



# DIGITAL DEMOGRAPHY: ANALYZING WEB AND SOCIAL MEDIA DATA

## BASICS OF DIGITAL DEMOGRAPHY

EDSD DECEMBER 2024

Tom Theile

Departement of digital and computational demography,

Max-Planck-Institute for Demographic Research, Rostock

## Department of digital and computational demography







## Departement of digital and computational demography





## Agenda For the week

- Block 1: Internet, Webscraping, APIs
- Block 2: Digital Demography
- Block 3: Kinship Microsimulation with rsocsim
- Block 4: Bibliometric Data for migration research
- Extra: Digital Datasets
- Thursday + Friday: assignment



## DIGITAL DEMOGRAPHY.

What is 'digital' demography?





# IS 'BIG DATA' NEW?

DEMOGRAPHER COLLECTING BIG  
DATA FOR THE 1925 US CENSUS



[https://upload.wikimedia.org/wikipedia/commons/6/6f/Volkstelling\\_1925\\_Census.jpg](https://upload.wikimedia.org/wikipedia/commons/6/6f/Volkstelling_1925_Census.jpg)



## NOT DIGITAL DEMOGRAPHY

- Demography is old
- It is your job to come up with novel methods to capture/analyze/draw-conclusions-from demographic data
- → Do the old thing on computers and you have a new field!



## NOT DIGITAL DEMOGRAPHY

- Demography is old
- It is your job to come up with novel methods to capture/analyze/draw-conclusions-from demographic data
- →Computers!
- No, Digital and computational demography is more than just using computers to do traditional demography





## LIMITATIONS OF TRADITIONAL DATA SOURCES

- **Costly**
- Outdated
- Time consuming
- Inconsistent
- Unavailable
- Lack of data on emigration
- Incomplete answers/misunderstanding of questions etc.
- Immigrants are often underrepresented in traditional data sources.
- limited in hard-to-reach contexts and societies.
- „observation effect“, social desirability bias



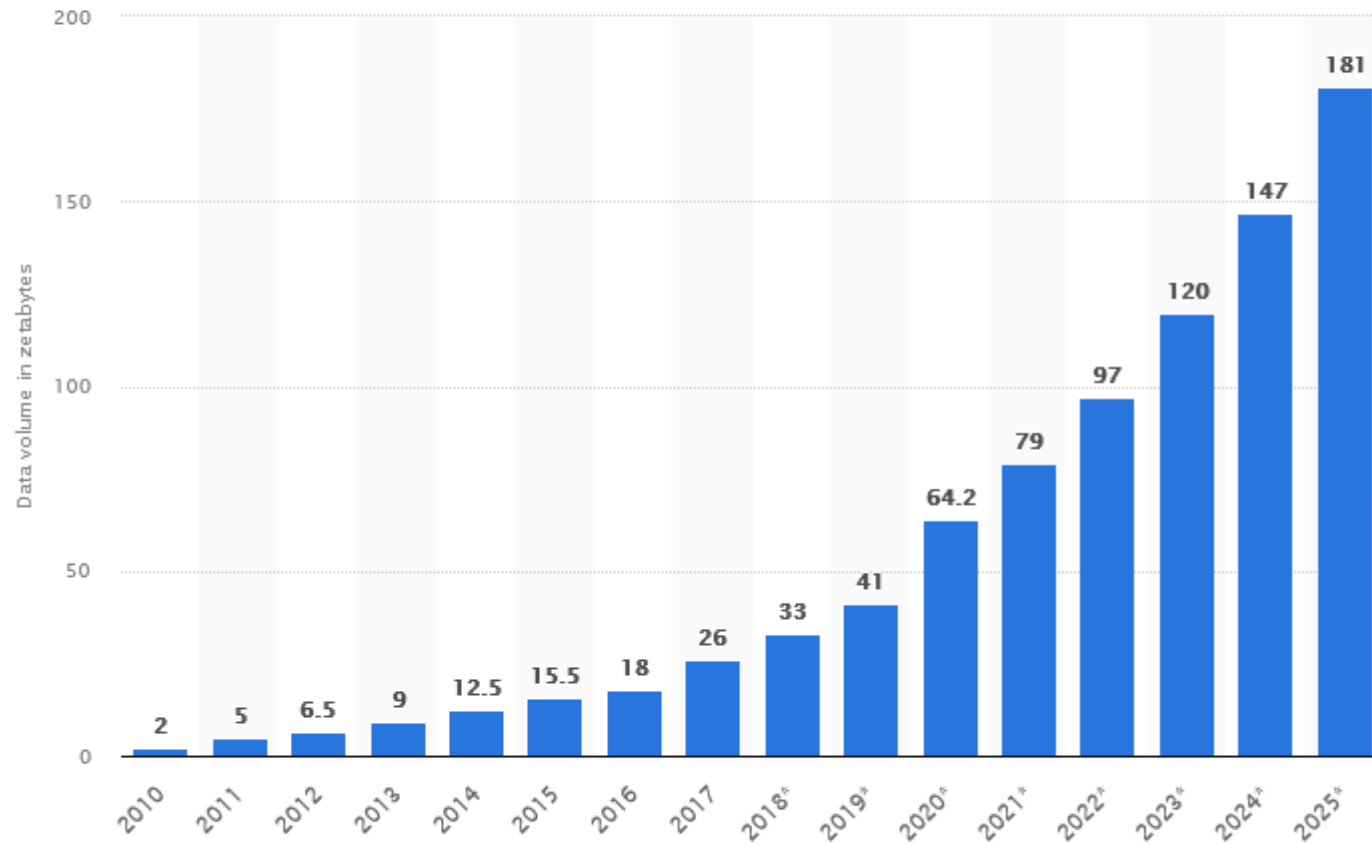
## DIGITAL DEMOGRAPHY

- Digital Data
- Digital methods

Use of data sources and methods that were not possible without computers

# Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025

(in zettabytes)



© Statista 2022

[Show source](#)

[Additional Information](#)

## DOWNLOAD



## Sources

- [→ Show sources information](#)
- [→ Show publisher information](#)
- [→ Use Ask Statista Research Service](#)

## Release date

June 2021

## Region

Worldwide

## Survey time period

2010 to 2020

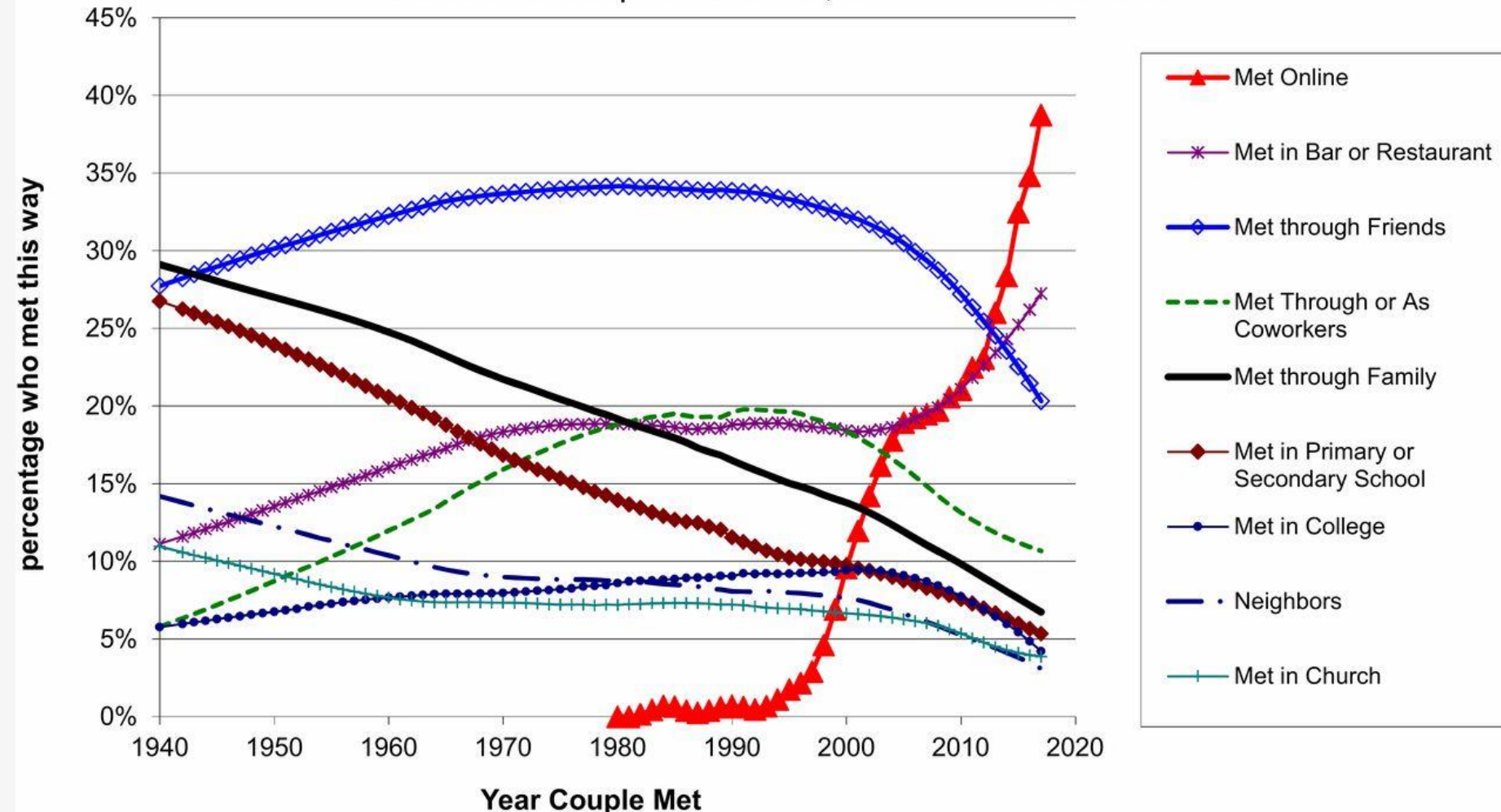
## Supplementary notes

\* The data was taken from various publications released over several years: Forecast for the years 2018 and 2019 as of 2018; Forecast for 2020 as of May 2021; Forecast for 2021 to 2025 as of March 2021.



# DIGITAL TRANSFORMATIONS HAVE CHANGED OUR LIVES

How heterosexual couples have met, data from 2009 and 2017



Source: <https://www.pnas.org/doi/10.1073/pnas.1908630116> - Rosenfeld 2019

# DIGITAL DATA SOURCES FOR DEMOGRAPHIC RESEARCH

1. Digital Trace Data (online and offline)
  1. Social media
  2. Mobile phones
  3. Wearable devices, etc...
2. Crowd-sourced online data
  1. Wikipedia and Wikidata
  2. DNA and online genealogies
  3. Petitions, etc...
3. Online Surveys
4. Simulations (made-up data, based on known rules and assumptions)

# WHAT IS DIGITAL TRACE DATA?

***What makes digital trace data a type of ‘new’ big data for population research that is different from ‘old’ big population data sources?***

- *by-product of digital activity*
- *always collected, not once in a while*
- *Unlike rectangular dataframes with rows and columns, many digital traces data are unstructured, messy, and come in formats unfamiliar to many demographers*
- *The variety of formats, units of analyses, and also sizes of these data sets, which may contain*
- *millions of records, often require computational approaches for data management, retrieval, and analysis that are not yet a part of mainstream demo-graphic training.*
- *not comparable over longer time periods - platforms, users and algorithms change fast*
- *these data often come from and are owned by private companies*

Ridhi Kashyap: Has demography witnessed a data revolution? Promises and pitfalls of a changing data ecosystem? 2021





# DIGITAL TRACES ARE BY-PRODUCTS OF OUR ONLINE PRESENCE

Digital breadcrumbs are unavoidable

- ▶ Pre-GDPR: largely unchecked
- ▶ Marketing-led
- ▶ Not collected for social-scientific research



## BIG DATA FOR SOCIAL RESEARCH: THE GOOD

Twitter, Bluesky, Facebook, Google, LinkedIn, Instagram, TikTok, ...

- Big
- Free (sometimes) or cheap (often, not always)
- Granular data
- Large scale data
- Continuously generated, always-on
- Non-reactive
- Information/opinion shared by users from an uncontrolled environment
- Various forms of data: video, image, text, audio etc.

## BIG DATA FOR SOCIAL RESEARCH: THE BAD

- Incomplete
- Inaccessible
- Biased sample, nonrepresentative (within- and out-of sample)
- drifting (population, behavioural, system)
- algorithmically confounded
- Sometimes inaccessible
- dirty
- sensitive

Salganik, M. (n.d.). Bit by Bit: Social Research in the Digital Age. Princeton, NJ: Princeton University Press.

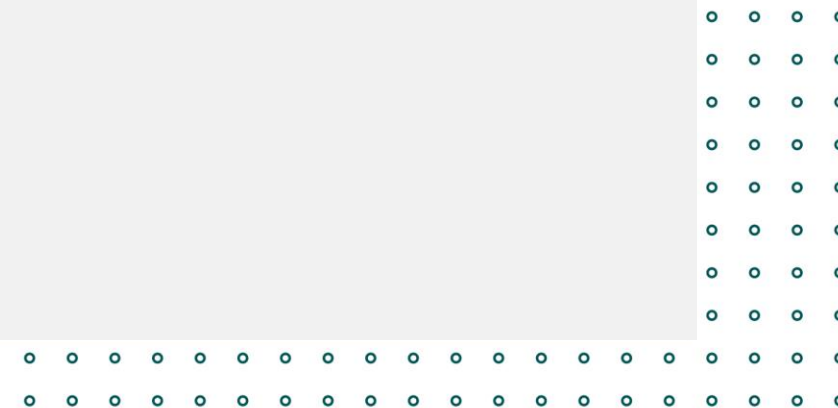


# CURRENT TOPICS IN DIGITAL DEMOGRAPHY

1. Methodological developments
  1. Inference from non-representative samples
  2. Understand and address online bias
  3. Nowcasting demographic processes
2. Understanding internet users and online use
  1. Infer demographics (age, sex, location, SE status, etc) from image and text
  2. Track inequalities in online access
  3. Consequences of platform use for users
3. Migration (internal and external)
  1. Estimate flows and stocks
  2. Mobility by subgroups (e.g. undocumented, highly-skilled)
  3. Cultural assimilation of immigrants
4. Mortality and morbidity
5. Online and offline fertility dynamics
6. Time use and well-being



# SURVEYS



# ONLINE SURVEYS ON SOCIAL MEDIA

Relative advantage: scale

Important for:

- Targeting specific subgroups or regions
- Conducting research outside of Western countries







## ISSUES WITH TRADITIONAL SURVEYS

Efficacy: traditional sampling methods are outdated and less feasible

Coverage: decreased response rates, difficult to sample from hard-to-reach populations

Resources: expensive and time consuming

Recency: quickly become outdated, long period between new data collections

Comparability: lack of common definitions across countries

# ONLINE SURVEYS



Cost-effectiveness: less expensive than traditional surveys

Coverage: targeted sub-populations, hard-to-reach populations

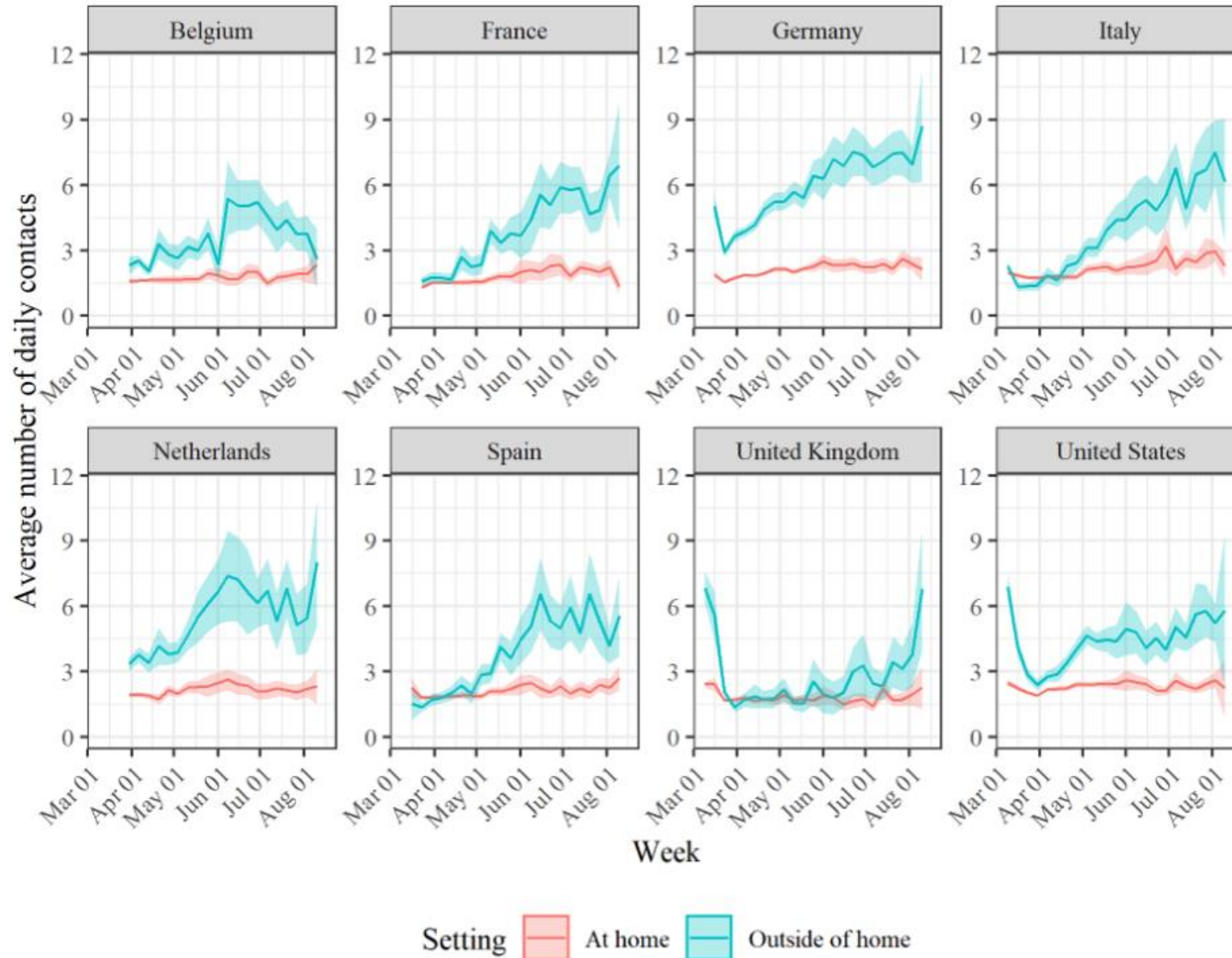
Timeliness: easy and timely implementation, data collection & analysis in near real-time

Flexibility: less burdensome, user-friendly interfaces, easy to manage

Recency: continuous data collection, easy to make edits

Comparability: cross-national surveys, comparative data collections, common definitions

# ONLINE SURVEYS ON SOCIAL MEDIA



Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S., & Zagheni, E. (2020). Addressing public health emergencies via Facebook surveys: advantages, challenges, and practical considerations. *Journal of medical Internet research*, 22(12), e20653.

# ONLINE SURVEYS ON SOCIAL MEDIA

**Table 4.** Comparison of Sample Sizes at Different Stages of the Sampling and Survey Process.

	Number of Users Belonging to the Target Population (According to Facebook) <sup>a</sup>	Unique Users Reached with Ads <sup>a</sup>		Paid Link Clicks <sup>a</sup>		Completed Questionnaires <sup>b</sup>	
		Percentage of Targeted FB Users in this Country		Percentage of Targeted FB Users in this Country		Percentage of Targeted FB Users in this Country	
		<i>n</i>		<i>n</i>		<i>n</i>	
Austria	15,000	7,918	52.79	408	2.72	117	0.78
Ireland	54,000	28,107	52.05	1,314	2.43	425	0.79
Switzerland	9,000	3,432	38.13	215	2.39	62	0.69
United Kingdom	410,000	50,979	12.43	1,257	0.31	424	0.10

*Note.* <sup>a</sup>Source of absolute figures: Facebook (FB) advertisement statistics. Relative values: own calculation. <sup>b</sup>Based on paradata. Only respondents who reached the questionnaire via the FB advertisements.

Budget of €500, and 96% of the 1,028 respondents belonged to target population.

- Pötzschke, S., & Braun, M. (2017). Migrant sampling using Facebook advertisements: A case study of Polish migrants in four European countries. *Social Science Computer Review*, 35(5), 633-653.

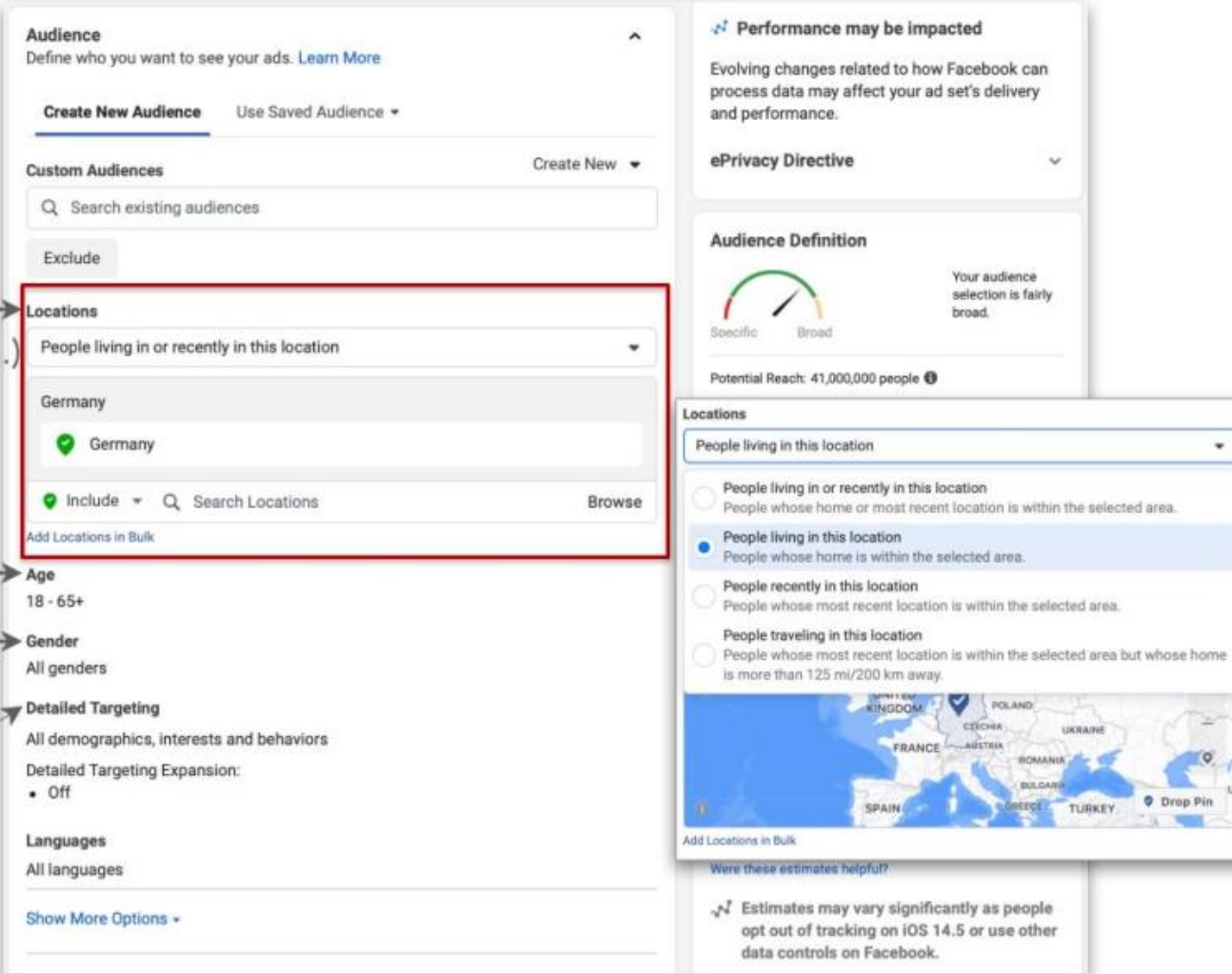
# ONLINE SURVEYS ON SOCIAL MEDIA

**Location**  
(country, region, city, ...)

**Age (ranges)**

**Gender**  
(men, women, both)

**Other matching criteria**  
(interests, industry, behaviours, ...)



**Audience**  
Define who you want to see your ads. [Learn More](#)

**Create New Audience** Use Saved Audience ▾

**Custom Audiences** Create New ▾

Q Search existing audiences

Exclude

**Locations**

People living in or recently in this location ▾

Germany

✓ Germany

✓ Include ▾ Q Search Locations Browse

Add Locations in Bulk

**Age**  
18 - 65+

**Gender**  
All genders

**Detailed Targeting**  
All demographics, interests and behaviors  
Detailed Targeting Expansion:  
• Off

**Languages**  
All languages

[Show More Options ▾](#)

**Performance may be impacted**  
Evolving changes related to how Facebook can process data may affect your ad set's delivery and performance.

**ePrivacy Directive** ▾

**Audience Definition**

Your audience selection is fairly broad.

Potential Reach: 41,000,000 people ⓘ

**Locations**


People living in this location ▾

☐ People living in or recently in this location  
People whose home or most recent location is within the selected area.

☒ People living in this location  
People whose home is within the selected area.

☐ People recently in this location  
People whose most recent location is within the selected area.

☐ People traveling in this location  
People whose most recent location is within the selected area but whose home is more than 125 mi/200 km away.



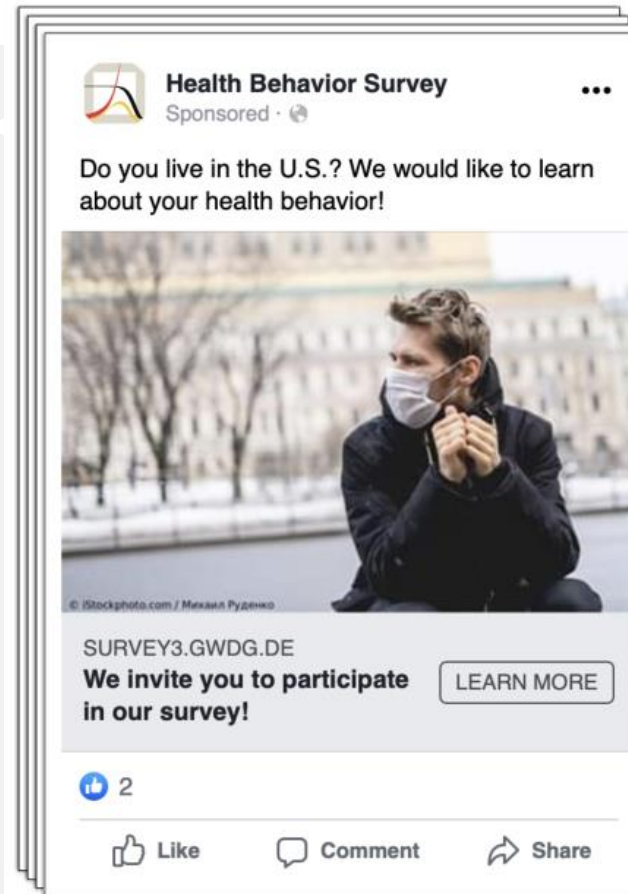
Add Locations in Bulk

Were these estimates helpful?

Estimates may vary significantly as people opt out of tracking on iOS 14.5 or use other data controls on Facebook.



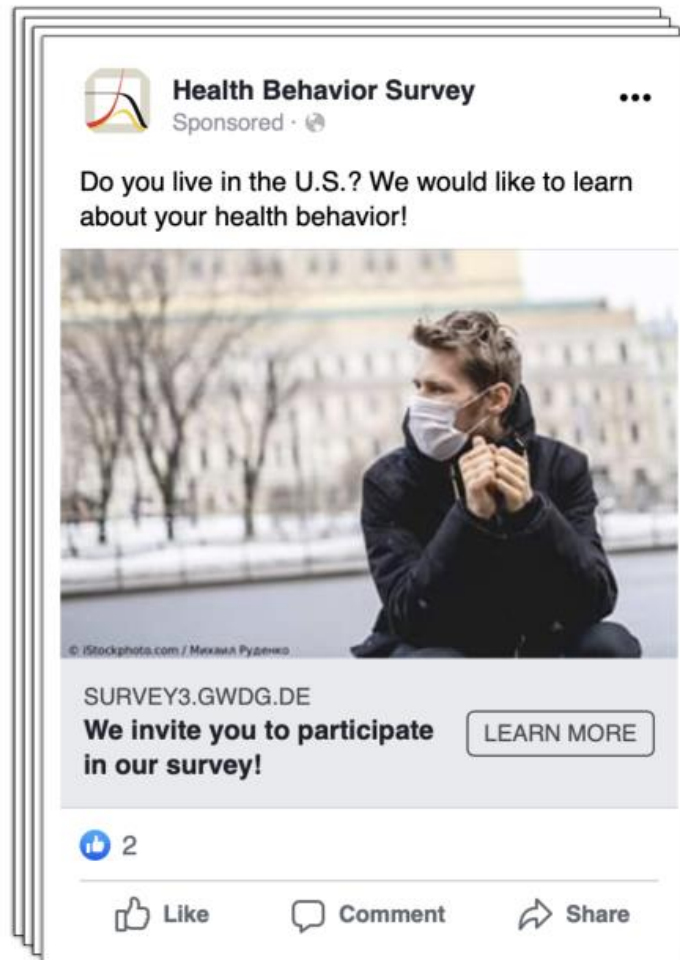
# ONLINE SURVEYS



Example of FB ad in the US.



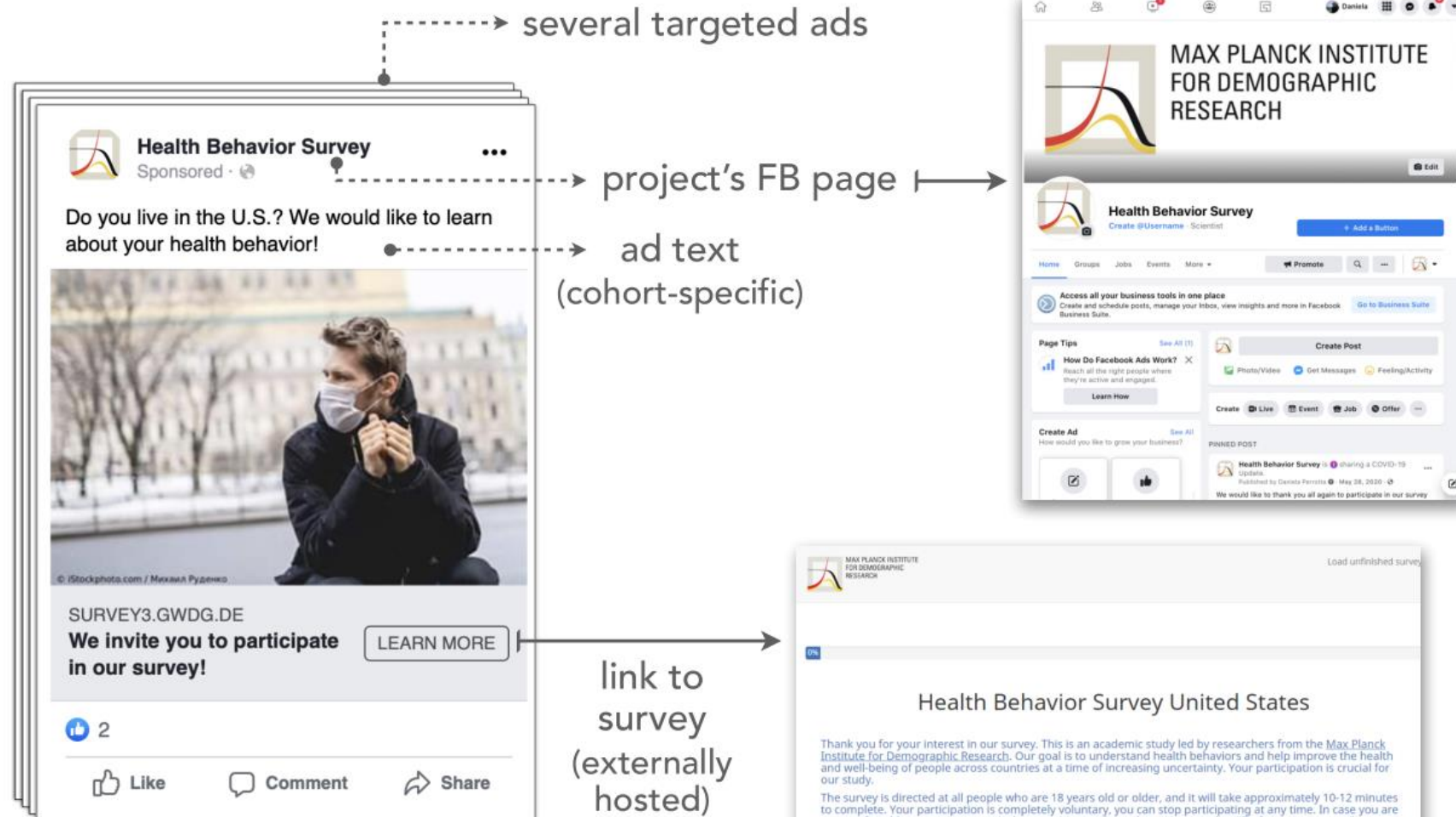
# RESPONDENTS RECRUITMENT VIA FACEBOOK



Example of FB ad in the US.

Grow A, Perrotta D, Del Fava E, Cimentada J, Rampazzo F, Gil-Clavel S, Zagheni E. Addressing Public Health Emergencies via Facebook Surveys: Advantages, Challenges, and Practical Considerations. JMIR, 2020

# RESPONDENTS RECRUITMENT VIA FACEBOOK



Example of FB ad in the US.

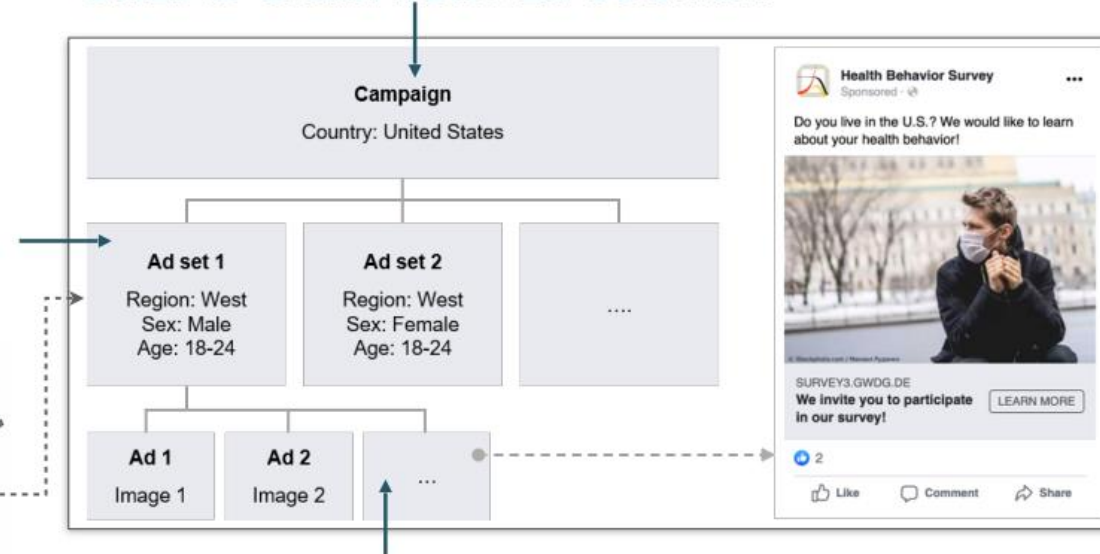
# RESPONDENTS RECRUITMENT VIA FACEBOOK

## ONE AD SET PER DEMOGRAPHIC GROUP

- ▶ sex (M, F)
- ▶ age (18-24, 25-44, 44-64, 65+)
- ▶ region of residence (NUTS1/US Census regions)



## ONE AD CAMPAIGN PER COUNTRY



## SIX AD IMAGES WITHIN EACH AD SET



1 – Male athlete  
©Adobe Stock/grki



2 – Group of athletes  
©Adobe Stock/nd3000



3 – Woman blowing nose  
©iStockphoto/Goodboy Picture Company



4 – Couple blowing noses  
©iStockphoto/Goodboy Picture Company



5 – Woman wearing mask  
©Adobe Stock/shintartanya



6 – Man wearing mask  
©iStockphoto/Михаил Руденко

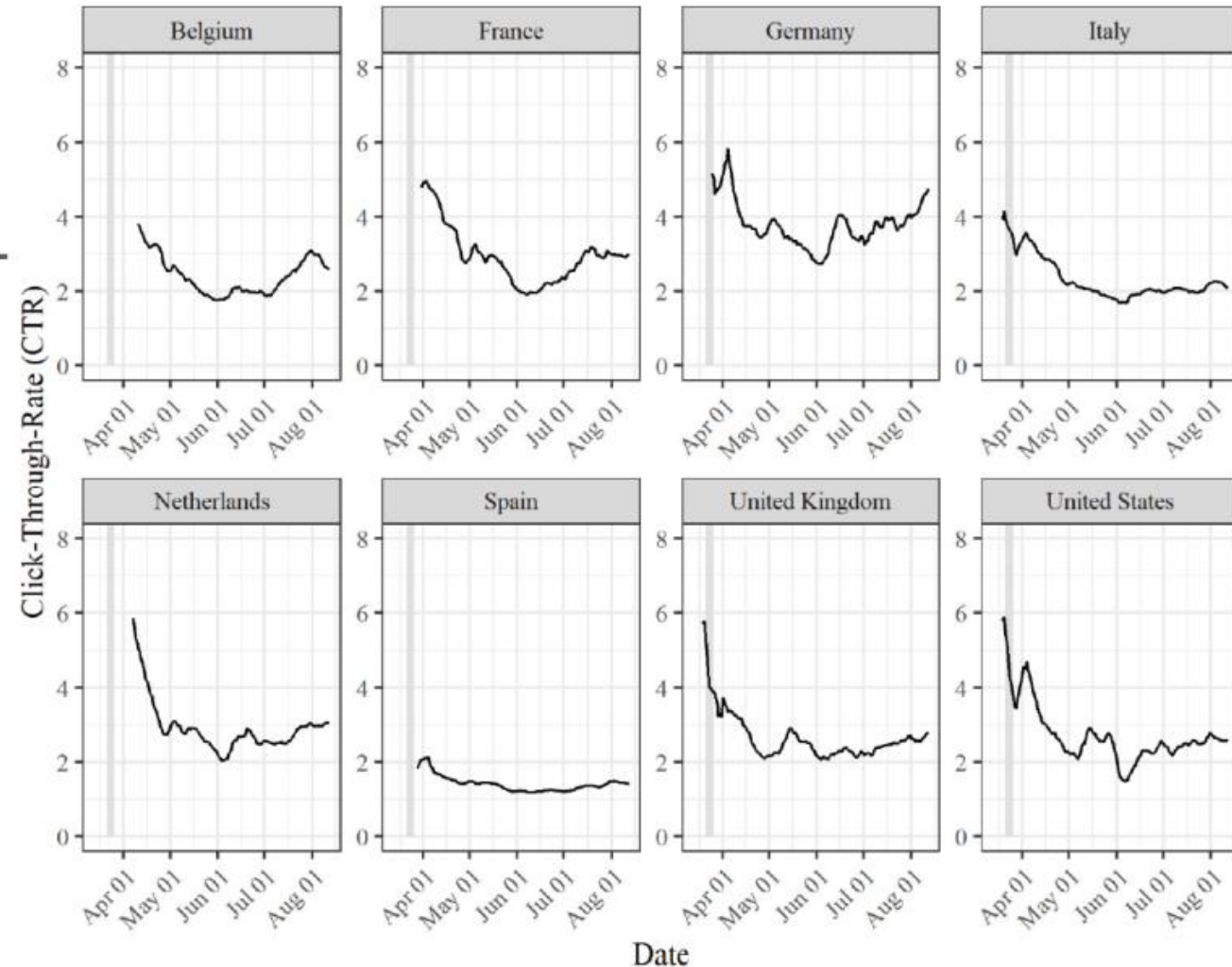
Grow A, Perrotta D, Del Fava E, Cimentada J, Rampazzo F, Gil-Clavel S, Zagheni E. *Addressing Public Health Emergencies via Facebook Surveys: Advantages, Challenges, and Practical Considerations*. JMIR, 2020



# RESPONDENTS RECRUITMENT VIA FACEBOOK

Click-through rate (CTR) =  
click-throughs / impressions

Facebook users were  
more likely to click on  
our ads in the early  
phases of the survey

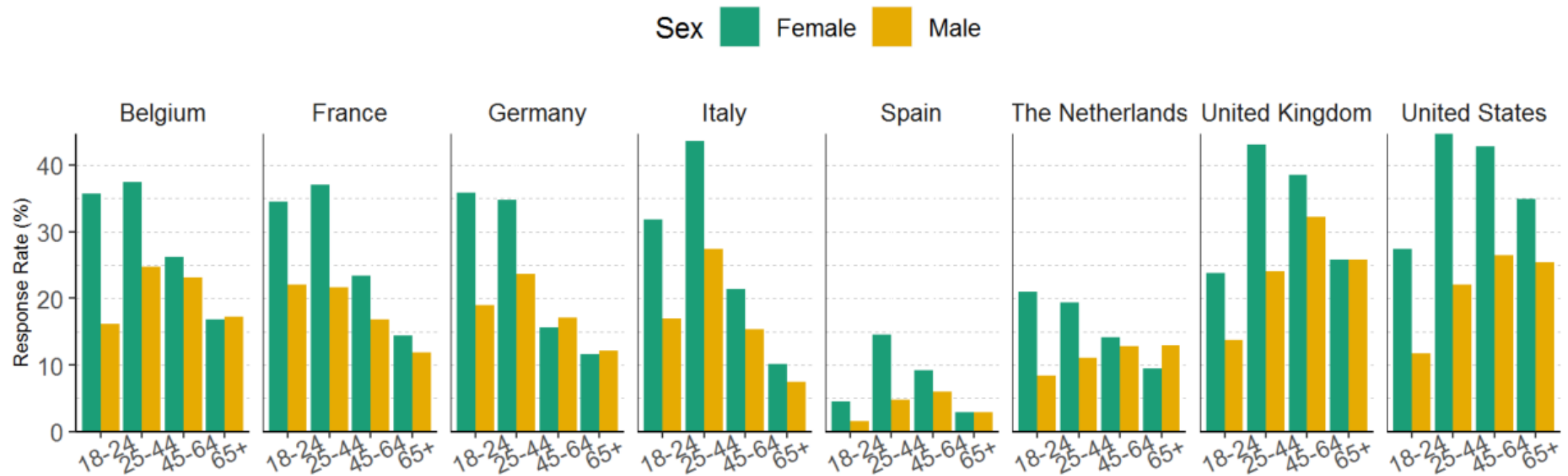




# RESPONDENTS RECRUITMENT VIA FACEBOOK

Response rate (%) = completed questionnaires / click-throughs

- ▶ response rate (overall): from 6% in Spain to 31% in UK and US
- ▶ response rate higher for women



# ONLINE SURVEYS ON SOCIAL MEDIA

## Challenges of recruitment via Facebook:

- Facebook is a “black box”: Facebook will “optimize” your advertisement in ways that skew participant demographics.
  - Also: Self-selection based on interest in the survey topic
  - Partial solution: run lots of advertisements targeting specific demographic groups. (Grow et al. 2020)
- Facebook will sometimes cause delays, either by reviewing your advertisements or claiming problems with your payment method.
  - Delays are unpredictable, but they are more common if your ad is related to social or political topics or if you offer to pay participants.
- Facebook has no built-in methods for paying participants or advertising specifically to people who previously completed your survey.
- Lack of attention of respondents

Rinderknecht, Gordon, PHDS 2022

## CROWDSOURCED PLATFORMS

### Relative advantage: convenience

- They facilitate participant payment, re-recruitment, and messaging without requesting personally identifying information. These features allow for complex research designs.



## Prescreen participants 1

### YOUR CRITERIA

#### Age

Minimum Age: 25, Maximum Age: 60

[Edit](#) [Remove](#)

#### COVID-19 Vaccination

Yes (at least one dose)

[Edit](#) [Remove](#)

[+ Add screener](#)

We've found **28,650** matching participants who have been active in the past 90 days

### STUDY COST

How long will your study take to complete?

1 Max. time: 30 mins

Participants are paid according to your estimated study completion time. If the median completion time exceeds your estimate we will ask you to make additional payments. [Read more about study completion time](#) [↗](#)

🕒 5 minutes

How much do you want to pay them?

£ 1.00

12.00/hr

Hourly rate

£6.00

£12.00 Great!

#### Cost

Participant payments	£500.00
<a href="#">Service fee</a>	£166.66
VAT (0% on service fee)	£0.00
<b>Total</b>	<b>£666.66</b>



Aguinis, H., Villamor, I., &  
Ramani, R. S. (2021). MTurk  
research: Review and  
recommendations. Journal of  
Management, 47(4), 823-837.

# CROWDSOURCED PLATFORMS

## Relative disadvantages

- Population size is relatively small. While researchers were able to recruit ~1,000 Polish migrants in four countries with Facebook, and Facebook reported ~500,000 such users on their platform, Prolific reports only ~400 such respondents available for recruitment in these countries.
- Participants tend to be concentrated in the West, especially the U.S.
- You cannot use these platforms to recruit voluntary samples





# SOCIAL MEDIA AND CROWDSOURCED PLATFORMS

**What do you think are potential problems?**



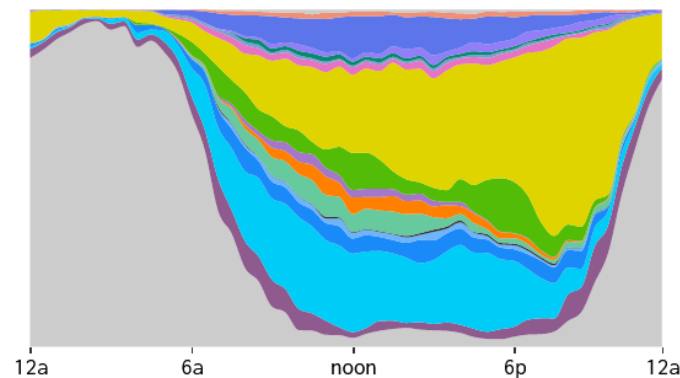
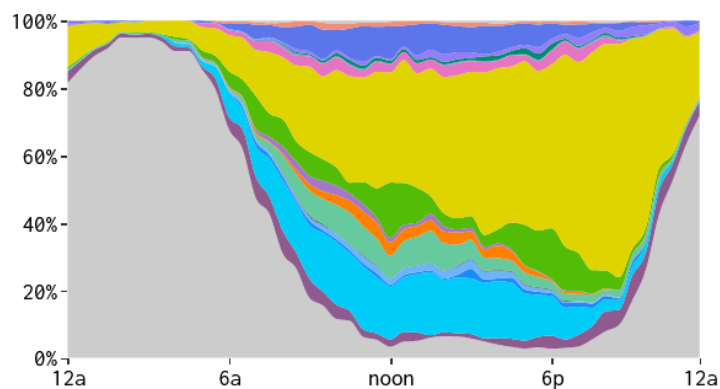
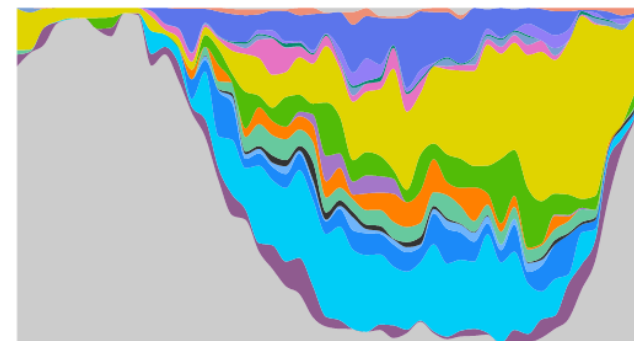
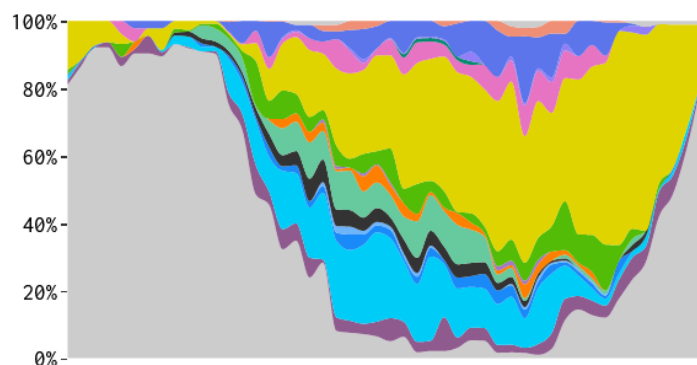
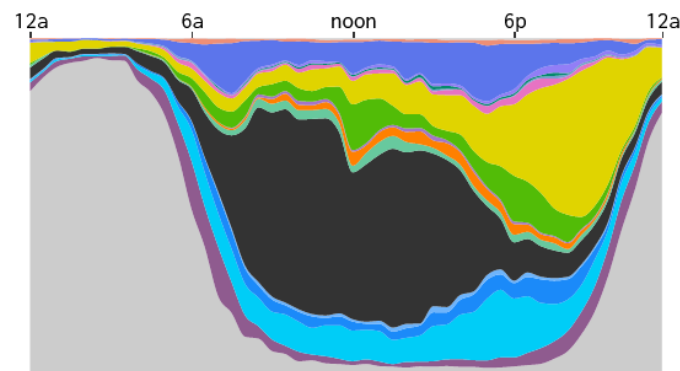
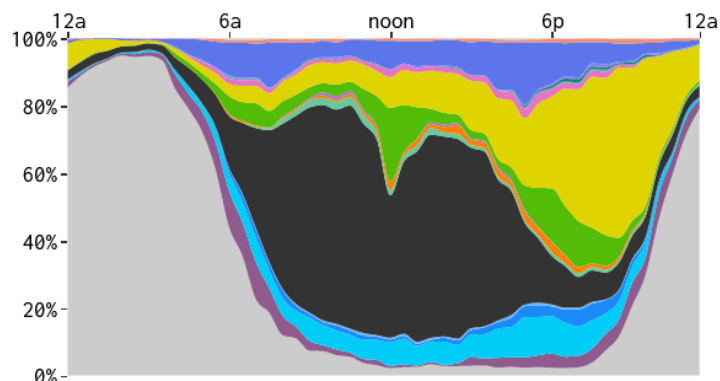
EMPLOYED

UNEMPLOYED

NOT IN LABOR FORCE

MEN

WOMEN



- Sleeping
- Personal Care
- Household Activities
- Caring for and Helping Household Members
- Caring for and Helping Non-Household Members
- Working
- Education
- Consumer Purchases
- Professional and Personal Care Services
- Eating and Drinking
- Socializing, Relaxing, and Leisure
- Sports, Exercise, and Recreation
- Religious and Spiritual Activities
- Volunteer Activities
- Telephone Calls
- Traveling
- Other
- Gap/can't remember

## Who participated in this activity with you?

<input type="checkbox"/>	Spouse/partner(s)
<input checked="" type="checkbox"/>	Own child/children
<input type="checkbox"/>	Other family member(s)
<input type="checkbox"/>	Co-worker/colleague(s)
<input checked="" type="checkbox"/>	Friend(s)
<input type="checkbox"/>	Other People
<input type="checkbox"/>	Pet(s)
<input type="checkbox"/>	No one

Select anyone you primarily engaged with via the internet or phone:

<input type="checkbox"/>	Own child/children
<input checked="" type="checkbox"/>	Friend(s)

- Rinderknecht, R. G., Doan, L., & Sayer, L. C. 2022. “MyTimeUse: An Online Implementation of the Day-Reconstruction Method” *Journal of Time Use Research*.



## PREPARATION

Please read:

- KASHYAP, Ridhi. Has demography witnessed a data revolution? Promises and pitfalls of a changing data ecosystem. *Population Studies*, 2021, 75. Jg., Nr. sup1, S. 47-75.
- You can find it in the folder “Readings”
- A long paper, but a good overview. Please distribute the work of reading the paper in your group. Tomorrow each person will explain/present their part to their group.



THANK YOU FOR  
YOUR ATTENTION!

**Tom Theile**

Research Software Engineer

theile@demogr.mpg.de