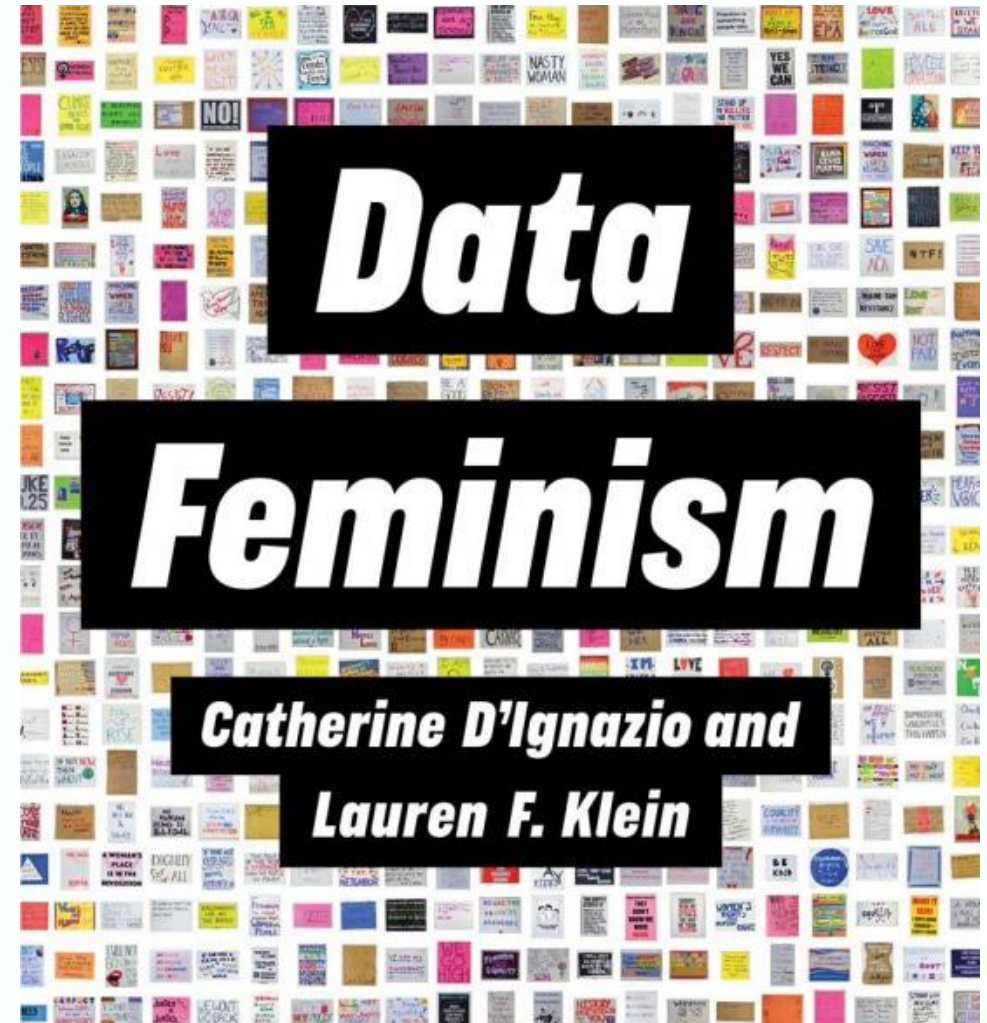


# **Errors in digital trace data collection**

## **Lecture 4**

# Literature:

- **Chapter 3:** “Asking questions” in *Bit by bit: Social science in the digital age (2017)*, Matthew Salganik **3.1 – 3.4**
- **Chapter 4:** “Errors of nonobservation: Sampling and coverage” in *Data collection with Wearables, Apps and Sensors (2023)*, Florian Keusch, Bella Struminskaya, Stephanie Eckman & Heidi Guyer
- **Introduction:** “Why data science needs feminism” and **Chapter 1:** “The power chapter” in *Data Feminism (2020)*, Catherine D’Ignazio & Lauren F. Klein

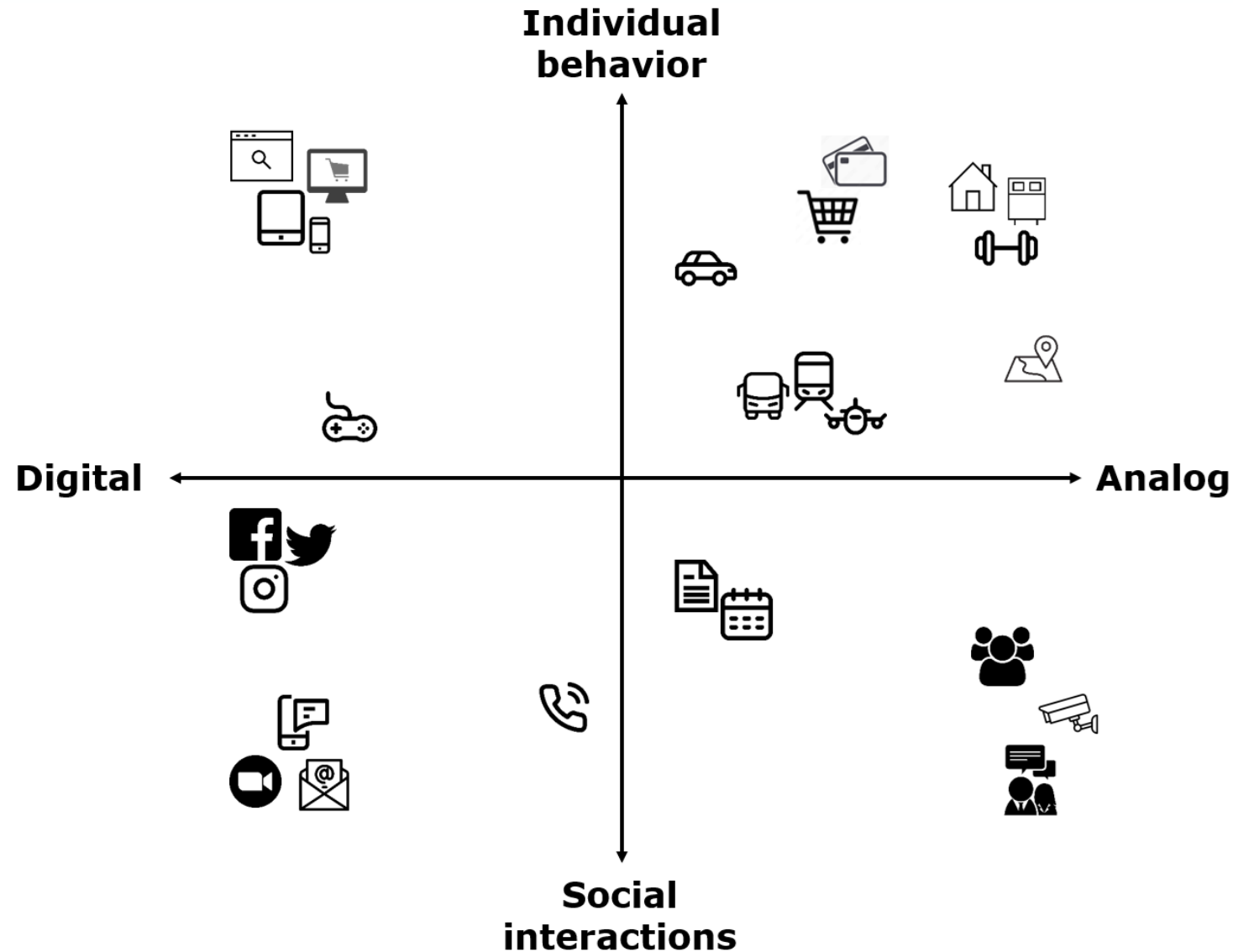


# Goals of this lecture

1. Understand the concept of an error and the purpose of error frameworks.
2. Distinguish between different types of errors.
3. Identify different errors in a specific study design.
4. Identify "external" errors.

# Recap week 1

# Where and when do you leave digital traces?



# Recap bit by bit chapter 1

**Readymade:** Repurpose big data sources that were originally created by companies and governments.

**Custommade:** A researcher started with a specific question and then used the tools of the digital age to create the data needed to answer that question.



Readymade

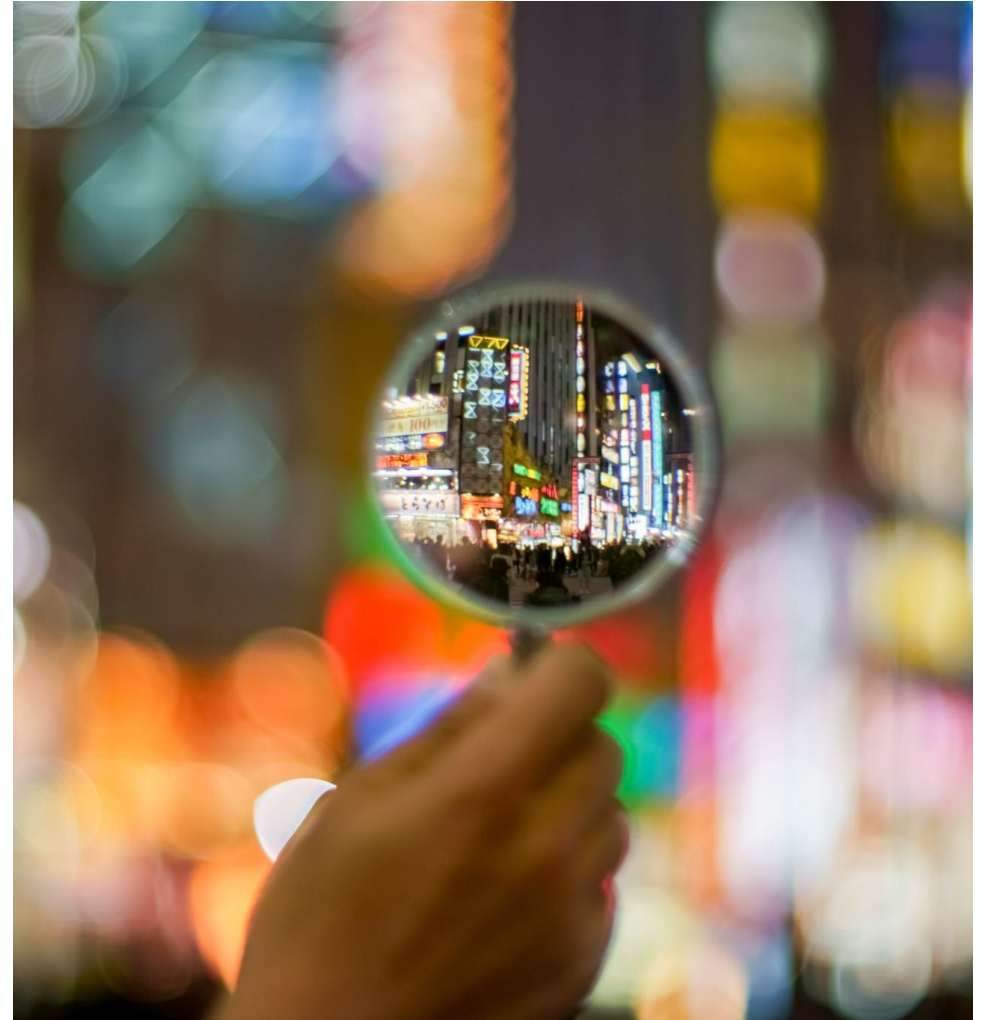


Custommade

# Recap bit by bit chapter 1

**Found data:** From the perspective of researchers, big data sources are “found”. However, they are designed by someone.

**Designed data:** Data designed specifically for a specific research purpose (experiment, survey or administrative).





# Survey methodology

 [32 languages](#) 

[Article](#) [Talk](#)

[Read](#) [Edit](#) [View history](#) [Tools](#) 

From Wikipedia, the free encyclopedia

*For the publication, see [Survey Methodology](#).*

**Survey methodology** is "the study of [survey methods](#)".<sup>[1]</sup> As a field of [applied statistics](#) concentrating on [human-research surveys](#), [survey methodology studies the sampling of individual units from a population and associated techniques of survey data collection](#), such as [questionnaire construction](#) and methods for improving the number and accuracy of responses to surveys. Survey methodology targets instruments or procedures that ask one or more questions that may or may not be answered.

Researchers carry out **statistical surveys** with a view towards making [statistical inferences](#) about the population being studied; [such inferences depend strongly on the survey questions used](#). [Polls](#) about [public opinion](#), public-health surveys, [market-research](#) surveys, government surveys and [censuses](#) all exemplify [quantitative research](#) that uses survey methodology to answer questions about a population. Although censuses do not include a "sample", they do include other aspects of survey methodology, like questionnaires, interviewers, and non-response follow-up techniques. Surveys provide important information for all kinds of [public-information](#) and research fields, such as [marketing](#) research, [psychology](#), [health-care provision](#) and [sociology](#).



# Why be interested in survey methodology?

In the “traditional” survey approach

# Some history

Table 3.1: Three Eras of Survey Research Based on [Groves \(2011\)](#)

	Sampling	Interviewing	Data environment
First era	Area probability sampling	Face-to-face	Stand-alone surveys
Second era	Random-digit dialing (RDD) probability sampling	Telephone	Stand-alone surveys
Third era	Non-probability sampling	Computer-administered	Surveys linked to big data sources

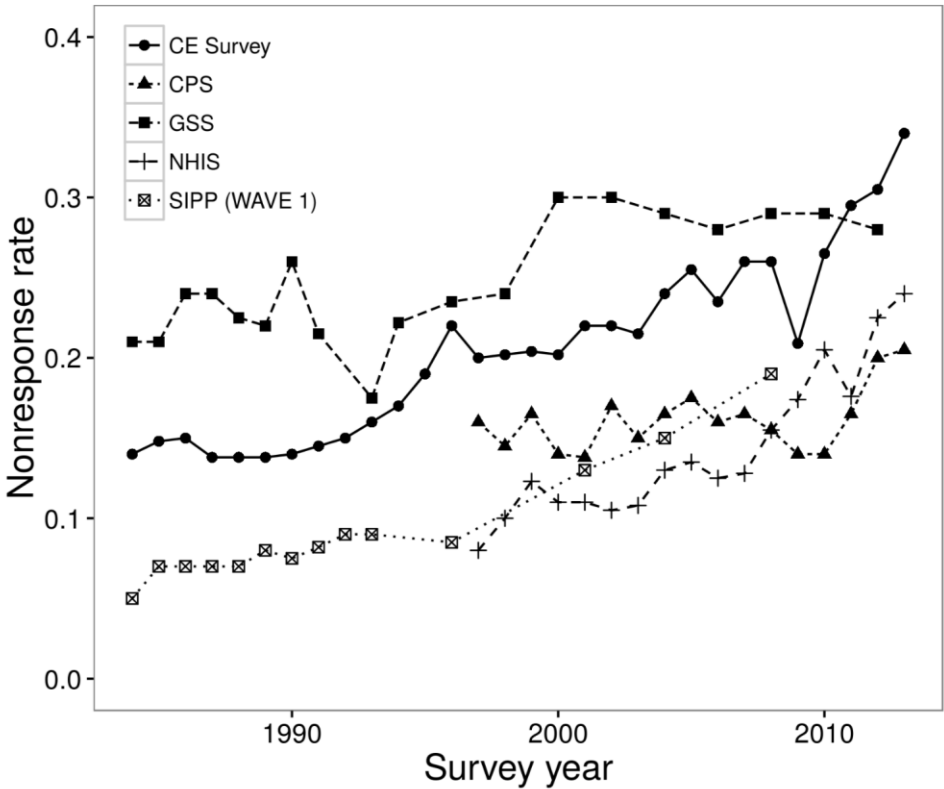
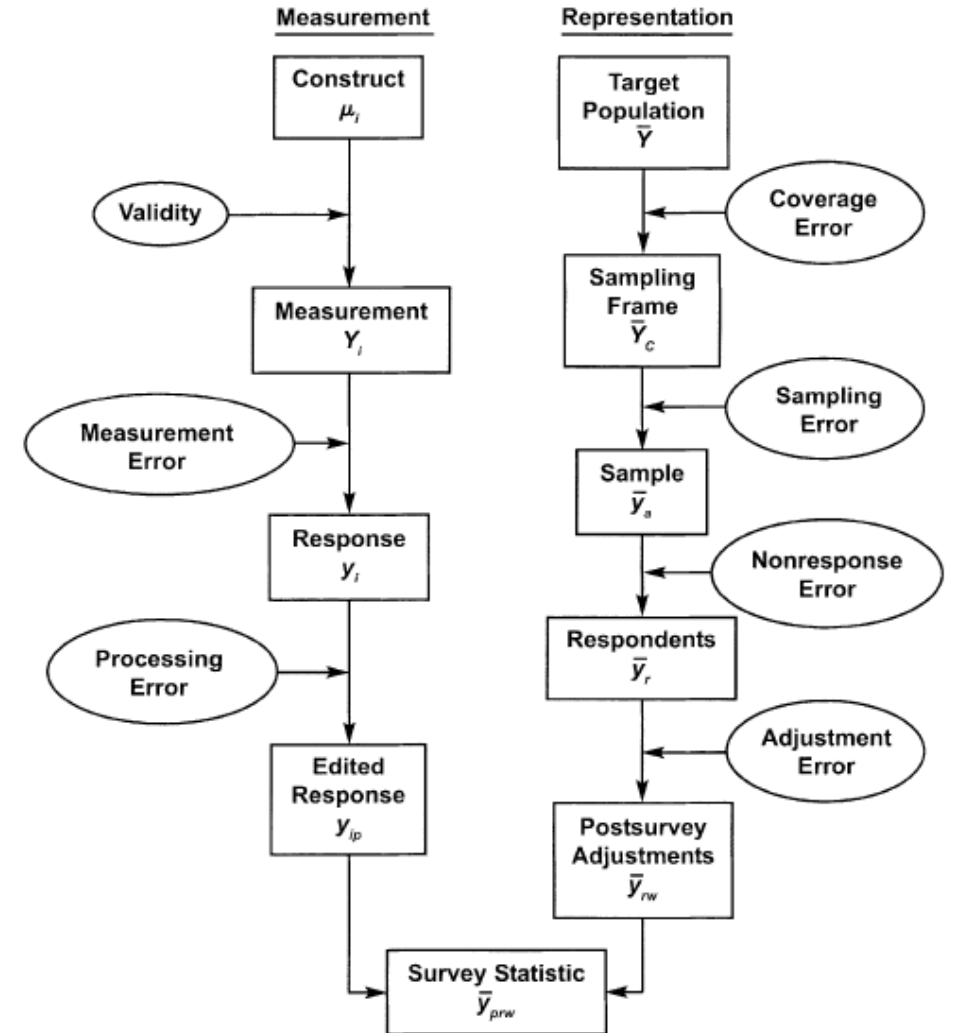


Figure 3.5: Nonresponse has been increasingly steadily, even in high-quality expensive surveys ([National Research Council 2013](#); [B. D. Meyer, Mok, and Sullivan 2015](#)). Nonresponse rates are much higher for commercial telephone surveys, sometimes even as high as 90% ([Kohut et al. 2012](#)). These long-term trends in nonresponse mean that data collection is more expensive and estimates are less reliable. Adapted from [B. D. Meyer, Mok, and Sullivan \(2015\)](#), figure 1.

# Total Survey Error Framework

In each step of the design and analysis phase of a survey, errors can arise that affect the quality of the final statistic of interest.



# What is measurement error?

In the “traditional” survey approach

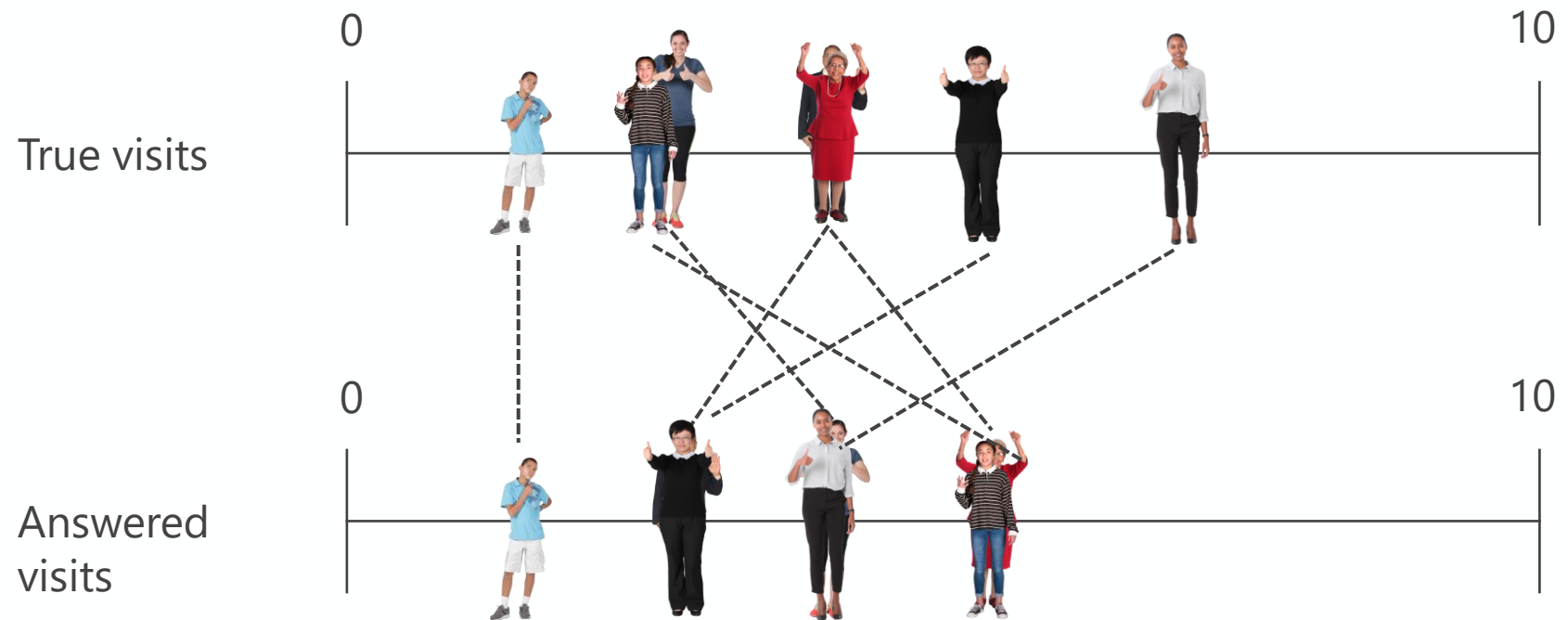
# Why study measurement error?

- Measurement error can bias means, but it especially biases relationships.
- This includes studying change over time.
- That is why it is important to know how much there is!

# Definition

- **Measurement error:** Difference between response and true value.
- **Example:** Mr. Jones says he went to the doctor three times, but actually went four times. Perhaps he forgot he had to go back for his test results. The measurement error is -1.
- **Answered visits** = True visits + Error
- **Observed value** = True value + Error
- **Measurement error:** The answer you have is influenced by things other than the value you are after. These things are defined as "errors".

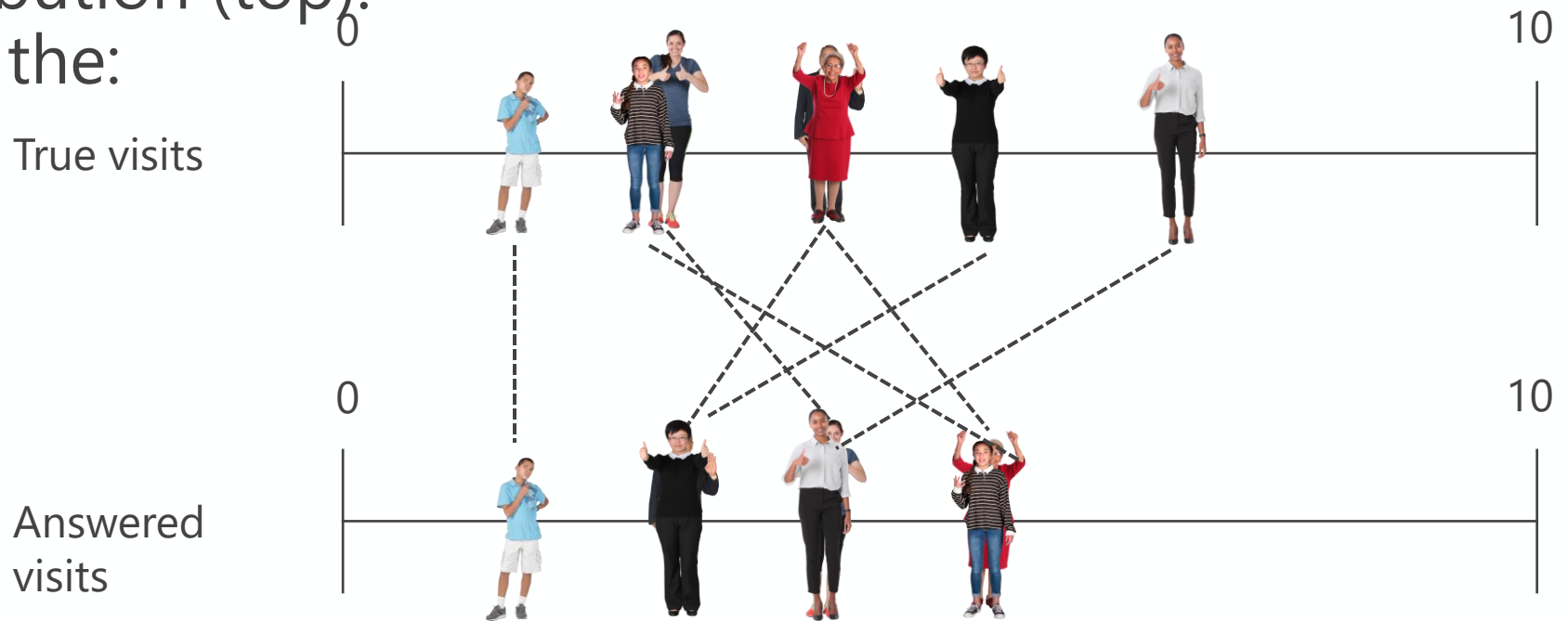
# How often have you visited the doctor in the past month?





# How often have you visited the doctor in the past month?

- The observed distribution of doctor's visit (bottom) is different from the true distribution (top).
- Resulting in bias in the:
  - **Mean**
  - **Variance**



# What is representation error?

In the “traditional” survey approach

# How often have you visited the doctor in the past month?

- Imagine a situation where the answer to this question is related to something else.
- For example: old people visit the doctor more often.
- If we ask this question in an internet survey, some of the old people have no computer and will not fill in the survey.



# How often have you visited the doctor in the past month?

- Imagine a situation where the answer to this question is related to something else.
- For example: old people visit the doctor more often.
- If we ask this question in an internet survey, some old people have no computer and will not fill in the survey.

- Resulting in bias in the **mean**
- And **variance**.

True visits

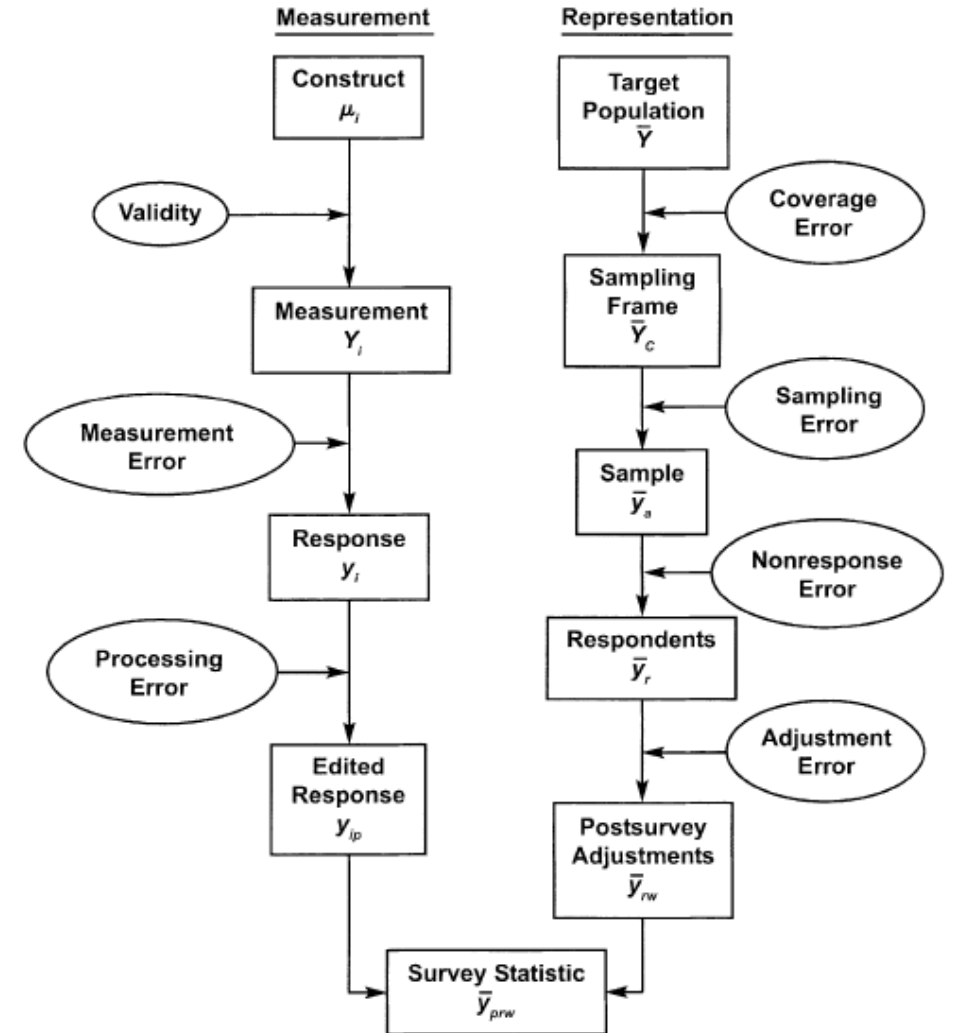


# Errors in digital trace data

Error frameworks

# Total Survey Error Framework

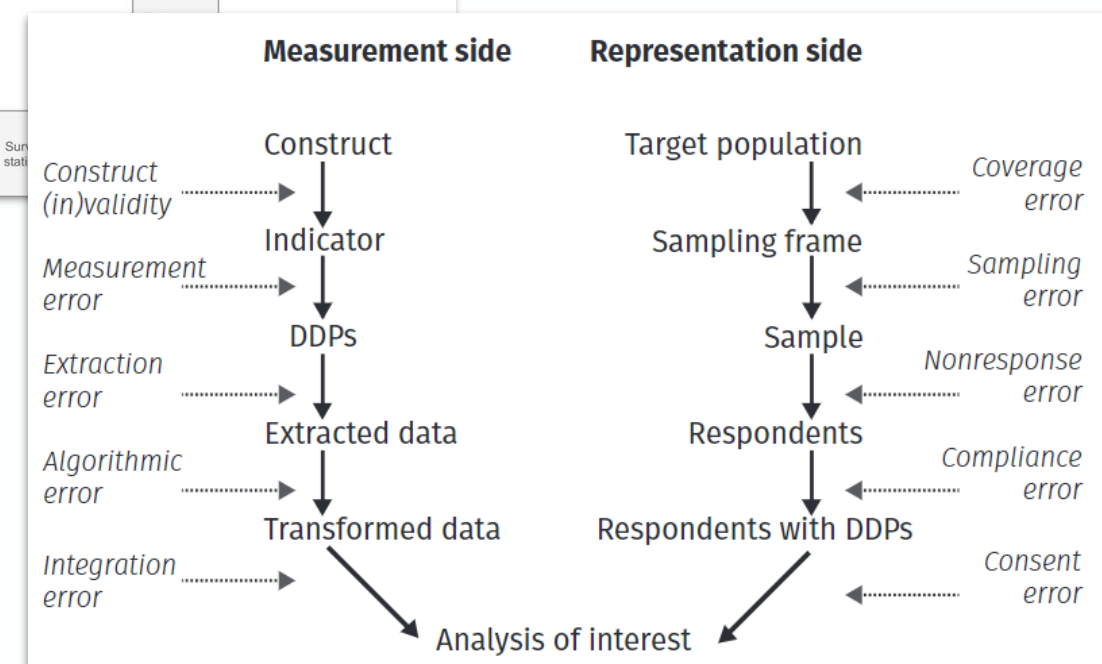
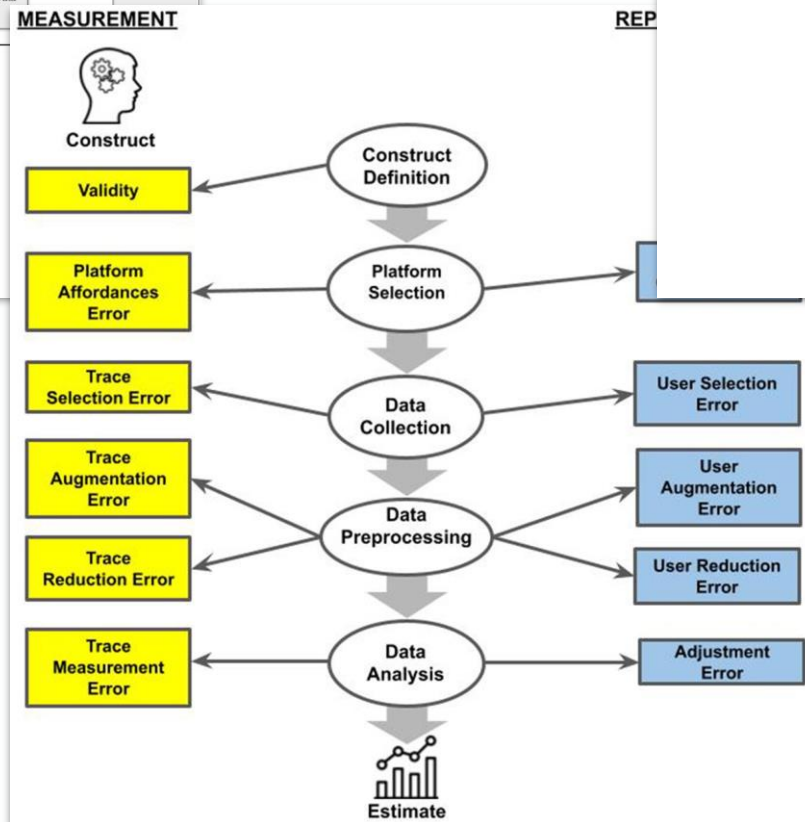
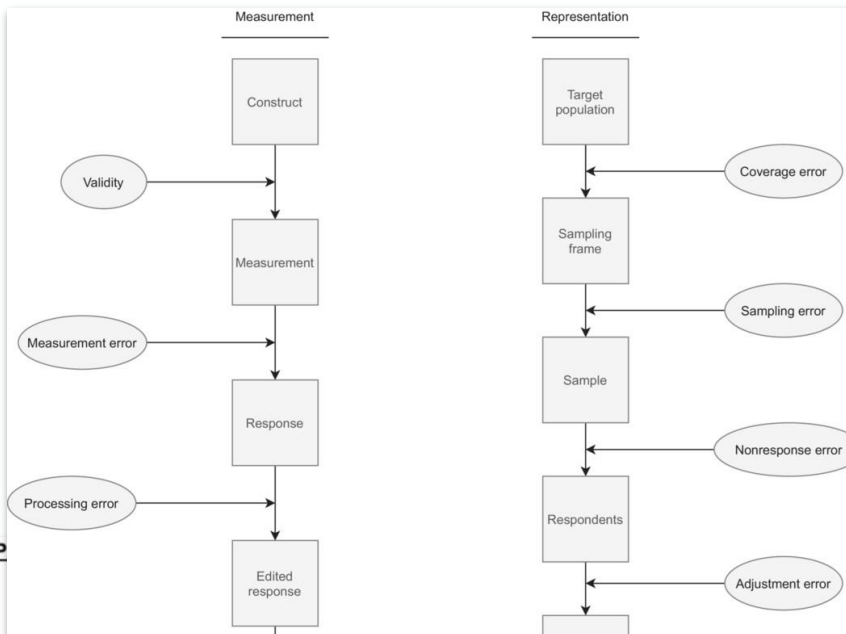
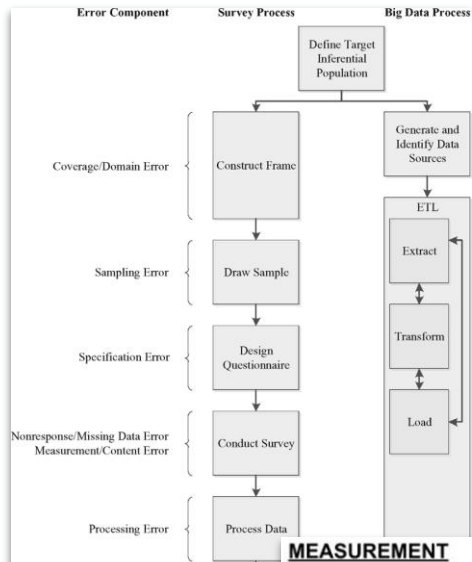
In each step of the design and analysis phase of a survey, errors can arise that affect the quality of the final statistic of interest.



# Error frameworks for digital trace data

- Total Error Framework (TEF) (Amaya et al. 2020)
- Total Error Framework for Digital Traces of Human Behavior on Online Platforms (TED-On) (Sen et al. 2021)
- Total Error Framework for Digital Traces Collected with Meters (TEM) (Bosch & Revilla 2022)
- Total Error for Social Scientific Data Collection with DDPs (Boeschoten, Ausloos, et al. 2022)





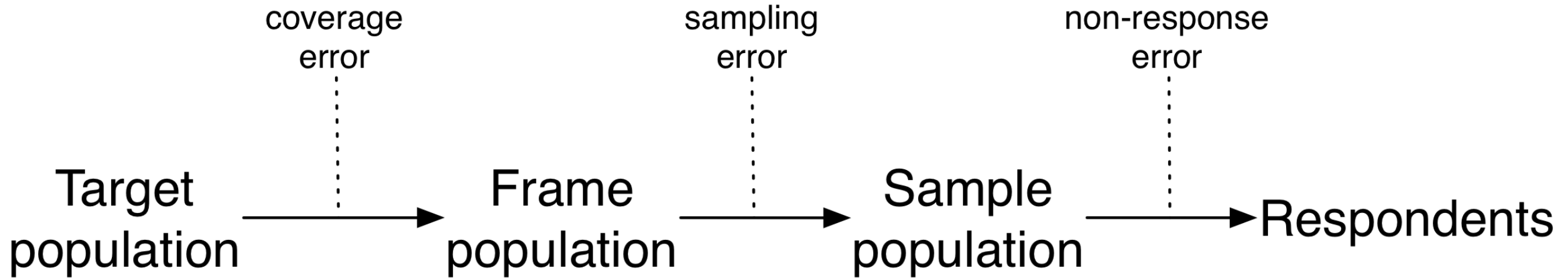
# Error frameworks for digital trace data

- Total Error Framework (TEF) (Amaya et al. 2020)
- Total Error Framework for Digital Traces of Human Behavior on Online Platforms (TED-On) (Sen et al. 2021)
- Total Error Framework for Digital Traces Collected with Meters (TEM) (Bosch & Revilla 2022)
- Total Error for Social Scientific Data Collection with DDPs (Boeschoten, Ausloos, et al. 2022)

**Representation & Measurement**

# Problems with representation

# Problems with representation in **surveys**



# Problems with representation **in DTD**

- Who uses a platform? → Coverage
- Who is willing to share their data and who not? → Non-participation
- Who is able to do everything that is required? → Non-compliance

# Coverage

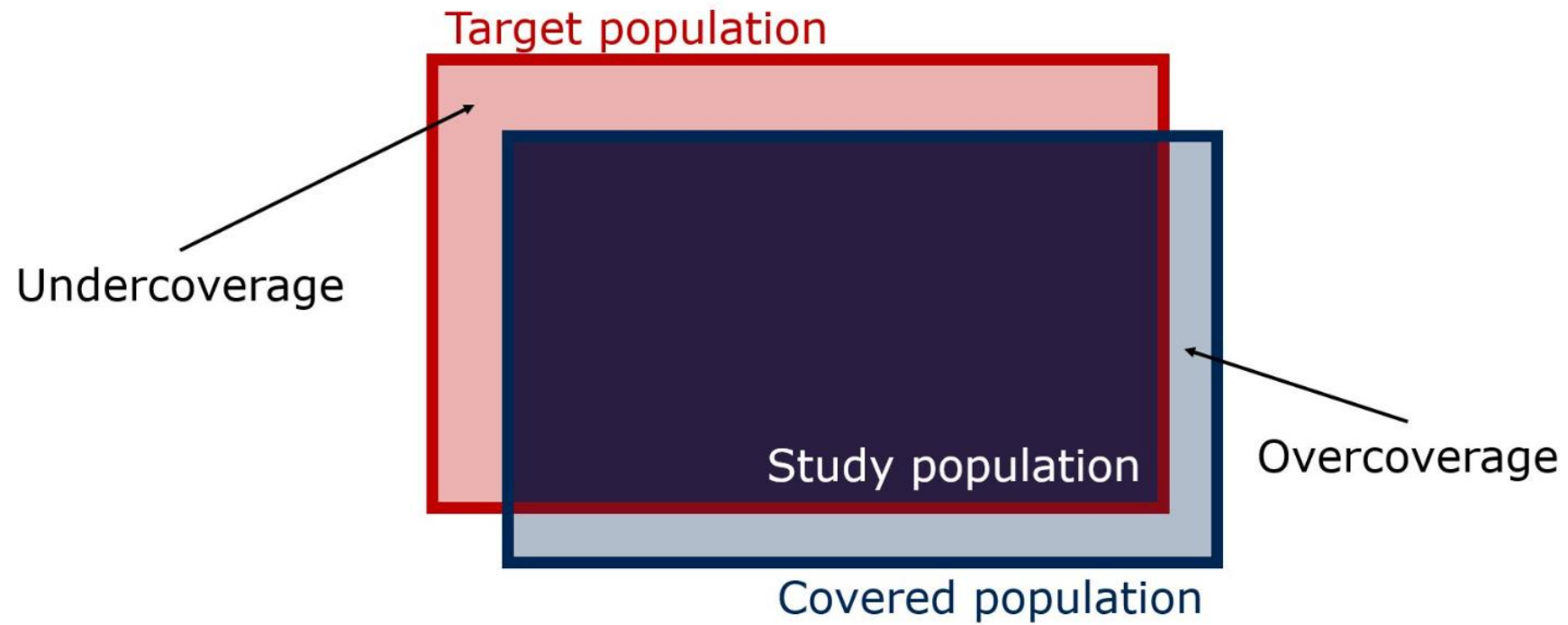
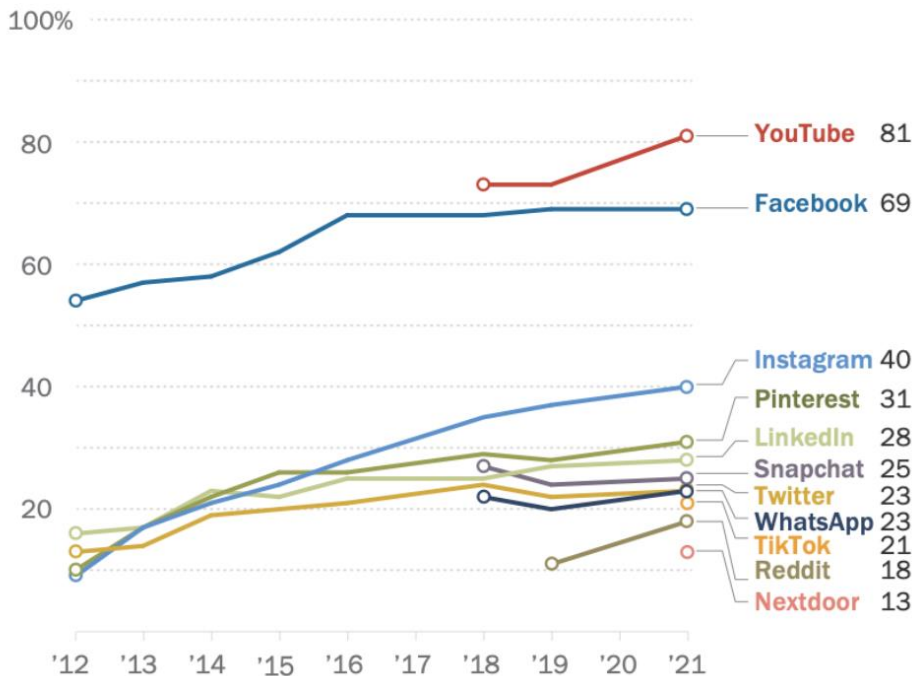


Figure 4.3: Target population, covered population, and study population

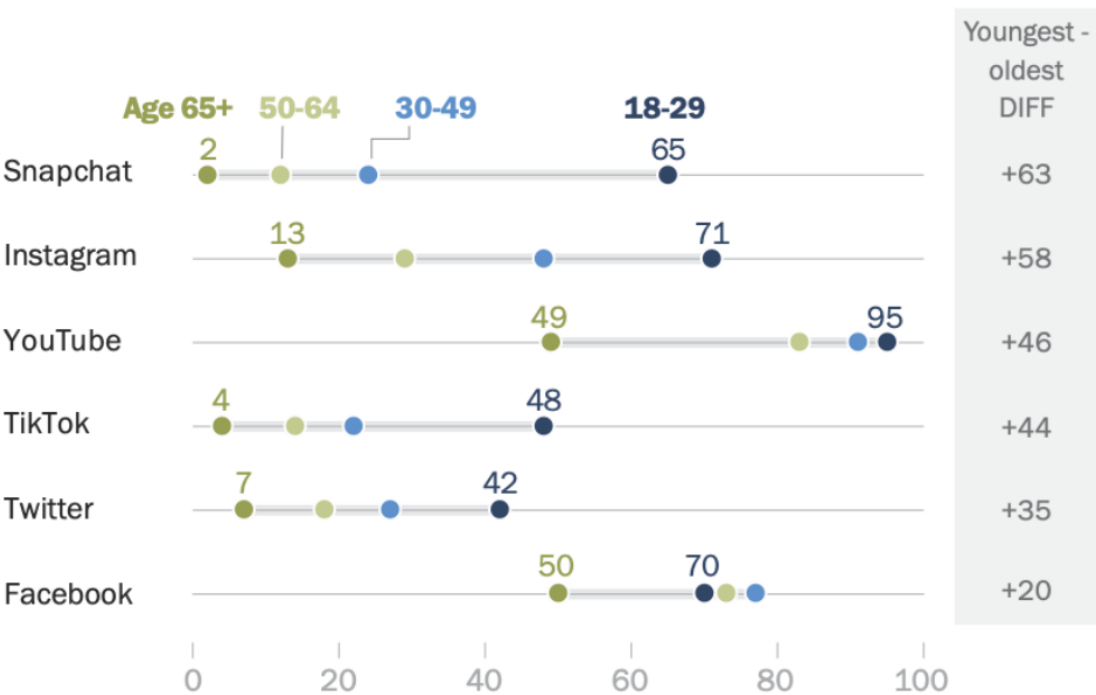
# Social media platform use

% of U.S. adults who say they ever use ...



Note: Respondents who did not give an answer are not shown. Pre-2018 telephone poll data is not available for YouTube, Snapchat and WhatsApp; pre-2019 telephone poll data is not available for Reddit. Pre-2021 telephone poll data is not available for TikTok. Trend data is not available for Nextdoor.  
Source: Survey of U.S. adults conducted Jan. 25-Feb. 8, 2021.  
"Social Media Use in 2021"

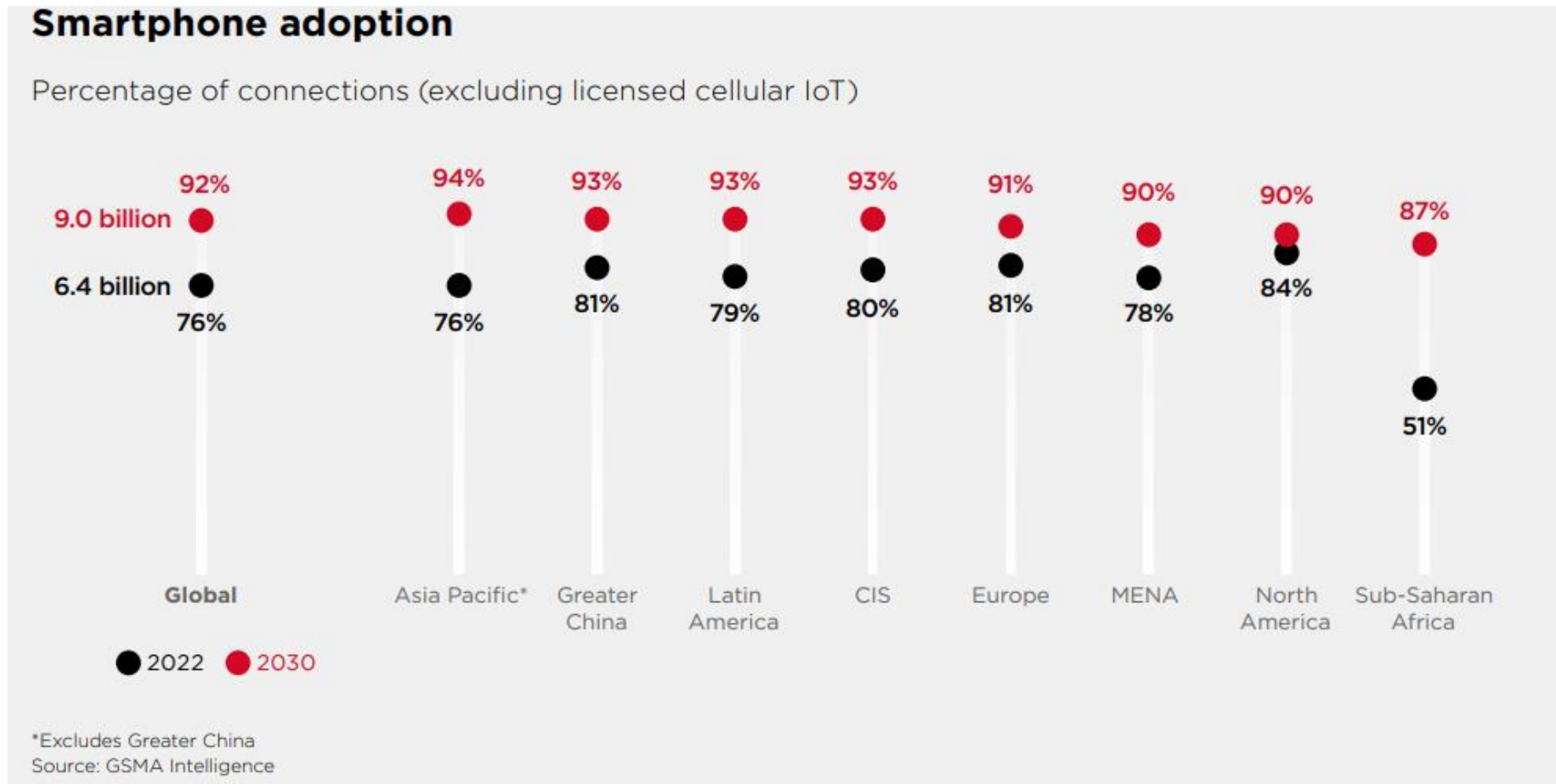
% of U.S. adults in each age group who say they ever use ...



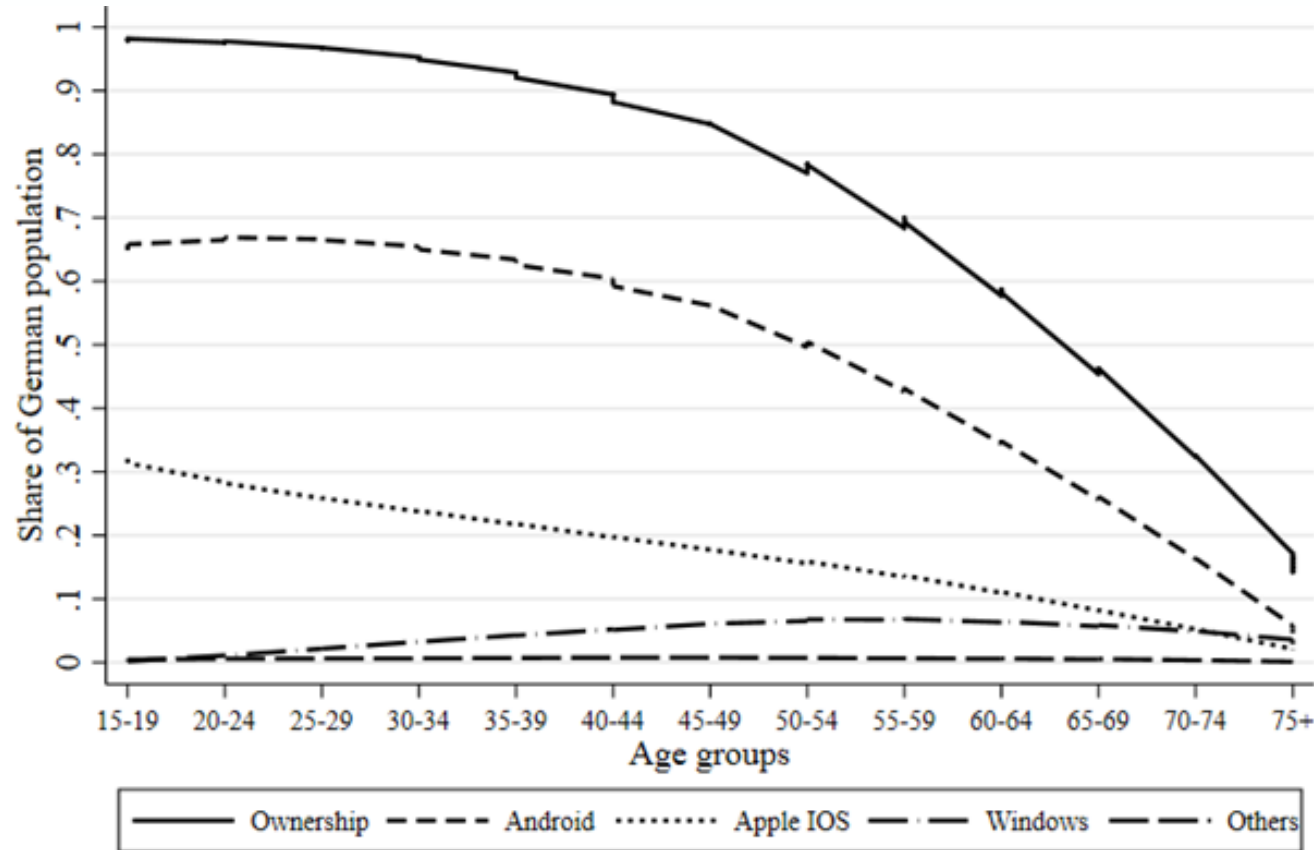
Note: All differences shown in DIFF column are statistically significant. The DIFF values shown are based on subtracting the rounded values in the chart. Respondents who did not give an answer are not shown.



# Coverage smartphones



# Smartphone coverage bias (Keusch et al. 2020; Data for Germany)

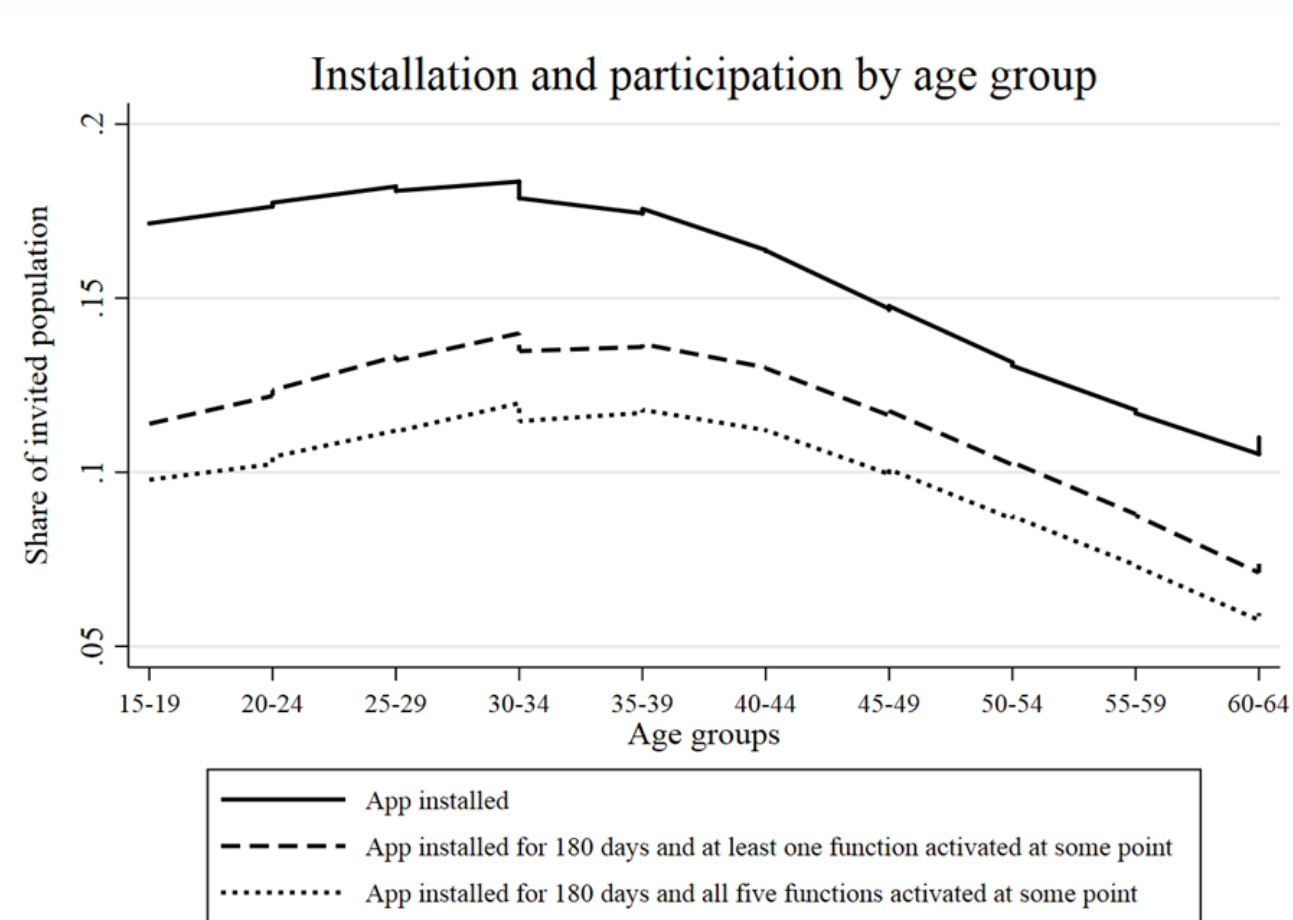


- Smartphone ownership also correlates with...
  - Educational attainment
  - Nationality
  - Region
  - Community size
- Bias of ownership rel. small for many substantive measures
  - But substantial bias for iPhone ownership

# Problems with representation **in DTD**

- Who uses a platform? → Coverage
- Who is willing to share their data and who not? → Non-participation
- Who is able to do everything that is required? → Non-compliance

# Nonparticipation in research app studies



Keusch, Bähr, et al. (2022)

Reasons not to participate	
Privacy, data security concerns	44%
No incentive, incentive too low	17%
Not enough information provided	12%
Do not download any apps	8%
Not interested	6%
Not enough time/Study too long	5%
Don't use smartphone enough	5%
Not enough storage	1%
Other reasons	6%

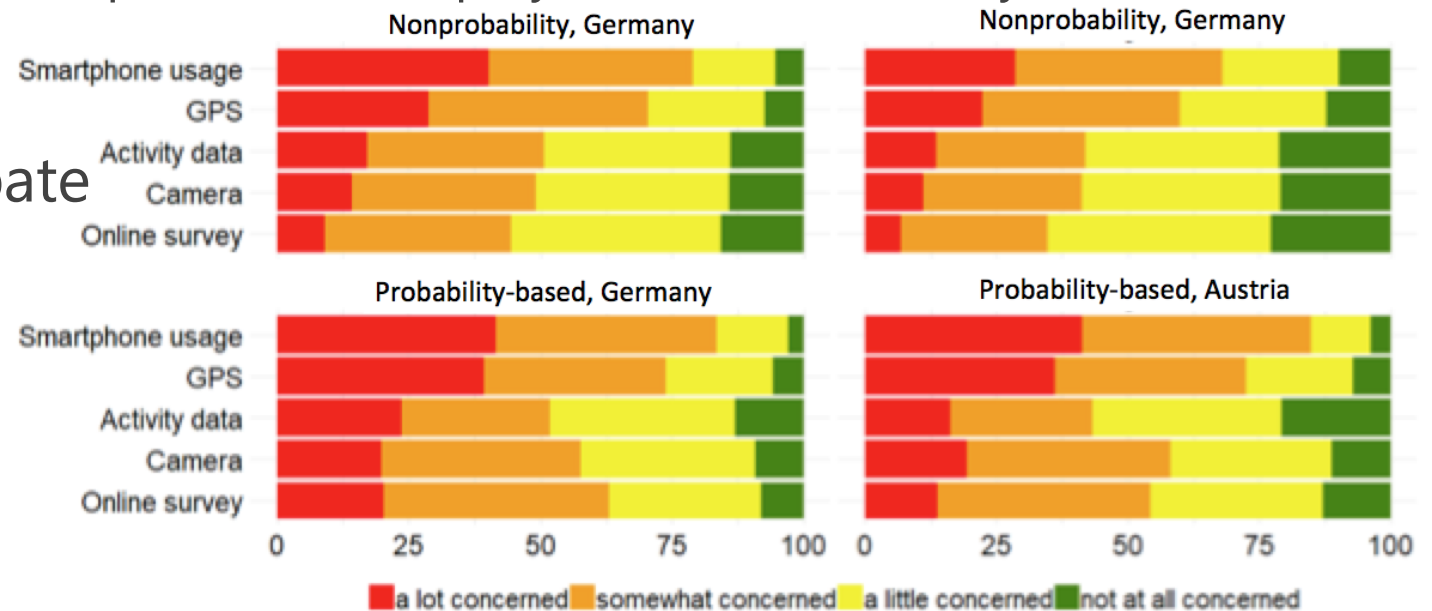
n=1,154

Keusch et al. (2019)

# Mechanisms of (non-)participation: Privacy concern

- Participants might have concerns about potential risks related to sensor data
  - Data streams could be intercepted by unauthorized party
  - Connecting multiple streams of data could re-identify previously anonymous users
  - Information could be used to impact credit, employment, or insurability
- Higher privacy & security concerns correlate with lower willingness to participate

(Keusch, et al. 2019; Revilla et al. 2019; Struminskaya et al. 2020; 2021; Wenz et al. 2019; Wenz & Keusch in press)



# Other Mechanisms of (non-)participation

- **Agency:** WTP higher for tasks where participants have agency over data collection (Revilla et al. 2019; Keusch et al. 2019; Struminskaya et al. 2020; 2021; Wenz & Keusch in press)
- **Sponsor:** WTP higher for university sponsor vs. market research and statistical office (Keusch et al. 2019; Struminskaya et al. 2020)
- **Framing:** emphasizing benefits does not influence WTP (Struminskaya et al. 2020; 2021)
- **Smartphone skills:** more activities on smartphone (e.g., using GPS, taking pictures, online banking, etc.) correlates with higher WTP (Keusch et al. 2019; Struminskaya et al. 2020; 2021; Wenz et al. 2019; Wenz & Keusch in press)
- **Experience:** prior research app download increases WTP (Keusch et al. 2019; Struminskaya et al. 2020; 2021)
- **Sociodemographics:** educational attainment (Jäckle et al. 2019; Keusch et al. 2021, 2022; McCool et al. 2021; Wenz & Keusch in press) and age (Jäckle et al. 2019; McCool et al. 2021; Keusch et al. 2022; Wenz & Keusch in press) correlated with WTP

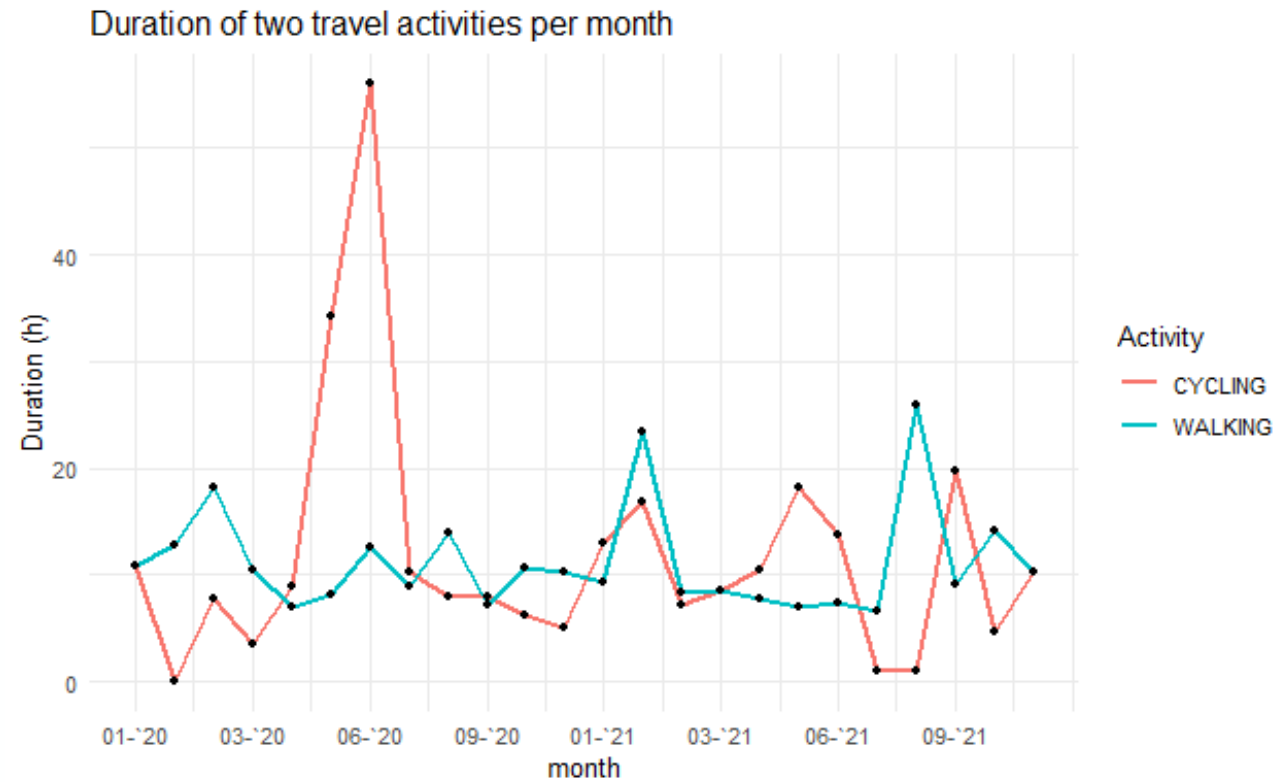
# Problems with representation **in DTD**

- Who uses a platform? → Coverage
- Who is willing to share their data and who not? → Non-participation
- Who is able to do everything that is required? → Non-compliance



# Willingness and compliance

Google Semantic Location History



# Willingness and compliance

- Study in Dutch online panel (CentERpanel)
- Google Semantic Location History data from DDP
- N=1,035 (75% AAPOR RR1)
- Integration of survey and data donation software (PORT)
- 30% willing, 14% eventually donated
  - Understanding of consent request sign. increased willingness and successful donation
  - Male, higher educated, and more technologically savvy more likely to donate

## Cycling

		Duration (hours)	Distance (km)
Year	Month		
2021	8	1.14	6.32

## In Bus

		Duration (hours)	Distance (km)
Year	Month		
2021	8	1.97	28.23

## In Passenger Vehicle

		Duration (hours)	Distance (km)
Year	Month		
2021	8	23.31	375.84

# Understanding of request to share

Statements asked to respondents	Correct %	Incorrect %	Don't know %
You are asked to download information from Google. TRUE	48.8	19.8	31.4
The software implemented in the survey will extract the information on the number of hours you cycle, walk, take public transport, travel by car. TRUE	62.3	6.1	31.2
Information on all the locations you visited will be shared with Centerdata. FALSE	39.2	31.4	29.4
Google collects information on location about everyone. FALSE	24.8	46.6	28.5
From the data you will provide, the information can be traced back to you. FALSE	45.3	22.2	32.5
You will be able to inspect the data before sending it to Centerdata. TRUE	59.0	7.8	33.1
It is impossible to identify you as an individual from the data that you provide. TRUE	43.4	19.6	37.0

# Understanding the consent request

- **5.5% had everything correct**
- Mean correct: 3.23, median = 4 (out of 7 questions)
- **People with more correct answers more likely to be willing & to donate:**
  - 4.54 correct statements for willing
  - 2.56 correct statements for non-willing
  - OR = 1.572,  $p < .001$
  - 5.33 correct statements for donated
  - 3.94 correct statements for not donated
  - OR = 1.795,  $p < .001$

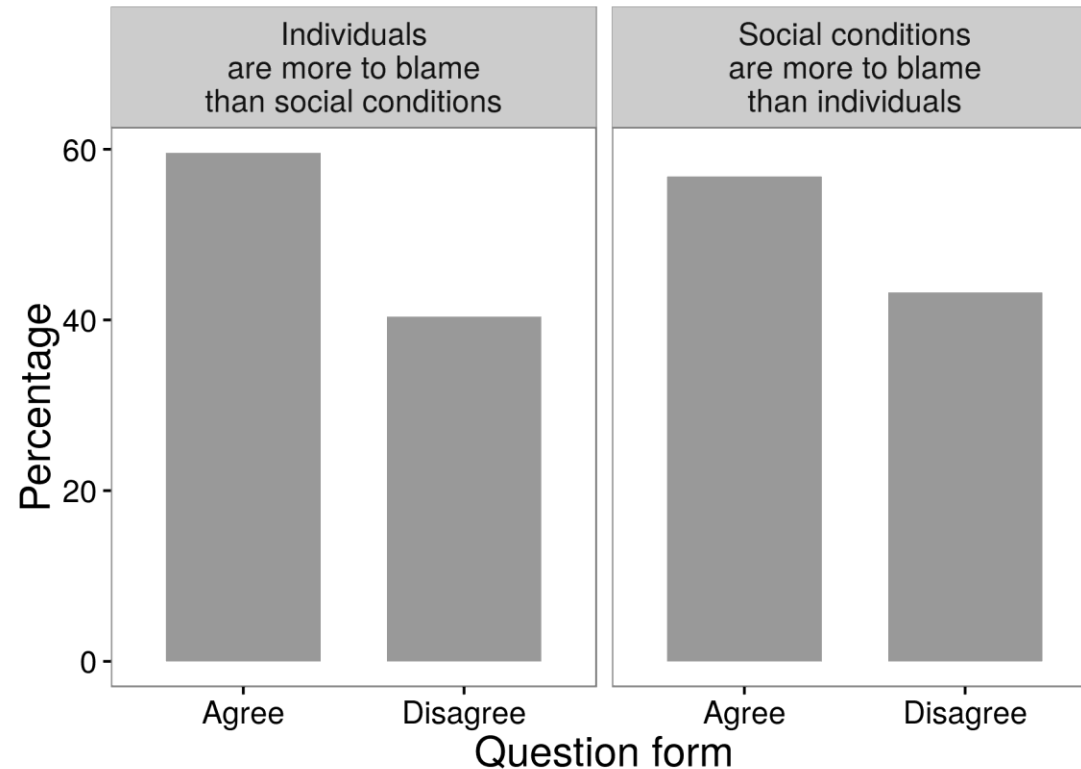
Coffee break



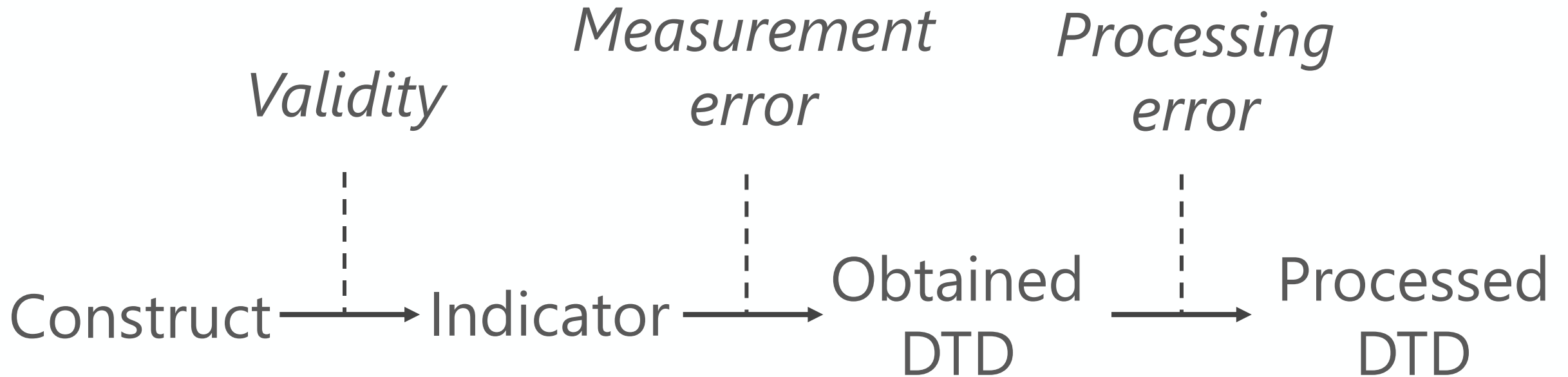
# Problems with measurement

# Measurement problems in surveys

How you ask a question matters!



# Measurement problems in **DTD**



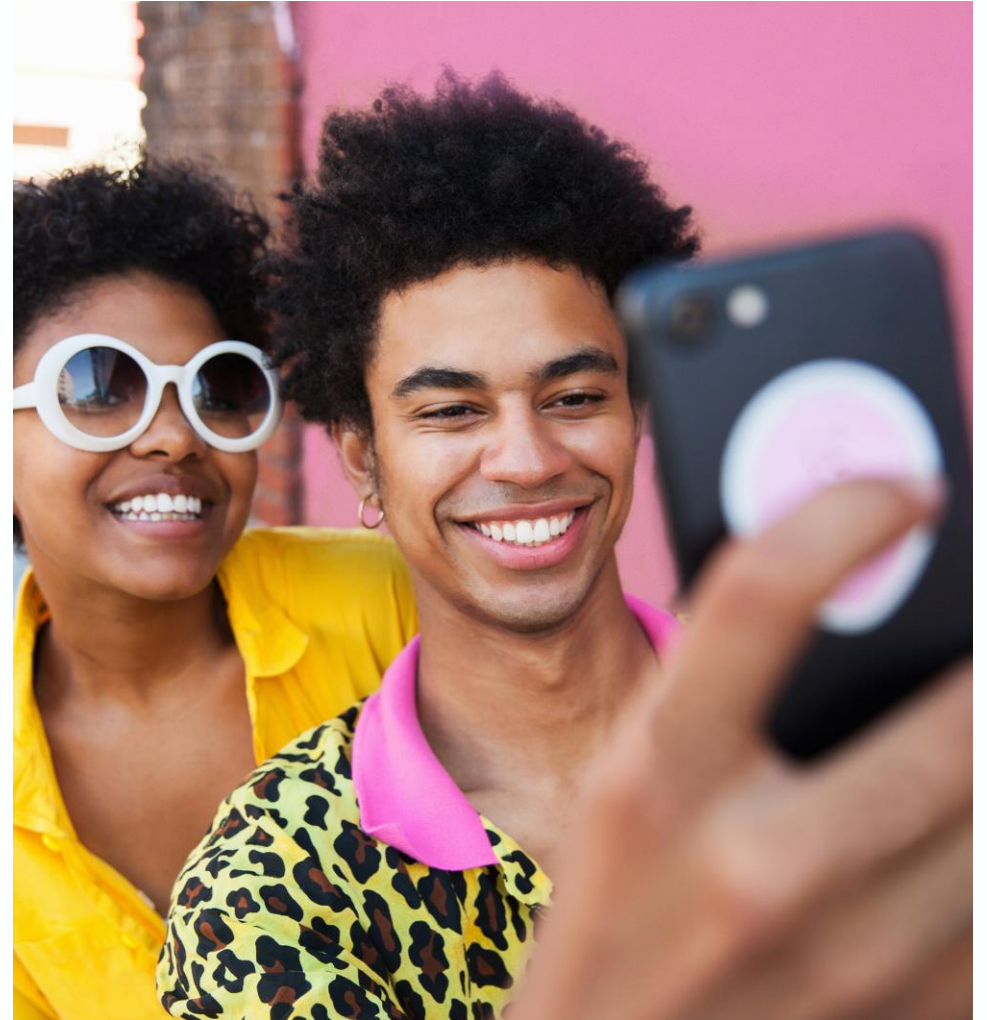
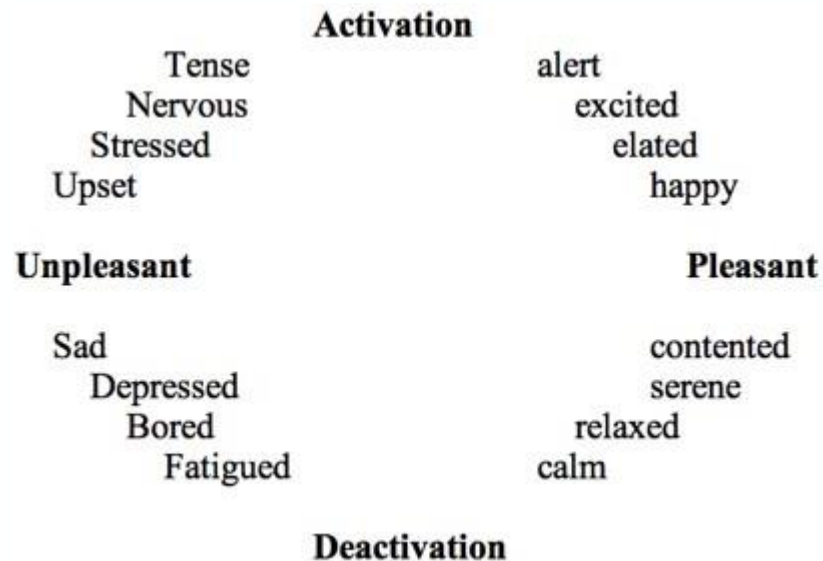


# Measurement problems **in DTD**

- Are you measuring what you want to measure? → Validity
- Are your measurements correct? → Measurement error
  - Can be an error on the platform
  - Or an error in your app/plug-in/tool!
- What do you need to do to get your measurements of interest from the data? → Processing error

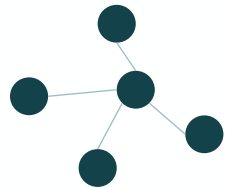
# Validity

- Example:
- We are interested in someone's mood.
- We use “facial expression on photo” as an indicator and collect photos through data donation.
- Are we measuring the concept appropriately?

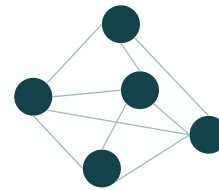


# Validity

- Facebook uses the “clustering coefficient” to recommend friends: e.g., if you have two friends, Sanne and Joep, that are not Facebook friends, Facebook will suggest Sanne and Joep to add each other as friends.
- Your measurement of social closure (clustering coefficient) is measuring *both* social closure and the effect of the algorithm → *it is algorithmically infused*



Low clustering



High clustering

# Measurement problems **in DTD**

- Are you measuring what you want to measure? → Validity
- Are your measurements correct? → Measurement error
  - Can be an error on the platform (wrong/incomplete)
  - Or an error in your app/plugin/tool!
- What do you need to do to get your measurements of interest from the data? → Processing error

# Measurement error

- We are still interested in someone's mood.
- Imagine using "facial expression on photo" is a good indicator.
- We use the Screenomics app.
- **Is this a correct representation of all your facial expressions on all photo's?**



# What devices are tracked?

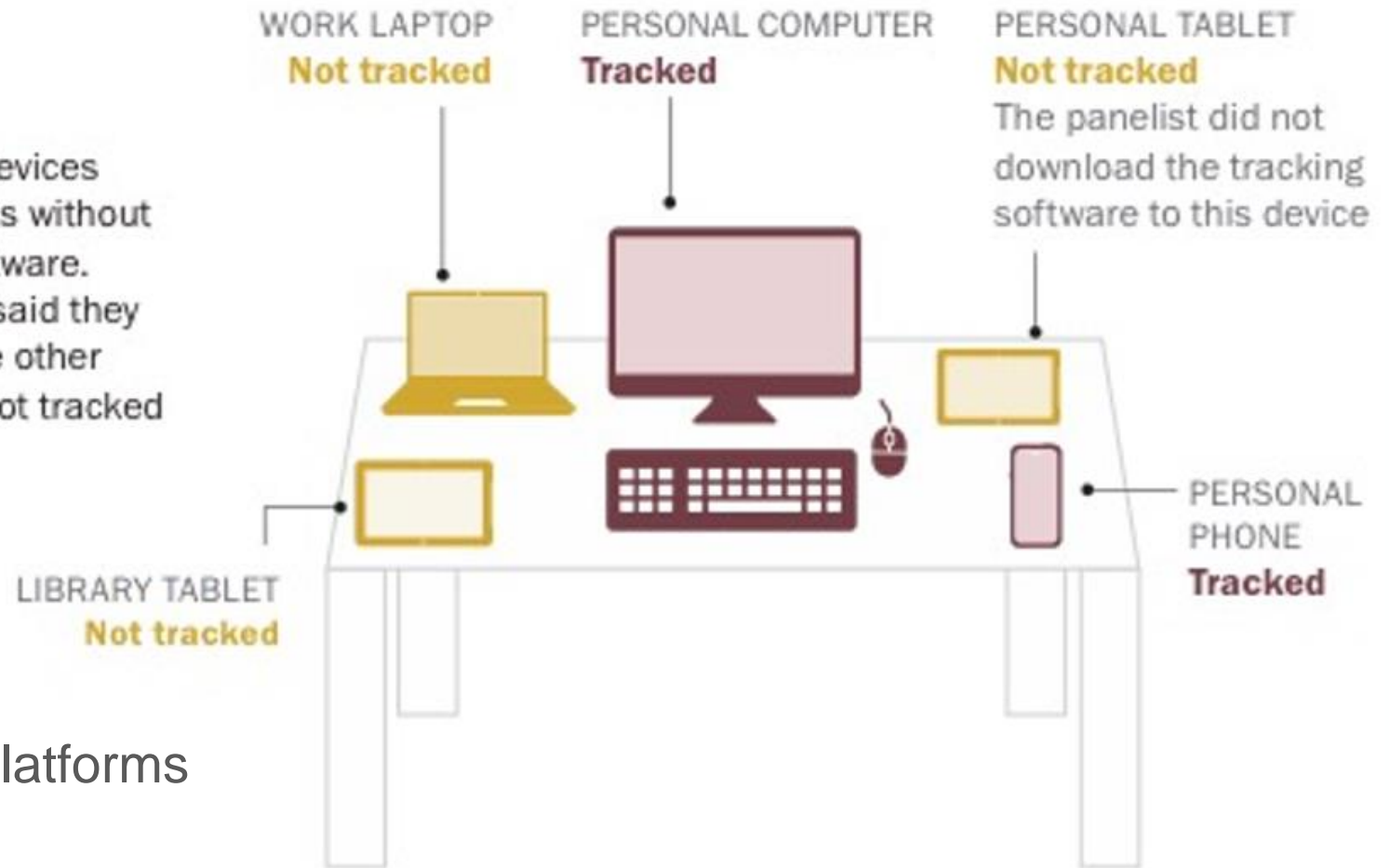
## DEVICE TRACKING

### Tracked:

Devices tracked in this study included computers, tablets and mobile phones used by panelists that they downloaded tracking software onto

### Not tracked:

Any additional devices used by panelists without the tracking software. Many panelists said they had one or more other digital devices not tracked

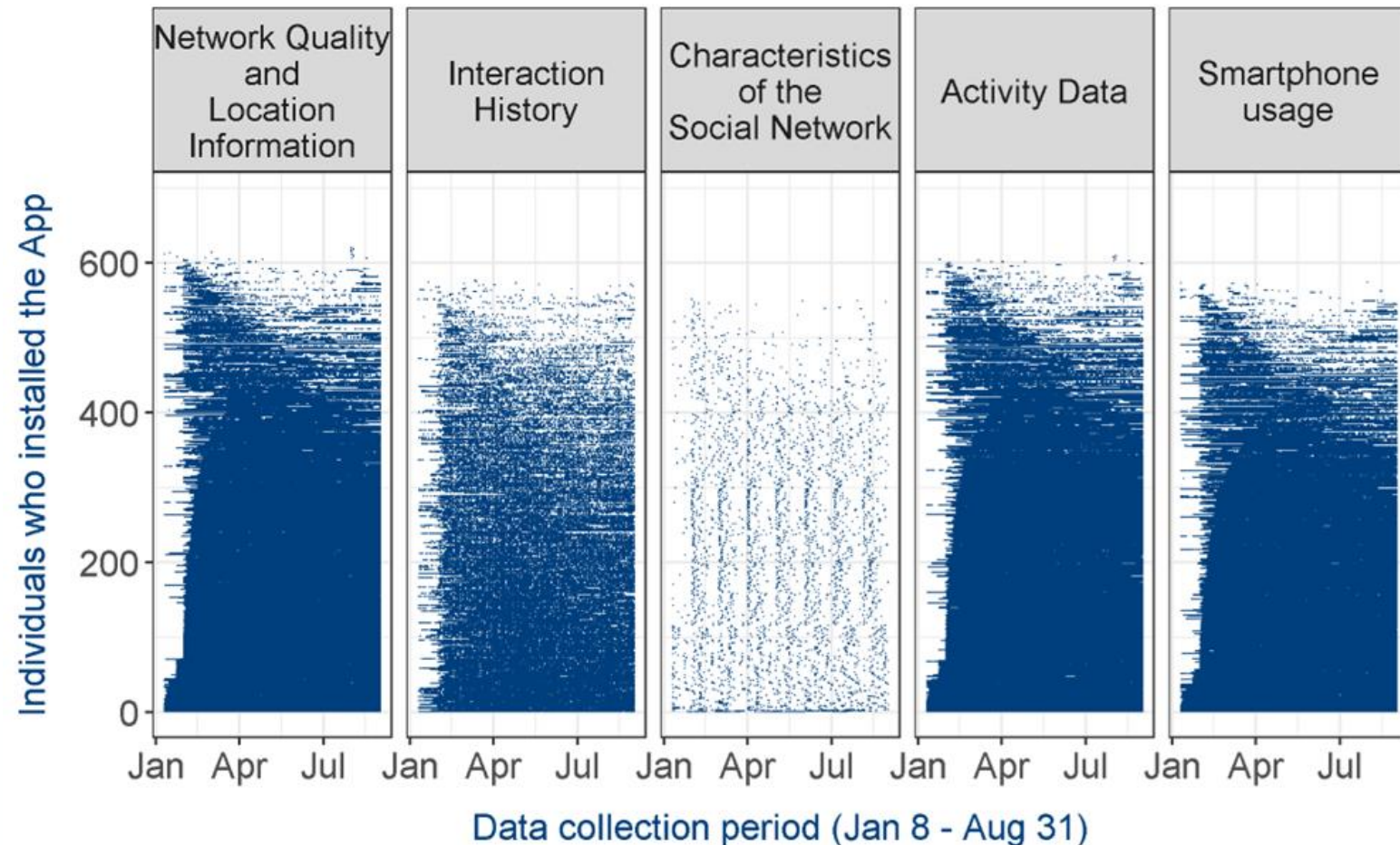


Also holds for use of different platforms  
(e.g., WhatsApp vs. Facebook)



# Measurement problems in DTD

- Missing data

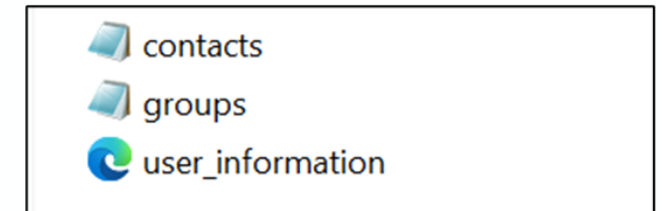
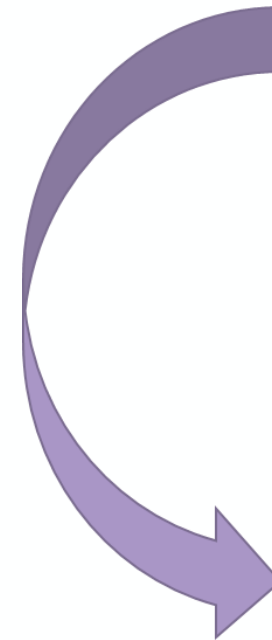


# Volatility of Data Download Packages (DDP)

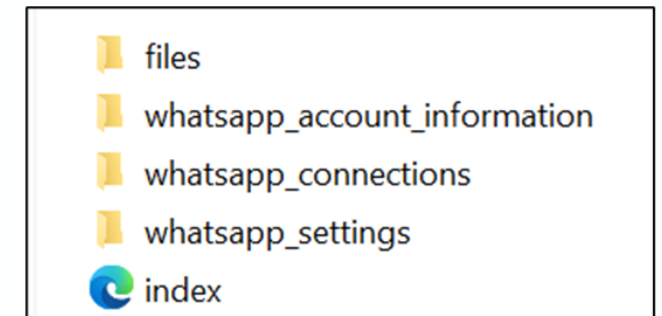
Example:

## WhatsApp DDP Folder structure

- Diversity within DDPs of the same platform (content and structure)
- Volatility complicates extraction of data donation
- Examples:
  - Change over time
  - Change over operating systems (Android, Apple)
  - Differences over languages



July 2022



August 2022





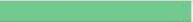
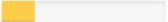





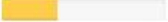








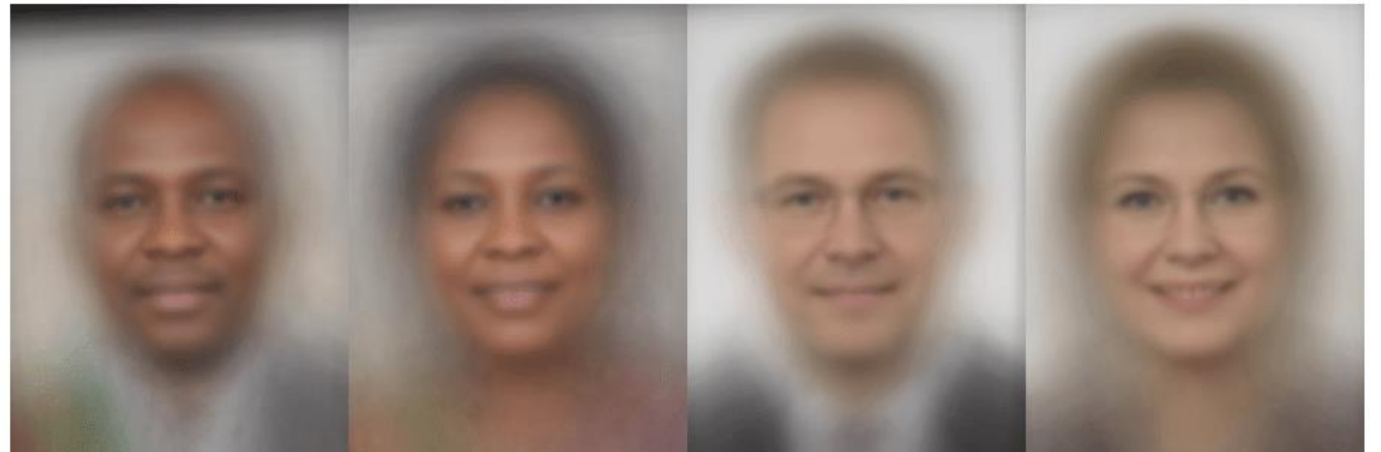
# Measurement problems in DTD

- Are you measuring what you want to measure? → Validity
- Are your measurements correct? → Measurement error
  - Can be an error on the platform
  - Or an error in your app/plug-in/tool!
- What do you need to do to get your measurements of interest from the data? → Processing error

# Processing error

- Studies show facial recognition software almost works perfectly – if you're a white male.

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0% 	79.2% 	100% 	98.3% 	20.8% 
 FACE++	99.3% 	65.5% 	99.2% 	94.0% 	33.8% 
 IBM	88.0% 	65.3% 	99.7% 	92.9% 	34.4% 



**“External” errors**

# POWER OF DATA



- Total Error Frameworks evaluate what can go wrong on a study level.
- If we zoom out, what can go wrong, or do we miss on a societal level when using (existing) digital traces for research purposes?

# Why data feminism?

- Feminism refers to the diverse and wide-ranging projects that name and challenge sexism and other **forces of oppression**, as well as those which seek to **create** most **just, equitable** and **livable** futures.
- Data feminism is a way of thinking about data, both their **uses** and their **limits**, that is informed by **direct experience**, by commitment to **action**, and by **intersectional** feminist thoughts.

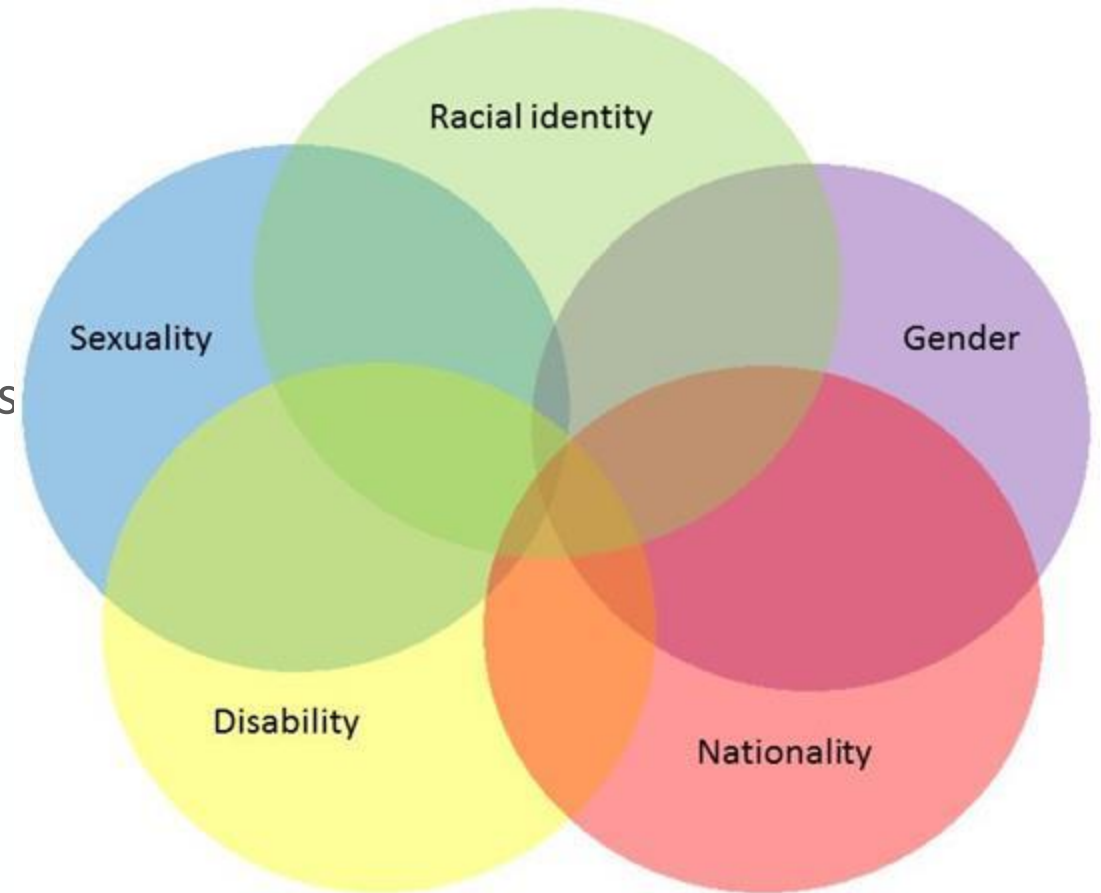
# Some definitions

**Oppression:** The systematic mistreatment of certain groups of people by other groups.

**Intersectionality:** The study of overlapping or intersecting social identities and related systems of oppression, domination, or discrimination

**Privilege hazard:** Forces of oppression can be difficult to detect when you benefit from them.

**Co-liberation:** Oppressive systems of power harm us all, they undermine the quality and validity of our work, and hinder us from creating true and lasting social impact with data science.



# Data feminism

Governments and corporations have employed data and statistics as management techniques to preserve an unequal status quo.

Data feminism:

- Acknowledges this history.
- Uses data to overcome these inequalities.

# Seven principles:



**Examine power**



Challenge power



Elevate emotion  
and embodiment



Rethink binaries  
and hierarchies



Embrace  
pluralism



Consider context



Make labor  
visible



# Examining power

- **Complexity of intersections:** Name and explain the forces of oppression that are so baked in our daily lives and our datasets, our databases and algorithms, that we often don't even see them.
- **Power:** The current configuration of structural privilege and structural oppression, in which some groups experience unearned advantages and other groups experience systematic disadvantages.



Serena Williams

January 15, 2018 · Facebook Creator ·

I didn't expect that sharing our family's story of Olympia's birth and all of complications after giving birth would start such an outpouring of discussion from women — especially black women — who have faced similar complications and women whose problems go unaddressed.


These aren't just stories: according to the CDC, (Center for Disease Control) black women are over 3 times more likely than White women to die from pregnancy- or childbirth-related causes. We have a lot of work to do as a nation and I hope my story can inspire a conversation that gets us to close this gap.

Let me be clear: EVERY mother, regardless of race, or background deserves to have a healthy pregnancy and childbirth. I personally want all women of all colors to have the best experience they can have. My personal experience was not great but it was MY experience and I'm happy it happened to me. It made me stronger and it made me appreciate women -- both women with and without kids -- even more. We are powerful!!!

I want to thank all of you who have opened up through online comments and other platforms to tell your story. I encourage you to continue to tell those stories. This helps. We can help others. Our voices are our power.

   27K

1.6K Comments 1.4K Shares 840K Views

 Share

# What do we mean by power?

Table 1.1: The four domains of the matrix of domination <sup>14</sup>	
<b>Structural domain</b>  Organizes oppression: laws and policies.	<b>Disciplinary domain</b>  Administers and manages oppression. Implements and enforces laws and policies.
<b>Hegemonic domain</b>  Circulates oppressive ideas: culture and media.	<b>Interpersonal domain</b>  Individual experiences of oppression.

# Ask the questions:

- Data science by whom?
- Data science for whom?
- Data science with whose interests and goals?



# Ask the questions:

- **Data science by whom?**
- Data science for whom?
- Data science with whose interests and goals?

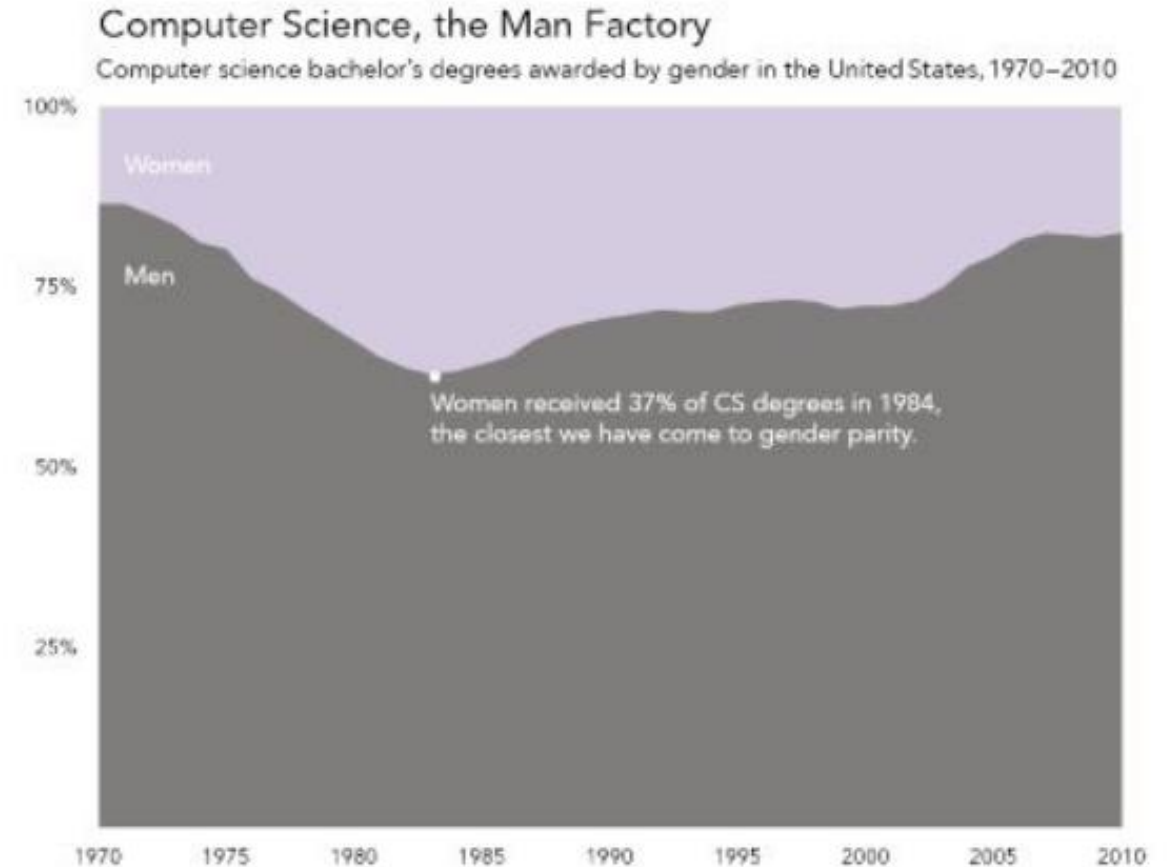


Figure 1.2: Computer science has always been dominated by men and the situation is worsening (even while many other scientific and technical fields have made significant strides toward gender parity). Women awarded bachelor's degrees in computer science in the United States peaked in the mid-1980s at 37 percent, and we have seen a steady increase in the ratio of men to women in the years since then. This particular report treated gender as a binary, so there was no data about nonbinary people. Graphic by Catherine D'Ignazio. Data from the National Center for Education Statistics. *Source:* Data from Christianne Corbett and Catherine Hill, *Solving the Equation: The Variables for Women's Success in Engineering and Computing* (Washington, DC: American Association of University Women, 2015). *Credit:* Graphic by Catherine D'Ignazio.



# Ask the questions:

- Data science by whom?
- **Data science for whom?**
- Data science with whose interests and goals?



# Ask the questions:

- Data science by whom?
- Data science for whom?
- **Data science with whose interests and goals?**

1. **Whose goals are prioritized in data science (and whose are not)?**
2. **What goals and purposes are going underserved?**
3. **Who is in charge of the institutions?**
4. **Who benefit the most from the status quo?**

# Exercise

Answer the questions on the previous slides for the Serena Williams case.

1. Whose goals are prioritized in data science (and whose are not)?
2. What goals and purposes are going underserved?
3. Who is in charge of the institutions?
4. Who benefit the most from the status quo?



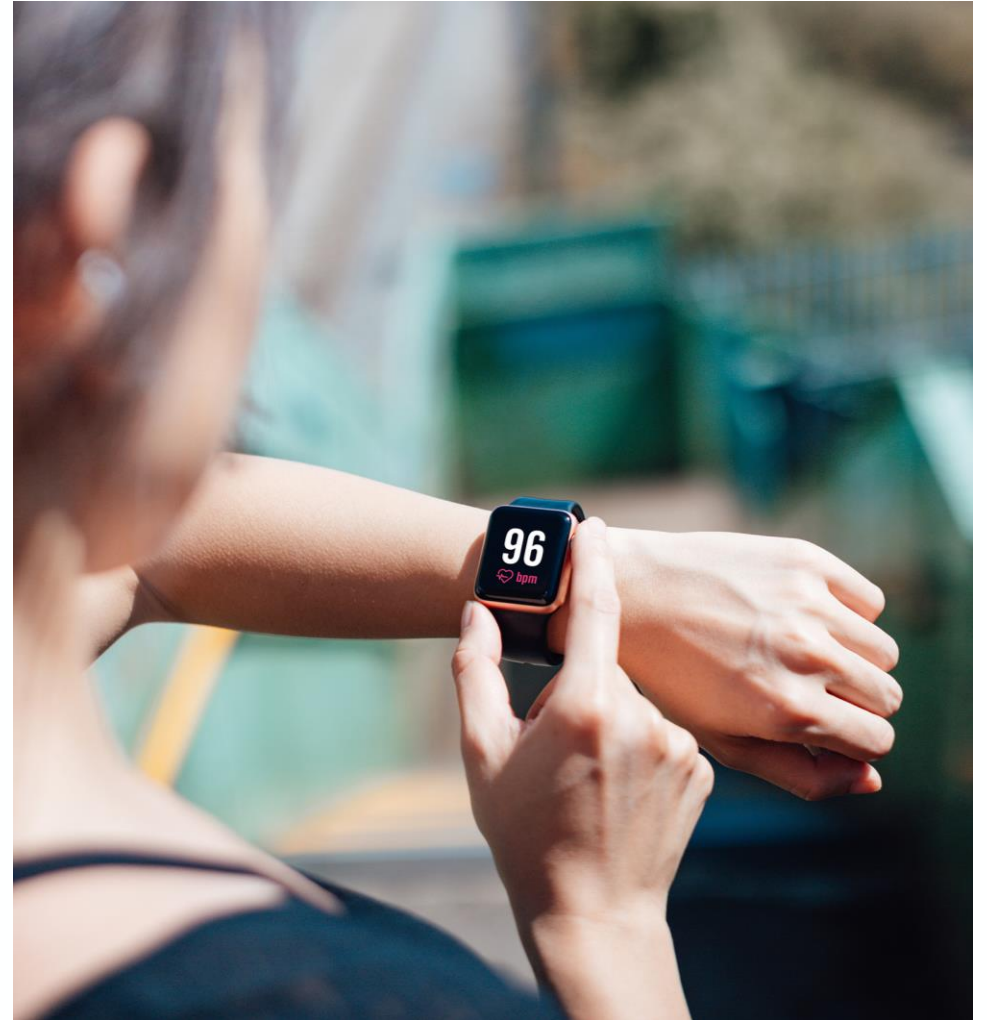


# Moving to digital trace data collection

What can we do?

- Examine power
- Challenge power
- Change power

Ask questions about other studies, but also your own!



A top-down view of a restaurant table. Two burgers with sesame seed buns and fries are served on wooden trays. A glass of iced coffee with a thick white foam is on the right, and a glass of orange juice is on the left. There are also condiment containers, a pepper mill, and a salt shaker. The table is set with white napkins, forks, and knives. The text "See you after lunch!" is overlaid in the center.

**See you after lunch!**