Reproducibilidad con Impacto

Lars Vilhuber

2026-01-23

Follow along

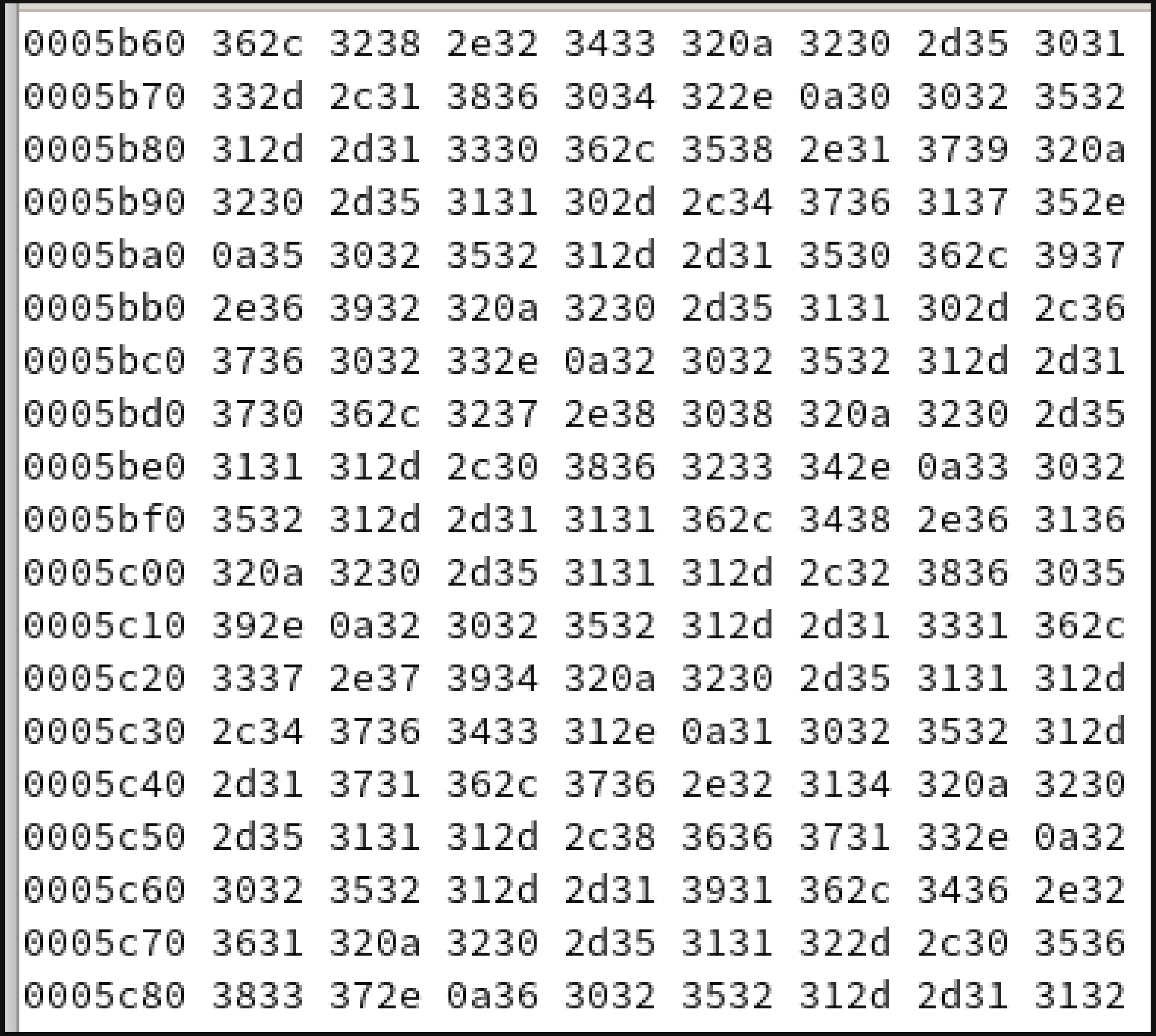
larsvilhuber.github.io/presentation-2026-01/ (PDF)

Reproducibilidad con Impacto

Cómo la Ciencia Abierta Puede Cambiar el Mundo

Lars Vilhuber
Cornell University

What is this?



```
0005b60 362c 3238 2e32 3433 320a 3230 2d35 3031
0005b70 332d 2c31 3836 3034 322e 0a30 3032 3532
0005b80 312d 2d31 3330 362c 3538 2e31 3739 320a
0005b90 3230 2d35 3131 302d 2c34 3736 3137 352e
0005ba0 0a35 3032 3532 312d 2d31 3530 362c 3937
0005bb0 2e36 3932 320a 3230 2d35 3131 302d 2c36
0005bc0 3736 3032 332e 0a32 3032 3532 312d 2d31
0005bd0 3730 362c 3237 2e38 3038 320a 3230 2d35
0005be0 3131 312d 2c30 3836 3233 342e 0a33 3032
0005bf0 3532 312d 2d31 3131 362c 3438 2e36 3136
0005c00 320a 3230 2d35 3131 312d 2c32 3836 3035
0005c10 392e 0a32 3032 3532 312d 2d31 3331 362c
0005c20 3337 2e37 3934 320a 3230 2d35 3131 312d
0005c30 2c34 3736 3433 312e 0a31 3032 3532 312d
0005c40 2d31 3731 362c 3736 2e32 3134 320a 3230
0005c50 2d35 3131 312d 2c38 3636 3731 332e 0a32
0005c60 3032 3532 312d 2d31 3931 362c 3436 2e32
0005c70 3631 320a 3230 2d35 3131 322d 2c30 3536
0005c80 3833 372e 0a36 3032 3532 312d 2d31 3132
```

If you I tell you that…

If I give you a name...

`...`

`SP500.csv`

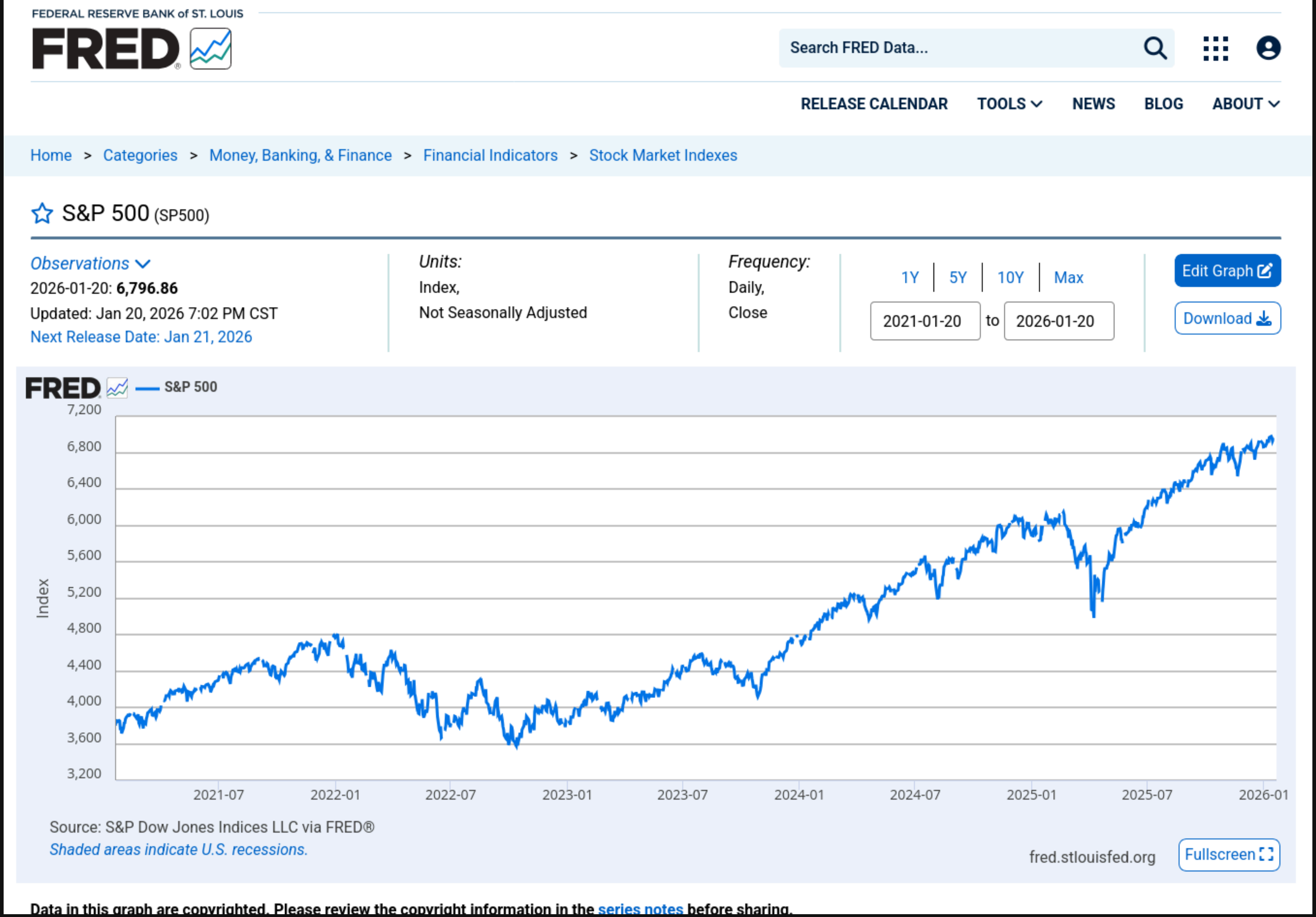
`...`

If I show you its contents...

```r
# read in the data
sp500 <- read.csv(here::here("presentation","SP500.csv"))
head(sp500)
```

```
  observation_date   SP500
1       2021-01-20 3851.85
2       2021-01-21 3853.07
3       2021-01-22 3841.47
4       2021-01-25 3855.36
5       2021-01-26 3849.62
6       2021-01-27 3750.77
```

If I tell you where I got it from...



Home > Categories > Money, Banking, & Finance > Financial Indicators > Stock Market Indexes

S&P 500 (SP500)

Observations
2026-01-20: **6,796.86**
Updated: Jan 20, 2026 7:02 PM CST
Next Release Date: Jan 21, 2026

Units:
Index,
Not Seasonally Adjusted

Frequency:
Daily,
Close

2021-01-20 to 2026-01-20

Source: S&P Dow Jones Indices LLC via FRED®
Shaded areas indicate U.S. recessions.

fred.stlouisfed.org

Data in this graph are copyrighted. Please review the copyright information in the series notes before sharing.

But I cannot give it to you!

> Copyright © 2016, S&P Dow Jones Indices LLC. All rights reserved. Reproduction of S&P 500 in any form is **prohibited** except with the prior written permission of S&P Dow Jones Indices LLC (“S&P”).

So how can you verify my results?

- By obtaining the file again
- By running my code again
- By verifying that the results are the same!

Trust but verify!

- Reproducibility is key!
- But there are weaknesses…

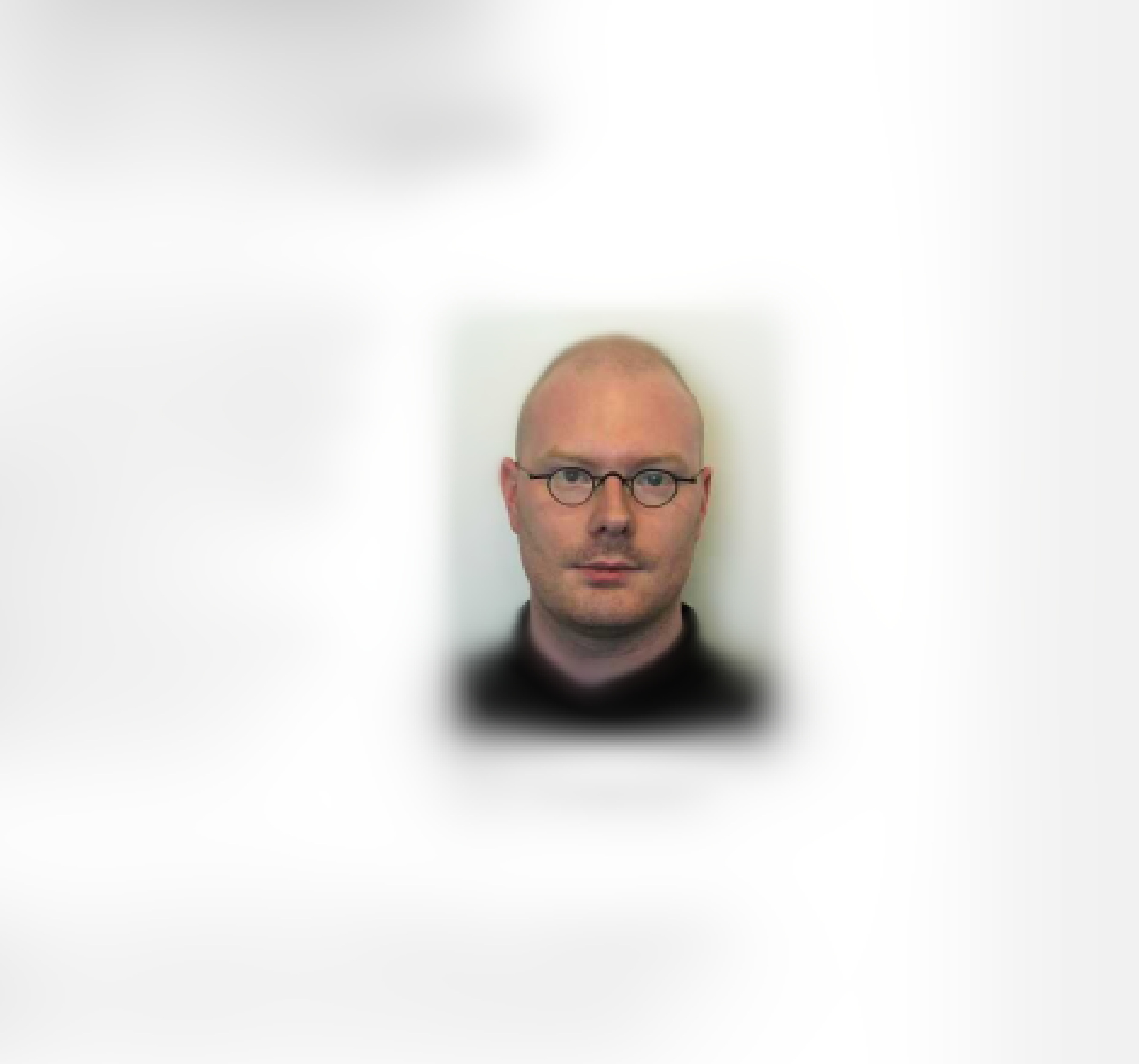
The issue of provenance

SP500

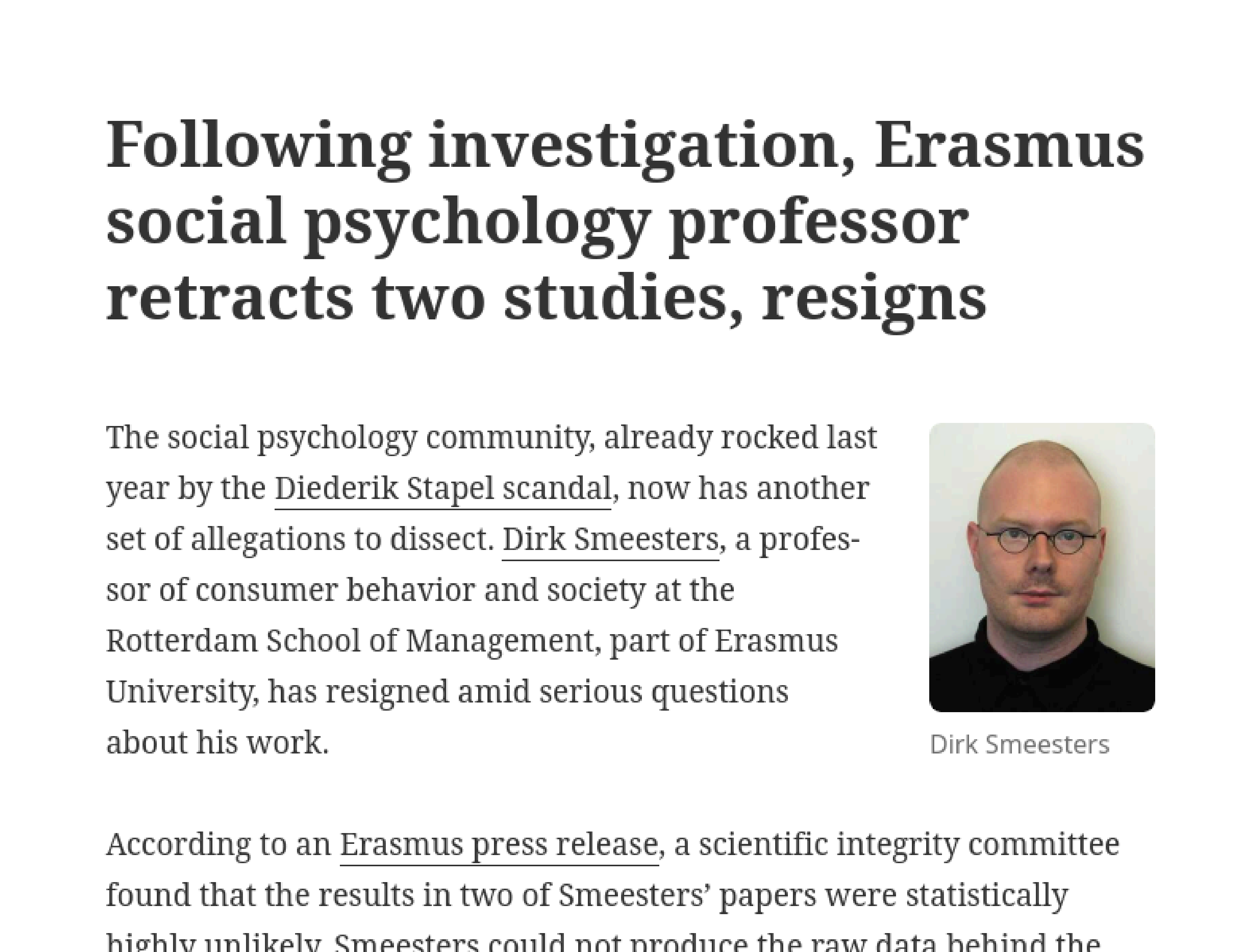
- Created by S&P Dow Jones Indices LLC
- Widely used as a market benchmark
- Commercial product, cannot be redistributed
- S&P considered **reliable**
- **Me** getting the file is the ***weak link***

What if I am the collector of data?

Who is this person?



Dirk Smeesters



Following investigation, Erasmus social psychology professor retracts two studies, resigns

The social psychology community, already rocked last year by the Diederik Stapel scandal, now has another set of allegations to dissect. Dirk Smeesters, a professor of consumer behavior and society at the Rotterdam School of Management, part of Erasmus University, has resigned amid serious questions about his work.

Dirk Smeesters

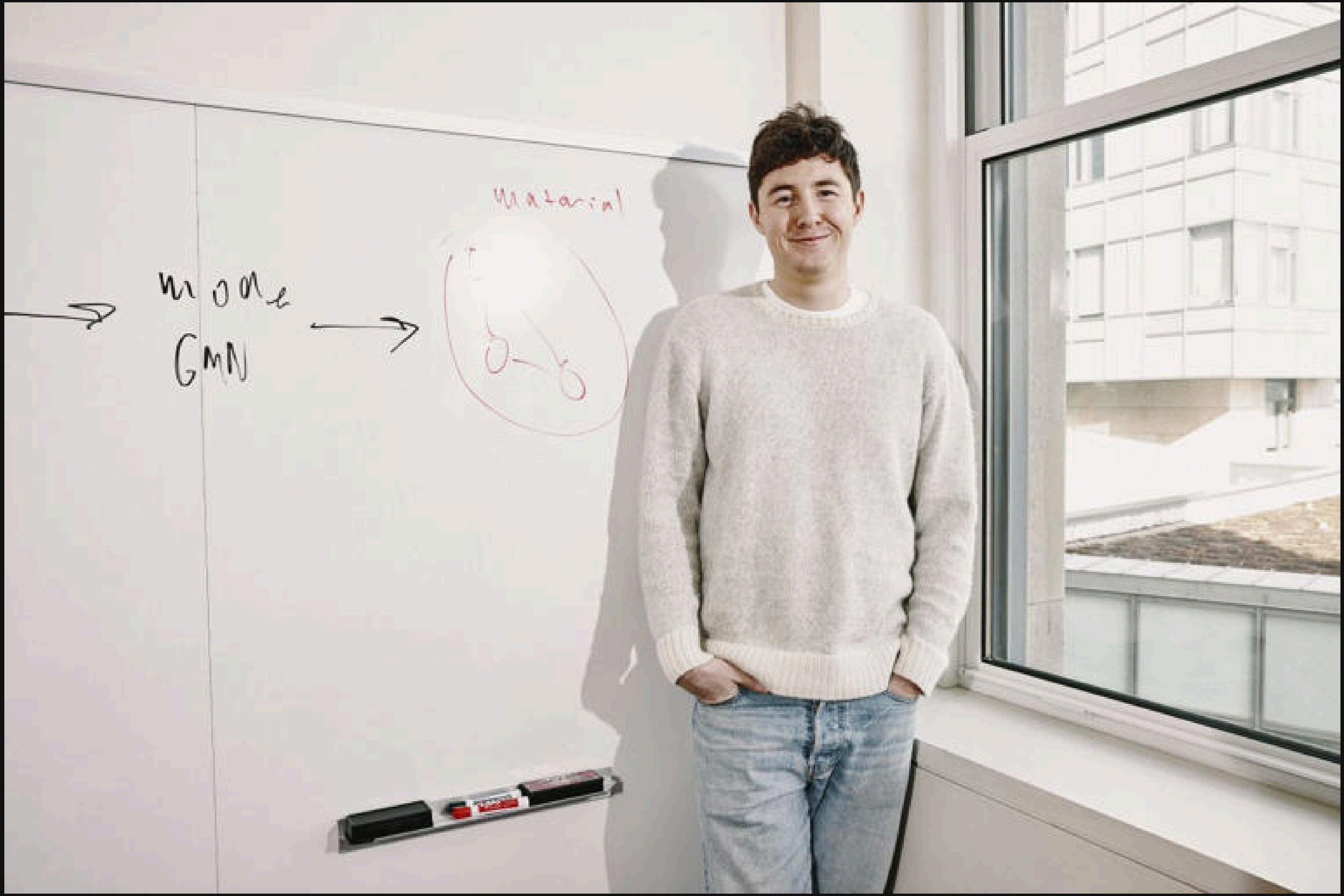
According to an Erasmus press release, a scientific integrity committee found that the results in two of Smeesters' papers were statistically highly unlikely. Smeesters could not produce the raw data behind the

Retraction

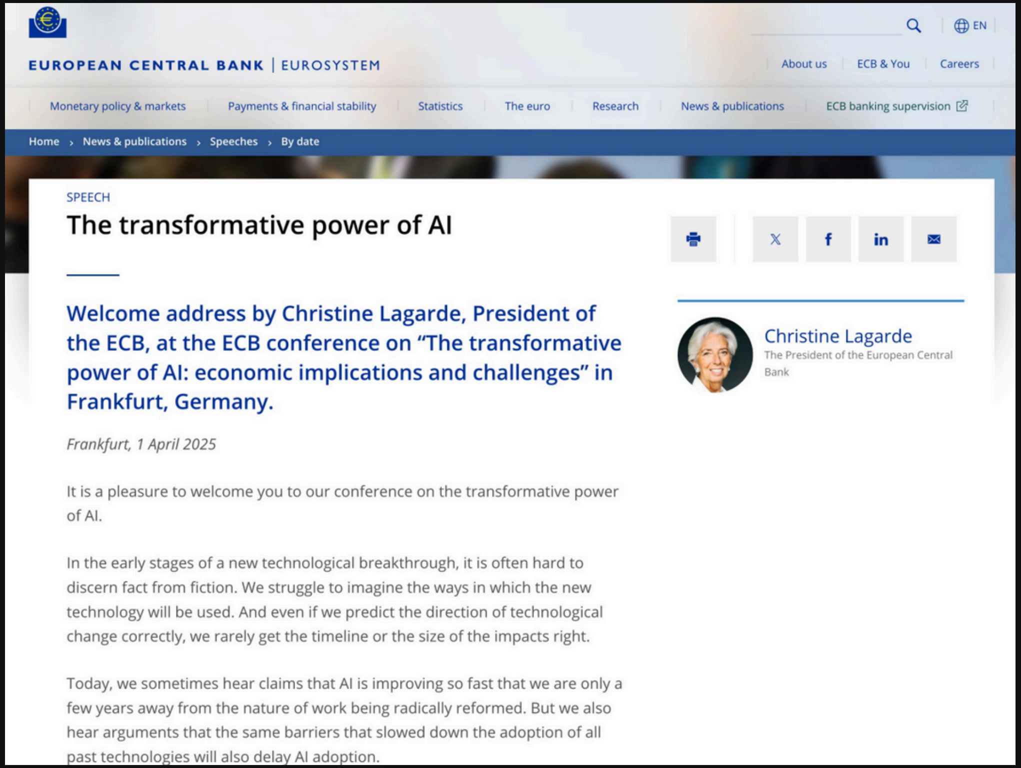
“a scientific integrity committee found that the results in two of Smeesters’ papers were **statistically highly unlikely**. Smeesters ***could not produce the raw data*** behind the findings, and told the committee that he **cherry-picked** the data to produce a statistically significant result. Those two papers are being retracted, and the university accepted Smeesters’ ***resignation*** on June 21.”

Who is this person? (2)

Aidan Toner-Rodgers



Maybe you've heard about him



EUROPEAN CENTRAL BANK | EUROSYSTEM
EN
About us
ECB & You
Careers
Monetary policy & markets
Payments & financial stability
Statistics
The euro
Research
News & publications
ECB banking supervision
Home › News & publications › Speeches › By date

SPEECH

The transformative power of AI

Welcome address by Christine Lagarde, President of the ECB, at the ECB conference on "The transformative power of AI: economic implications and challenges" in Frankfurt, Germany.

Christine Lagarde
The President of the European Central Bank

Frankfurt, 1 April 2025

It is a pleasure to welcome you to our conference on the transformative power of AI.

In the early stages of a new technological breakthrough, it is often hard to discern fact from fiction. We struggle to imagine the ways in which the new technology will be used. And even if we predict the direction of technological change correctly, we rarely get the timeline or the size of the impacts right.

Today, we sometimes hear claims that AI is improving so fast that we are only a few years away from the nature of work being radically reformed. But we also hear arguments that the same barriers that slowed down the adoption of all past technologies will also delay AI adoption.

ECB speech

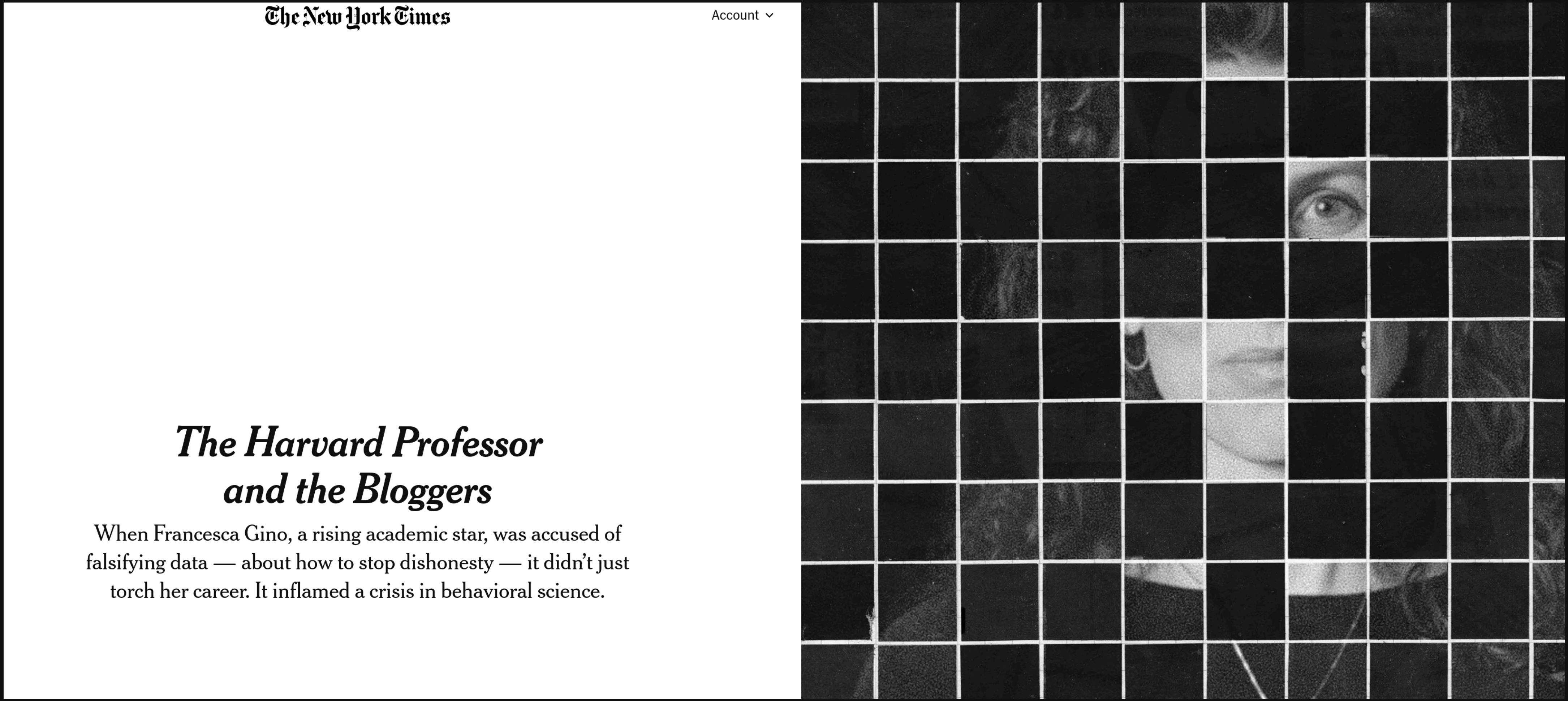
"AI-assisted researchers discover 44% more materials, resulting in a 39% increase in patent filings."

Now

“MIT now declares “**no confidence in the provenance**, reliability or validity of the data and…in the veracity of the research”. Mr Toner-Rodgers’s paper has been ***withdrawn*** from the pre-print repository on which it first appeared [arXiv]; … The lab at the heart of his findings remains **unknown**.”

How can we know that a data source is reliably obtained?

Consider the case of Gino



The New York Times

The Harvard Professor and the Bloggers

When Francesca Gino, a rising academic star, was accused of falsifying data — about how to stop dishonesty — it didn't just torch her career. It inflamed a crisis in behavioral science.

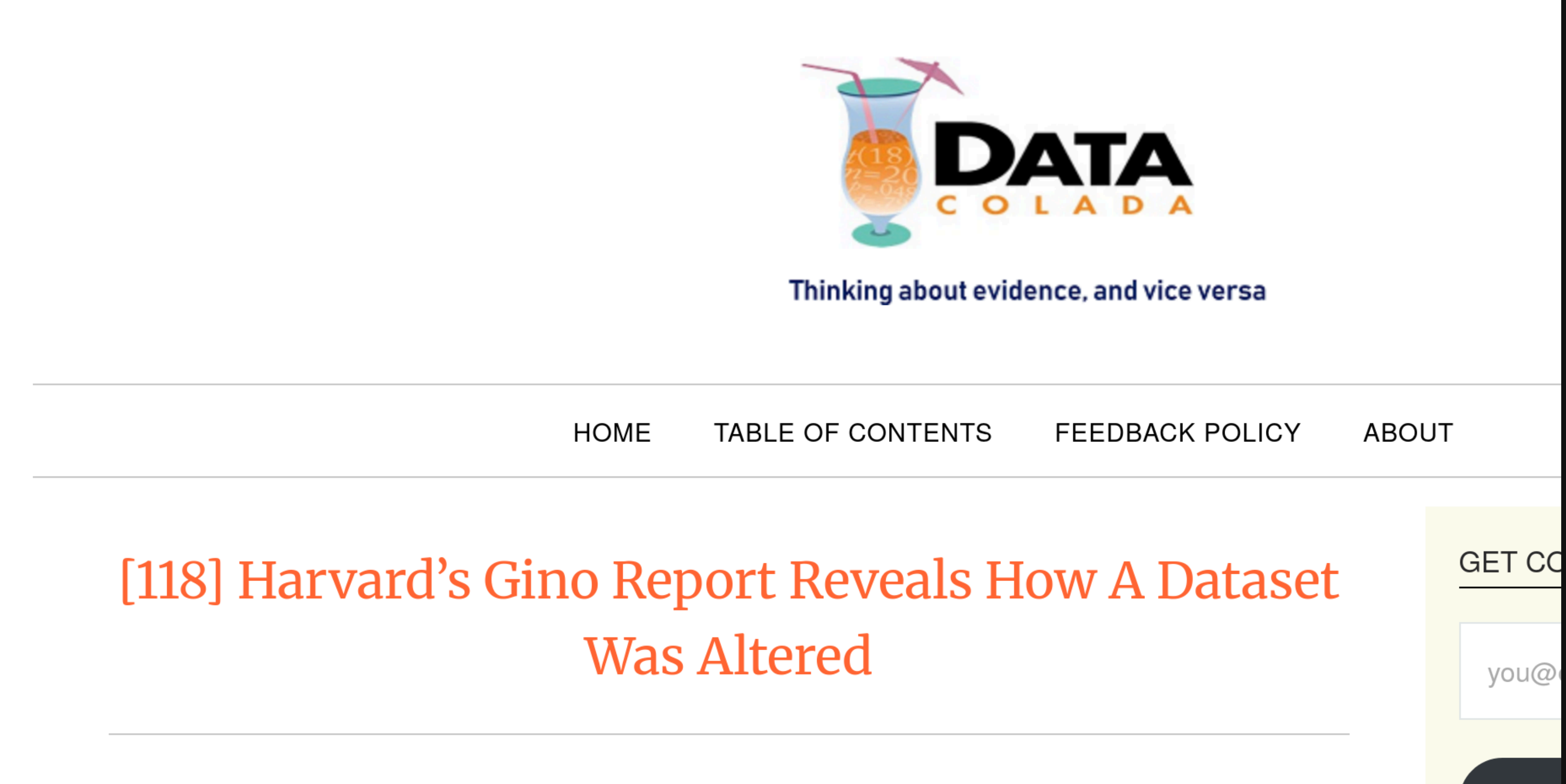
Francesca Gino

The case of Gino

- Francesca Gino was a ***tenured*** professor at Harvard Business School, writing on **honesty** (!)

The case of Gino

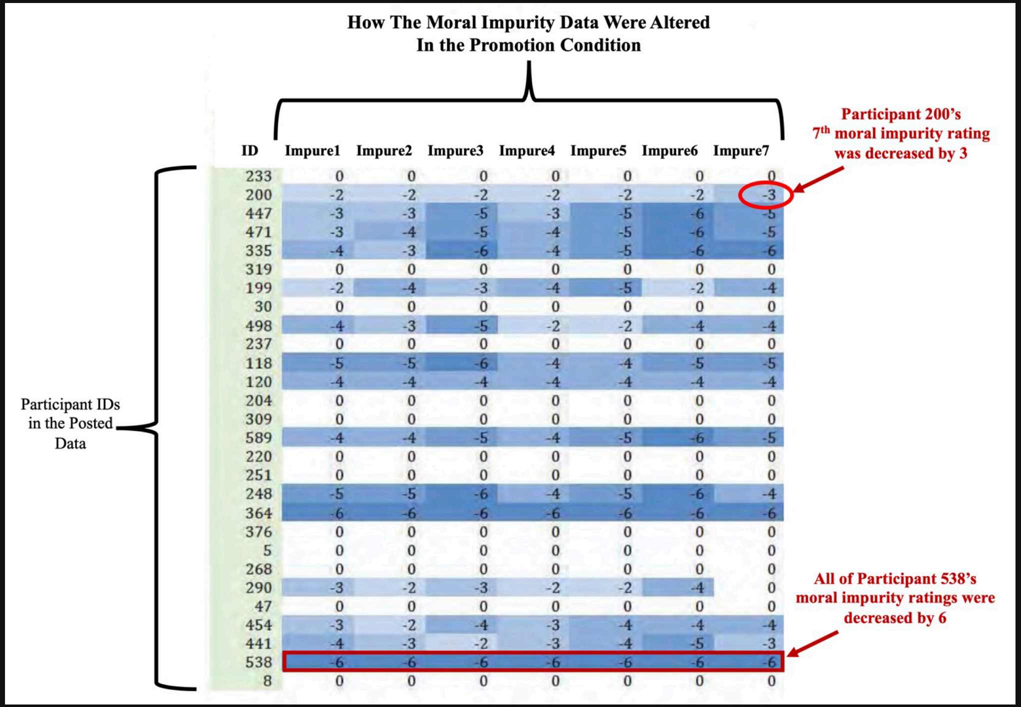
- Several articles were investigated by third parties (***Data Colada***, in particular [1]), and found to be **problematic**



[118] Harvard's Gino Report Reveals How A Dataset Was Altered

The case of Gino

- At least one of them had manipulated data **AFTER** it had been collected, **BEFORE** it had been analyzed.



How The Moral Impurity Data Were Altered In the Promotion Condition

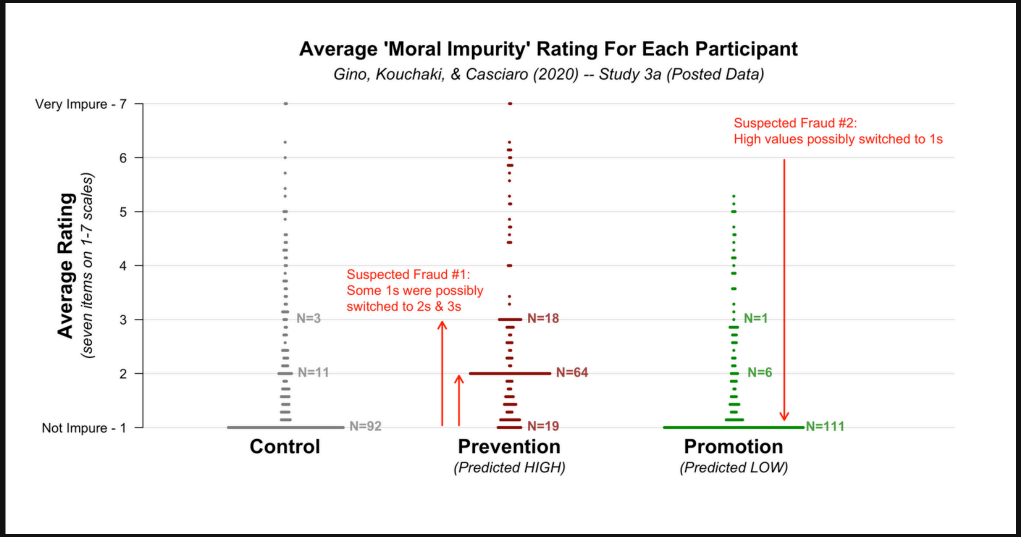
ID	Impure1	Impure2	Impure3	Impure4	Impure5	Impure6	Impure7
233	0	0	0	0	0	0	0
200	-2	-2	-2	-2	-2	-2	-3
447	-3	-3	-5	-3	-5	-6	-5
471	-3	-4	-5	-4	-5	-6	-5
335	-4	-3	-6	-4	-5	-6	-6
319	0	0	0	0	0	0	0
199	-2	-4	-3	-4	-5	-2	-4
30	0	0	0	0	0	0	0
498	-4	-3	-5	-2	-2	-4	-4
237	0	0	0	0	0	0	0
118	-5	-5	-6	-4	-4	-5	-5
120	-4	-4	-4	-4	-4	-4	-4
204	0	0	0	0	0	0	0
309	0	0	0	0	0	0	0
589	-4	-4	-5	-4	-5	-6	-5
220	0	0	0	0	0	0	0
251	0	0	0	0	0	0	0
248	-5	-5	-6	-4	-5	-6	-4
364	-6	-6	-6	-6	-6	-6	-6
376	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
268	0	0	0	0	0	0	0
290	-3	-2	-3	-2	-2	-4	0
47	0	0	0	0	0	0	0
454	-3	-2	-4	-3	-4	-4	-4
441	-4	-3	-2	-3	-4	-5	-3
538	-6	-6	-6	-6	-6	-6	-6
8	0	0	0	0	0	0	0

Participant IDs in the Posted Data

Participant 200's 7th moral impurity rating was decreased by 3

All of Participant 538's moral impurity ratings were decreased by 6

Data manipulation



Average 'Moral Impurity' Rating For Each Participant

Gino, Kouchaki, & Casciaro (2020) -- Study 3a (Posted Data)

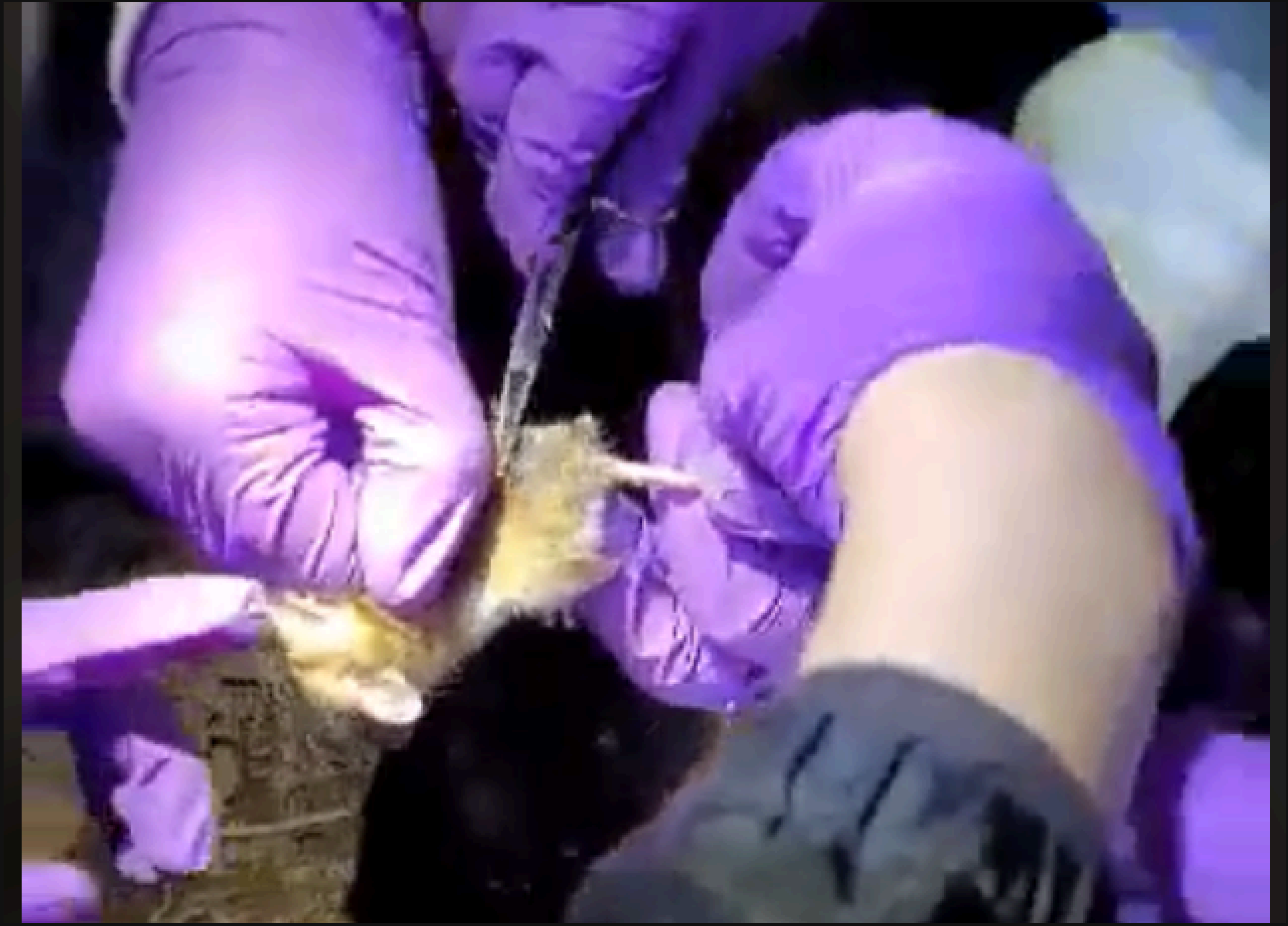
Suspected Fraud #1: Some 1s were possibly switched to 2s & 3s

Suspected Fraud #2: High values possibly switched to 1s

Results of manipulation

What can YOU do?

What is this?



Training

- Biology students learn key lab techniques
  - Pipetting
  - Capture-recapture of wild animals

Training

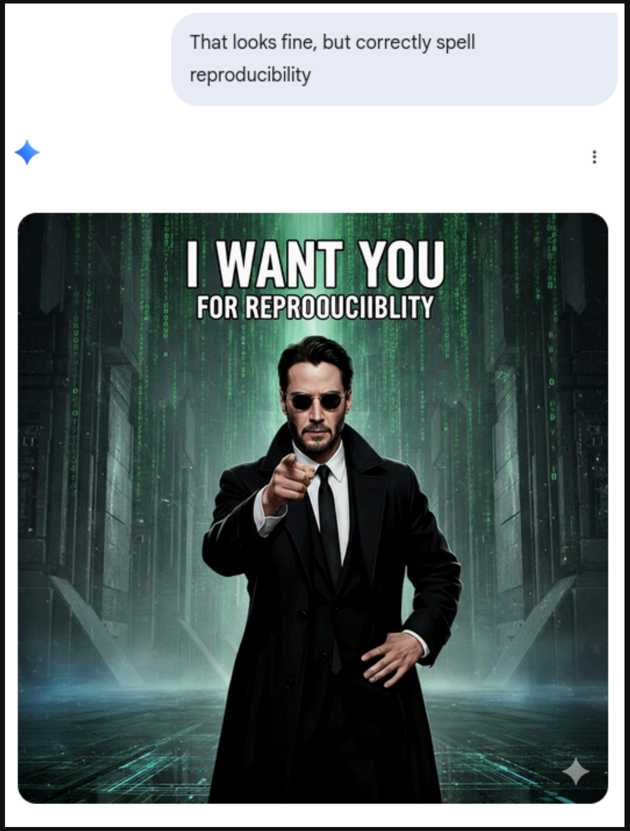
- Biology students learn key lab techniques
    - Pipetting
    - Capture-recapture of wild animals

That’s my daughter’s hands in there

YOU!

(BTW...)

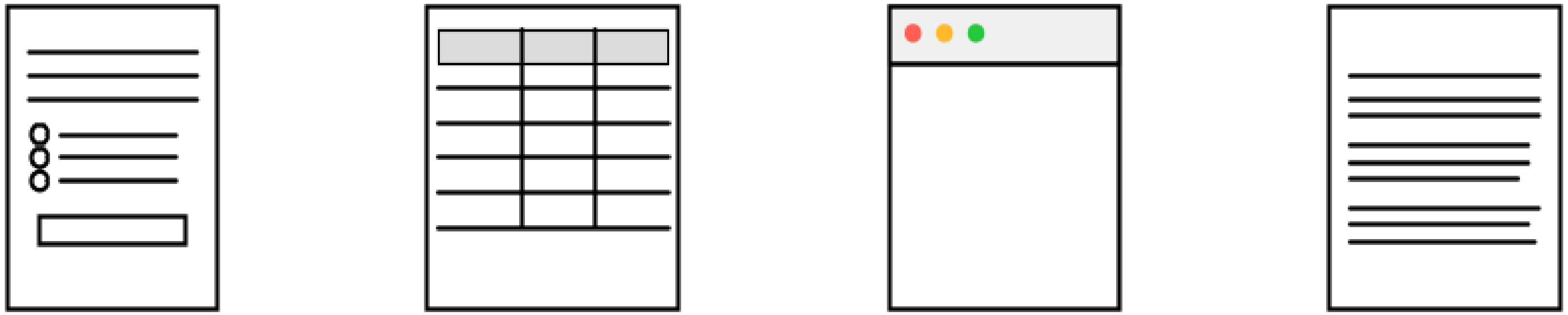
When I prompted Gemini to correct the spelling in "Reproducibility", it only made it worse!



That looks fine, but correctly spell reproducibility

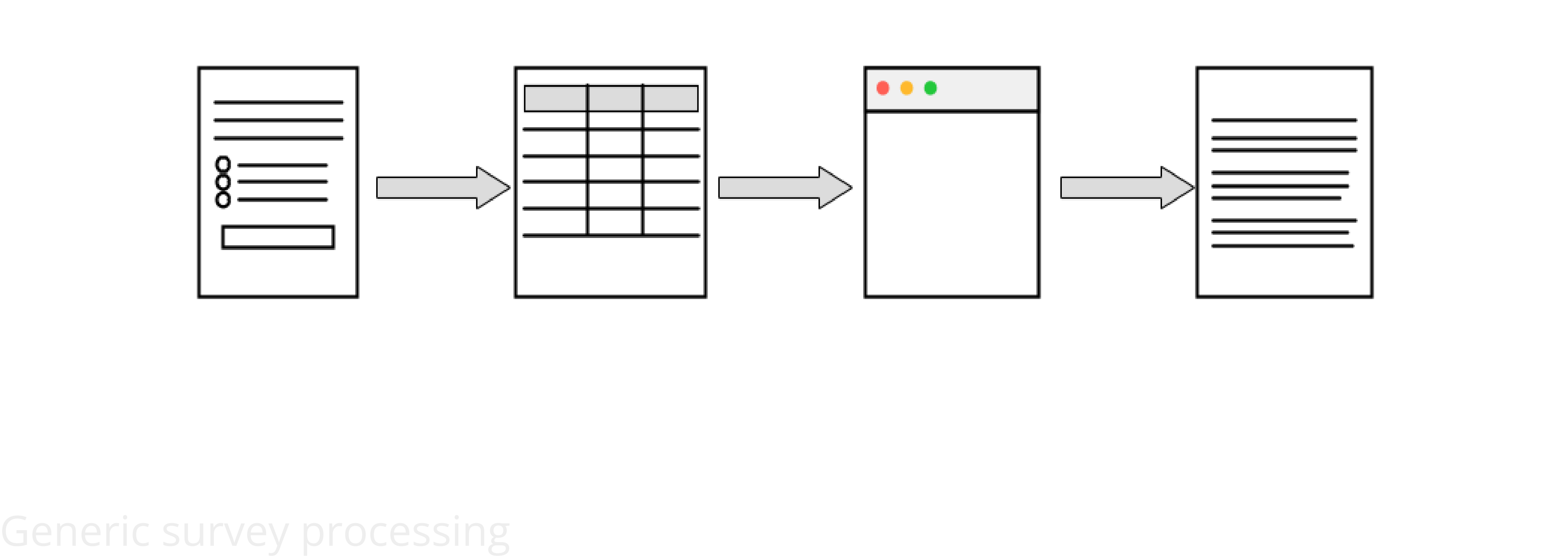
Back to the topic

Generic survey processing

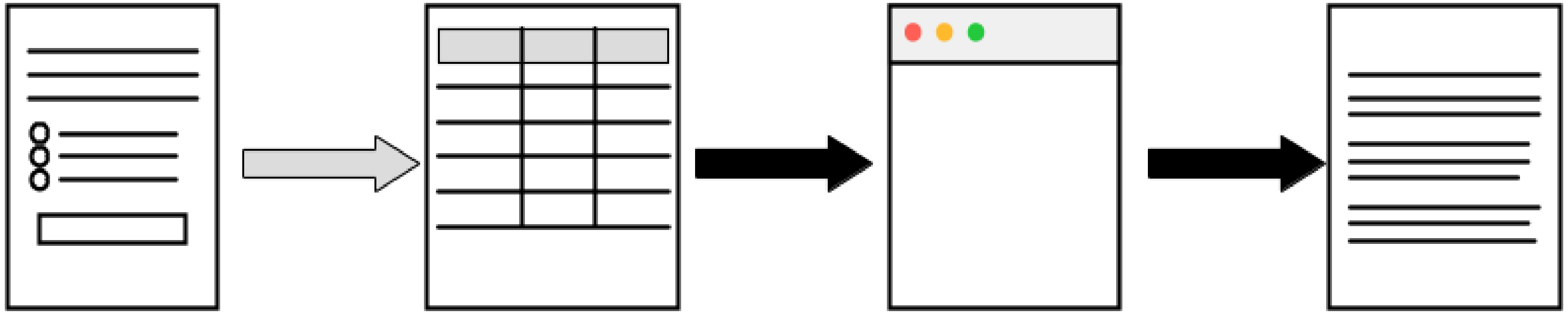


Generic survey processing

Generic survey processing

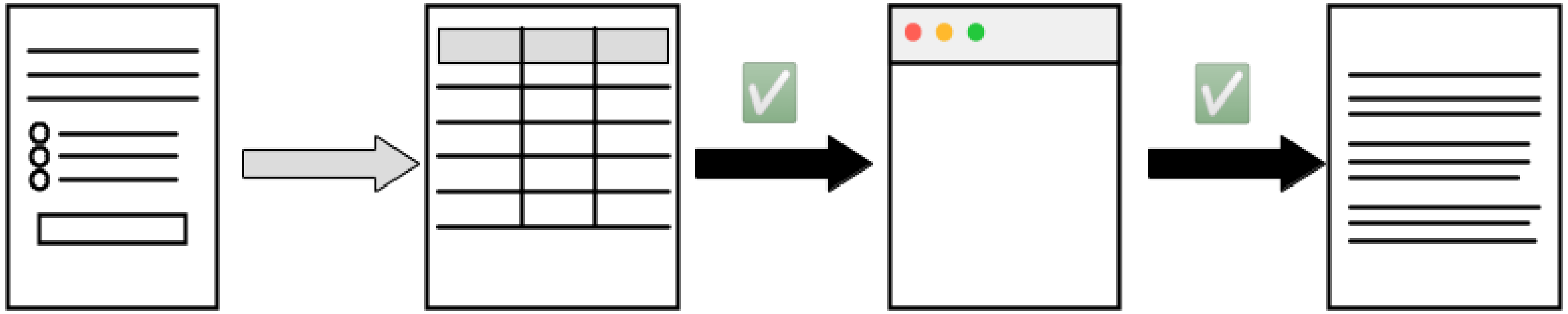


Requiring transparency in academia



Generic survey processing

Verifying transparency in academia



Generic survey processing

Verification by journals

- **Provision** (publication of materials) provides transparency
- **Verification** (running the analysis again - computational reproducibility) compensates for ***mistrust***/***absence of trust***

Which journals

- American Economic Association (8)
- Econometric Society (3)
- Canadian Journal of Economics (1)
- Royal Economic Society (2)
- Western Economic Association International (1)
- European Economic Association (1)
- Review of Economic Studies (1)
- Journal of the European Economic Association (1)
- Journal of Political Economy (3)
- American Journal of Political Science (1)
- American Political Science Review (1)

Verification by others

- Pre-publication: cascad

cascad

- Post-publication: Data Colada, Institute for Replication

I4R

Verification by institutions

- World Bank[2]

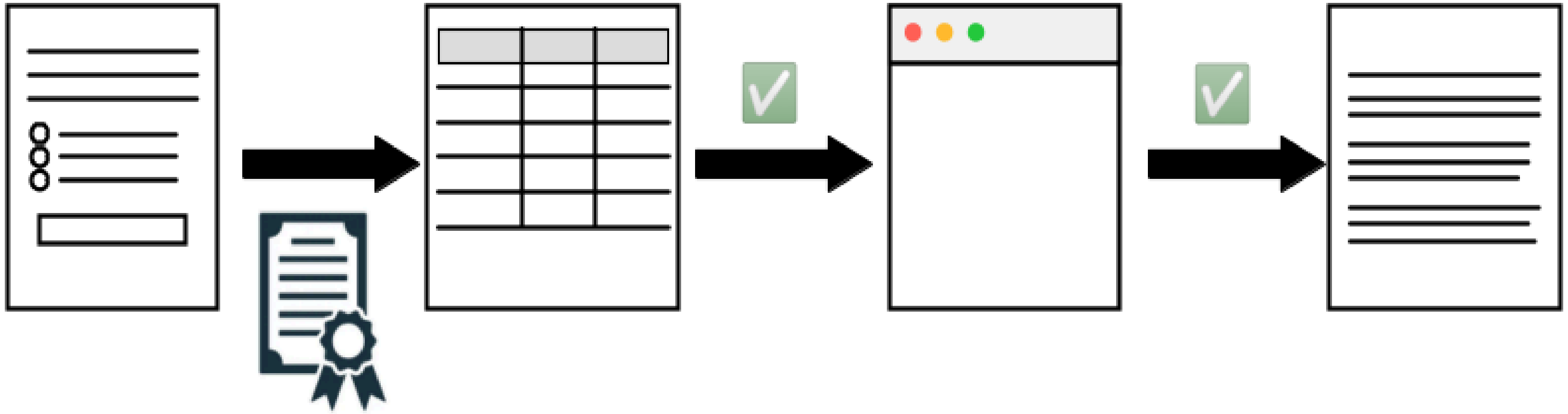
THE WORLD BANK
IBRD • IDA
Reproducible Research Repository

World Bank RRR

Verification by you

- Toner-Rodgers was caught b/c others scrutinized his work
  - Investigated the statistics (not too fishy)
  - Investigated the sources (huh?)
  - Investigated the coherence of the paper (much harder)

Taking it a step further



Survey flow

Taking it a step further

- Has been discussed by authors behind Data Colada
- Survey tool provider (Qualtrics, etc.) exports data, posts **checksum**
- Survey tool provider exports data only to institution directly into trusted repository, researchers obtain data from there (with privacy protections)
- Researcher can verify **checksum**

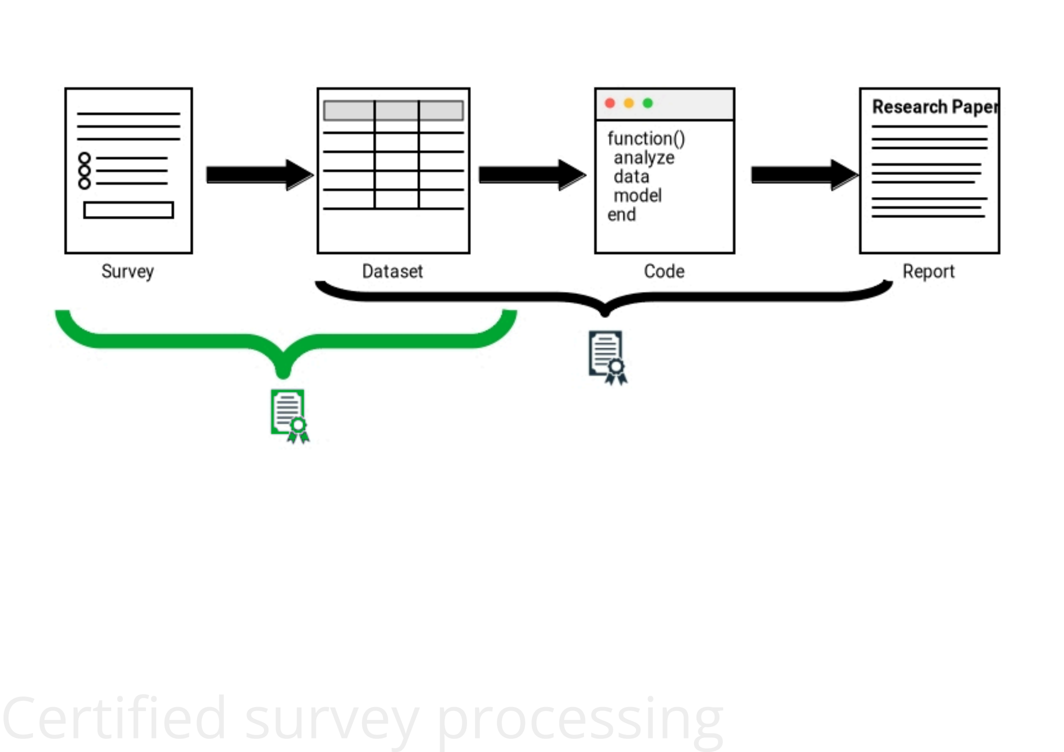
Does not work yet

What if you could verify the file?

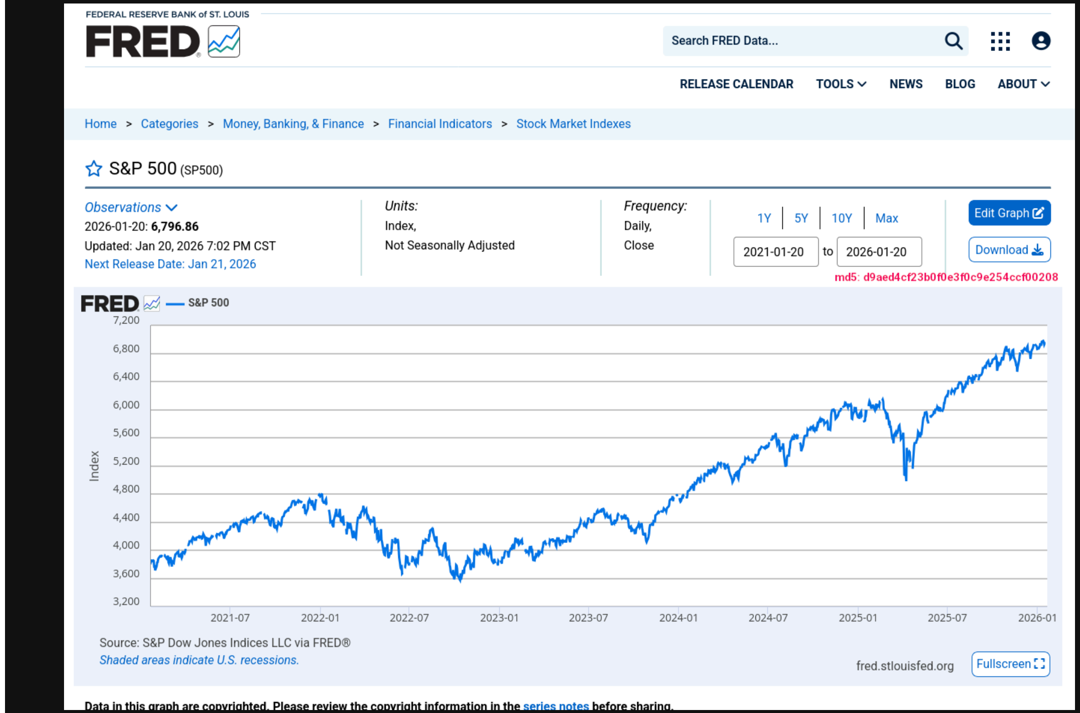
```r
# compute checksum of the file
tools::md5sum(here::here("presentation","SP500.csv"))
```

```
/home/runner/work/presentation-2026-01/presentation-2026-
01/presentation/SP500.csv

"d9aed4cf23b0f0e3f0c9e254ccf00208"
```



Certified survey processing



Home > Categories > Money, Banking, & Finance > Financial Indicators > Stock Market Indexes

S&P 500 (SP500)

Observations
2026-01-20: **6,796.86**
Updated: Jan 20, 2026 7:02 PM CST
Next Release Date: Jan 21, 2026

Units:
Index,
Not Seasonally Adjusted

Frequency:
Daily,
Close

2021-01-20 to 2026-01-20

md5: d9aed4cf23b0f0e3f0c9e254ccf00208

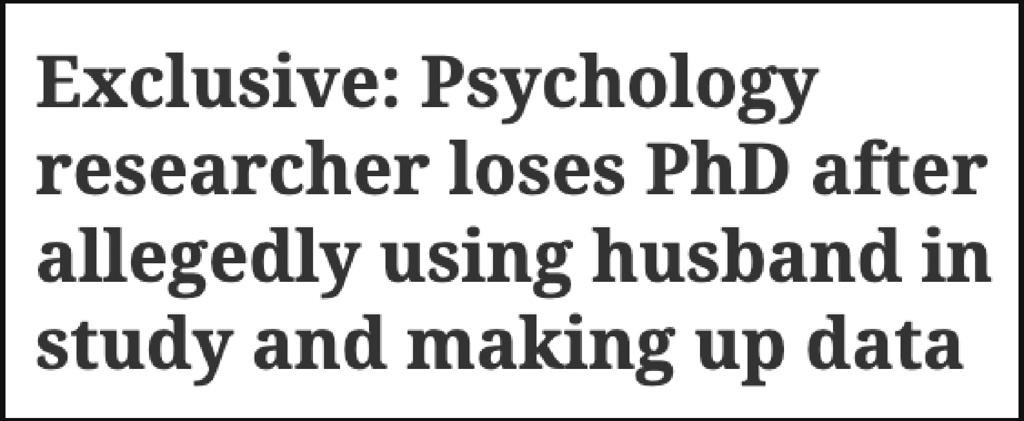
Source: S&P Dow Jones Indices LLC via FRED®
Shaded areas indicate U.S. recessions.

Data in this graph are copyrighted. Please review the copyright information in the series notes before sharing.

Not yet available!

Does not prevent all fraud

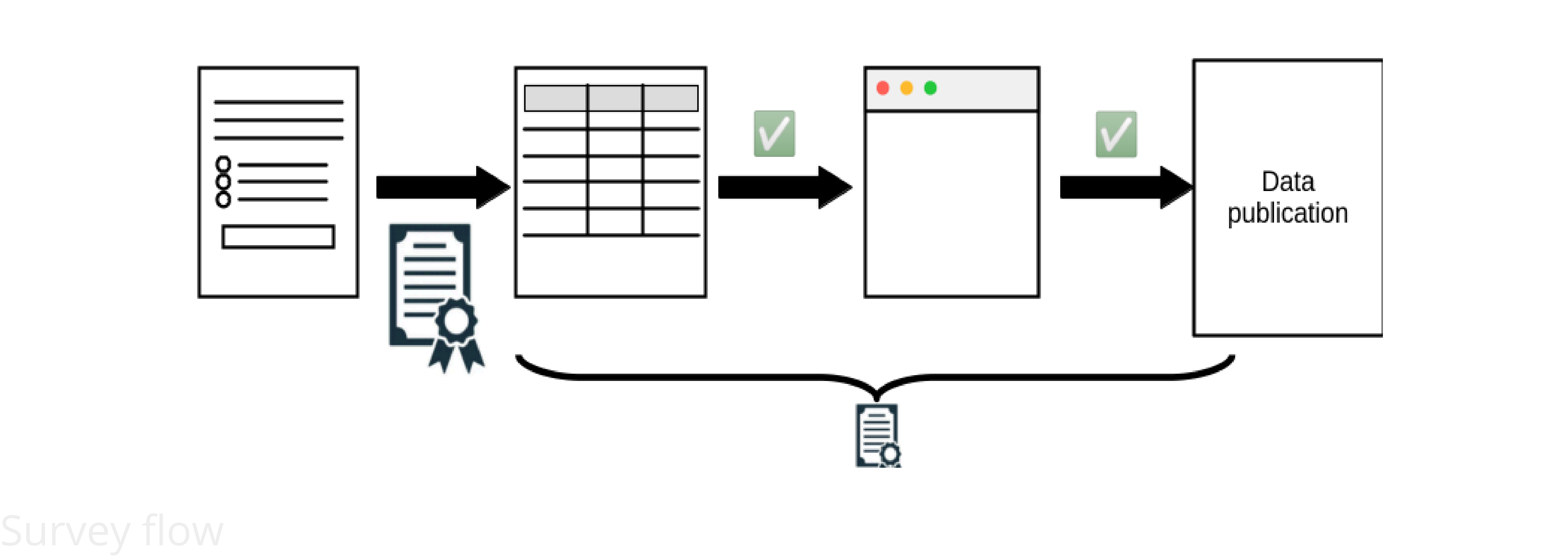
Toronto researcher loses Ph.D.



Exclusive: Psychology researcher loses PhD after allegedly using husband in study and making up data

Toronto case

How to document the full process?



Survey flow

How to document the full process?

- Identify all sources - very precisely!
- Document all processing steps - using tools you are learning here!
- Transparent about your process - show your code!
- Expect critique (if not criticism) - embrace it!

Churchill said...

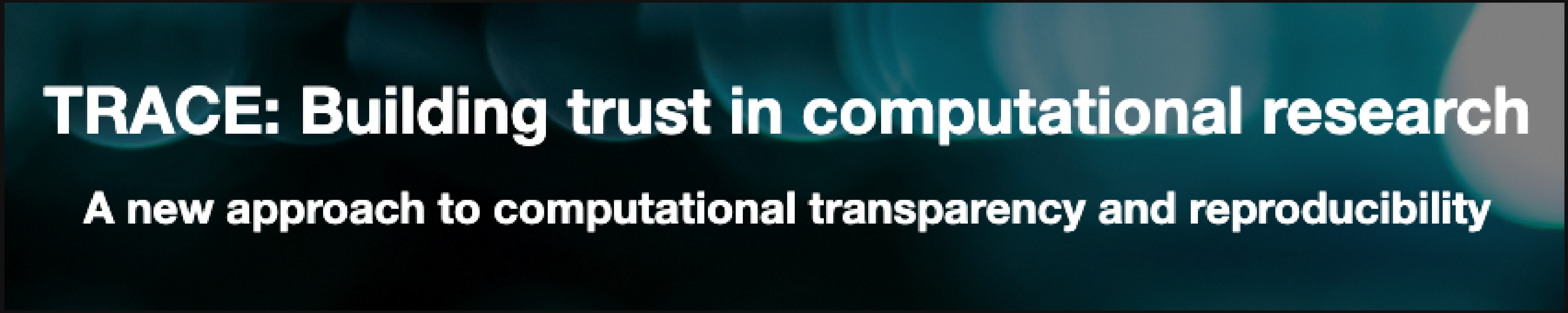
"Do not trust any statistics you did not fake yourself."

Or did he...

"No, of course, the British Prime Minister ... **never claimed such nonsense**. But putting his name in front of a quote gives it a more solemn, more imposing, more definitive appearance." [Source]

A sketch: Transparency Certified

https://transparency-certified.github.io/



TRACE: Building trust in computational research

A new approach to computational transparency and reproducibility

Transparency Certified

Work in progress

- Working with cascad, several INEXDA members, and others
- Relying on external certification of data inputs (data catalogs with metadata, checksums)

Who is this person? (3)

Who is this person? (3)

Alp Simsek, one of …

Transparency is the norm

nearly 5,000 authors who have had their work verified by the AEA Data Editor and team.

You can do it, too!

Bonus: You can help, too!

- Apply the reproducible tools ***not just for research***: Policy analysis, program evaluation, business analytics, etc.
- Push data providers to provide better access tools (**API access**, checksums, **DOI**)
- Be able to explain why this ***benefits*** them!

The end!

Source

- This document's source: https://github.com/larsvilhuber/presentation-2026-01

Footnotes

1.

https://datacolada.org/109, https://datacolada.org/110, https://datacolada.org/111, https://datacolada.org/112, https://datacolada.org/114, https://datacolada.org/118

2.

Jones, M. (2024). Introducing Reproducible Research Standards at the World Bank. Harvard Data Science Review, 6(4). https://doi.org/10.1162/99608f92.21328ce3