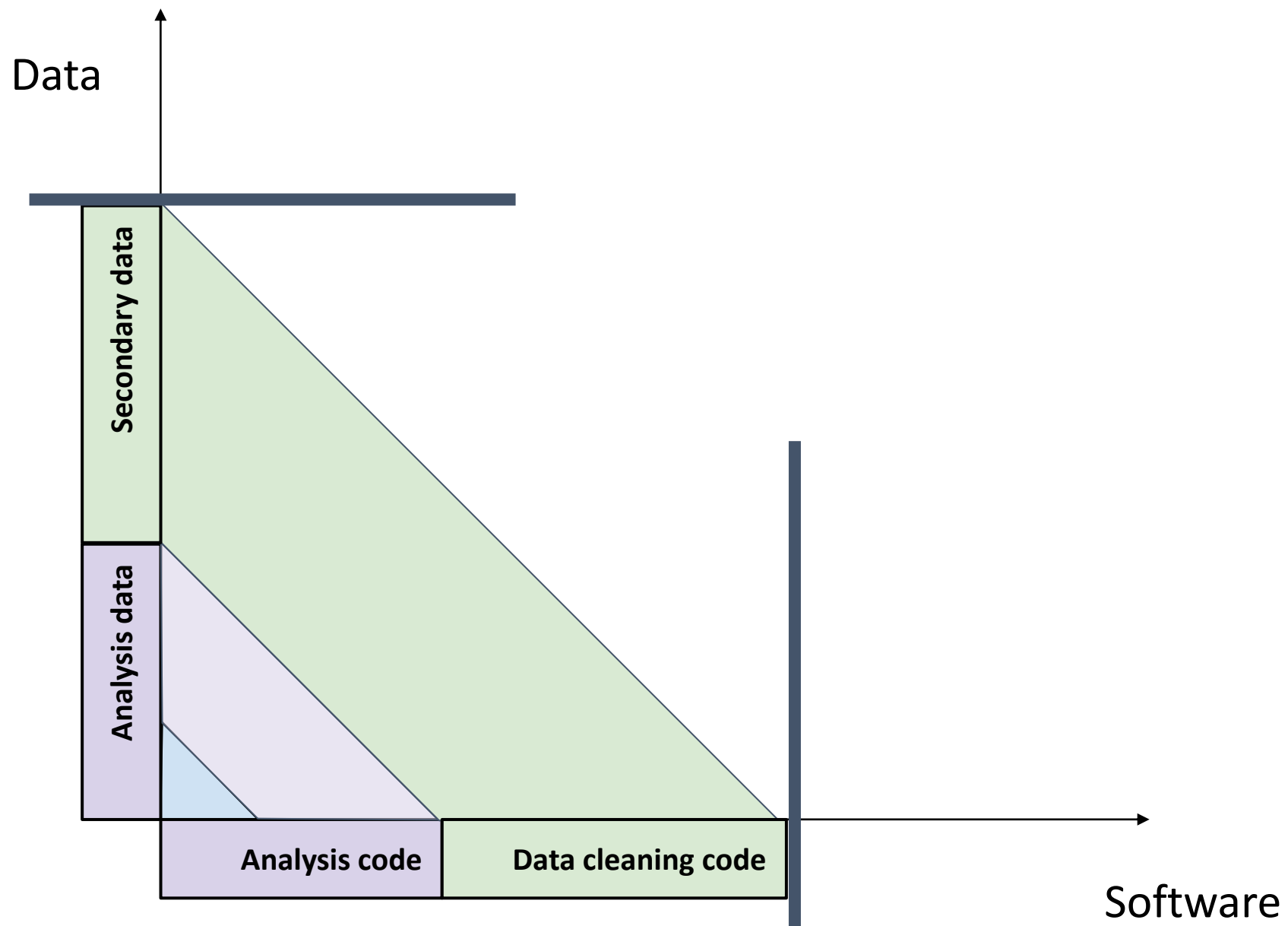
- 3 February 2023
- 25 January 2023  
Economic Persp
- 17 January 2023  
(2013) (Emp Ecc
- 19 October 2022
- 10 October 2022

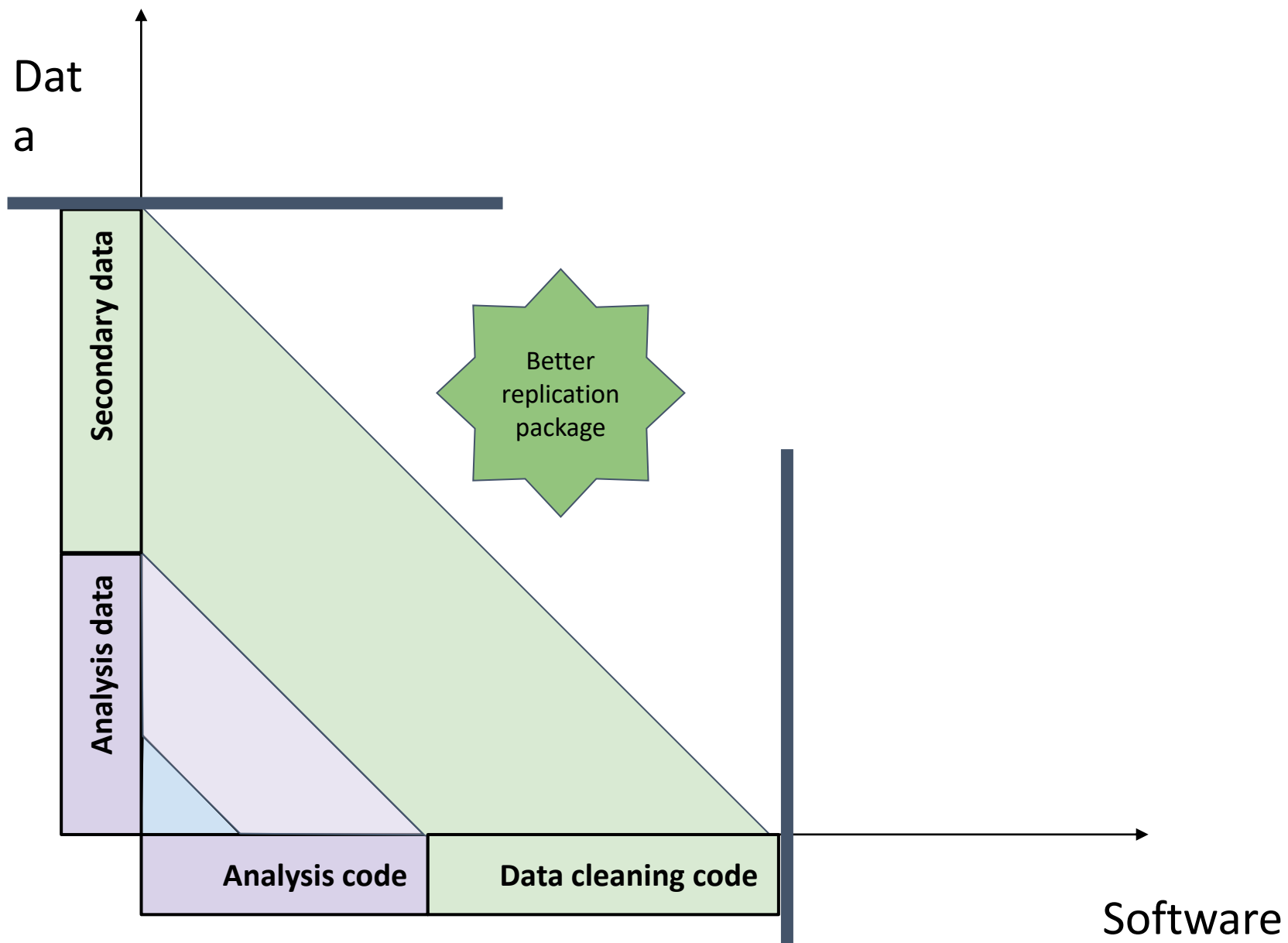
**News**  
Webinar: Replicating

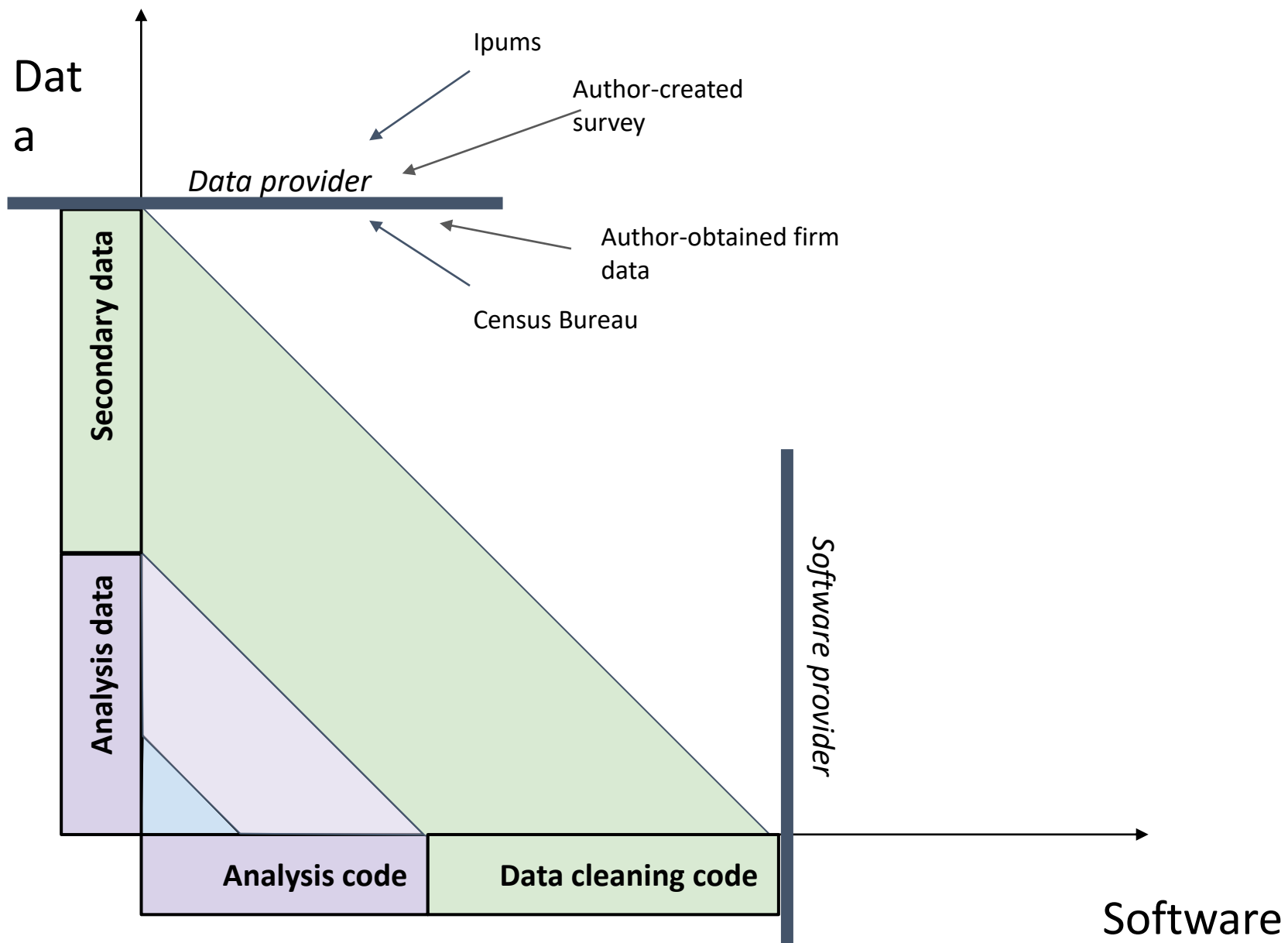
Pushing the  
boundaries

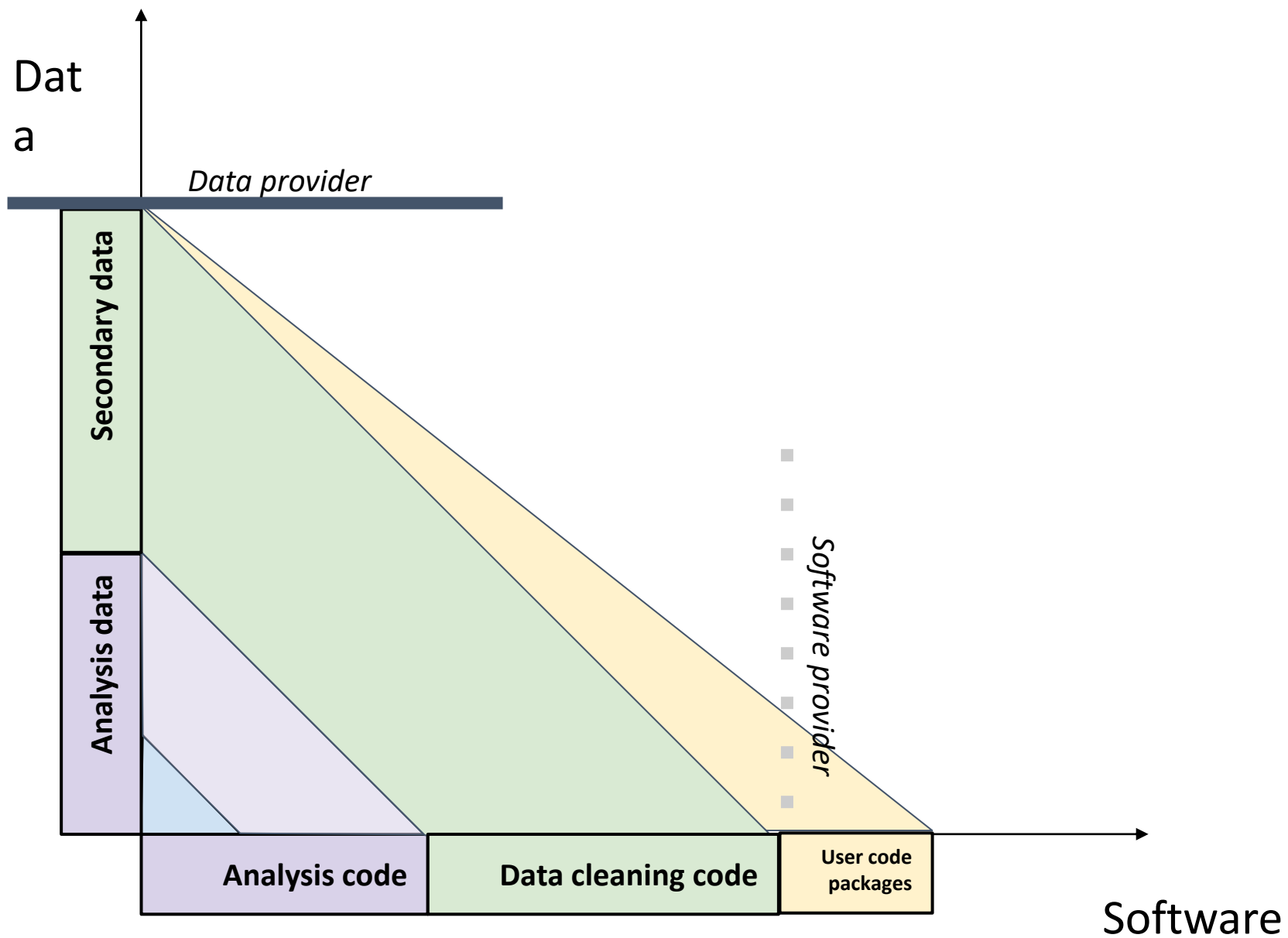




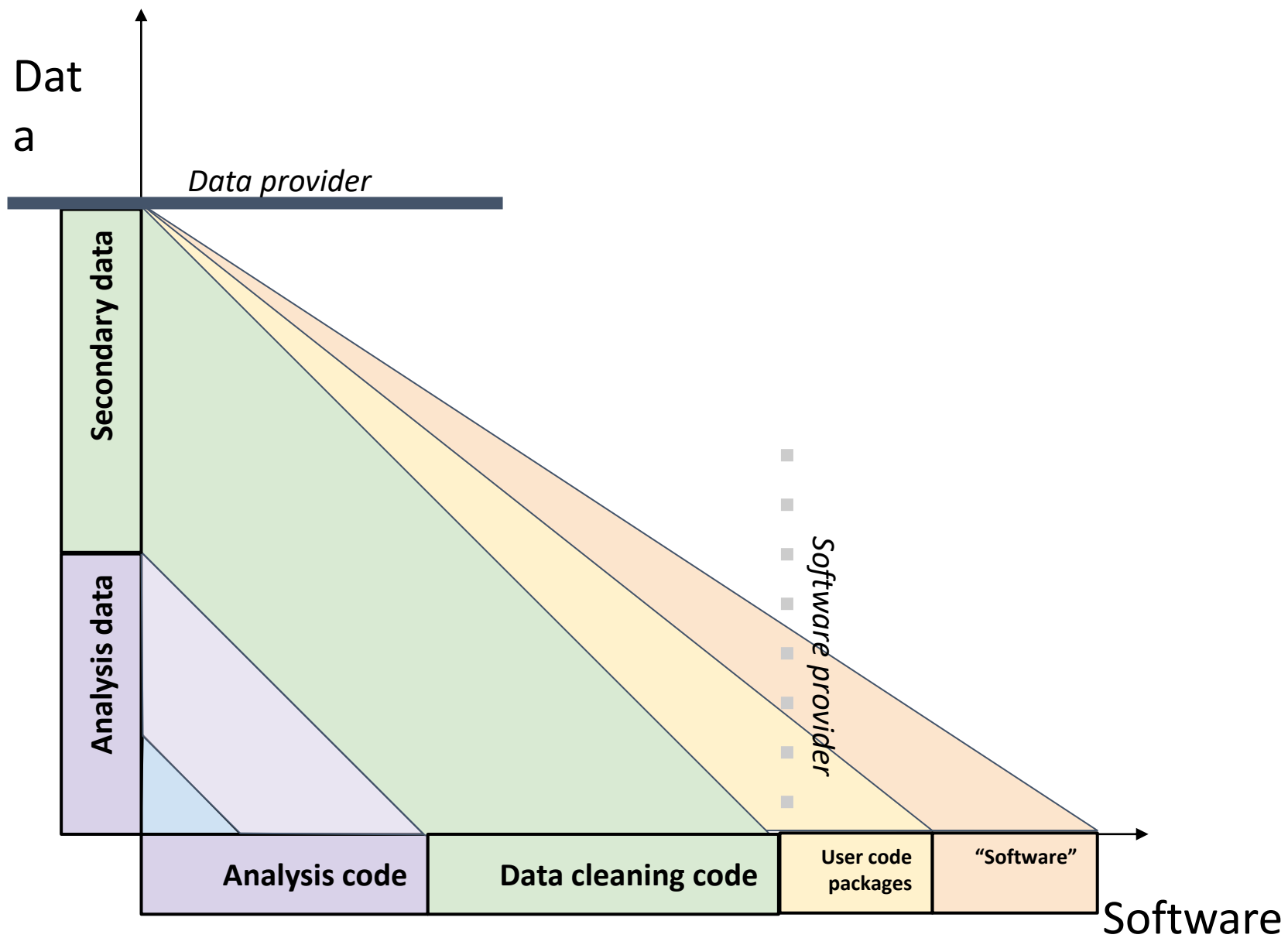


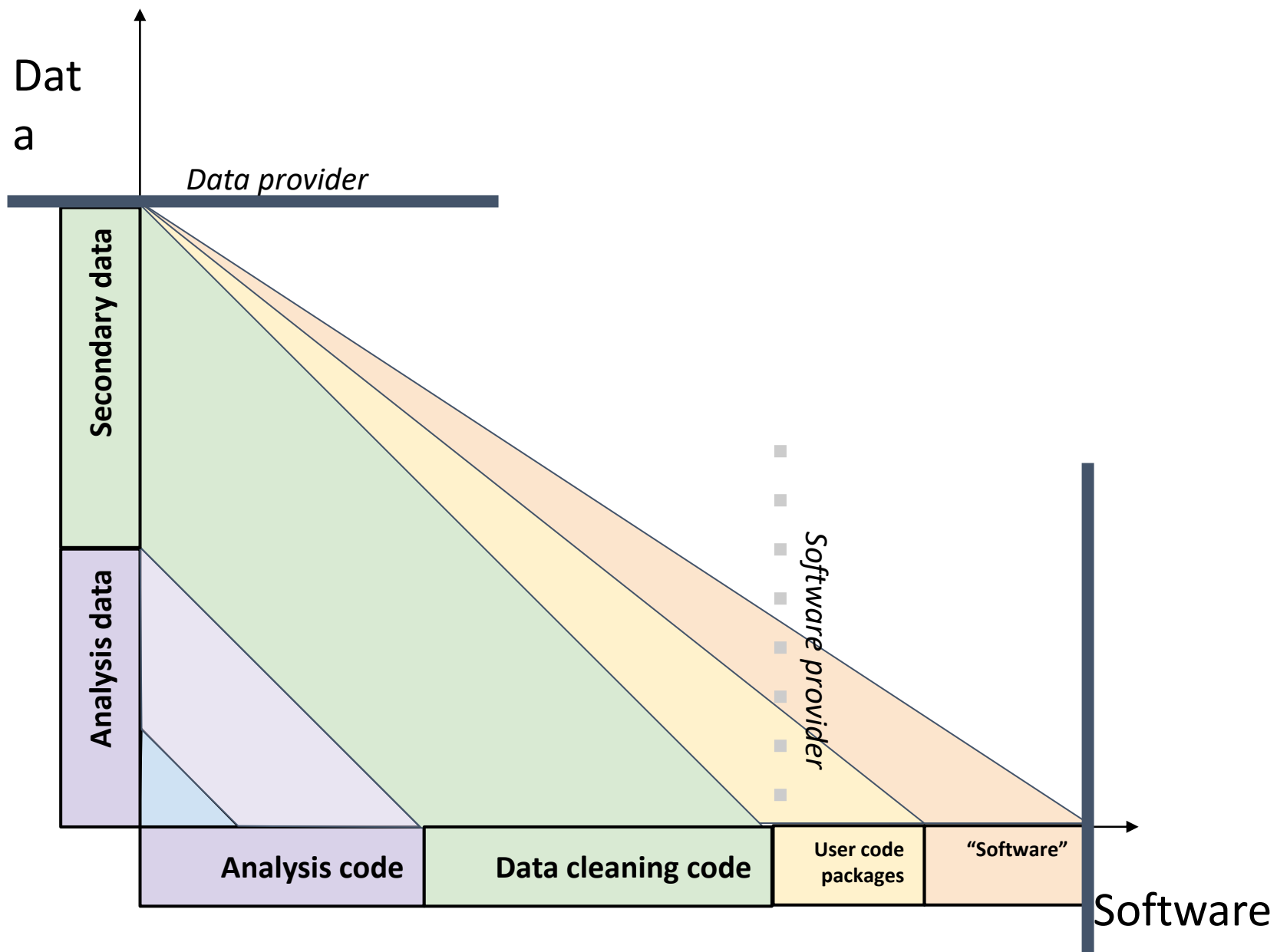


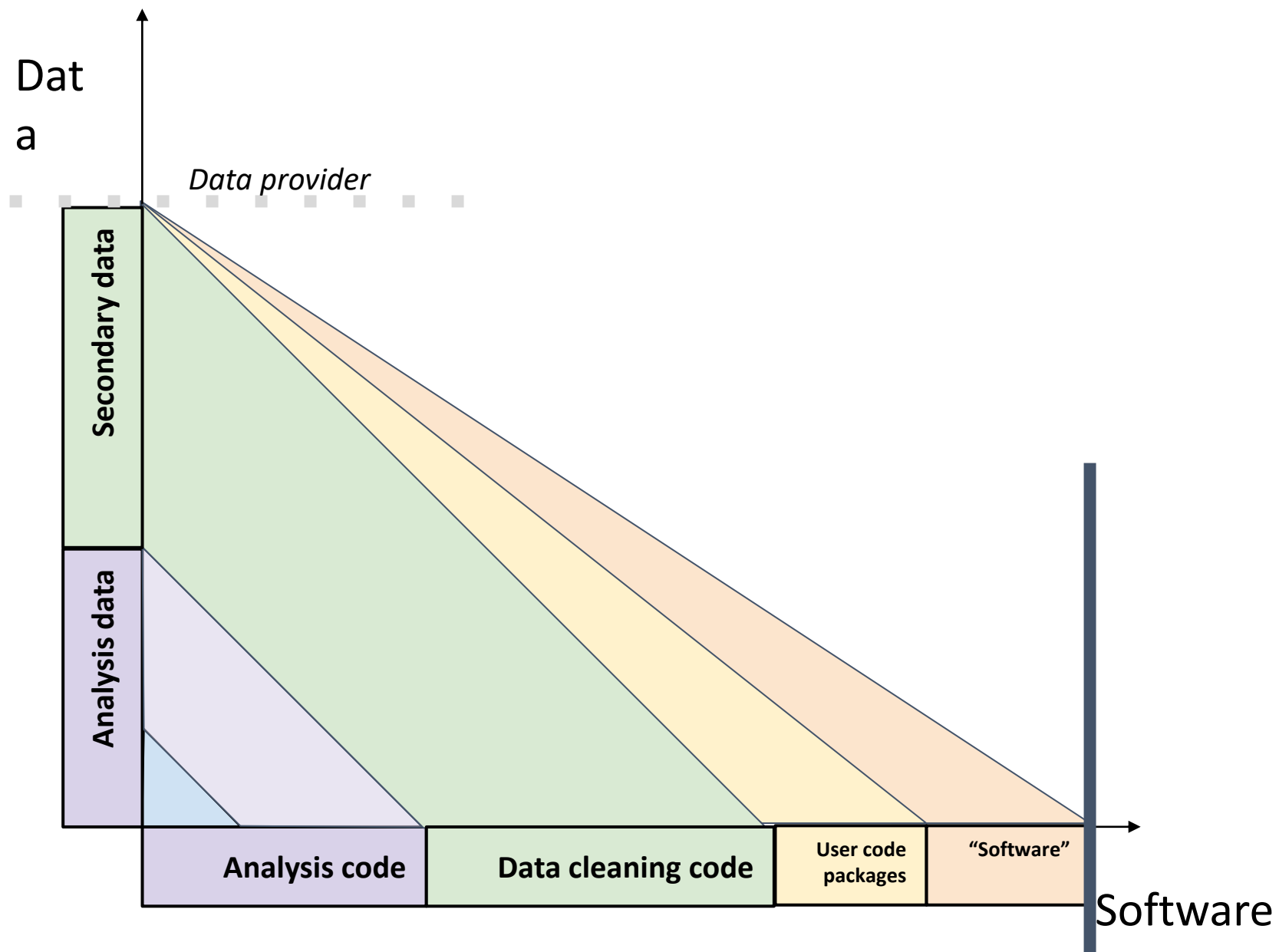


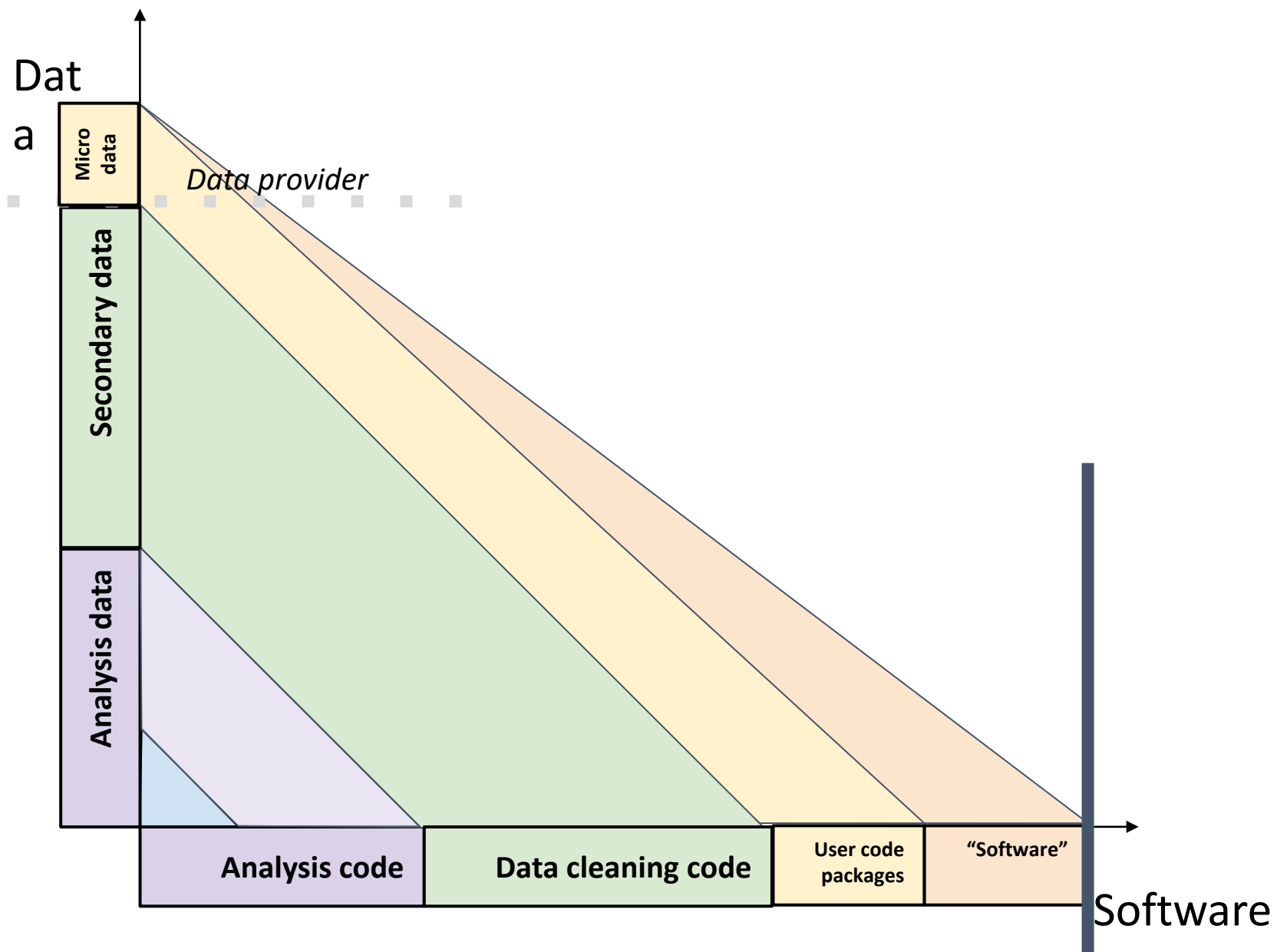


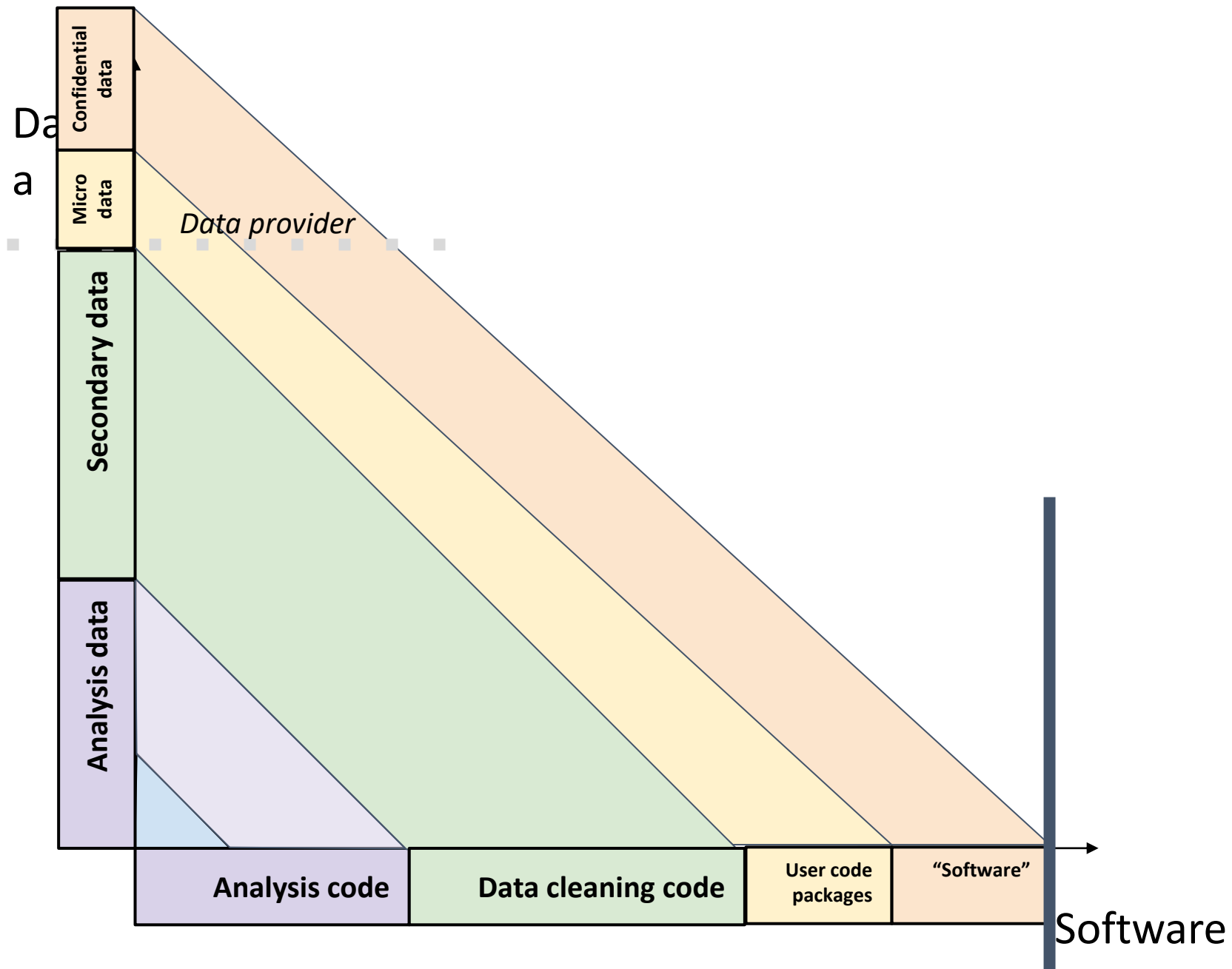


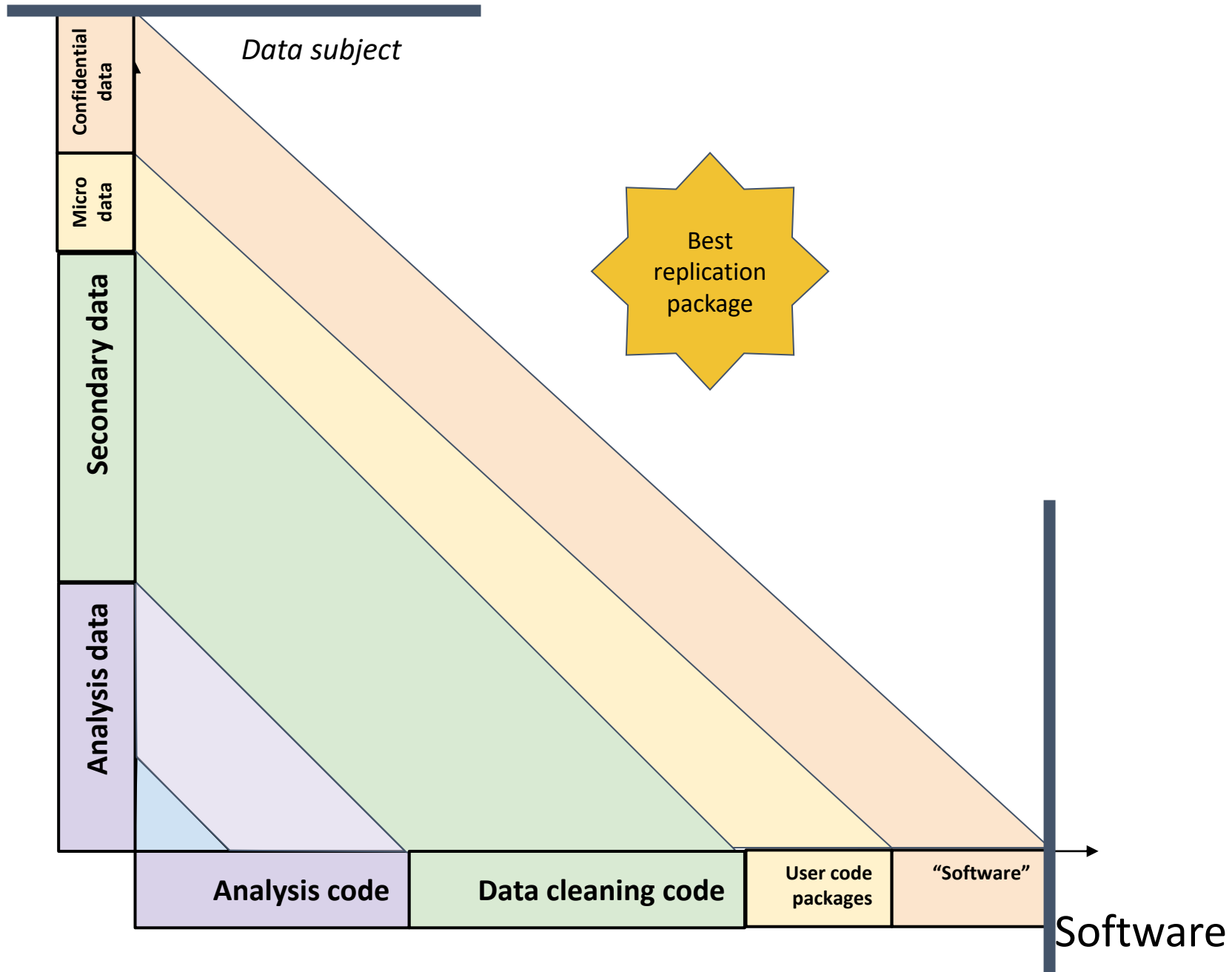


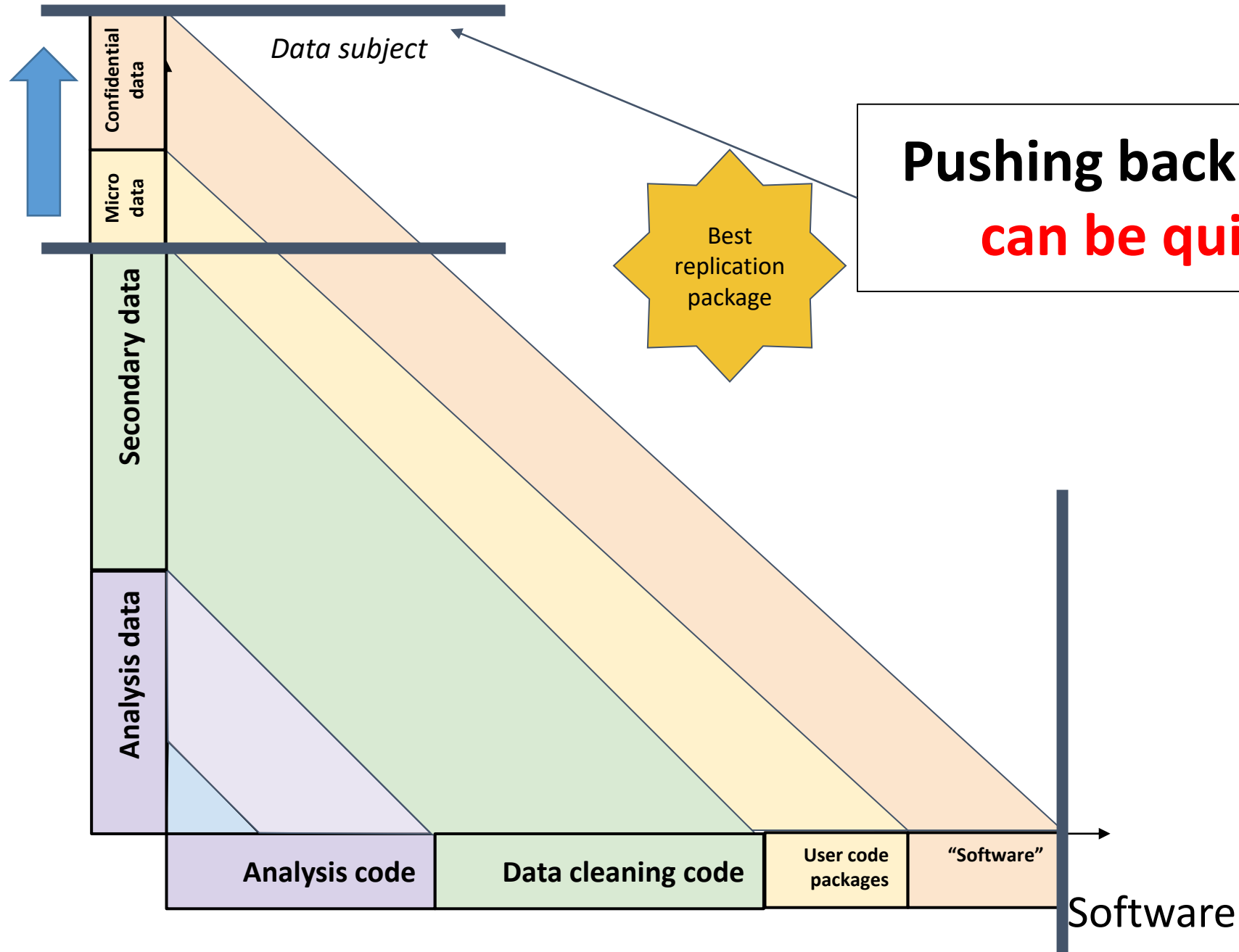












# Conclusion





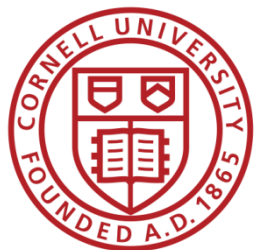
# Transparency and reproducibility...

- ... are critical foundations for future research
- ... do not end with the article publication
- ... need testing
- ... need dissemination platforms that allow for multiple versions



## Goal: Transportability

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



Social science “guild”



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



# Some resources

- <https://social-science-data-editors.github.io/guidance/>
  - template README
  - discussion of licensing
  - data citation guidance
- <https://aeadataeditor.github.io/>



Office of the AEA Data Editor

[START REPLICATION PACKAGE](#)

[FAQ](#)

[Blog Posts](#)

[Talks](#)

[Projects](#)

[Surveys](#)

[Publications](#)

The following steps outline what you should expect after conditional acceptance of your manuscript, in compliance with the [AEA Data and Code Availability Policy](#):

# 1

## Prepare

Prepare your data and code replication package (including data citations and provenance information). You can do this at any time, even before submitting to the AEA journals.

[Start](#)

# 2

## Upload

Provide metadata and upload the replication package. This step simultaneously prepares the materials for the verification process as well as for subsequent publication.

[Do it!](#)

# 3

## Submit

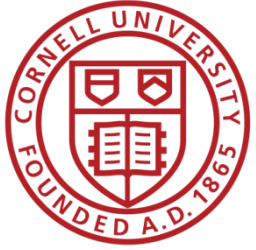
Submit the [Data and Code Availability Form](#) together with your manuscript native files as instructed, and as per guidelines at your journal (for example, [AER guidelines](#)). Only once these materials have been received by the editorial office are [verification checks started](#).

[Ready to submit?](#)

**Thank you!**



**Extra slides**



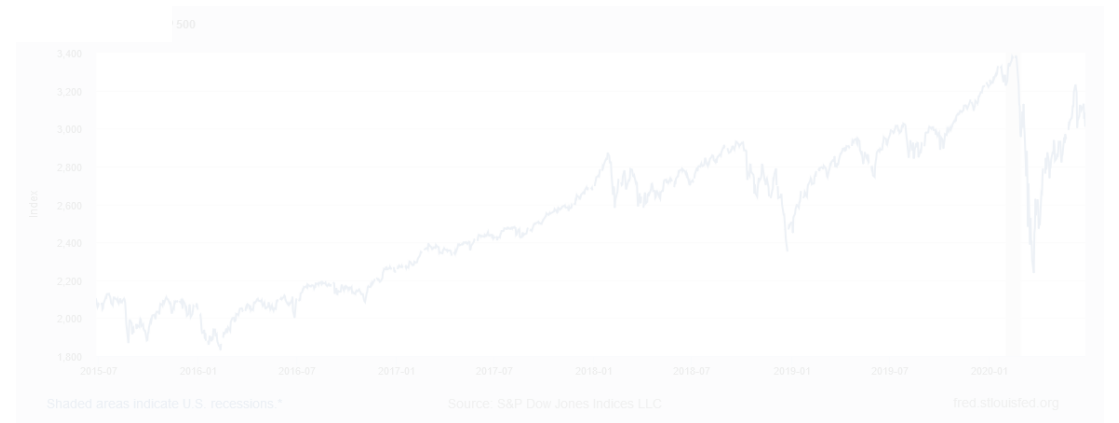
# Example of data provenance





# “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





# “It’s a file called SP500.xlsx”

SP500

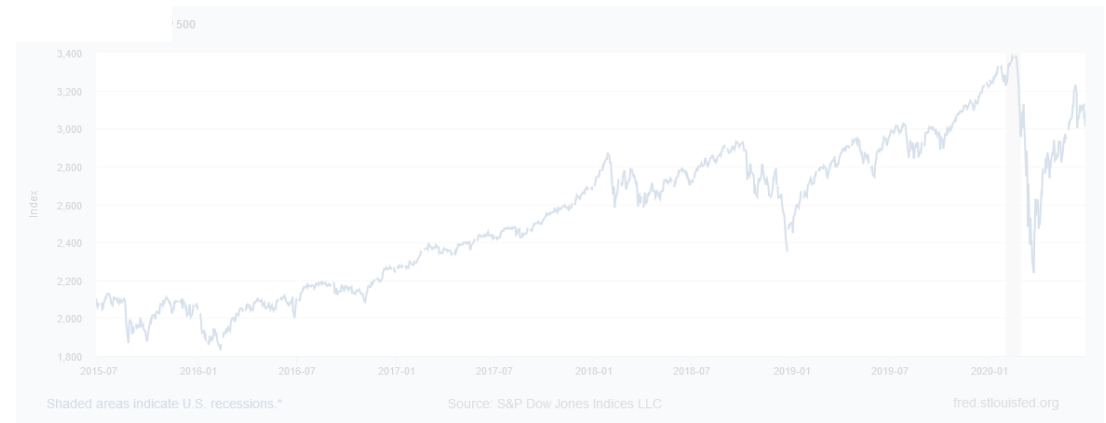
S&P 500, Index, Daily,  
Not Seasonally Adjusted

Frequency: Daily, Close

observation\_date

SP500

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





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

SP500

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

Frequency: Daily, Close

observation\_date

SP500

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





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

SP500	S&P 500, Index, Daily, Not Seasonally Adjusted
Frequency: Daily, Close	
observation_date	SP500
2015-06-26	2101.49
2015-06-29	2057.64
2015-06-30	2063.11
2015-07-01	2077.42
2015-07-02	2076.78
2015-07-03	0
2015-07-06	2068.76
2015-07-07	2081.34
2015-07-08	2046.68
2015-07-09	2051.31
2015-07-10	2076.62
2015-07-13	2099.60
2015-07-14	2108.95
2015-07-15	2107.40
2015-07-16	2124.29
2015-07-17	2126.64
2015-07-20	2128.28

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





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

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

Frequency: Daily, Close  
observation\_date

SP500

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

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





# Data Availability Statements



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

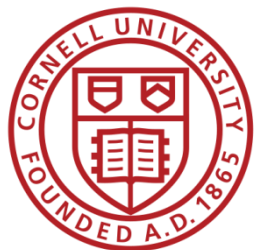
Describes data file, where to get it, how to get it, and any conditions of obtaining it

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

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



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



# Data Citations



# Data citations

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



## Data and Code Guidance by Data Editors

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

### Guidance on Data Citations

On this page:

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

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

### What is not a data citation

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

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





# Example 4: German Restricted-access



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

[Home](#) | [Newsletter](#) | [Jobs](#) | [Contact](#) | [Data Privacy](#) | [Imprint](#)



Data Version	DOI (Link to Description of Data Version)	Availability (yyyy-mm-dd)
<b>BHP 7518 v1 (current)</b>	<a href="#">10.5164/IAB.BHP7518.de.en.v1</a>	2020-01-13
<b>BHP 7517 v1</b>	<a href="#">10.5164/IAB.BHP7517.de.en.v1</a>	2018-12-12
<b>BHP 7516 v1</b>	<a href="#">10.5164/IAB.BHP7516.de.en.v1</a>	2018-04-11

## External data

[Data Archive](#)

[Data Access](#)

[Campus Files](#)

[Publications](#)

[Events](#)

[Projects of FDZ users](#)

[FDZ Projects](#)

[Complaint point of the  
RatSWD](#)

[Figures of the FDZ](#)

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

## Data Versions

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

Data Version	DOI (Link to Description of Data Version)	Availability (yyyy-mm-dd)
--------------	---	---------------------------

**BHP 7518 v1 (current)**

[10.5164/IAB.BHP7518.de.en.v1](#)

2020-01-13



# Data Citation



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

Attributes the file to the proper source

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

2015-07-08	2101.49
2015-07-09	2057.64
2015-07-10	2063.11
2015-07-13	2075.12
2015-07-14	2076.78
2015-07-15	2076.78
2015-07-16	2068.76
2015-07-17	2081.34
2015-07-20	2046.68
	2051.31
	2076.62
	2099.60
	2108.95
	2107.40
	2124.29
	2126.64
	2128.28

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





# Element of a (data) citation

ICPSR notes that a citation should include the following items:

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

## Suggested Citation:

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



# Element of a (data) citation

ICPSR notes that a citation should include the following items:

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

## **Constructed Citation:**

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



# Element of a (data) citation

ICPSR notes that a citation should include the following items:

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

## Constructed Citation:

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



# Try it out yourself

- Construct an (approximate) data citation
- <https://social-science-data-editors.github.io/guidance/addtl-data-citation-guidance.html#try-it-out>

## 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. 2022. "Guidance on Data Citations". *Data and Code Guidance by Data Editors*. Accessed at <https://social-science-data-editors.github.io/guidance/addtl-data-citation-guidance.html> on 2022-06-30.

Contributors: Lars Vilhuber

This project is maintained by [social-science-data-editors](#)

Disclaimer

In some cases, the data provider (often a firm) must remain anonymous. This does not prevent citation, and the provider should be mentioned in much the same way as when there is no formal access mechanism:

Anonymous Firm. 1999. "Personnel records of windowshield installers." Unpublished data. Accessed February 29, 2000.

### Try it out

Authors or Producer:	<input type="text" value="Author"/>
Title:	<input type="text" value="Title"/>
Date of publication:	<input type="text" value="2022"/>
Distributor:	<input type="text" value="Distributor"/>
Version:	<input type="text" value="V1"/>
Persistent identifier or URL:	<input type="text" value="https://doi.org/123/345"/>
Date of access:	<input type="text" value="2022-01-22"/>
Accessed or downloaded?	<input type="radio"/> Accessed <input type="radio"/> Downloaded
<input type="button" value="Compute citation"/>	



The Future of Research Communications and e-Scholarship

Search

English

ABOUT ▾ COMMUNITY ▾ CODE OF CONDUCT GROUPS RESOURCES ▾ NEWS + BLOGS ▾ EVENTS ▾ PUBLIC

perceived criteria of importance.

## 1. Importance

Data should be considered legitimate, citable products of research. Data should be accorded the same importance in the scholarly record as citable research objects, such as publications[1].

*DC<sup>1</sup>*  
Data Citation Principles

## 2. Credit and Attribution

Data citations should facilitate giving scholars credit and recognition and

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

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

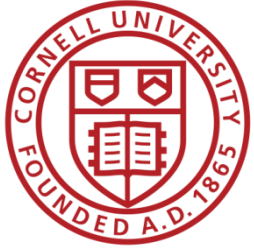
## 4. Unique Identification

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

## 5. Access

Data citations should facilitate access to the data themselves and to such related data, documentation, code, and other materials as are necessary for

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



# Rights to use data

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





## Rights to distribute the data

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

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



# Data: Citations, Access, Rights

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



## Data Availability Statements (DAS)

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

to publicly  
ated during

y providing a

I restrictions,

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



## Data Availa

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



## Solution 3: Data Citations

Cite every data source

(not only the paper that describes the source!)

(also: add them to the  
Social Science Data Editors' template README)