Errors in digital trace data collection

Lecture 4

Literature:

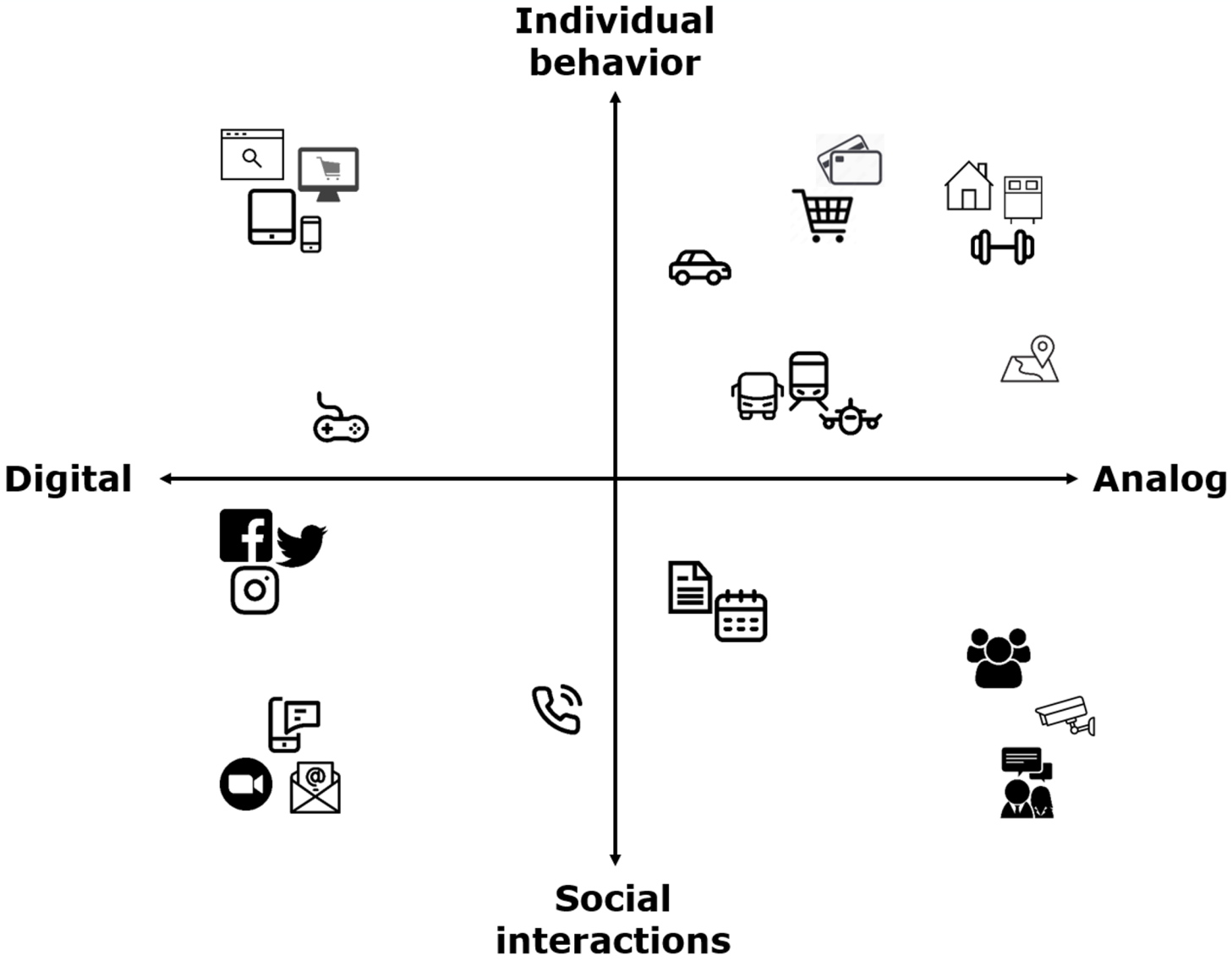
- **Chapter 3**: "Asking questions" in *Bit by bit: Social science in the digital age (2017)*, Matthew Salganik **3.1 – 3.4 & 3.6**
- **Chapter 4**: "Errors of nonobservation: Sampling and coverage" in *Data collection with Wearables, Apps and Sensors (2023)*, Florian Keusch, Bella Struminskaya, Stephanie Eckman & Heidi Guyer
- **Corten et al**., (2024) Assessing Mobile Instant Messenger Networks with Donated Data.
- **Chapter 3**: "Record Linkage" in Big data and Social Science **3.1 & 3.2**

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. Understand how enriched and amplified asking work and their differences.
5. Understand the role of record linkage in designed big data.

Recap week 1

Where and when do you leave digital traces?



Keusch & Kreuter (2022)

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

From Wikipedia, the free encyclopedia

For the publication, see Survey Methodology.

Survey methodology is "the study of survey methods".[1] 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?

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

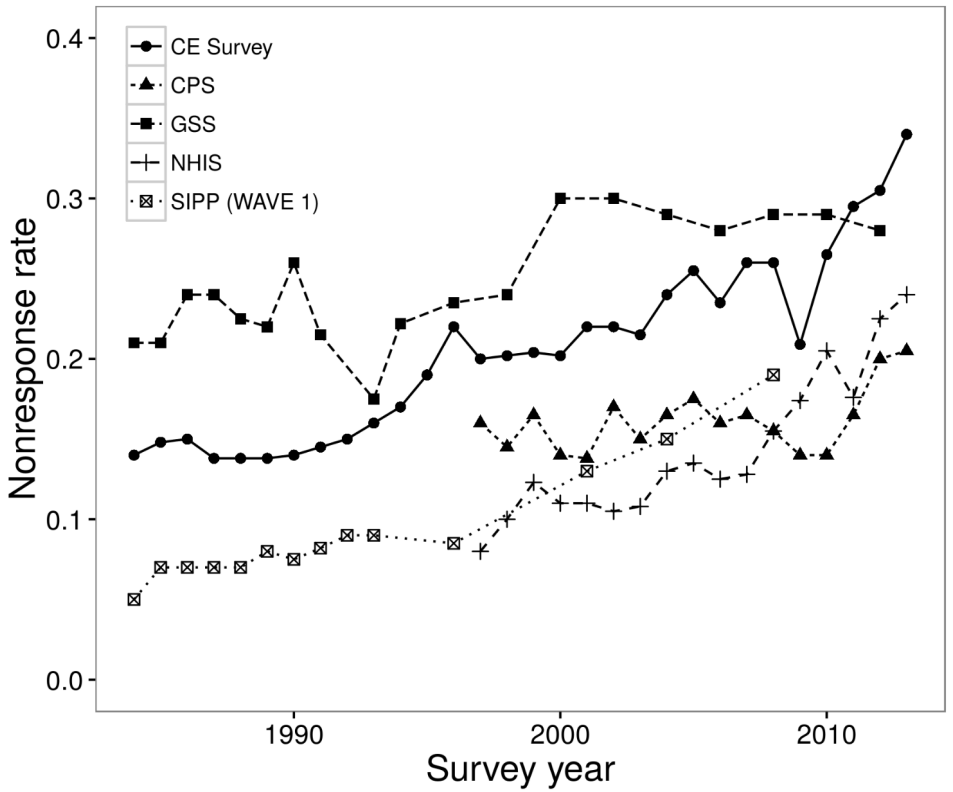
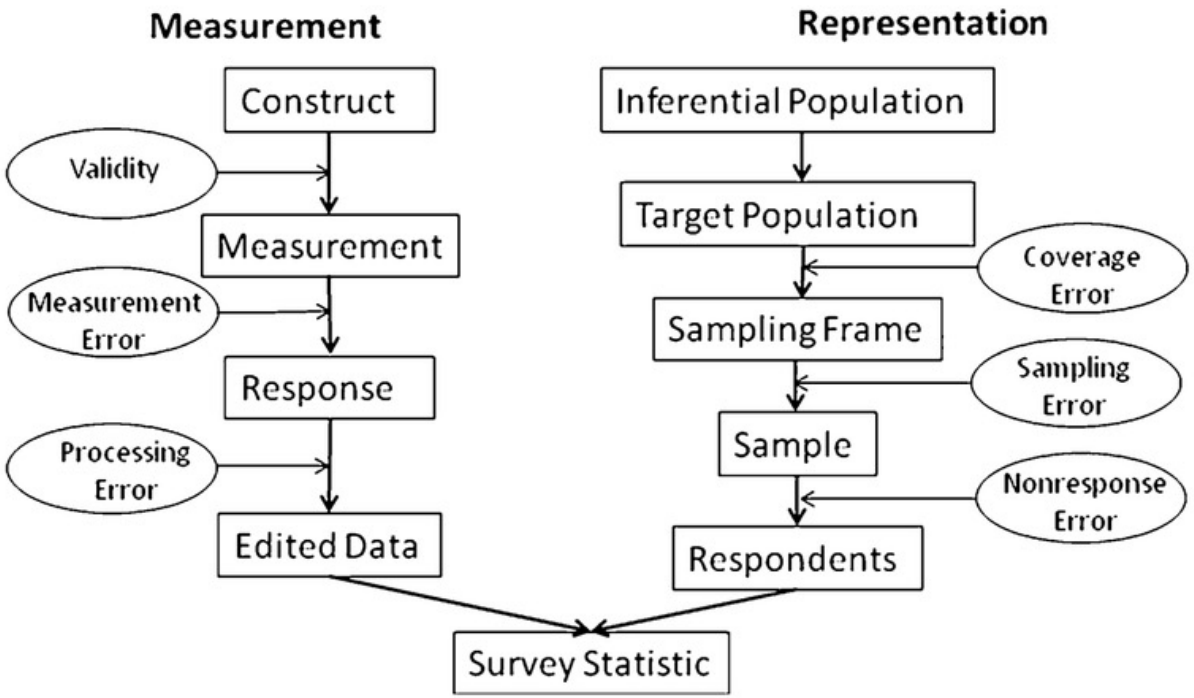


Figure 3.5: Nonresponse has been increasingly steadily, even in high-quality expensive surveys (National Research Council 2013; B. D. Meyer, Mok, and Sullivan 2015). Nonresponse rates are much higher for commercial telephones surveys, sometimes even as high as 90% (Kohut et al. 2012). These long-term 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.



Source: Groves et al. (2009)

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

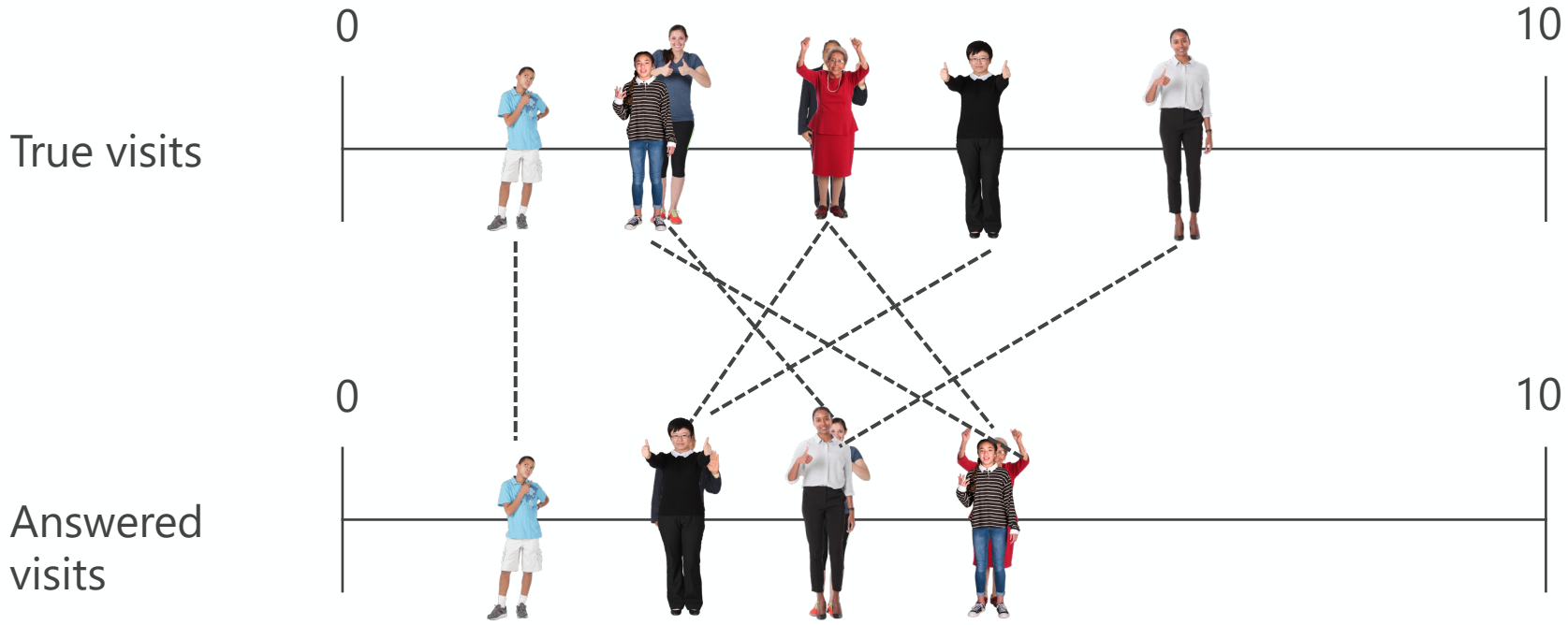
- **Measurement error**: Difference between response and true value.

- **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

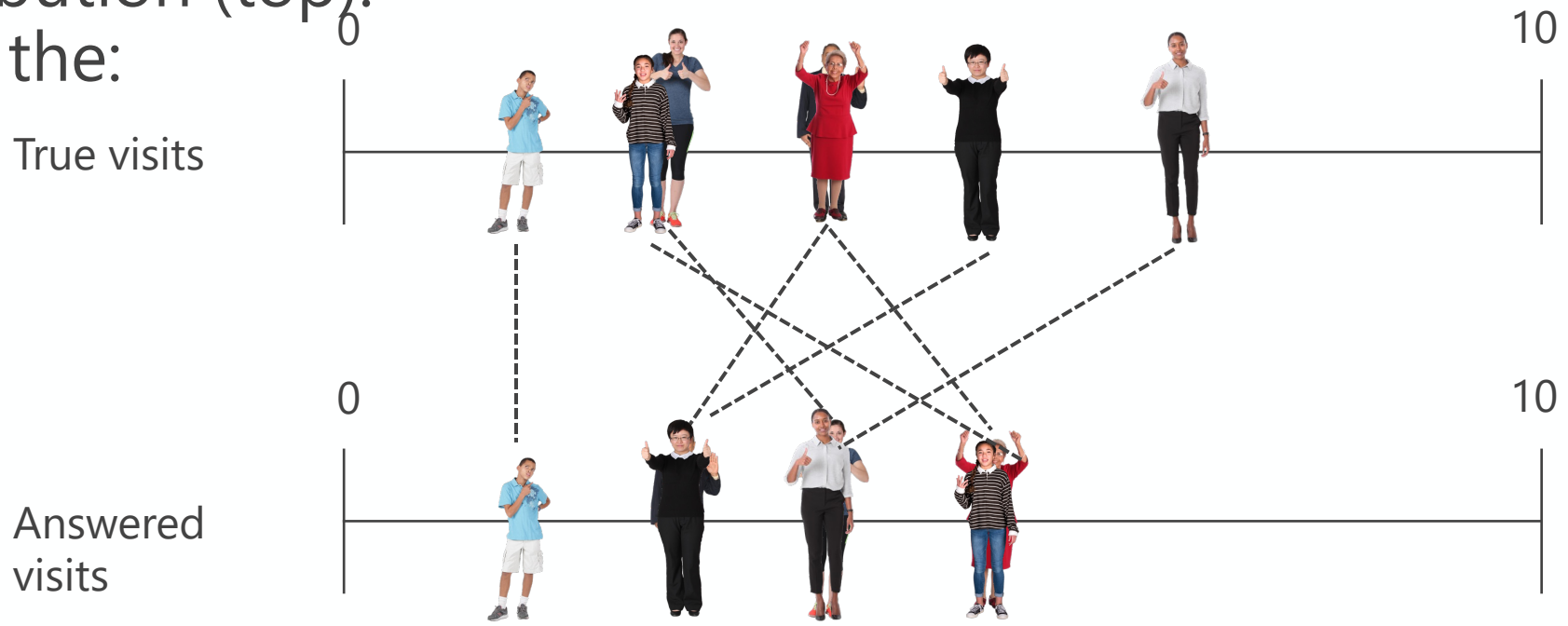
- **Measurement error:** The answer you have is influenced by things other than the value you are after. These things are defined as "errors".

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.

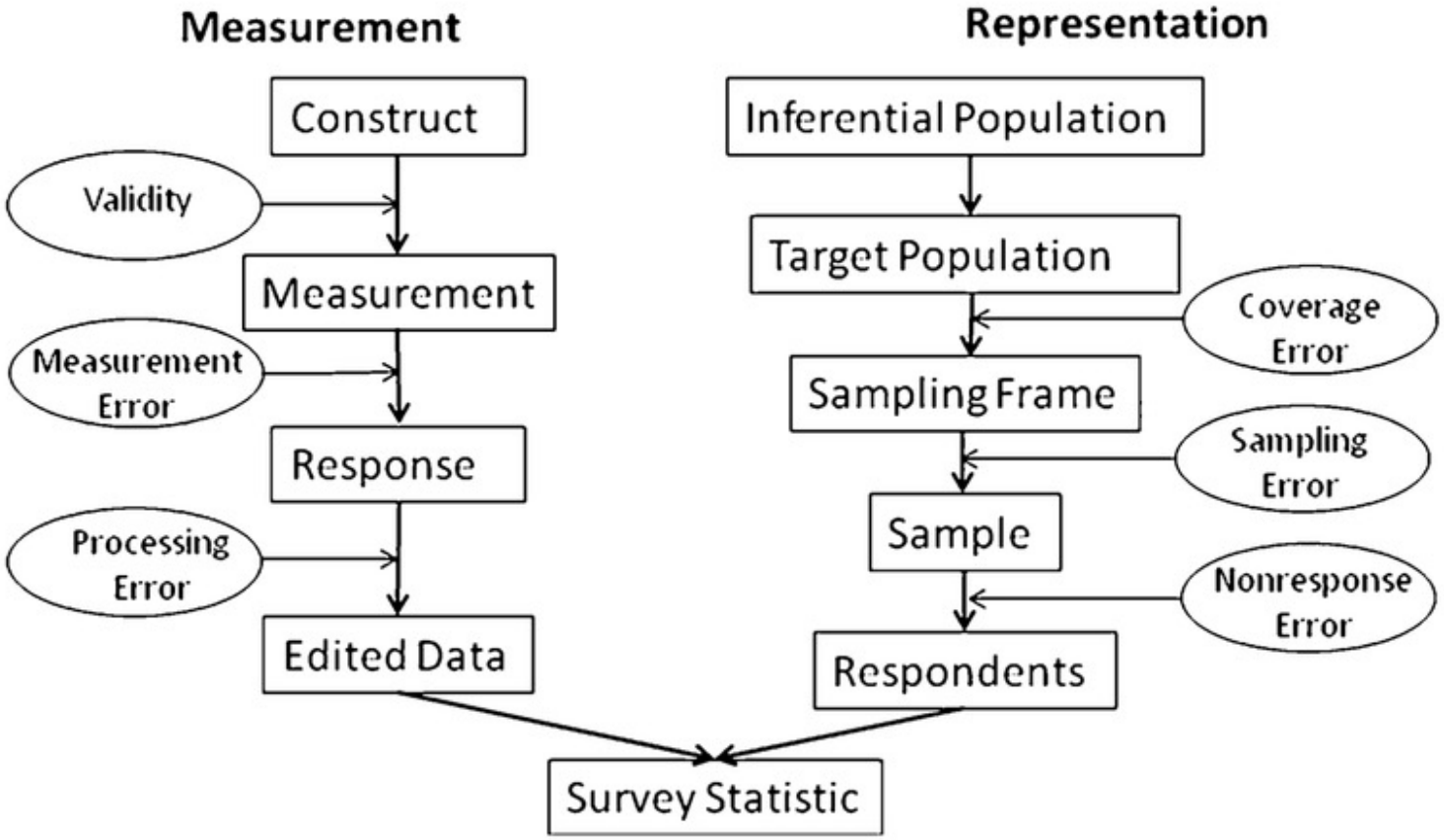
- Resulting in bias in the **mean**
- And **variance**.

Errors in digital trace data

Error frameworks

Total Survey Error Framework

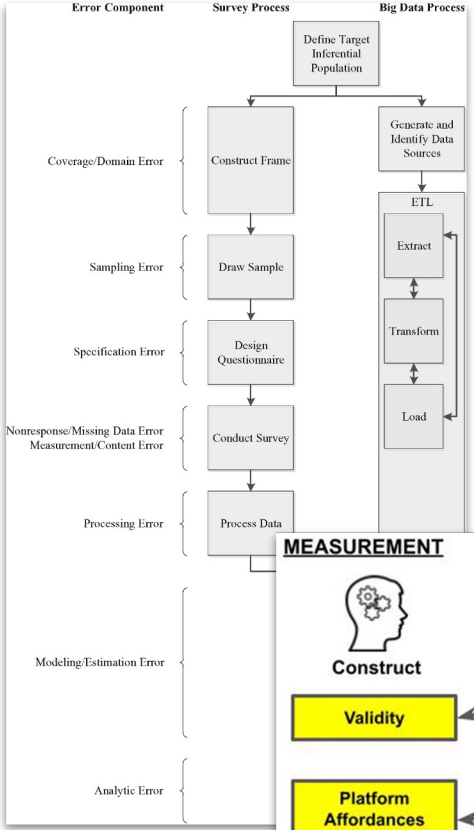
In each step of the design and analysis phase of a survey, errors can arise that affect the quality of the final statistic of interest.

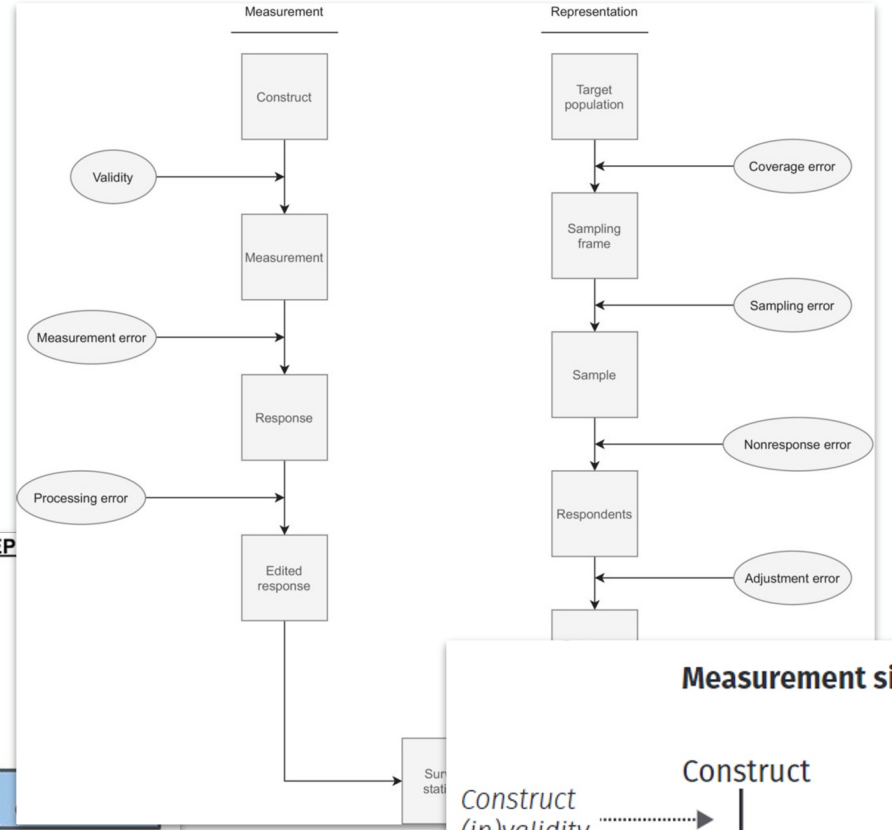


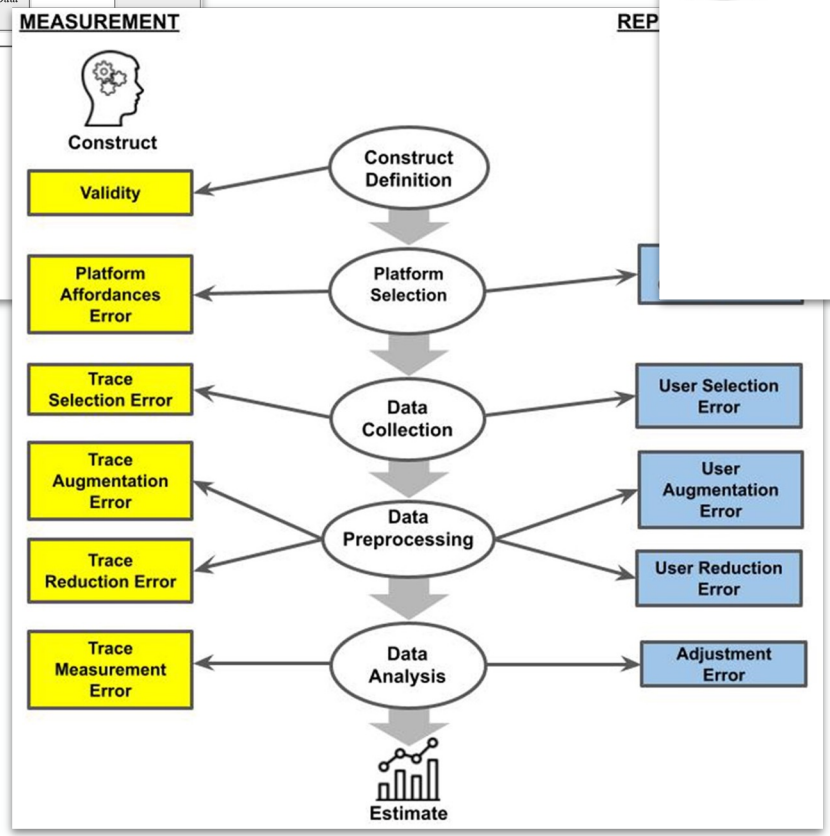
Source: Groves et al. (2009)

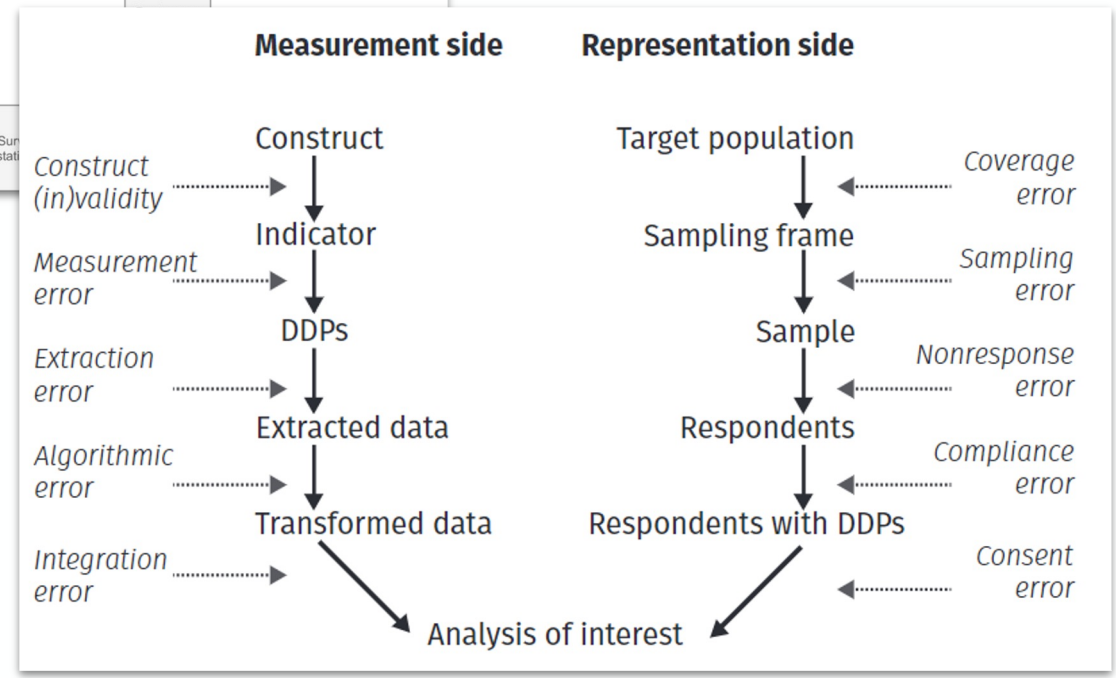
Error frameworks for digital trace data

- 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 Component
Survey Process
Big Data Process
Define Target Inferential Population
Coverage/Domain Error
Construct Frame
Generate and Identify Data Sources
ETL
Sampling Error
Draw Sample
Extract
Specification Error
Design Questionnaire
Transform
Nonresponse/Missing Data Error
Measurement/Content Error
Conduct Survey
Load
Processing Error
Process Data
Modeling/Estimation Error
Analytic Error


Measurement
Representation
Construct
Target population
Validity
Coverage error
Sampling frame
Measurement
Measurement error
Sampling error
Sample
Response
Nonresponse error
Processing error
Respondents
Edited response
Adjustment error


MEASUREMENT
REP
Construct
Construct Definition
Validity
Platform Affordances Error
Platform Selection
Trace Selection Error
Data Collection
User Selection Error
Trace Augmentation Error
User Augmentation Error
Data Preprocessing
Trace Reduction Error
User Reduction Error
Trace Measurement Error
Data Analysis
Adjustment Error
Estimate


Measurement side
Representation side
Construct
Target population
Construct (in)validity
Coverage error
Indicator
Sampling frame
Measurement error
Sampling error
DDPs
Sample
Extraction error
Nonresponse error
Extracted data
Respondents
Algorithmic error
Compliance error
Transformed data
Respondents with DDPs
Integration error
Consent error
Analysis of interest

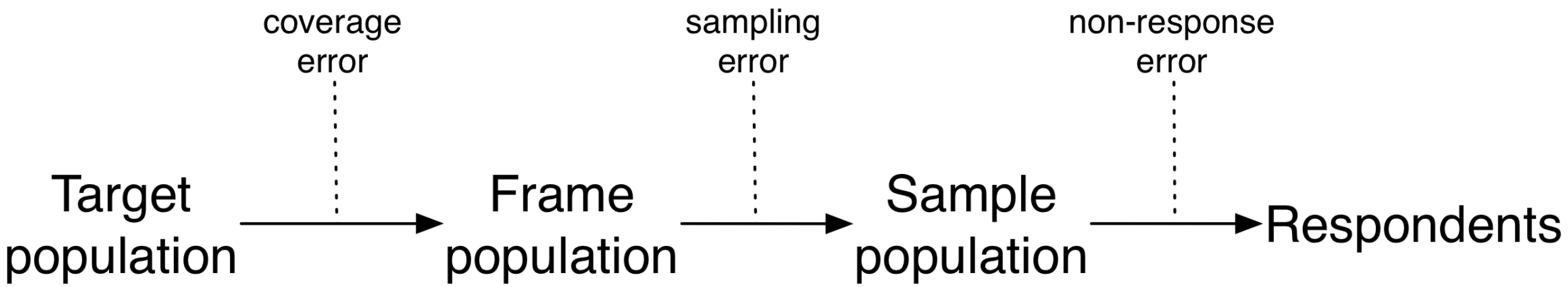
Error frameworks for digital trace data

- 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)

Representation & Measurement

Problems with representation

Problems with representation in surveys



Problems with representation in DTD

- Who uses a platform? → Coverage error
- Who will receive your invite to participate? → Sampling error
- Who is willing to share their data and who not? → Non-response error

Coverage error

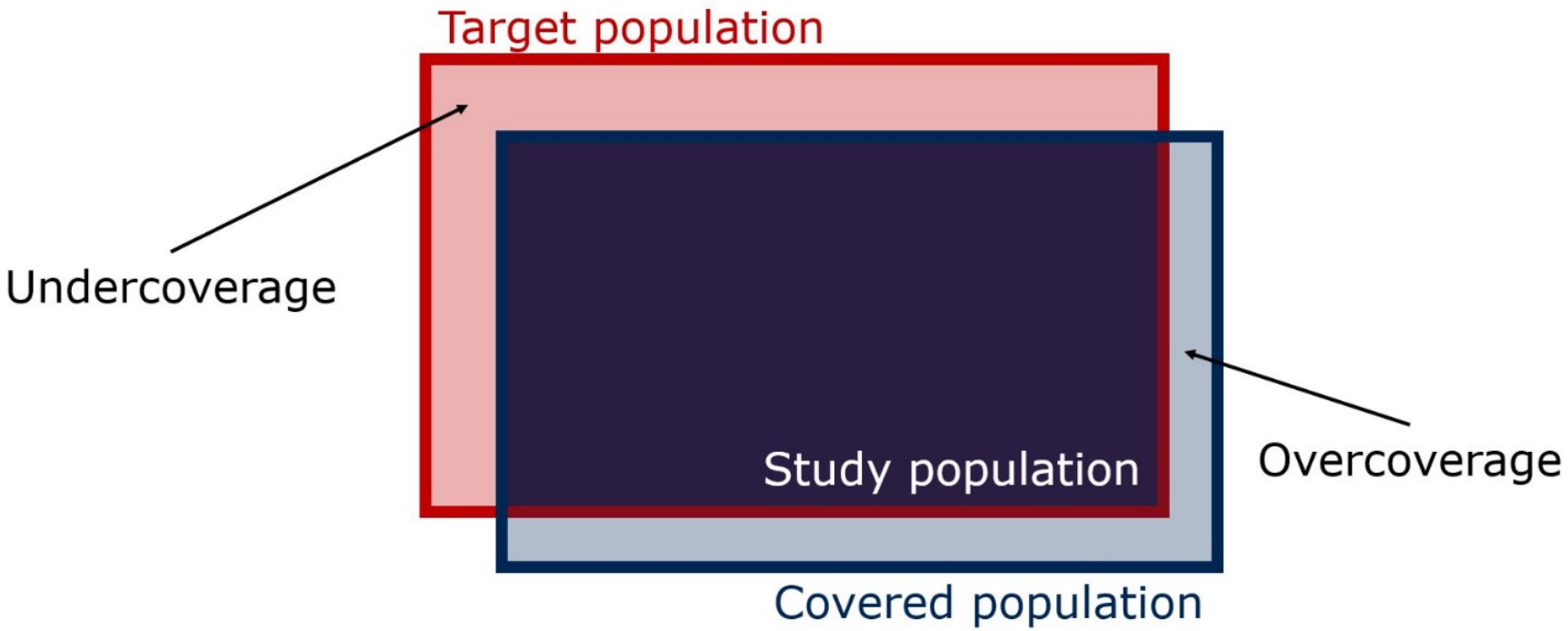
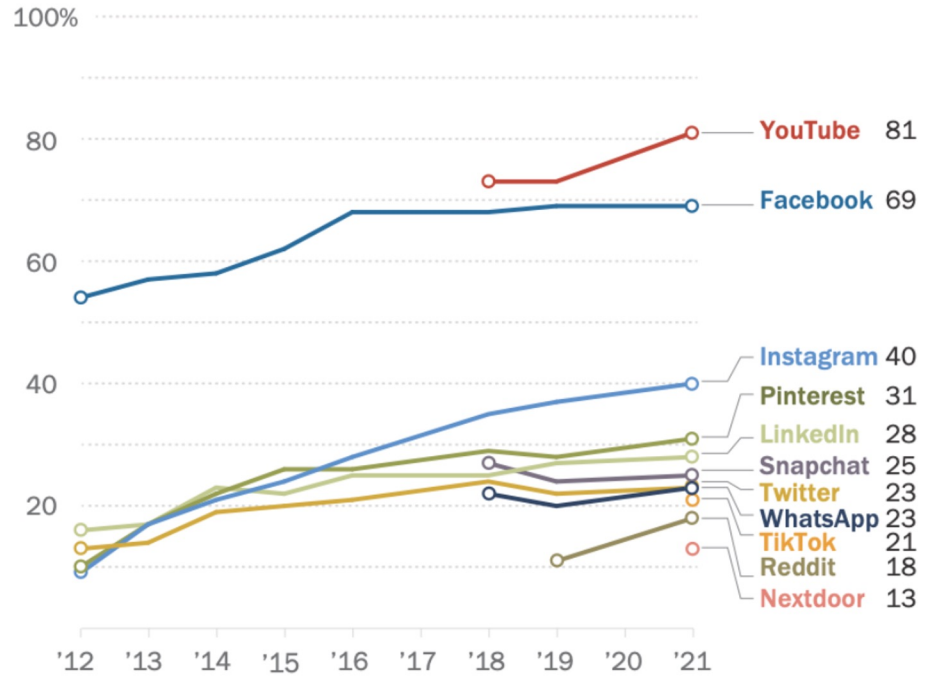


Figure 4.3: Target population, covered population, and study population

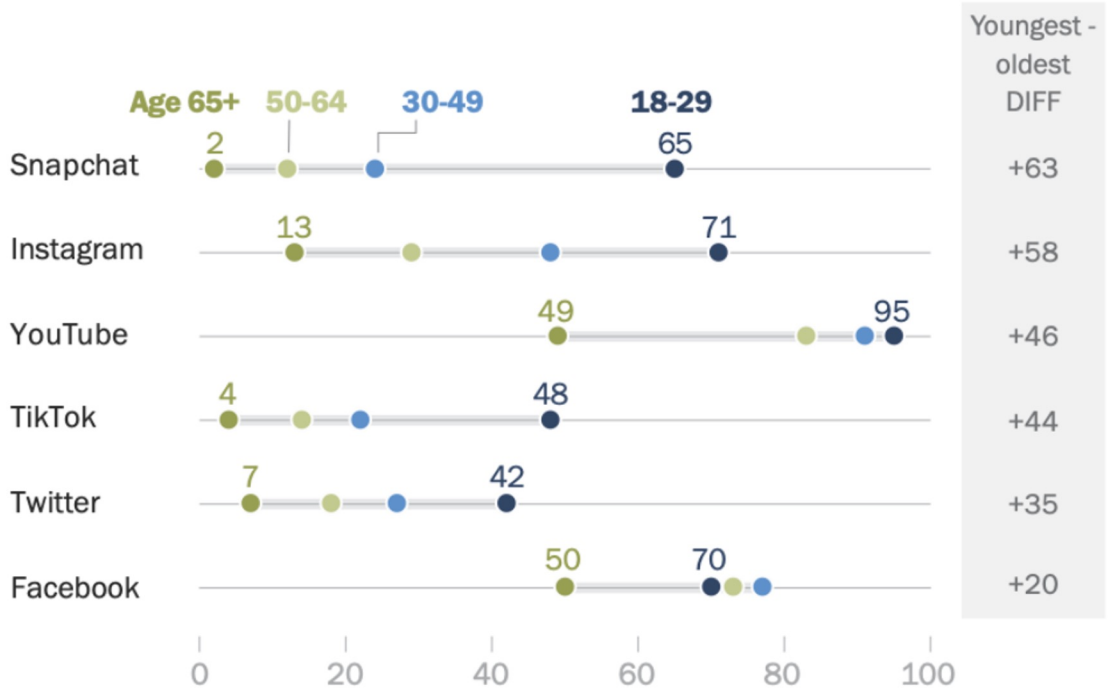
Social media platform use

% 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"

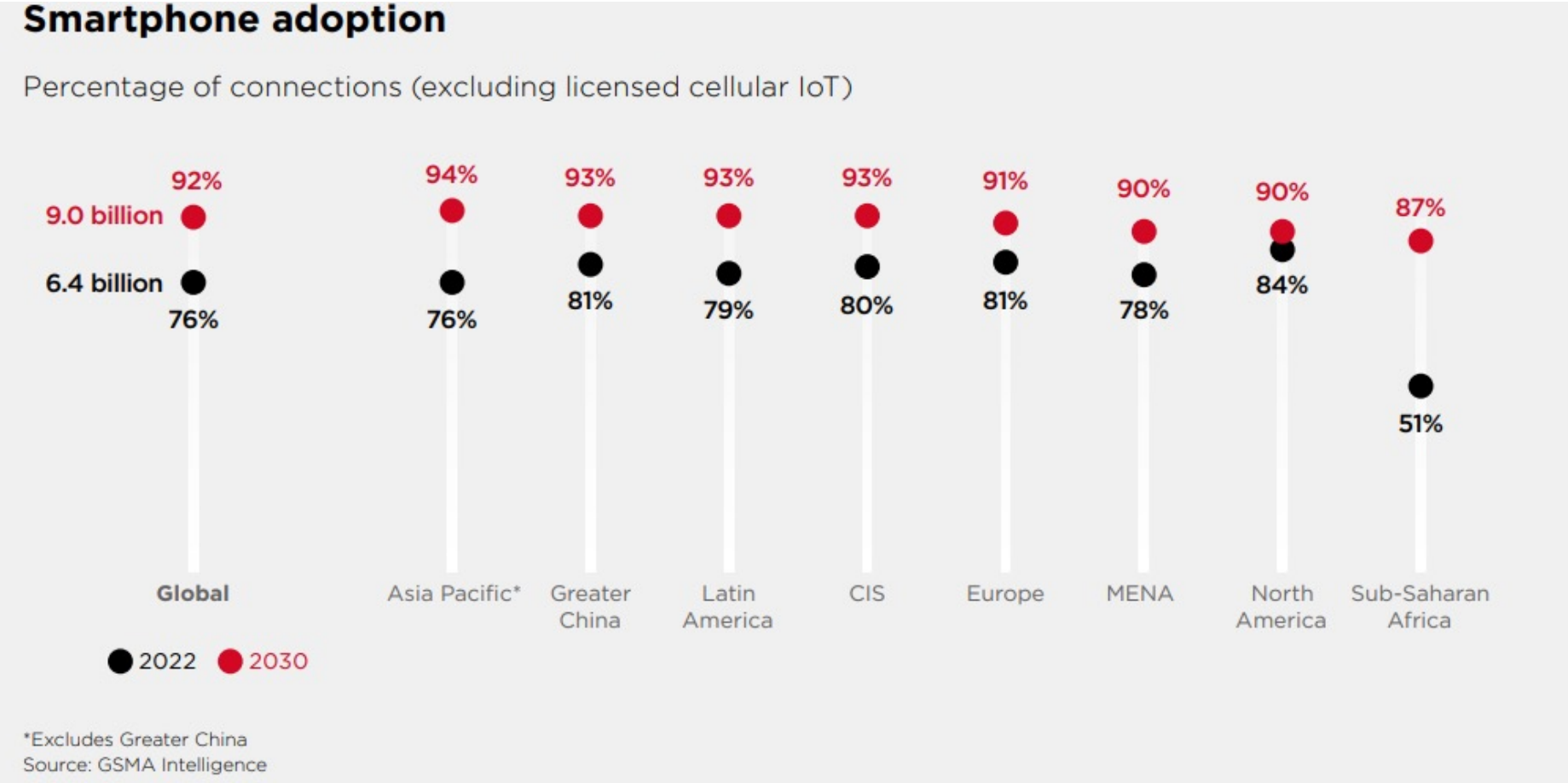
% of U.S. adults in each age group who say they ever use ...



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.

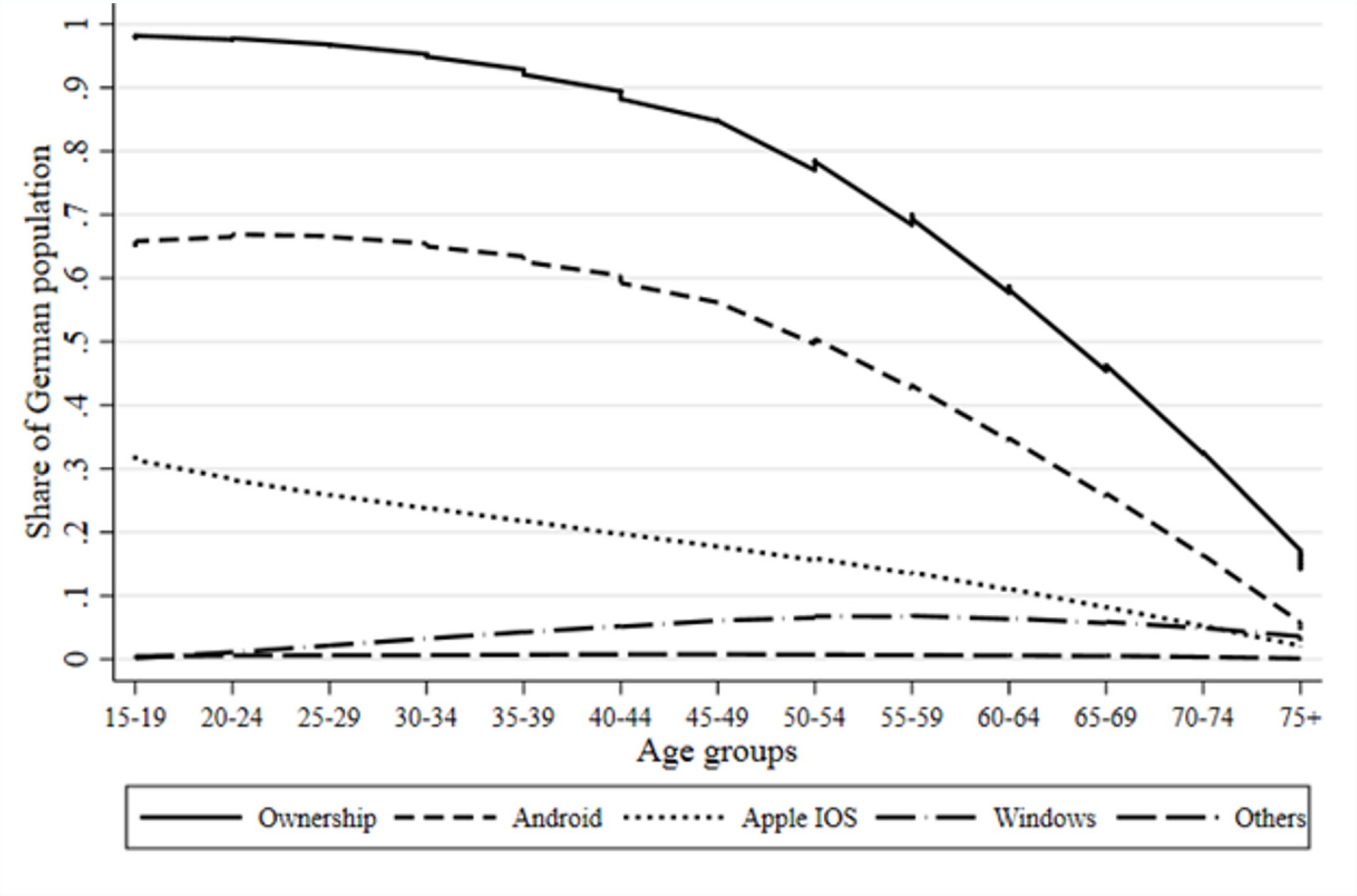
Source: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

Coverage smartphones



See GSMA (2023)

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

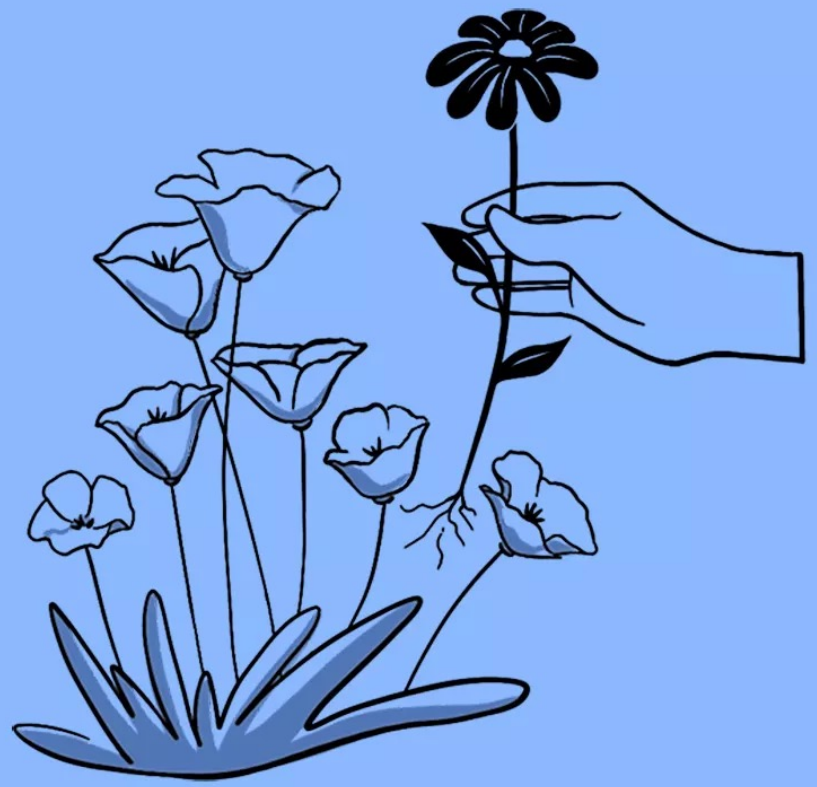


- Smartphone ownership also correlates with…
  - Educational attainment
  - Nationality
  - Region
  - Community size
- Bias of ownership rel. small for many substantive measures
  - But substantial bias for iPhone ownership

Source: PASS Wave 11; n = 13,703; Locally weighted scatter-plot smoother (LOWESS) regression

Problems with representation in DTD

- Who uses a platform? → Coverage
- Who will receive your invite to participate? → Sampling error
- Who is willing to share their data and who not? → Non-response error



Sampling Error

['sam-pliŋ 'er-ər]

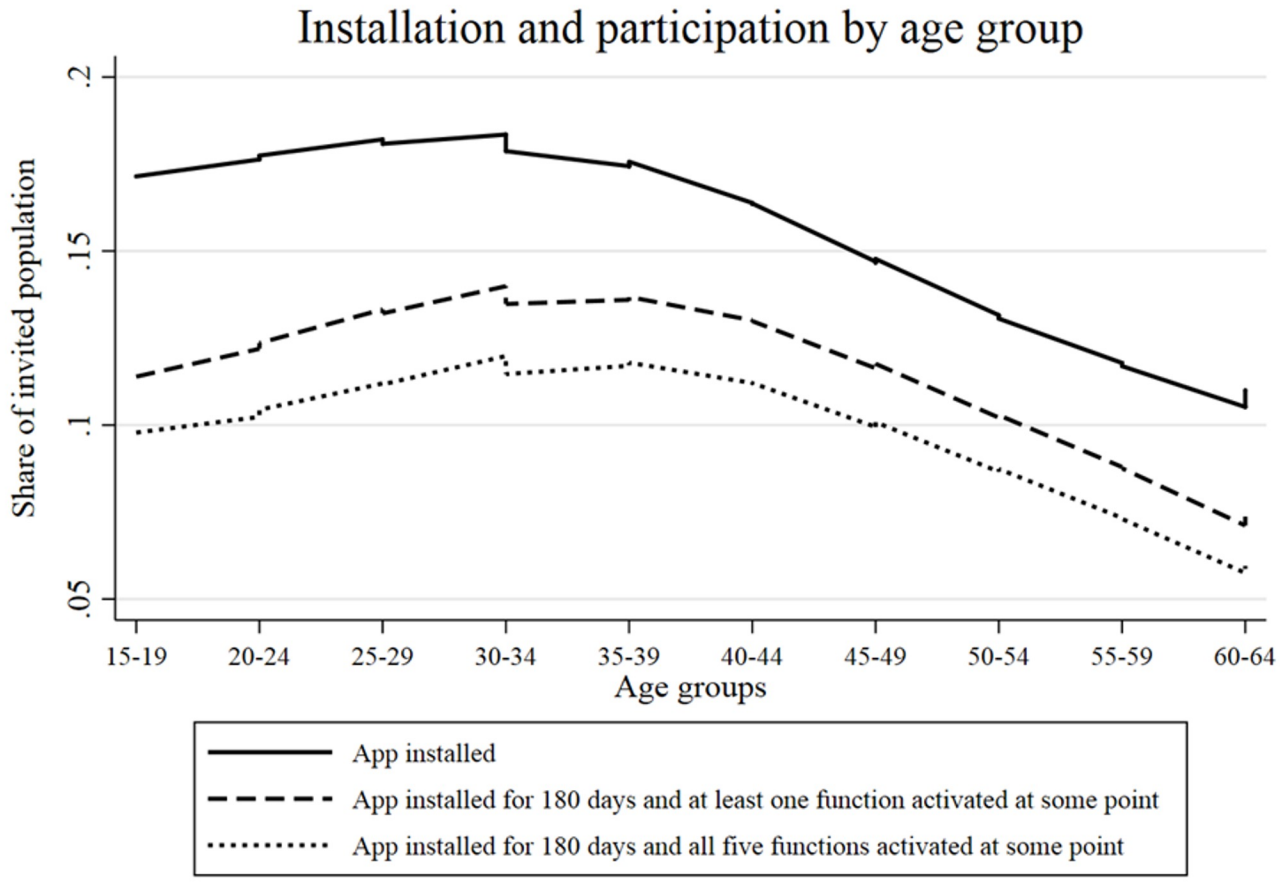
A statistical error that occurs when the analyst selects a sample that is not representative of the population being studied.

Investopedia

Problems with representation in DTD

- Who uses a platform? → Coverage
- Who will receive your invite to participate? → Sampling error
- Who is willing to share their data and who not? → Non-response error
    - We find the following terminology in the literature:
    - Non-participation
    - Non-willingness
    - Non-compliance

Non-response in research app studies



Keusch, Bähr, et al. (2022)

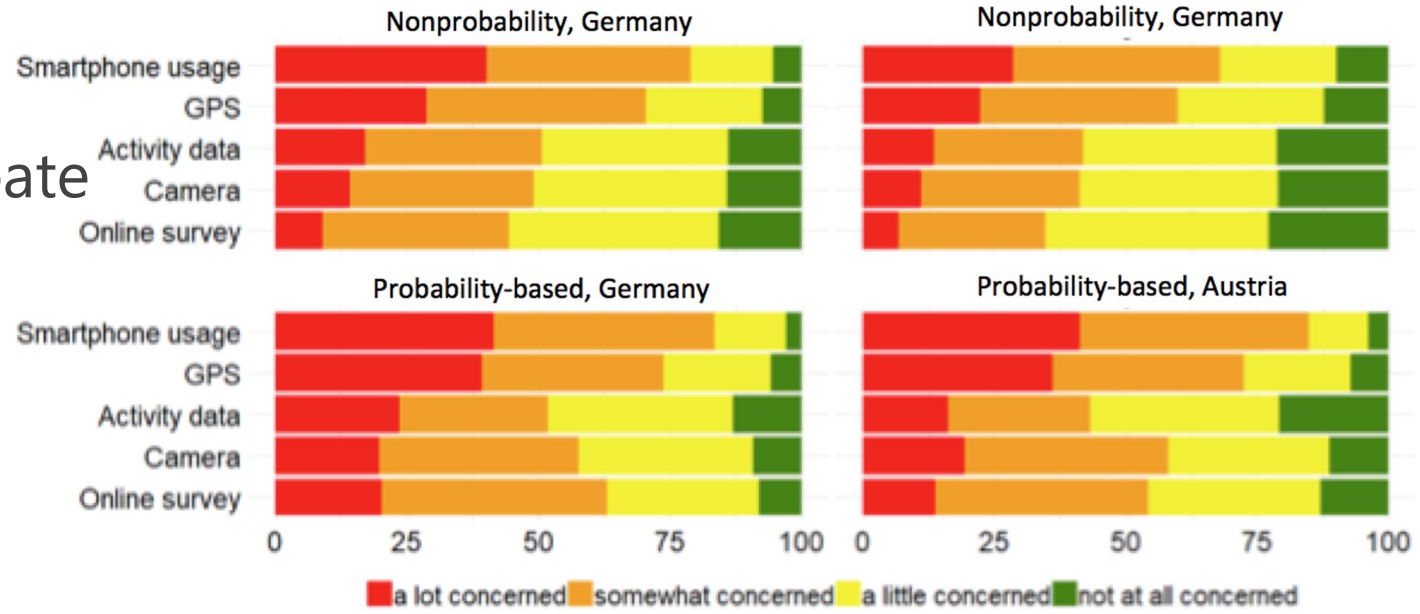
Reasons not to participate	
Privacy, data security concerns	44%
No incentive, incentive too low	17%
Not enough information provided	12%
Do not download any apps	8%
Not interested	6%
Not enough time/Study too long	5%
Don’t use smartphone enough	5%
Not enough storage	1%
Other reasons	6%
n=1,154	

Keusch et al. (2019)

Mechanisms of (non-)response: 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)



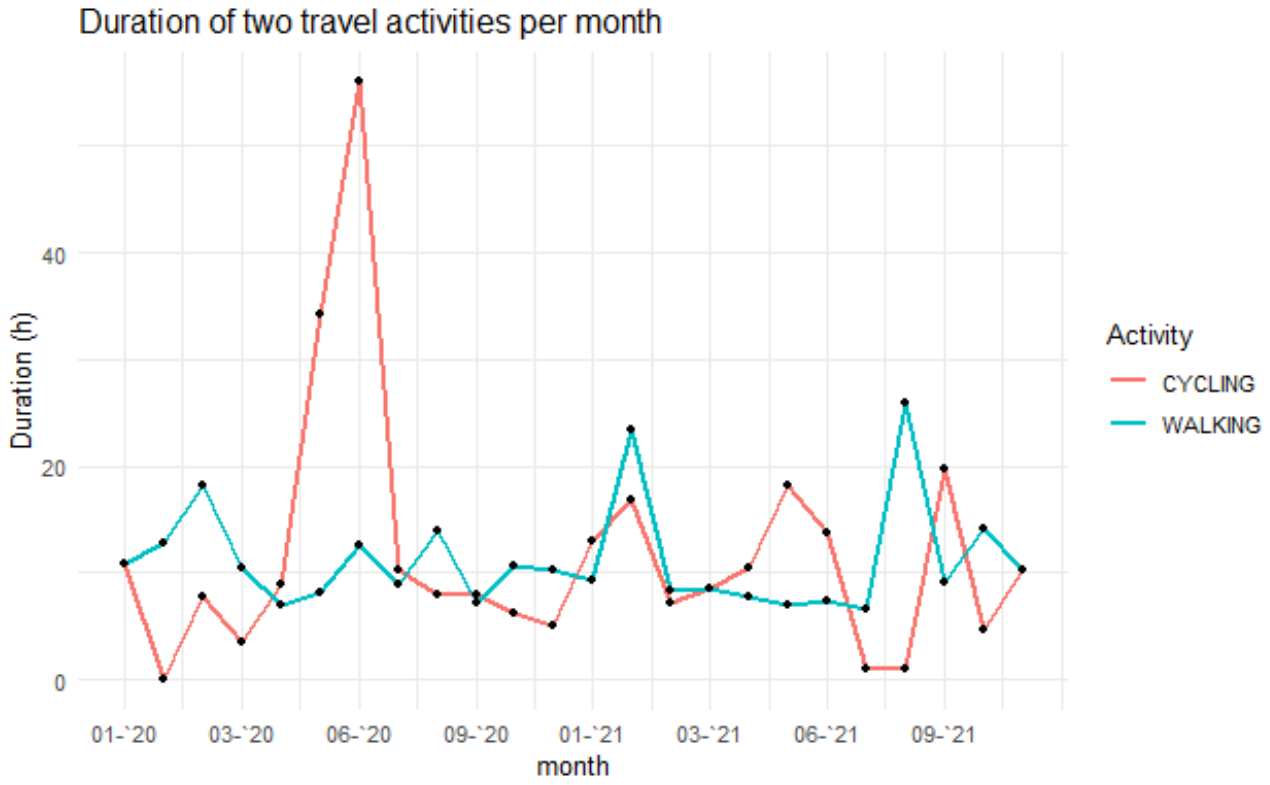
Keusch et al. (2021)

Other Mechanisms of (non-)response

- 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

More detail: Willingness and compliance

Google Semantic Location History



Struminskaya & Boeschoten (2022)

More detail: 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

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

In Bus

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

In Passenger Vehicle

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

Struminskaya & Boeschoten (2022)

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

Struminskaya & Boeschoten (2022)

Understanding the consent request

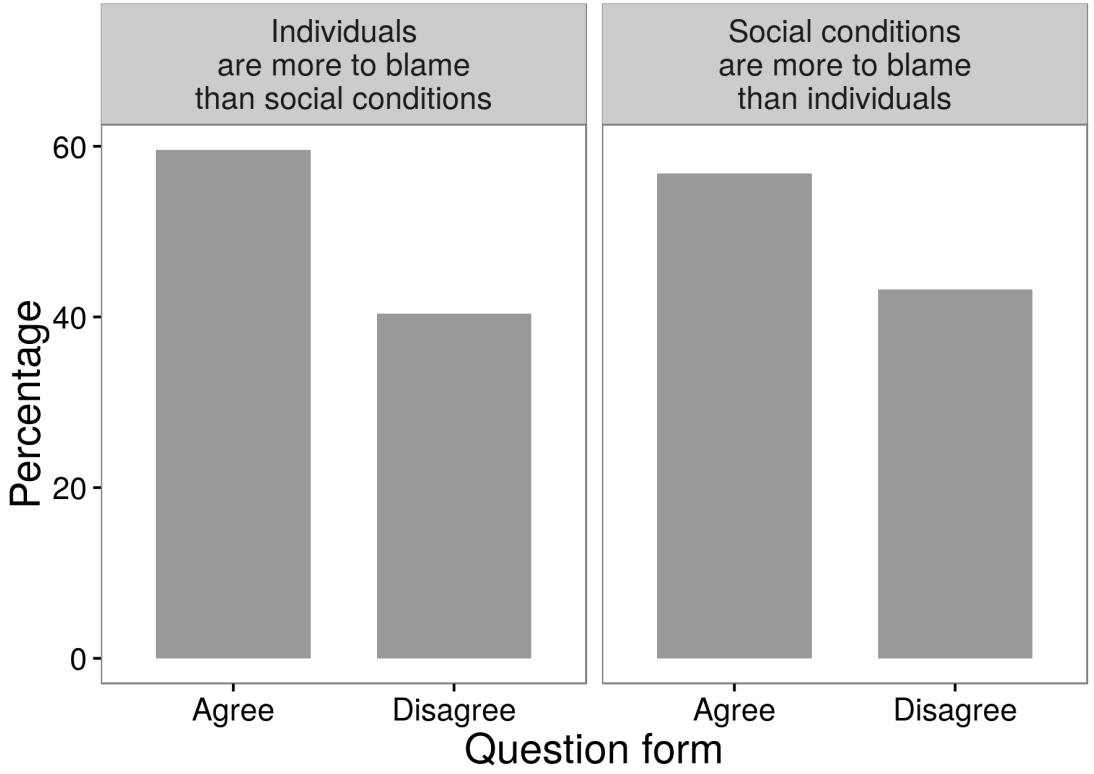
- **5.5% had everything correct**
- Mean correct: 3.23, median = 4 (out of 7 questions)
- **People with more correct answers more likely to be willing & to donate:**
  - 4.54 correct statements for willing
  - 2.56 correct statements for non-willing
  - OR = 1.572, p <.001

  - 5.33 correct statements for donated
  - 3.94 correct statements for not donated
  - OR = 1.795, p <.001

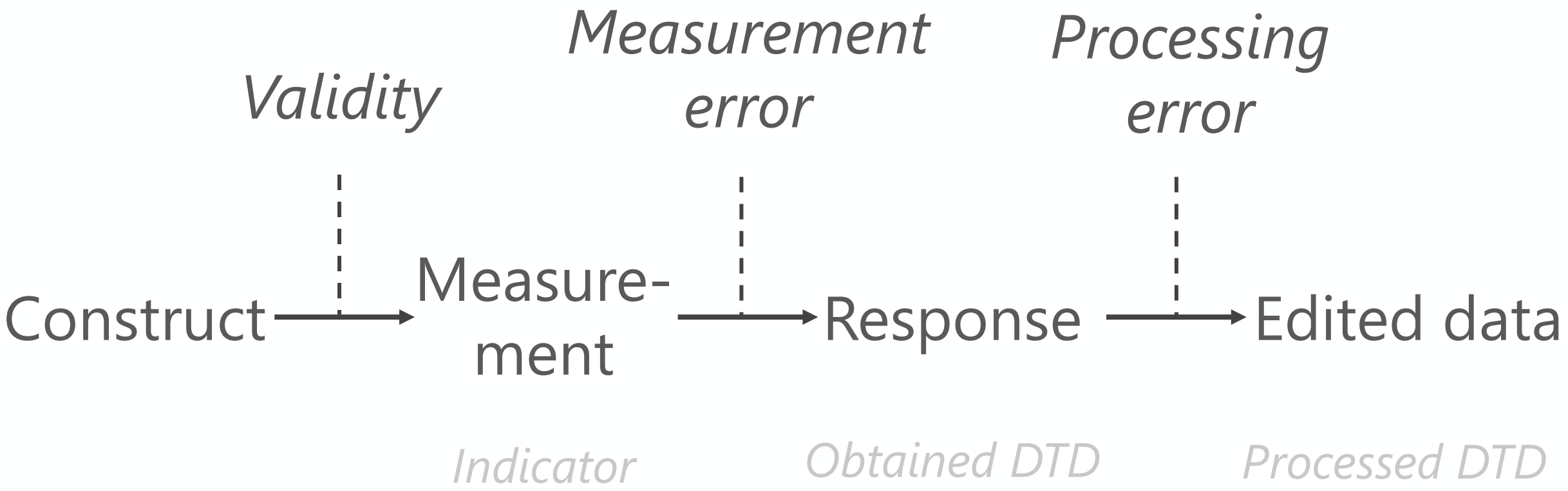
Problems with measurement

Measurement problems in surveys

How you ask a question matters!



Measurement problems in DTD

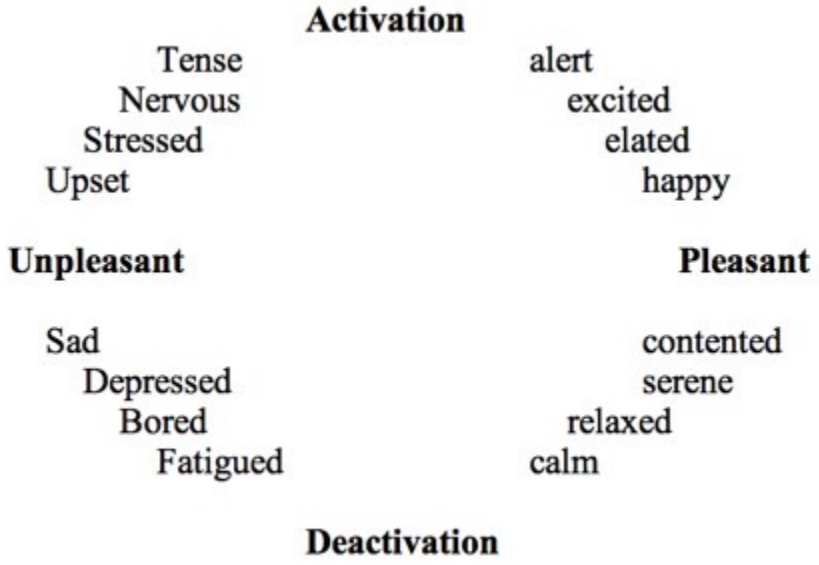


Measurement problems in DTD

- 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

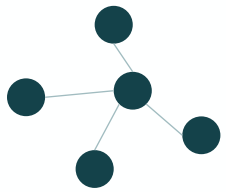
- Example:
- We are interested in someone's mood.
- We use "facial expression on photo" as an indicator and collect photos through data donation.
- **<u>Are we measuring the concept appropriately?</u>**



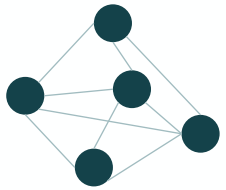
Validity

- 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/plug-in/tool!
- What do you need to do to get your measurements of interest from the data? → Processing error

Measurement error

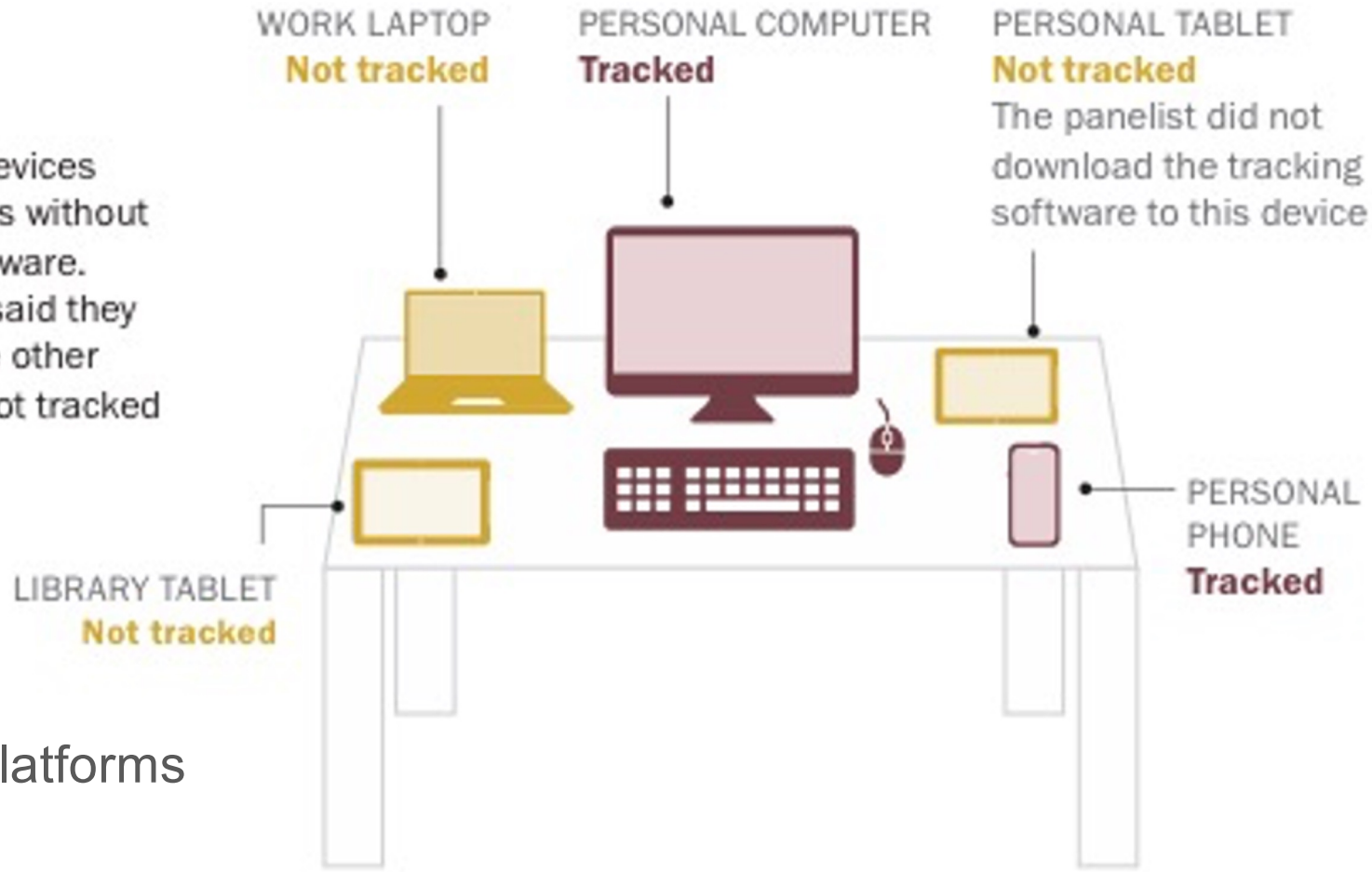
- We are still interested in someone's mood.
- Imagine using "facial expression on photo" is a good indicator.
- We use the Screenomics app.
- **Is this a correct representation of all your facial expressions on all photo's?**

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

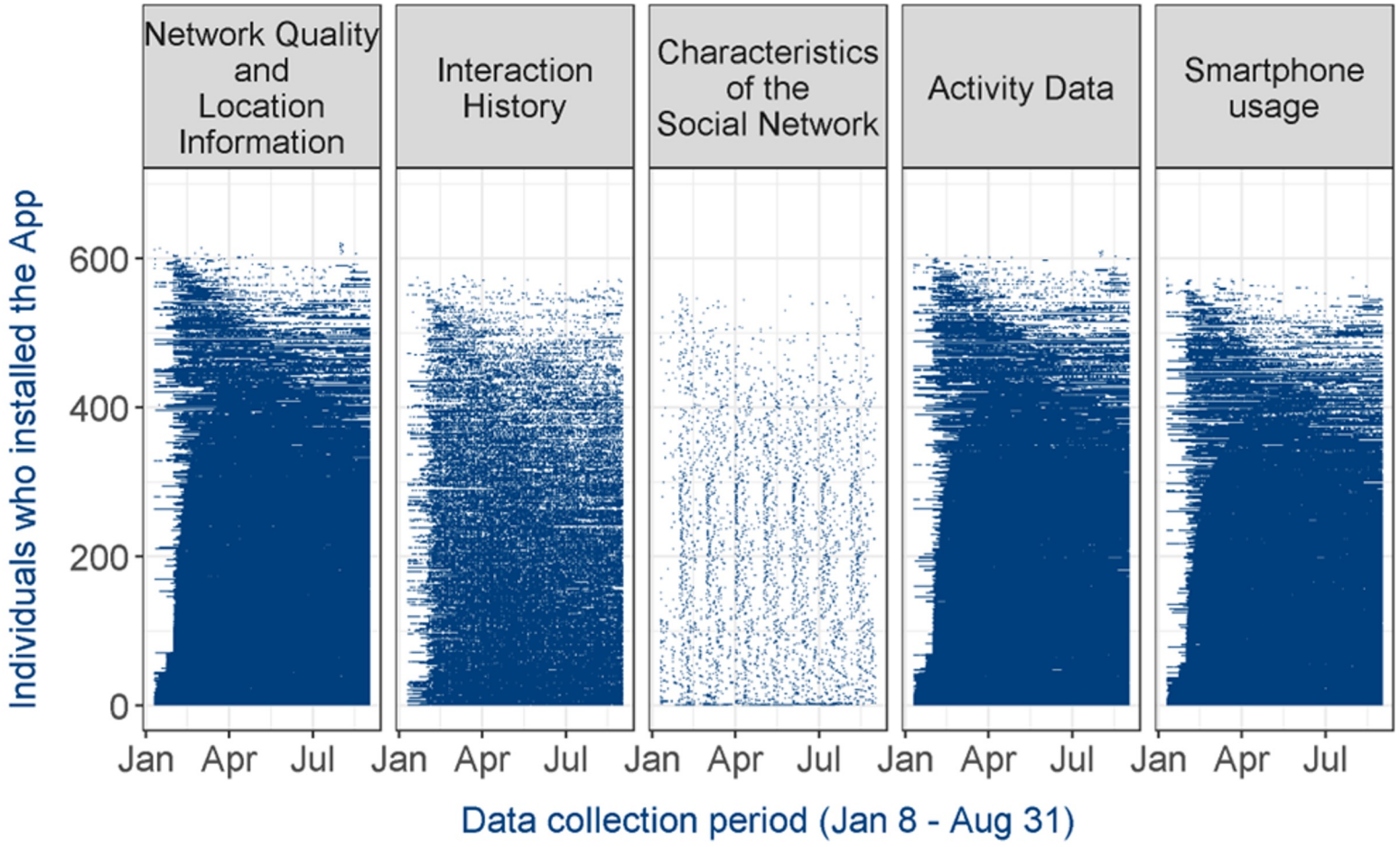


Also holds for use of different platforms (e.g., WhatsApp vs. Facebook)

Source: https://www.pewresearch.org/journalism/2020/12/08/measuring-news-consumption-in-a-digital-era/

Measurement problems in DTD

- Missing data

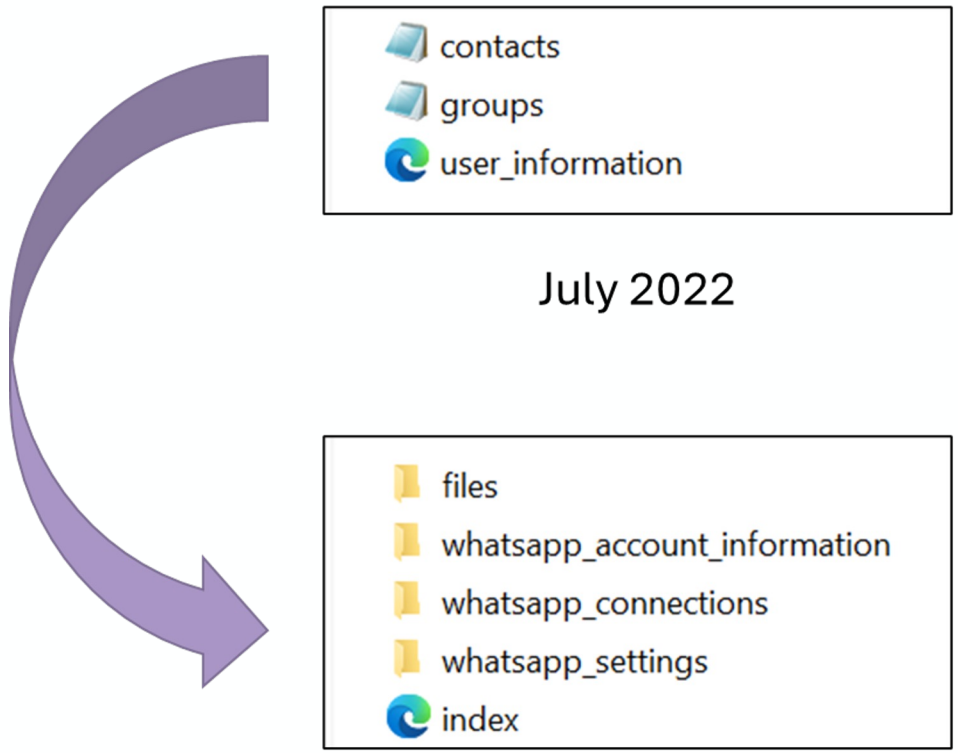


Bähr et al. (2022)

Volatility of Data Download Packages (DDP)

- 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

Example: WhatsApp DDP Folder structure



Example by T.C. Carriere

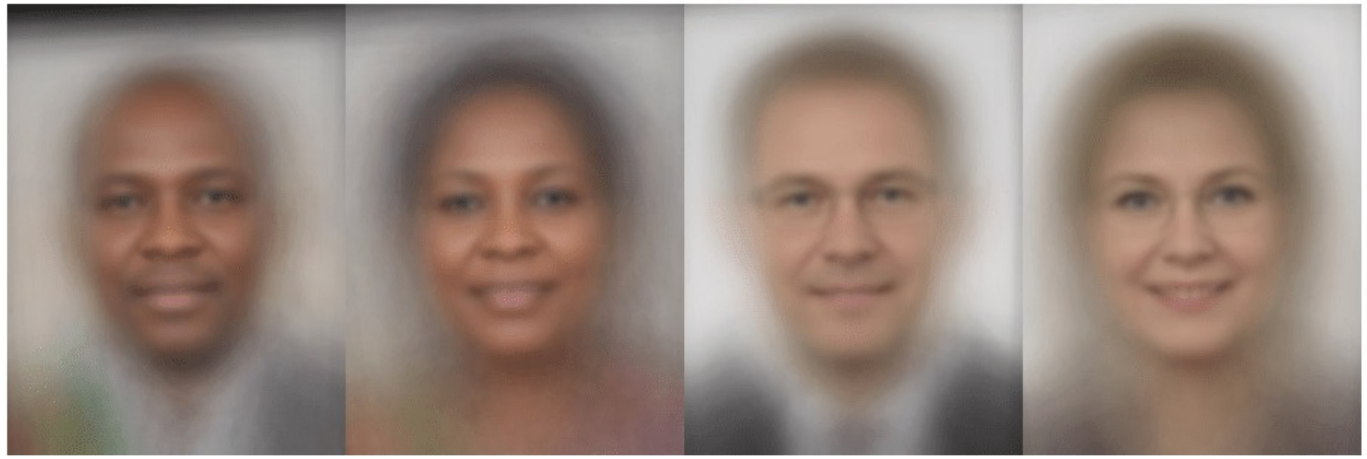
Measurement problems in DTD

- 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%



Source

Designed big data

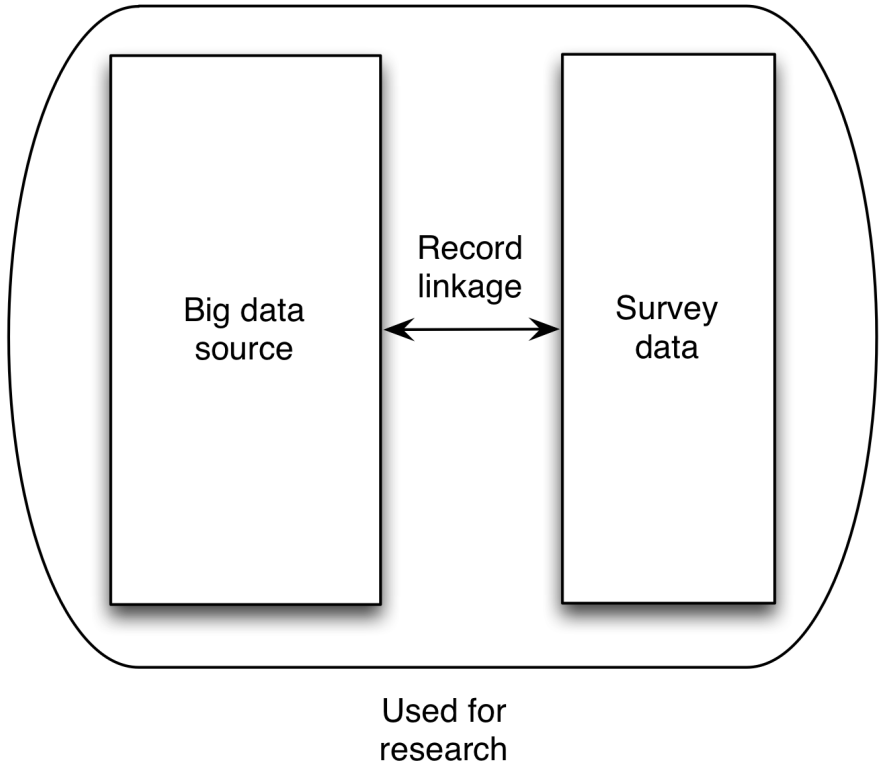
How to combine survey data and digital trace data?

Enriched asking

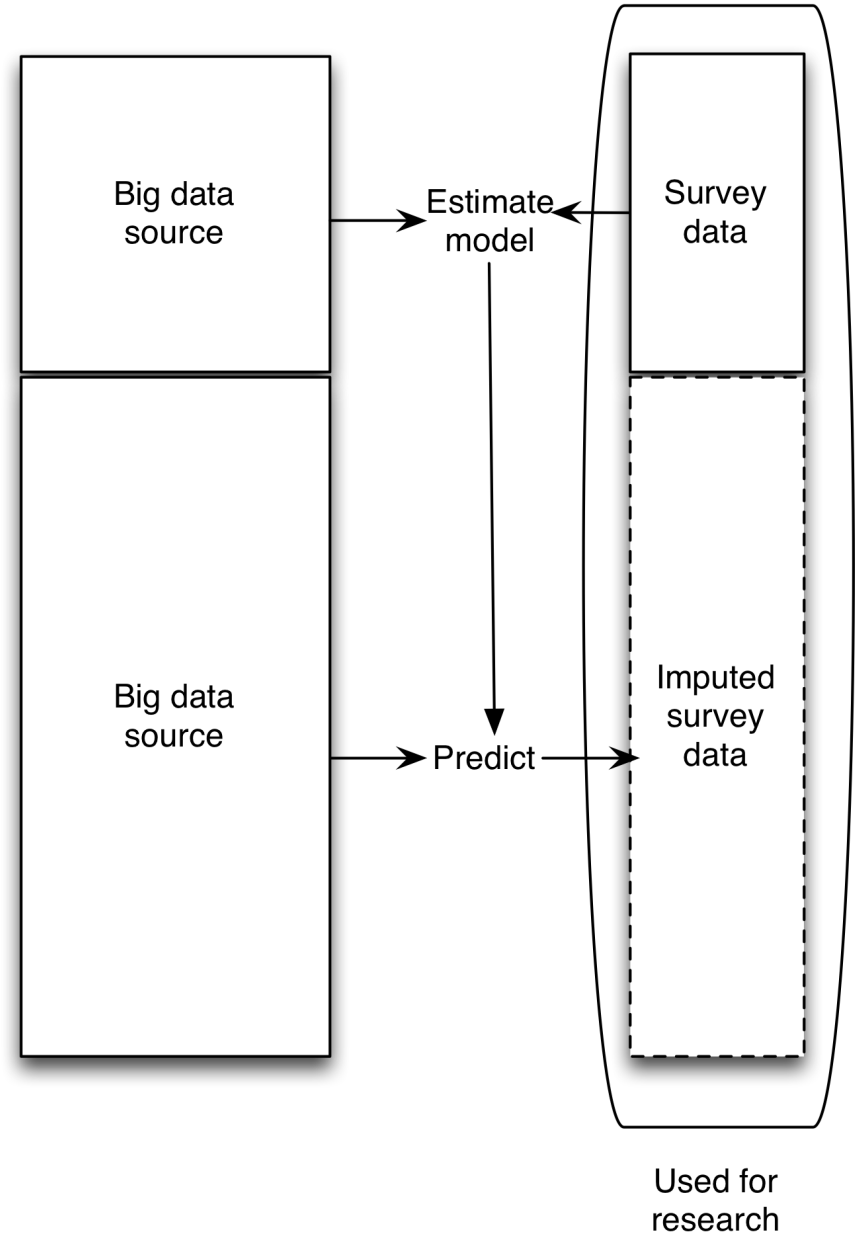
Amplified asking

Surveys and big data are complements and not substitutes!

Enriched asking



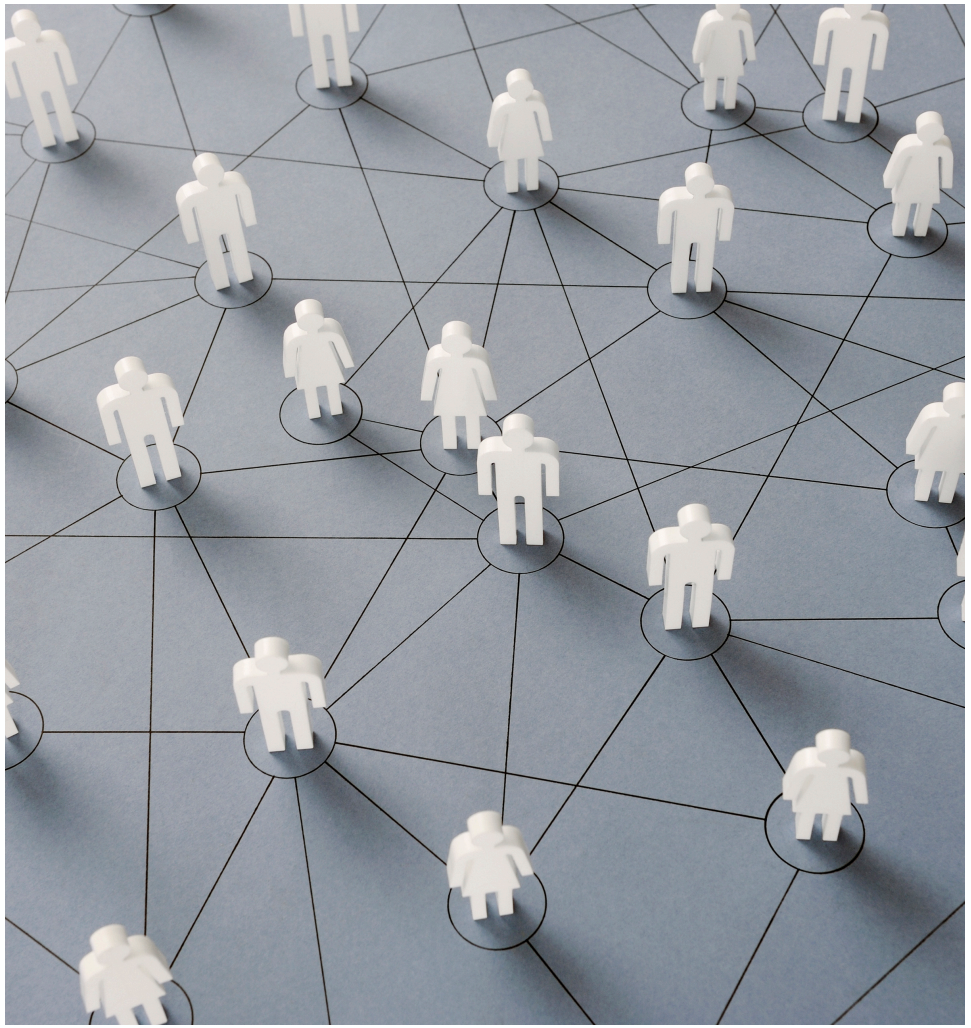
Amplified asking



Enriched asking

Example:

Communication via Facebook and feelings of closeness



The practical and fundamental limitations of big data sources, and how they can be overcome with surveys, are illustrated by Moira Burke and Robert Kraut's (2014) research on how the strength of friendships was impacted by interaction on Facebook. At the time, Burke was working at Facebook so she had complete access to one of the most massive and detailed records of human behavior ever created. But, even so, Burke and Kraut had to use surveys in order to answer their research question. Their outcome of interest—the subjective feeling of closeness between the respondent and her friend—is an internal state that only exists inside the respondent's head. Further, in addition to using a survey to collect their outcome of interest, Burke and Kraut also had to use a survey to learn about potentially confounding factors. In particular, they wanted to separate the impact of communicating on Facebook from communication through other channels (e.g., email, phone, and face to face). Even though interactions through email and phone are automatically recorded, these traces were not available to Burke and Kraut so they had to collect them with a survey. Combining their survey data about friendship strength and non-Facebook interaction with the Facebook log data, Burke and Kraut concluded that communication via Facebook did in fact lead to increased feelings of closeness.

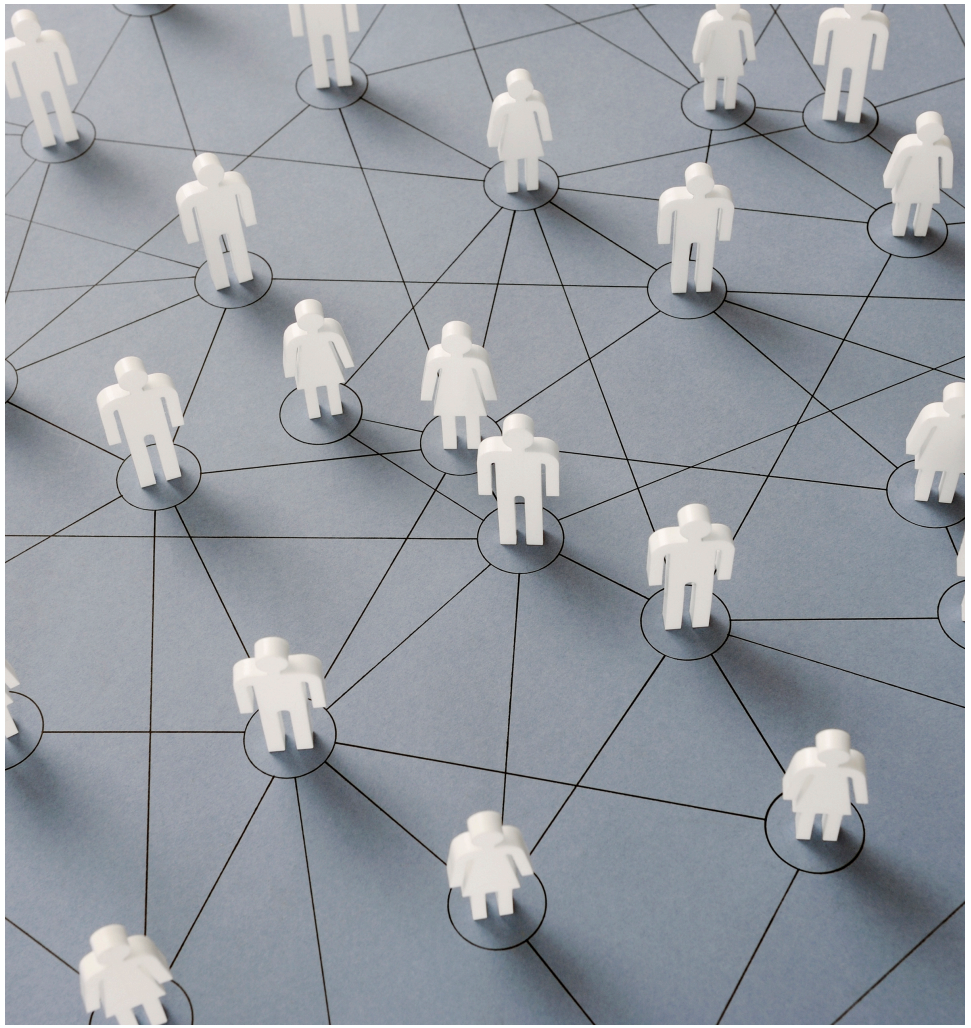
Enriched asking

Example:

Communication via Facebook and feelings of closeness

They did not have to deal with the following challenges:

- Data from the survey and Facebook need to be linked (record linkage, unique identifier needed).
- Quality of the big data source is often difficult or impossible to assess.



Project results

- Directed, composed communication is linked with increases in tie strength.
- So does passively reading a partner’s posts.
- Both broadcasting by yourself and by your partner is linked with declines in tie strength when those stories are not read.
- Family ties are less affected by Facebook activity than non-family.

Enriched asking

Example:

How does what people say about voting differ from their voting behavior?

Image Fredex8/iStock.com

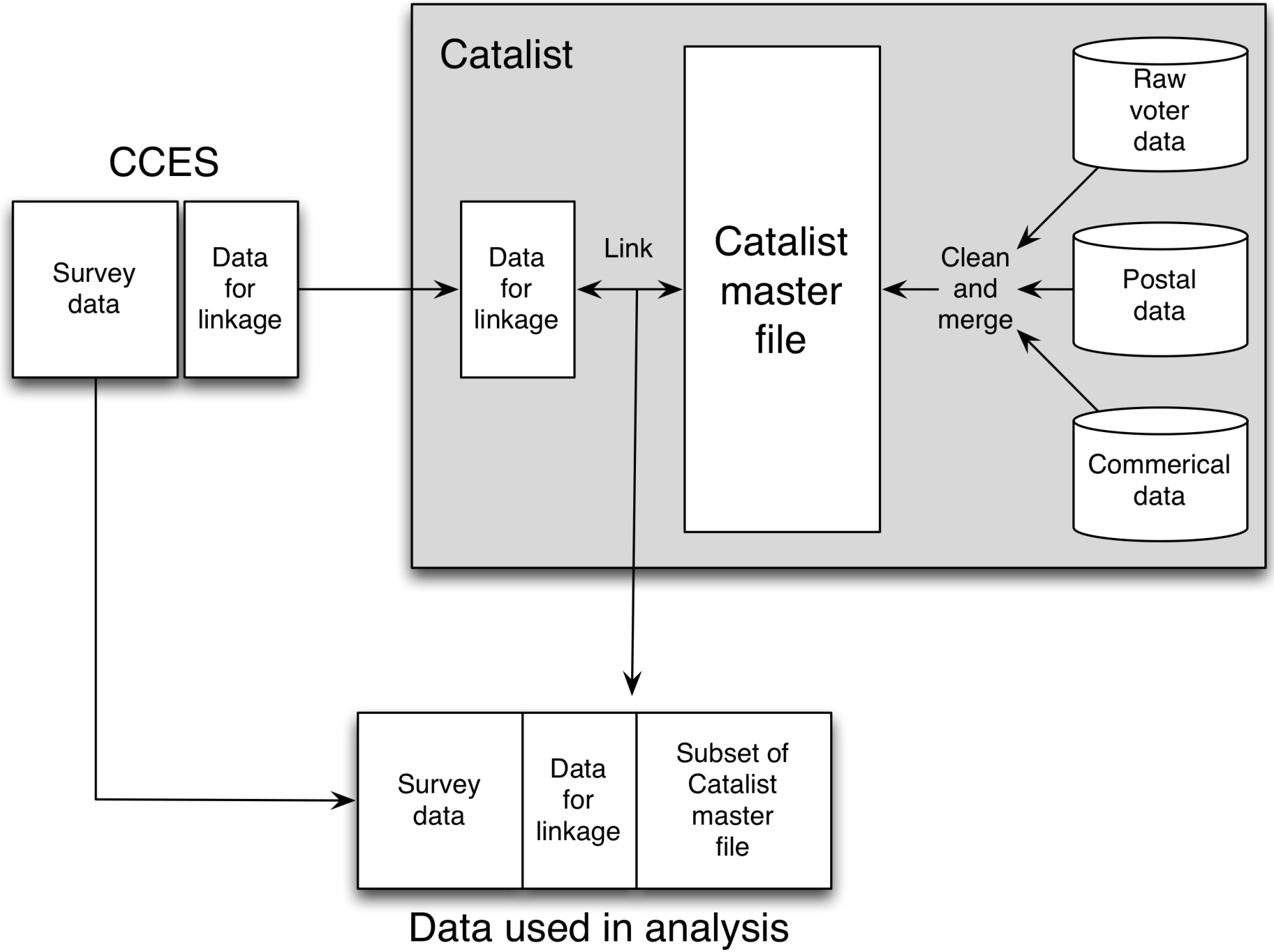
Enriched asking

Example:

How does what people say about voting differ from their voting behavior?

- US Government records whether each citizen has voted.
- This can be supplemented with attitudes of respondents from a large social survey

Image Fredex8/iStock.com


Catalist
CCES
Survey data
Data for linkage
Data for linkage
Link
Catalist master file
Clean and merge
Raw voter data
Postal data
Commerical data
Survey data
Data for linkage
Subset of Catalist master file
Data used in analysis

Enriched asking

Example 2:

How does what people say about voting differ from their voting behavior?

- US Government records whether each citizen has voted.
- This can be supplemented with attitudes of respondents from a large social survey
- **Allows to compare survey and admin voting behavior**

Sources of error:

1. The merging done to create the admin Masterfile
2. No unique identifiers, errors in record linkage

Image Fredex8/iStock.com

Project results

- Public opinion surveys overestimate voter turnout. If someone reported voting, there is only an 80% change that they actually did.
- Over-reporting is not random: A particular group consistently misreport: well-educated, high-income, partisan, politically active, church-attending.
- As a result, the differences between voters and non-voters are smaller than appears from literature.

How can enriched asking help?

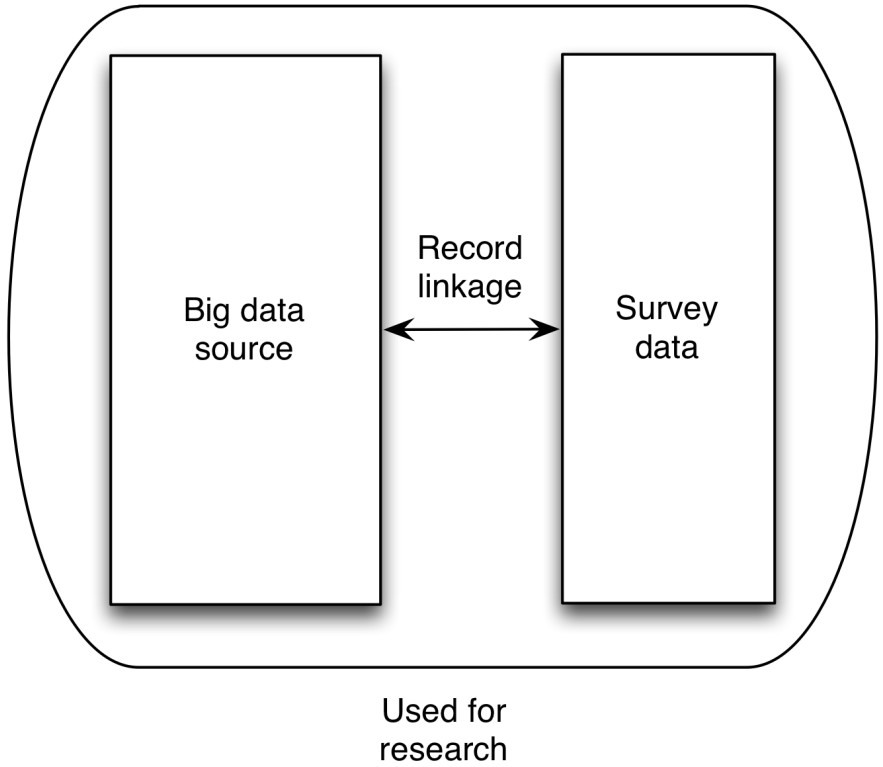
- There is big value in enriching big data sources and in enriching surveys.
- We can do things that are not possible to do with just one of them.
- Researchers can benefit from the efforts done by private companies.

- Administrative or commercial datasets cannot be considered a “ground truth”
- But surveys also not!

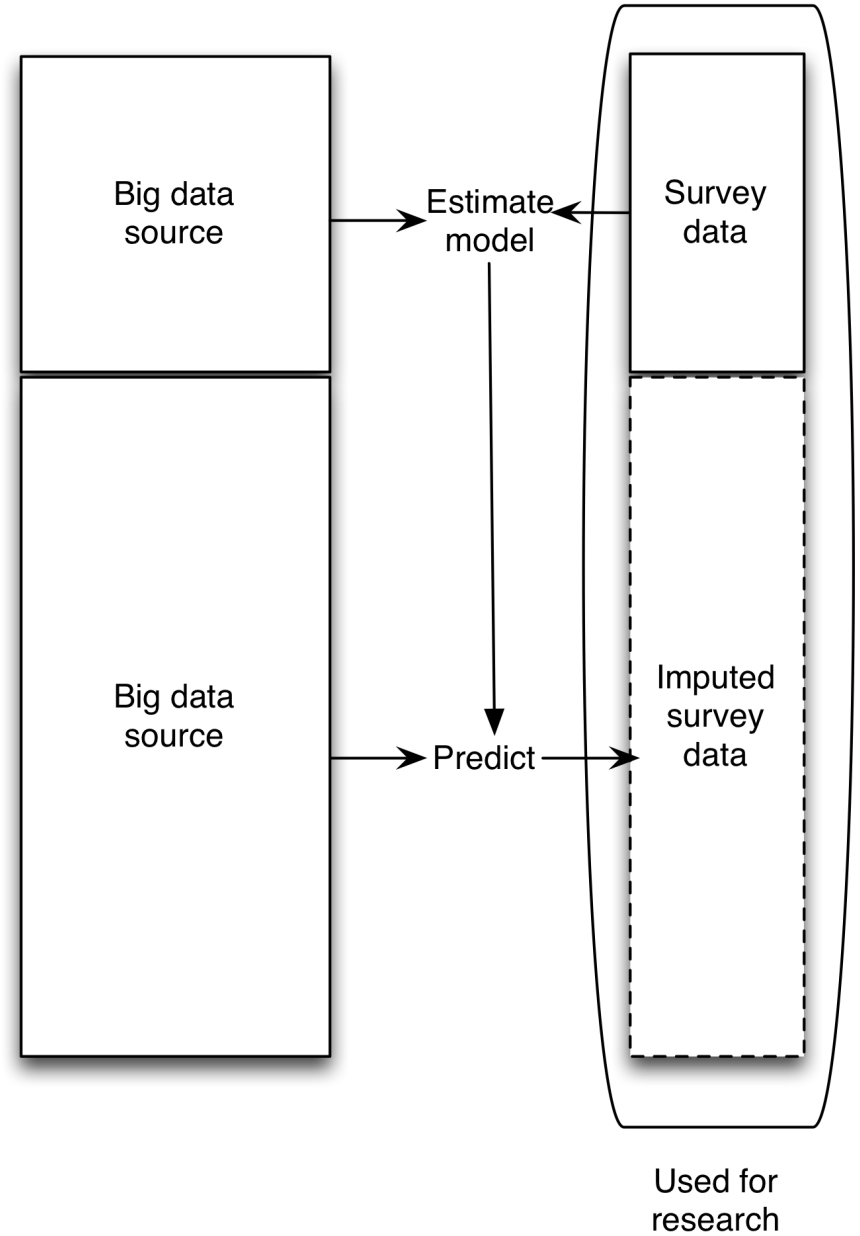
Crucial: record linkage

- Enriched asking assumes perfect linkage between data sources.
- If linkage is not perfect, this can be handled statistically.
- But …. This comes with assumptions.
- Record linkage is essential to make use of enriched asking.

Enriched asking



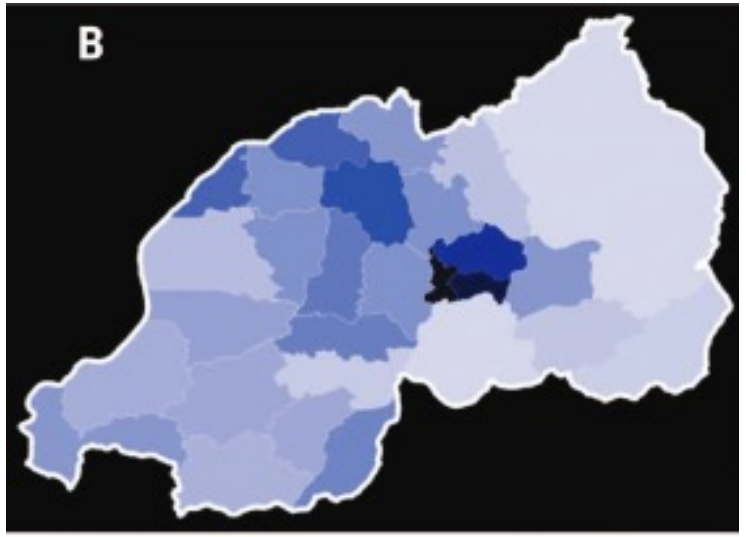
Amplified asking



Amplified asking

Use a predictive model to **combine** a small amount of **survey** data with a **big data** source to produce estimates at scale or granularity that would not be possible with either one of the two sources individually.

Example: Study by Blumenstock
Goal: help guide development in poor countries.
Concepts: measure wealth and wellbeing



(Blumenstock et al., 2015)

Traditional approaches

Survey:
- Hard to make estimates about specific geographical regions or demographic groups.

Census:
- Expensive
- Can only ask a few questions
- Does not happen often

Combine and get the best of both!

Collaboration with a phone company

Phone company:
- Has all phone record data

How to supplement with survey data:
- Select a random sample of phone numbers
- Call them
- Ask for consent (+ consent to link)
- Ask survey questions to measure wealth and well-being.

Two step procedure on phone data

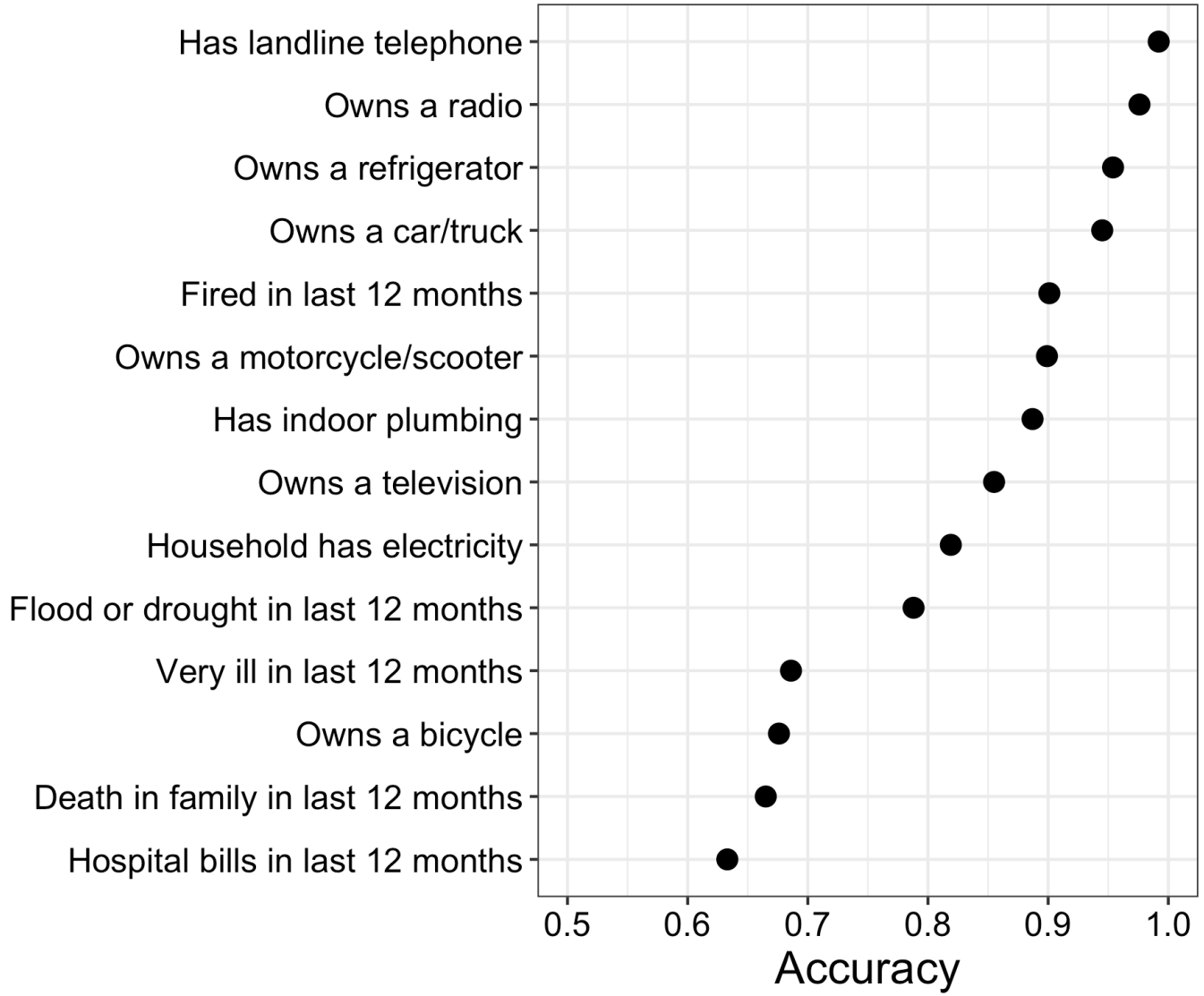
Step 1: Feature engineering step:

- Call records converted into a set of characteristics about each person.
- Features / variables
- Examples:
    - Number of days with activity
    - Number of distinct people a person has been in contact with
    - Amount of money spent on airtime

Two step procedure on phone data

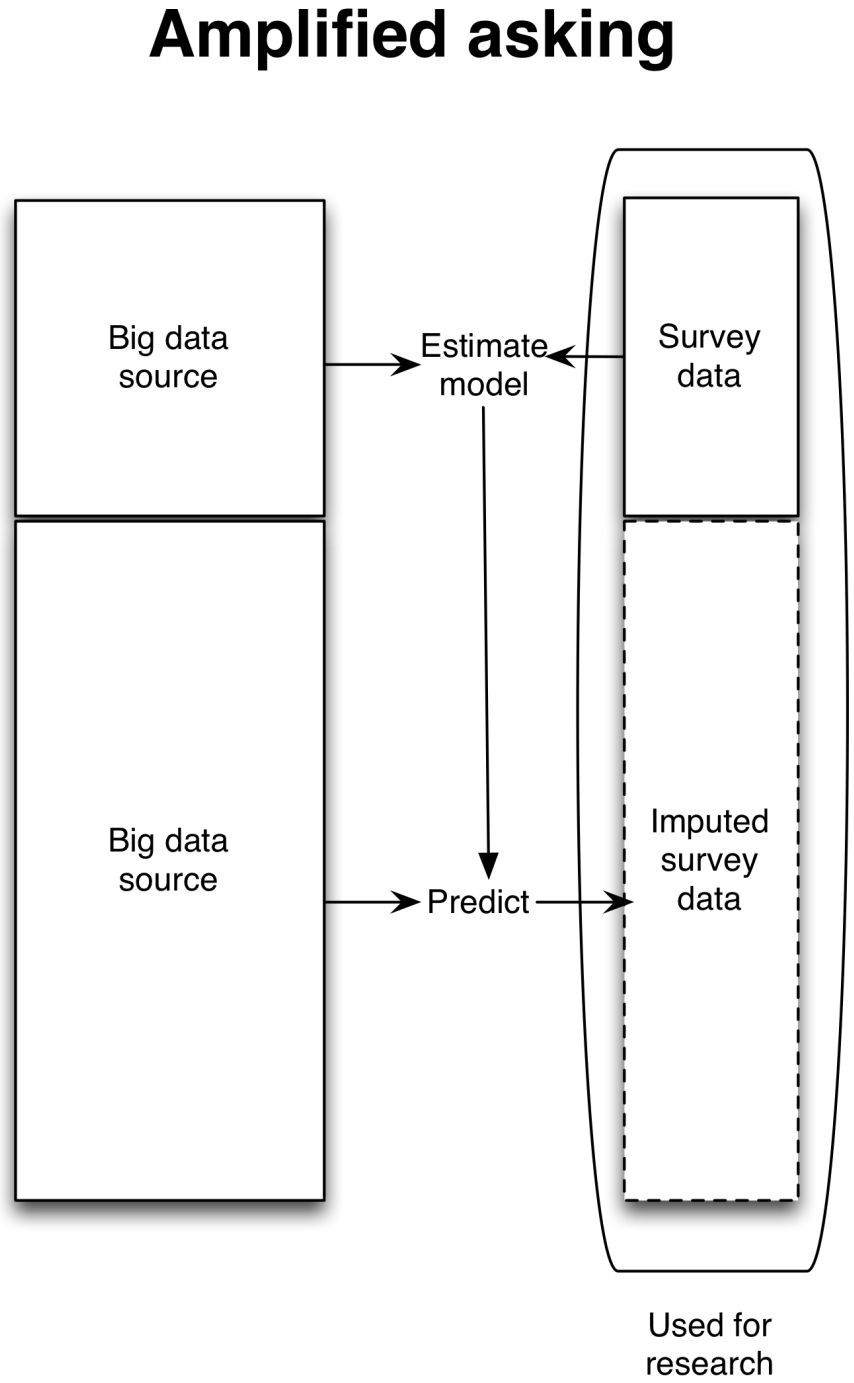
Step 2: Supervised learning step:

- Model to predict the survey response for each person based on their features.
- Used cross-validation to evaluate the performance of the model.
    - How well did the model perform beyond just making a baseline prediction?
    - Make 10 groups of persons in the data, train model on 9 and evaluate performance on 10, do this 10 times.



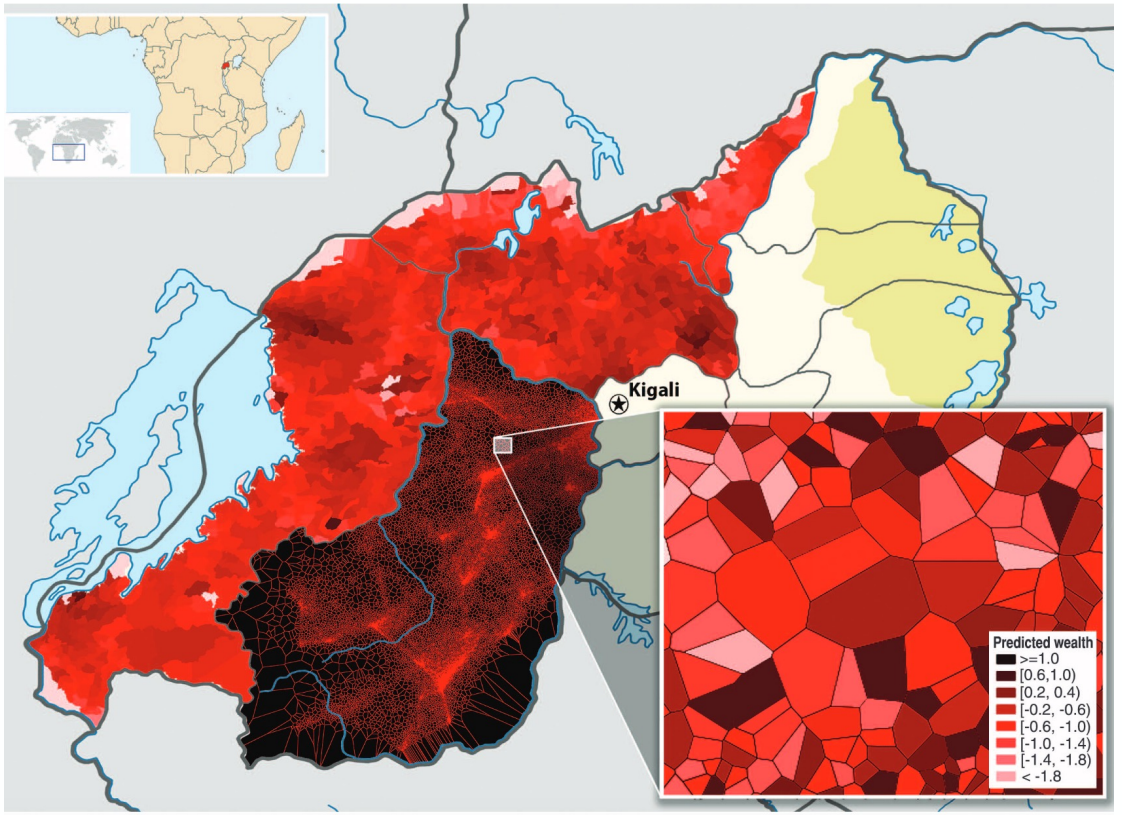
Next: Prediction model

- Instead of all separate survey variables, create composite wealth index.
- (again cross-validate)
- Predict the wealth index score of all 1.5 million people in the call records.



Next: Prediction model

- Predict the wealth of all 1.5 million people in the call records.
- Geospatial information from the call data gives an estimate of the geographic distribution of wealth at an extremely fine spatial granularity.
- Can estimate wealth of each of Rwanda's **neighborhood**.



Why be skeptical?

What types of errors are introduced by this procedure?

Why be skeptical?

- Predictions at individual level are noisy.
- People with mobile phones are systematically different from people without mobile phones.
    - Especially when it comes to wealth → **coverage** error
- Errors have been introduced during the feature selection and supervised learning step.
    - **Algorithmic** error for the measurement of wealth.

Comparison to high quality survey

- Survey considered the gold standard.
- Estimates were very similar.
- But this method: 10x faster and 50x cheaper.
- With the budget of the survey, which is done every couple of years, you can do this every month.

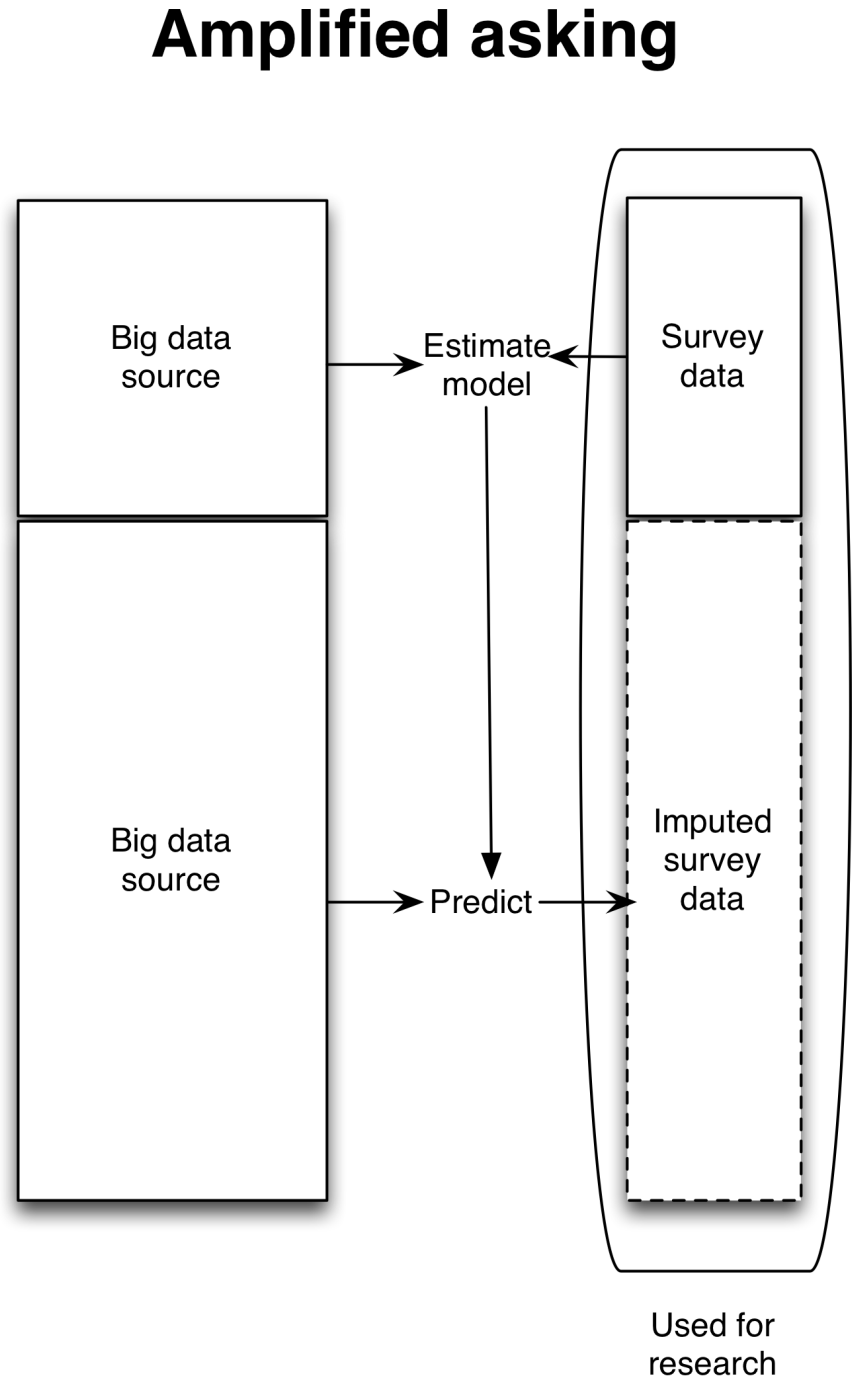
Trade-offs:

- No strong theoretical basis for this kind of approach.
- Do not know when this will work and when not.
- Especially coverage bias is/can be a big issue.

General conclusion

When you have
1. Data source has many variables but of few people
2. And one has few variables of many people

You can do:
1. For the people in both data sources, build a machine learning model that uses digital trace data to predict survey answers
2. Use that model to infer the survey answers of everyone in the big data source.

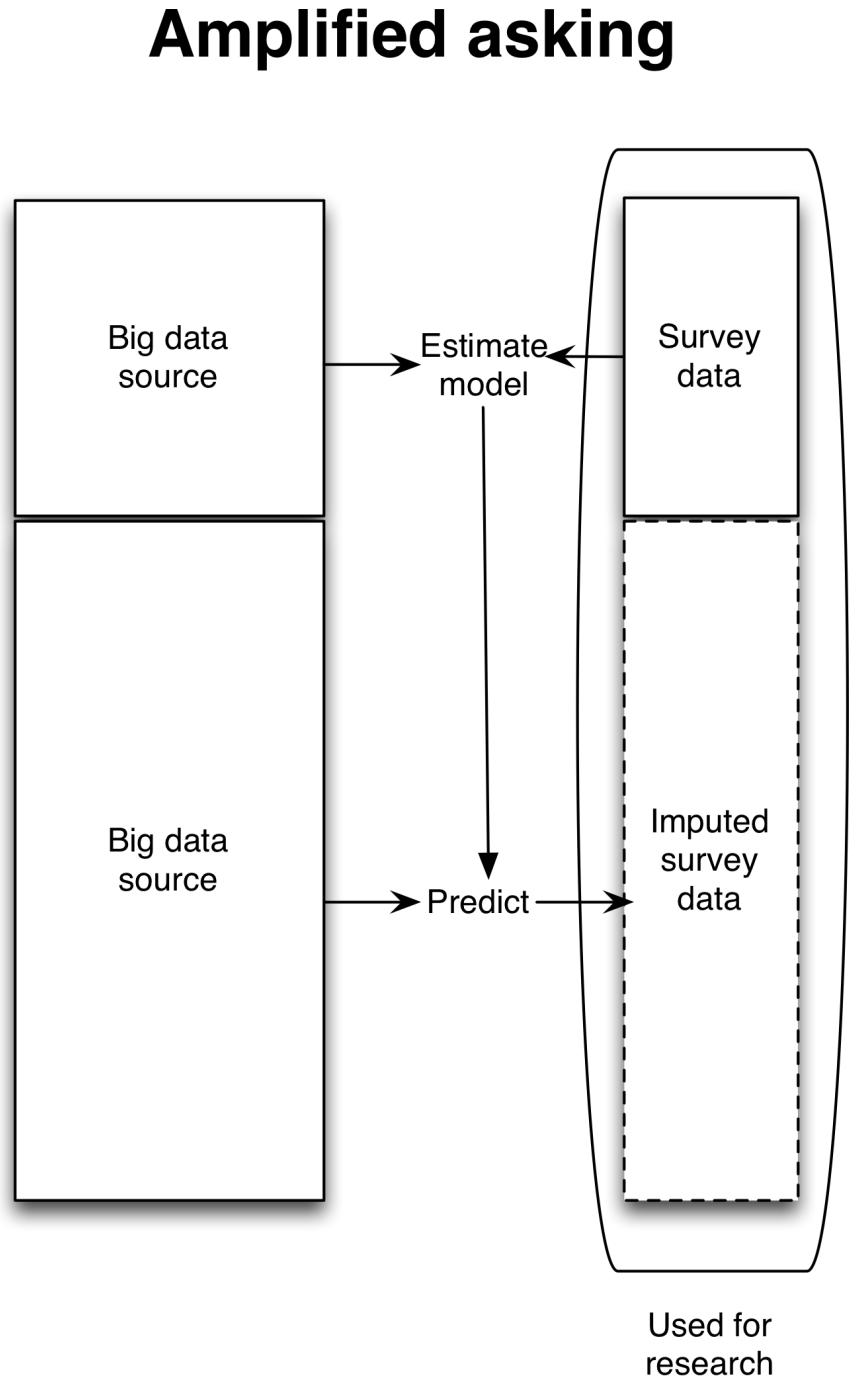


Generalize

Can use this approach for many different types of (combinations of) datasets

E.g. with a user-centric approach, we can:

- Have many people fill in a survey.
- A small group does the app/donation/etc.
- Make a model for the survey part.
- Use that to make predictions for the dtd part.
- Tailor your survey so that it can optimally inform the model.



Enjoy the lab meeting!