

The uncertainties of working with web tracking data, and on how to (maybe) tackle them

Oriol J. Bosch | University of Oxford & RECSM



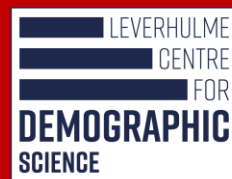
oriol.bosch-jover@demography.ox.ac.uk



orioljbosch



<https://orioljbosch.com/>



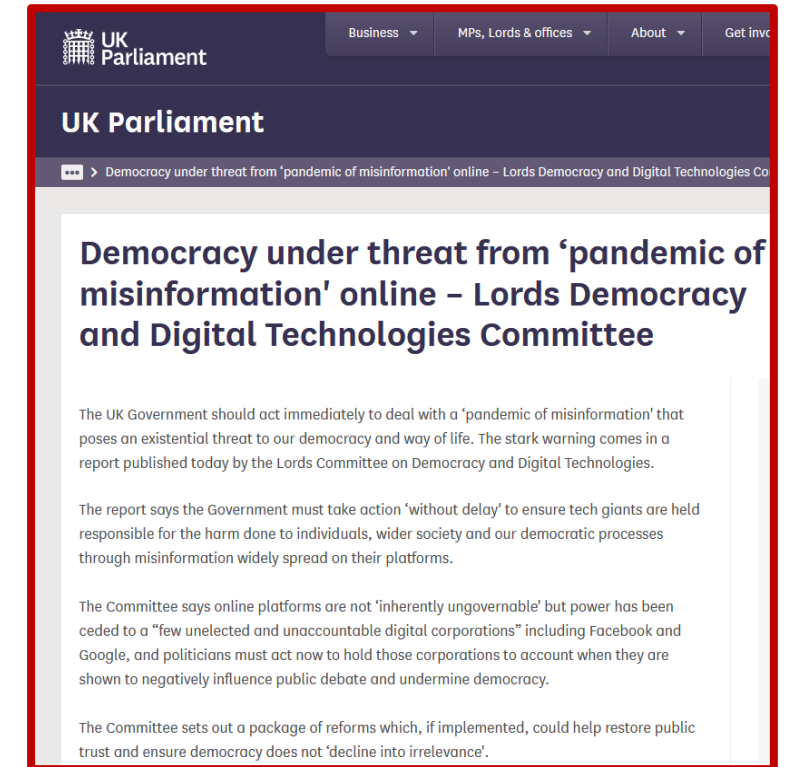
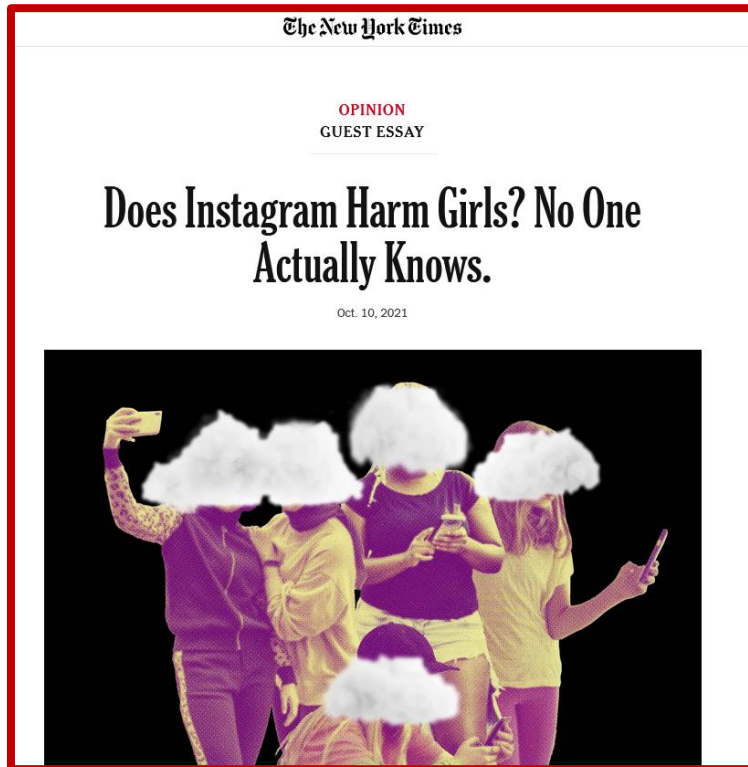
Universitat
Pompeu Fabra
Barcelona



Funding: This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 849165; PI: Melanie Revilla); the Spanish Ministry of Science and Innovation under the "R+D+i projects" programme (grant number PID2019-106867RB-I00 /AEI/10.13039/501100011033 (2020-2024), PI: Mariano Torcal); and the BBVA foundation under their grant scheme to scientific research teams in economy and digital society, 2019 (PI: Mariano Torcal).

The importance of measuring what people do online

- It is becoming vital to better understand what people do online and what impact this has on online and offline phenomena.



Web tracking data to understand online behaviours




Survey self-reports are still the **most common approach**

The Immensely Inflated News Audience: Assessing Bias in Self-Reported News Exposure [Get access >](#)

Markus Prior 

Public Opinion Quarterly, Volume 73, Issue 1, Spring 2009, Pages 130-143, <https://doi.org/10.1093/poq/nfp002>

Published: 18 March 2009

 Cite  Permissions  Share ▼

Abstract

Many studies of media effects use self-reported news exposure as their key independent variable without establishing its validity. Motivated by anecdotal evidence that people's reports of their own media use can differ considerably from independent assessments, this study examines systematically the accuracy of survey-based self-reports of news exposure. I compare survey estimates to Nielsen estimates, which do not rely on self-reports. Results show severe overreporting of news exposure. Survey estimates of network news exposure follow trends in Nielsen ratings relatively well, but exaggerate

But they might be affected by many errors



Web tracking data to understand online behaviours

Survey self-reports are still the **most common approach**

More and more availability of **digital traces to directly observe media exposure**

Web tracking data

Direct observations of online behaviours using tracking solutions, or *meters*.



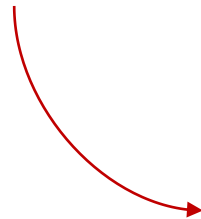
Group of tracking technologies (plug-ins, apps, proxies, etc)



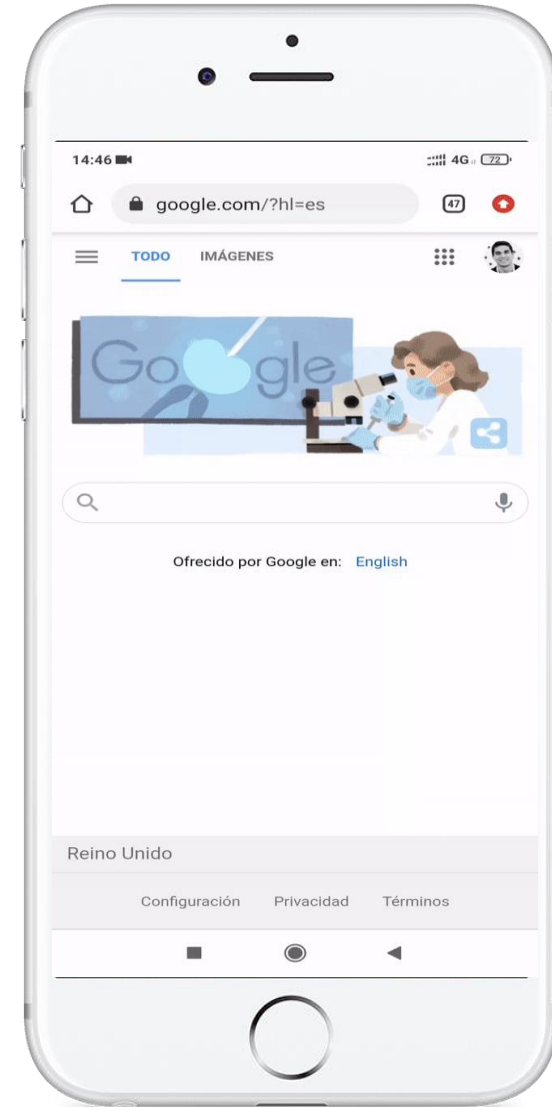
Installed on participants devices



Collect traces left by participants when interacting with their devices online: URLs, apps visited, cookies...



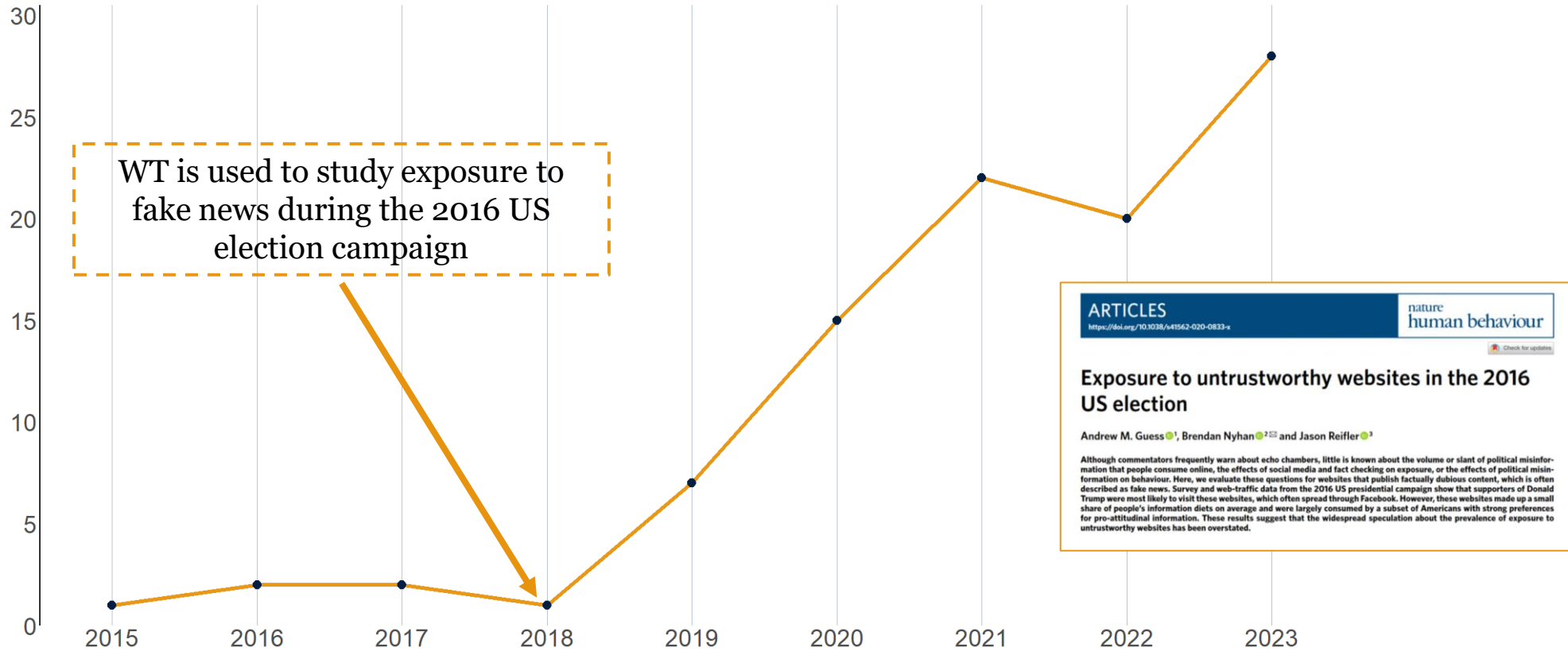
Great, we will get unbiased measures!



The rise of web tracking data

Web tracking data is becoming part of social scientists' toolkit

Number of papers using web tracking data



But... do we even know how to use web tracking data?

Let's say that we want to identify cheaters in online surveys

Let's say that we want to identify cheaters in online surveys

- **Political knowledge** is a **central construct** in political science, communications, and related fields
- Online surveys can harm the quality of this measures if participants search the answers online

PRE PREKNOW_CATCH

CASI PRE: KNOWLEDGE CATCH QUESTION

Survey Question	In what year did the Supreme Court of the United States decide Geer v. Connecticut? Type the year.
Logic	If type =video or phone, skip.
Response Order	No response options
Misc Spec	- Response Type: Numeric Entry



Google

In what year did the Supreme Court of the United States decide Geer v. Connecticut?

Images News Videos Books Finance

About 295,000 results (0,39 seconds)

1896

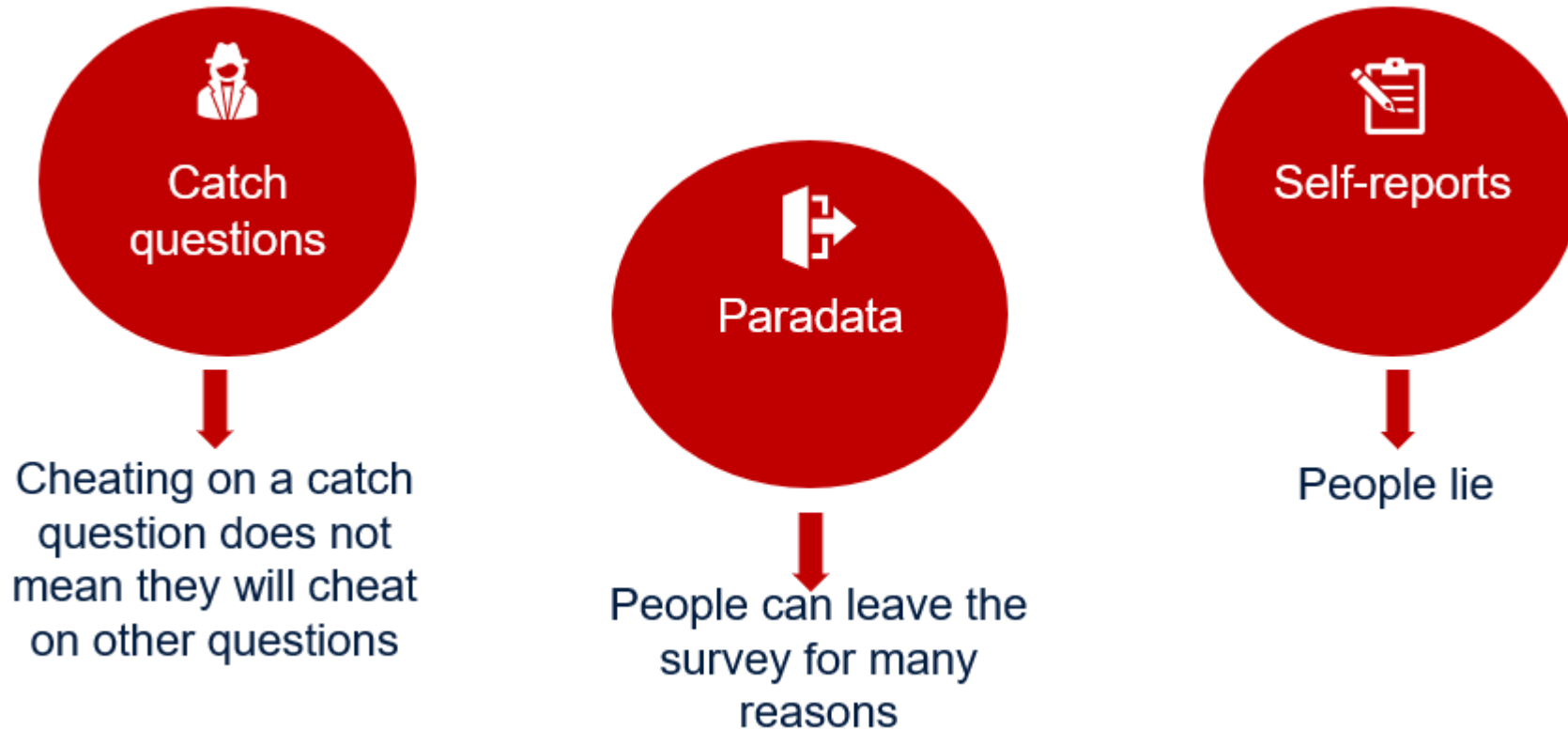
Connecticut, 161 U.S. 519 (1896)

Justia US Supreme Court Center
<https://supreme.justia.com/cases/federal>

[Geer v. Connecticut :: 161 U.S. 519 \(1896\)](#)

Let's say that we want to identify cheaters in online surveys

- How can we identify cheating?



Let's say that we want to identify cheaters in online surveys

Web tracking data allows to catch participants *in flagrante*

URL

<https://www.google.com/search?client=firefox-b-d&q=In+what+year+did+the+Supreme+Court+of+the+United+States+decide+Geer+v.+Connecticut%3F>

URL

<https://supreme.justia.com/cases/federal/us/161/519/>

Ok, it seems like cheating can be easily identified with web trackers

THE COMPLEXITIES OF WEB TRACKING DATA

Let's put this into test



Let's put this into test

Survey combined with **web tracking data** at the individual level

Netquest metered panel in Spain

- **Cross-quotas:** gender, age, and education
- **Sample size:** 1,200
- **Fieldwork:** Late May – Early June 2023

Tracking technologies installed in both **mobile and desktop devices**

Part of the ERC project **WEB DATA OPP**

What did we check?

5 legitimate political questions, plus one catch questions

- Is the Defense Minister in Spain Margarita Robles? (Yes/No)
- What political offices does Emmanuel Macron hold? (Open-ended, w/ picture)
- What percentage of the Spanish congress are women? (choose correct option)
- What was the date chosen for the upcoming general election? (choose correct option)
- What political party has decided not to run in this general election? (open-ended)
- Who was the first president of the Second Spanish Republic? (catch, open-ended)

What did we check?

5 legitimate political questions, plus one catch questions

- Is the Defense Minister in Spain Margarita Robles? (Yes/No)
- What political offices does Emmanuel Macron hold? (Open-ended, w/ picture)
- What percentage of the Spanish congress are women? (choose correct option)
- What was the date chosen for the upcoming general election? (choose correct option)
- What political party has decided not to run in this general election? (open-ended)
- Who was the first president of the Second Spanish Republic? (catch, open-ended)

Practically no one should know this

Our approach to catch cheaters

1. Identify the exact time in which participants answered the knowledge questions
2. Extract all the URLs that a participant visited during that identified period of time
3. Manually check all those URLs to see whether someone cheated



Web tracking
data

An honest depiction of how the process went

What were our expectations going into this project?

25%

ANES Survey 2019 pilot study

30%

Respondi Sample, Germany

14%

YouGov Sample, USA

13%

CCES Sample, USA

How many cheaters did we find?

How many cheaters did we find?



How many cheaters did we find?

What is
going on?



How many cheaters did we find?



Our bad!

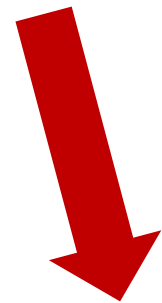
Ok, let's do this again... how many did we find?

Ok, let's do this again... how many did we find?

4.8%

Ok, let's do this again... how many did we find?

4.8%



20.6% self-report having cheated

Ok, let's do this again... how many did we find?

What is
going on?



Ok, let's do this again... how many did we find?



Everything is
alright!

Ok, let's do this again... how many did we find?

Who was the first president of the
Second Spanish Republic?

Ok, let's do this again... how many did we find?

Who was the first president of the
Second Spanish Republic?



0.3 % cheated

Ok, let's do this again... how many did we find?

Who was the first president of the
Second Spanish Republic?



0.3 % cheated

+20% left the page

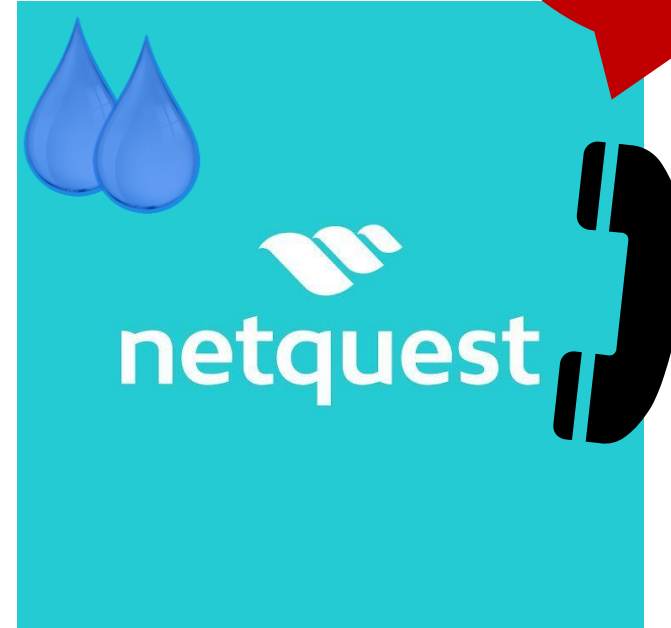
Ok, let's do this again... how many did we find?

Are you
kidding me?!

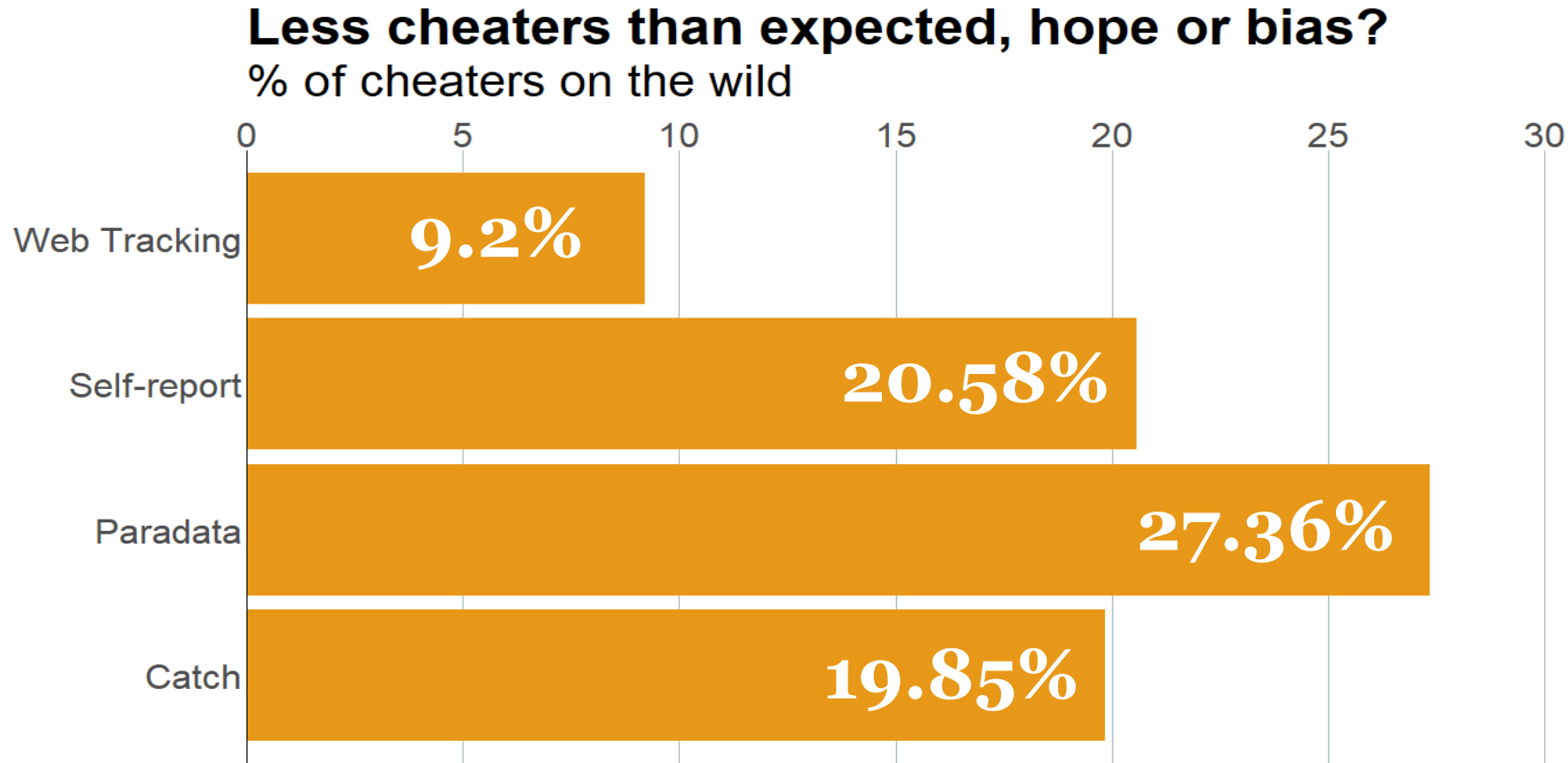


Ok, let's do this again... how many did we find?

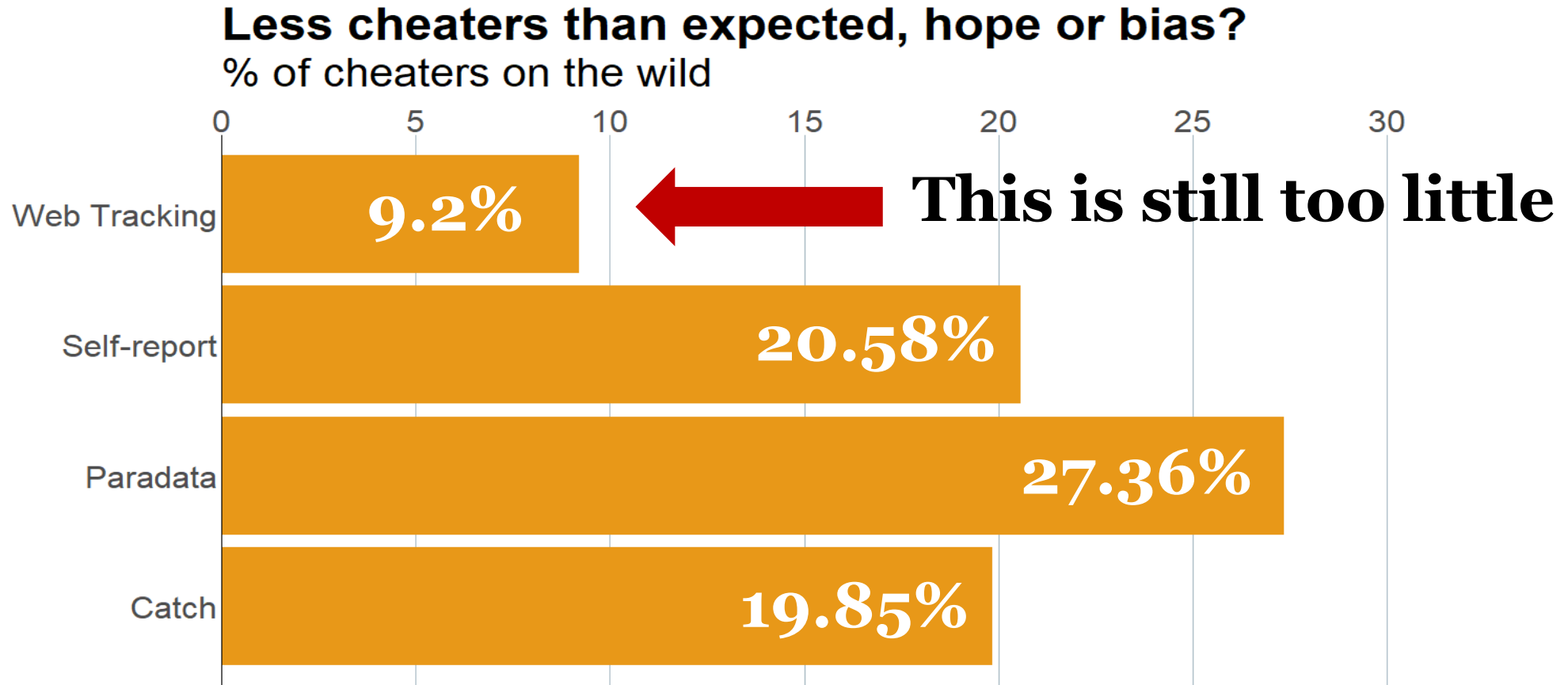
Ah, you are
right!



One more time...how many did we find?



One more time...how many did we find?

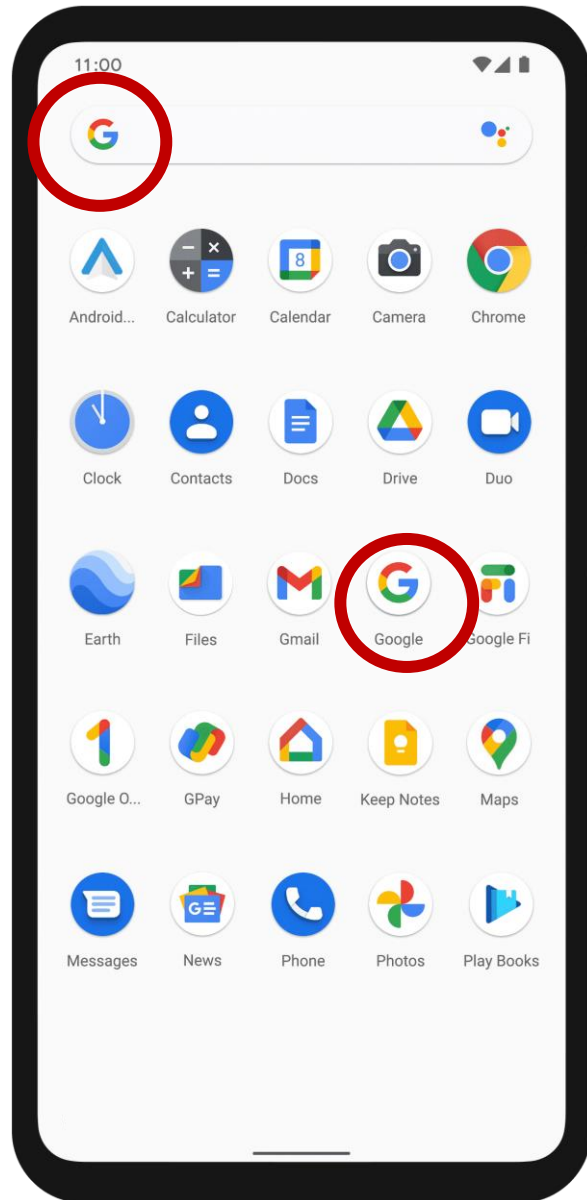


One more time...how many did we find?

I am losing my
mind



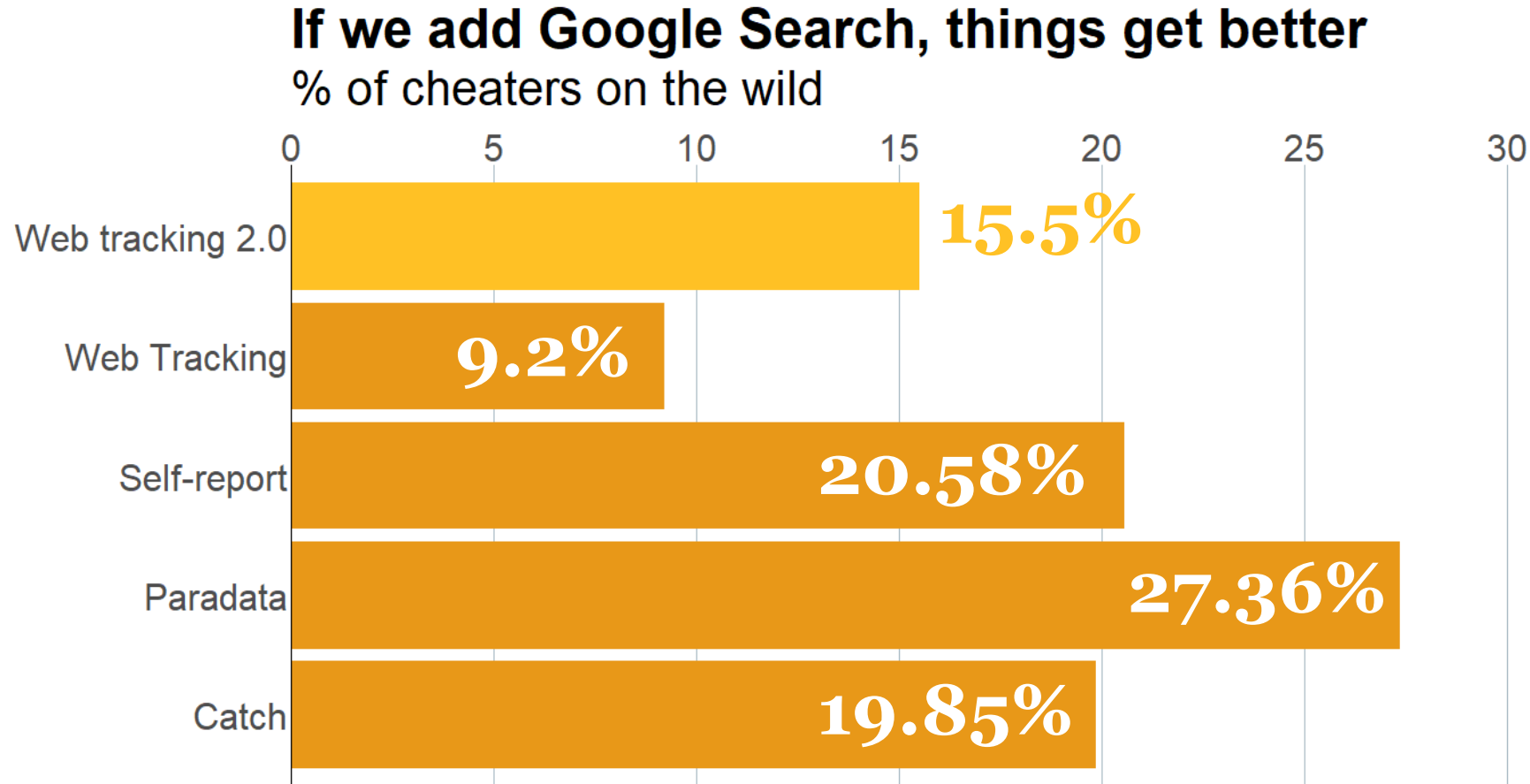




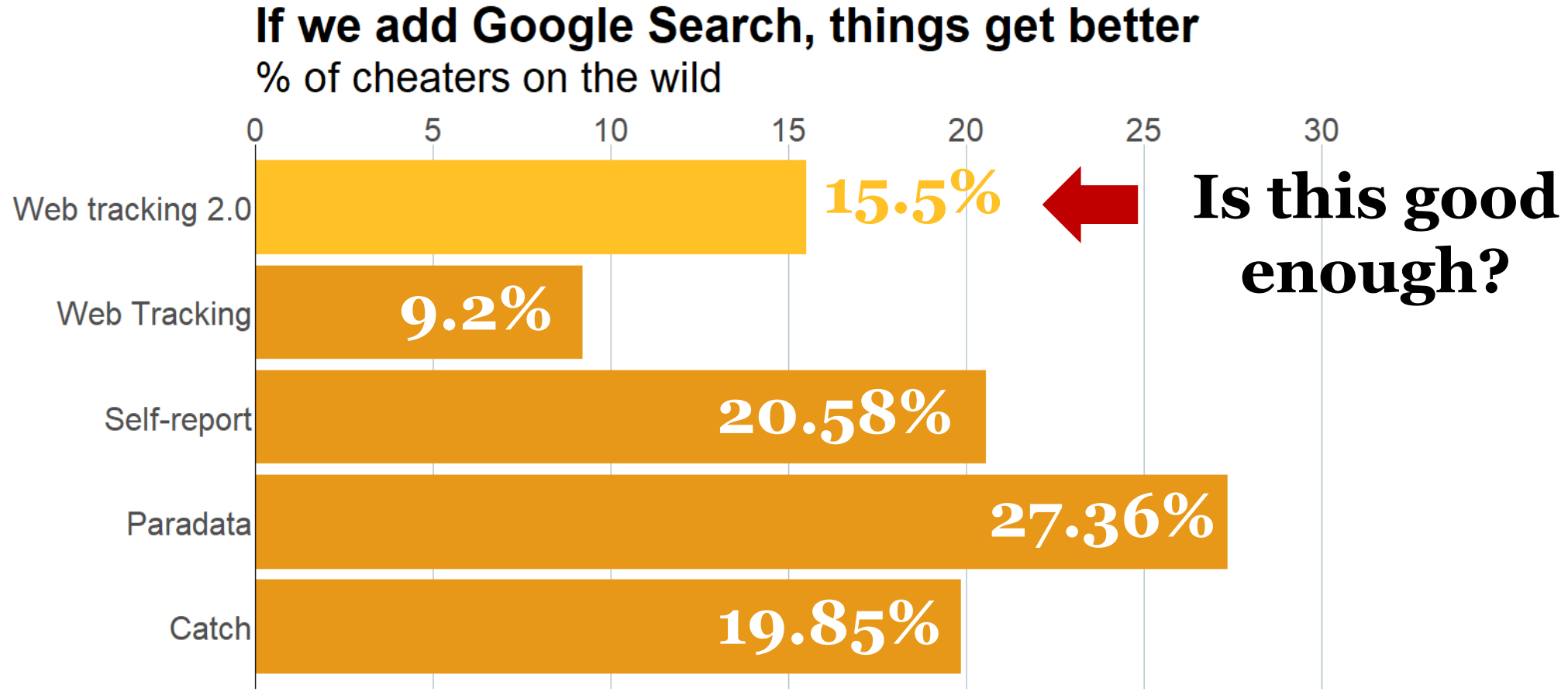
When we conducted
the study, our
technologies could
not see inside apps,
such as google search



Okey, accounting for Google Search things look better



Okey, accounting for Google Search things look better

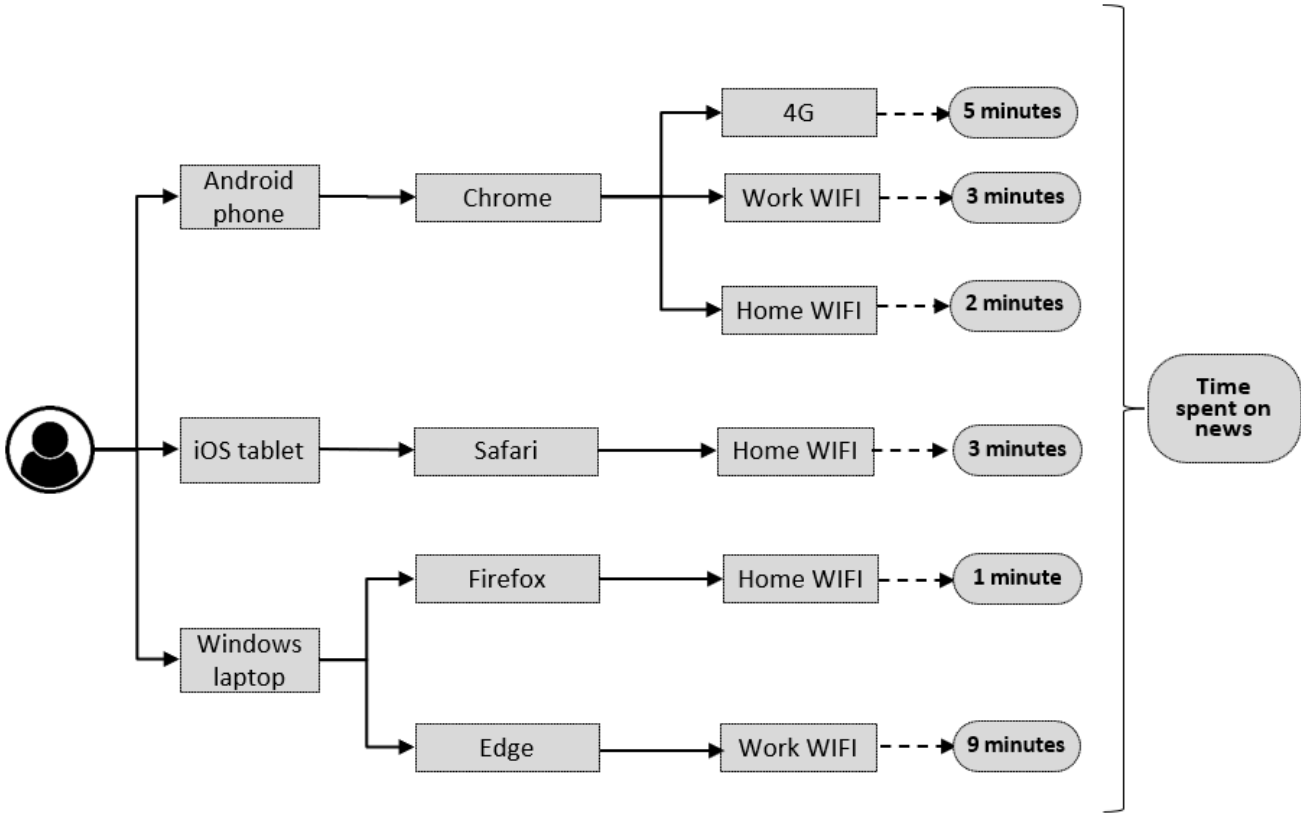


One more time...how many did we find?

Wait, I wrote a
paper about
this



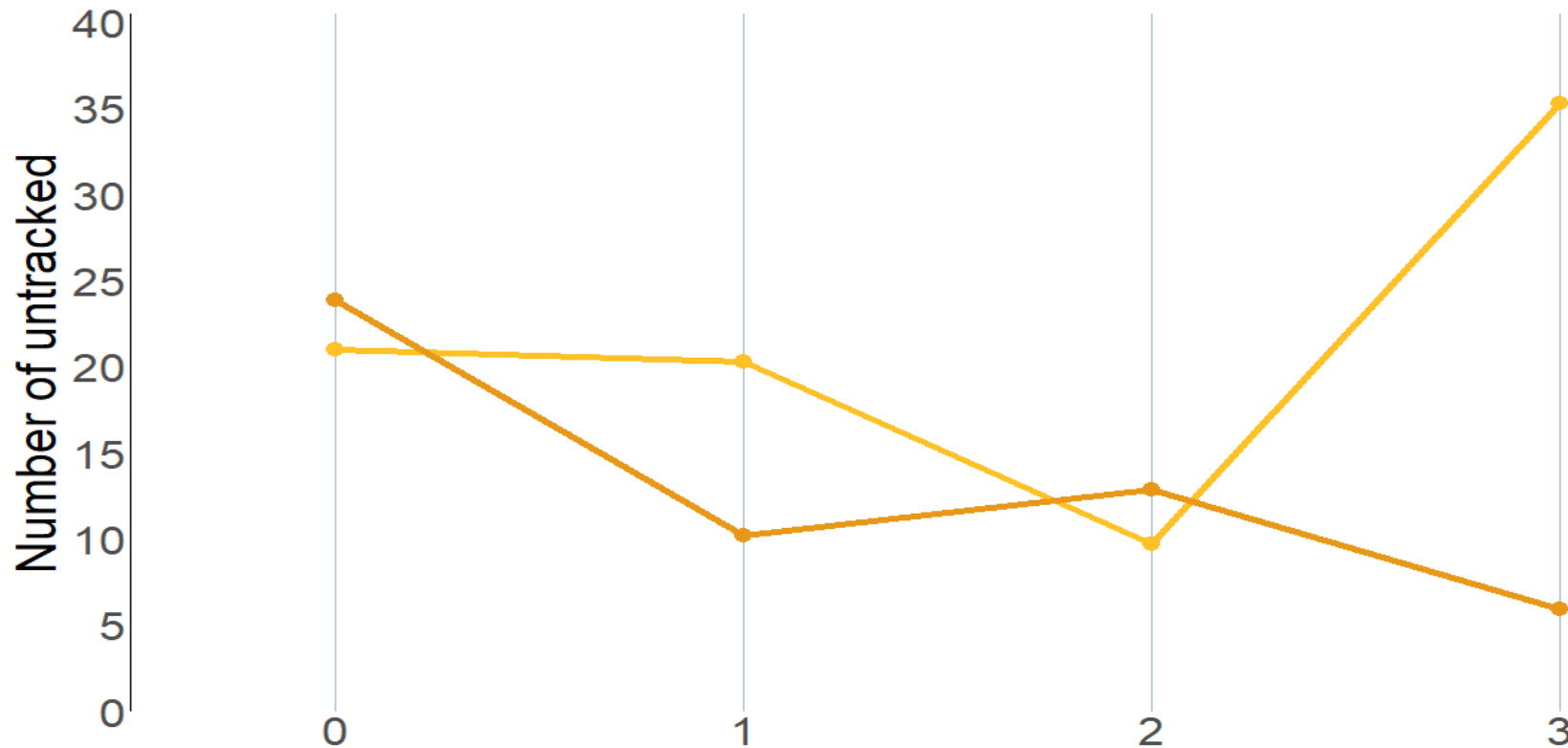
Participants might
have cheated on non-
tracked devices



Tracking undercoverage, as usual, messing with us

Undercoverage might explain everything else!

% of cheaters on the wild



What is this exemplifying?

Data extraction is complex,
even when you ask for URLs

Data extraction is complex,
even when you ask for URLs

1. You need complex guidelines
2. And tough validation strategies



The screenshot shows the top portion of a research article page. At the top left is the Elsevier logo, which includes a tree and the word 'ELSEVIER'. To the right of the logo, the text 'Data in Brief' is displayed in a large font, followed by 'Available online 9 May 2023, 109219' and 'In Press, Journal Pre-proof' with a question mark icon and a link 'What's this?'. On the far right, there is a small red icon with the text 'Data in Brief'. Below this header, the text 'Data Article' is visible. The main title of the article is 'The dynamics of political and affective polarisation: Datasets for Spain, Portugal, Italy, Argentina, and Chile (2019-2022)'. At the bottom, the authors are listed: 'Mariano Torcal¹  , Emily Carty², Josep Maria Comellas³, Oriol J. Bosch⁴, Zoe Thomson¹, Danilo Serani²'.

Technologies are not perfect,
and they fail in mysterious ways

Technologies are not perfect,
and they fail in mysterious ways

You need to be sceptical, question
all your results, and push panel
companies

Even when they work perfectly,
technologies are limited

Even when they work perfectly,
technologies are limited

		PC app	PC plug-ins			Android SDK	iOS proxy
			Chrome	Firefox	Safari		
Online tracking							
URLs	Http traffic	Yes	Yes	Yes	Yes	Yes	Yes
	Https traffic	No	Yes	Yes	Yes	Yes	No
	Incognito sessions	No	Yes	Yes	Yes	Yes	No
	HTML	No	Yes	Yes	Yes	No	No
	Time stamps	Yes	Yes	Yes	Yes	Yes	Yes
Apps	App name	-	-	-	-	Yes	Yes
	App usage start time	-	-	-	-	Yes	Yes
	App usage duration	-	-	-	-	Yes	Estimated
	Offline apps	-	-	-	-	Yes	No
	In-app behaviour	-	-	-	-	No	No
Search terms	Search terms	Yes	Yes	Yes	Yes	Yes	No
Device information							
Device type	E.g. desktop	Yes	Yes	Yes	Yes	Yes	Yes
Device brand	E.g. Xiaomi		No	No	No	Yes	Yes
Device model	E.g. S9	No	No	No	No	Yes	Yes
Operating system	E.g. iOS	Yes	Yes	Yes	Yes	Yes	Yes
OS version	E.g. 10.1.2	No	No	No	No	Yes	Yes
Internet provider	E.g. Voxi	No	No	No	No	Yes	Yes

Extremely rich data...for some
devices, for some people

Extremely rich data...for some
devices, for some people

Is content data important enough
to focus only on a few devices?

What we want to measure, and
what we measure, might be
different

What we want to measure, and
what we measure, might be
different

Consider how design decisions will
impact the reliability & validity of
what you measure

VALIDITY AND RELIABILITY OF DIGITAL TRACE DATA
IN MEDIA EXPOSURE MEASURES: A MULTIVERSE OF
MEASUREMENTS ANALYSIS

Oriol J. Bosch

Tracking undercoverage is
prevalent, and biases our data

Tracking undercoverage is prevalent, and biases our data

Introduce strategies to identify undercoverage, simulate the biases, and correct your results

UNCOVERING DIGITAL TRACE DATA BIASES: TRACKING UNDERCOVERAGE IN WEB TRACKING DATA

Oriol J. Bosch, Patrick Sturgis, Jouni Kuha, and Melanie Revilla

ORIOI BEING A SUPER NICE HOST

Restaurant recommendations



Thanks!

Oriol J. Bosch | Postdoctoral Researcher, University of Oxford

 oriol.bosch-jover@demography.ox.ac

 orioljbosch

 <https://orioljbosch.com/>



European Research Council
Established by the European Commission