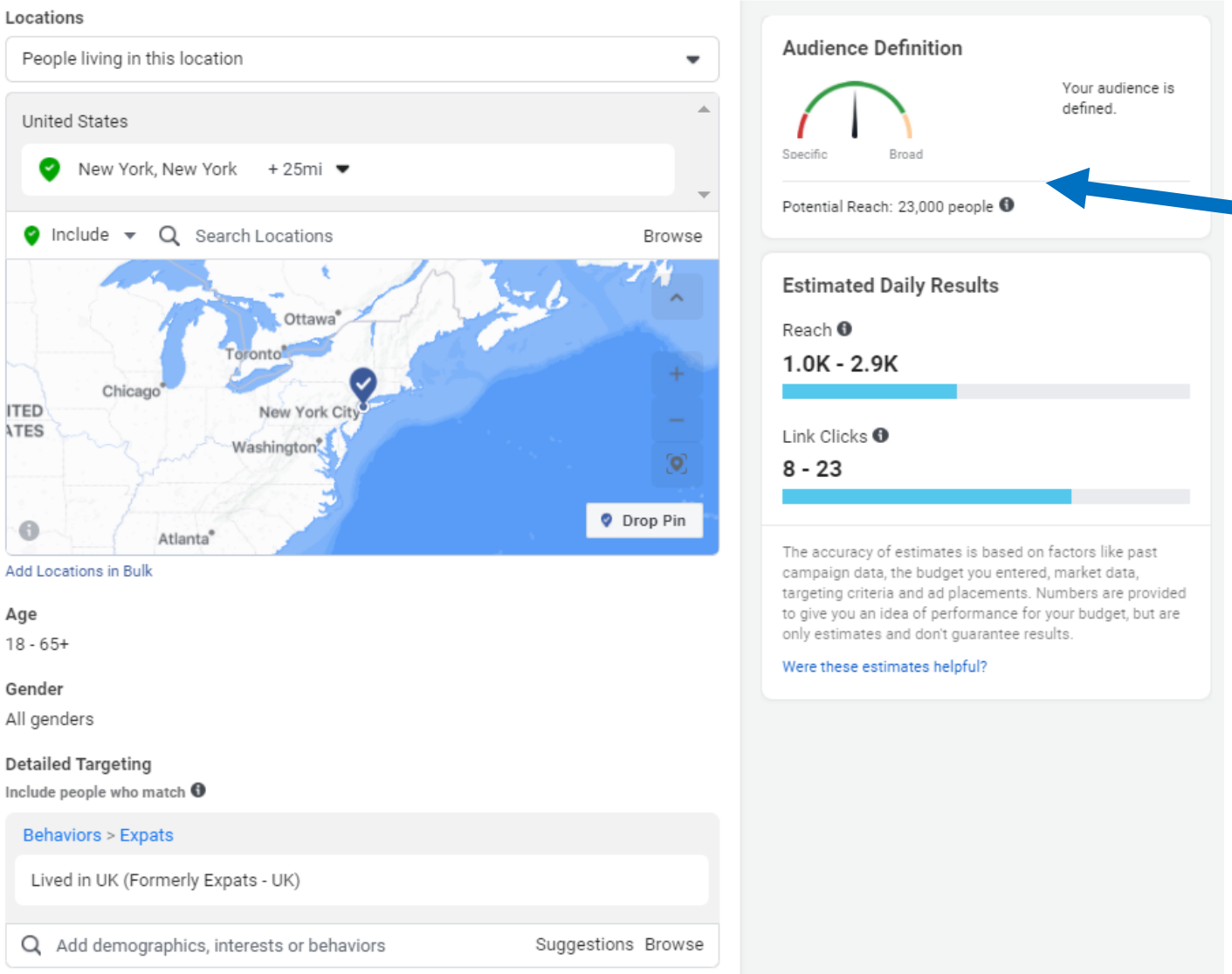
MEASURING AND UNDERSTANDING BIAS IN DIGITAL TRACE DATA

Emilio Zagheni & Tom Theile

Max Planck Institute for Demographic Research

SICSS Saarbruecken – Sept 10, 2025

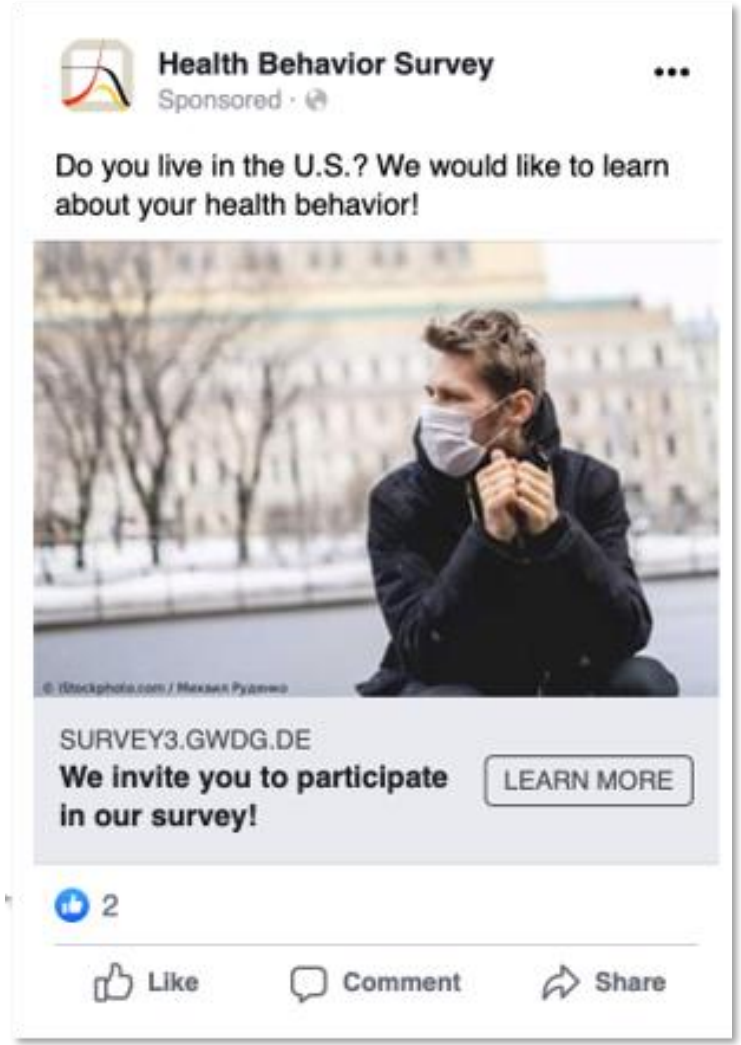
ADVERTISING PLATFORMS LIKE THOSE OF FACEBOOK, LINKEDIN, ETC.

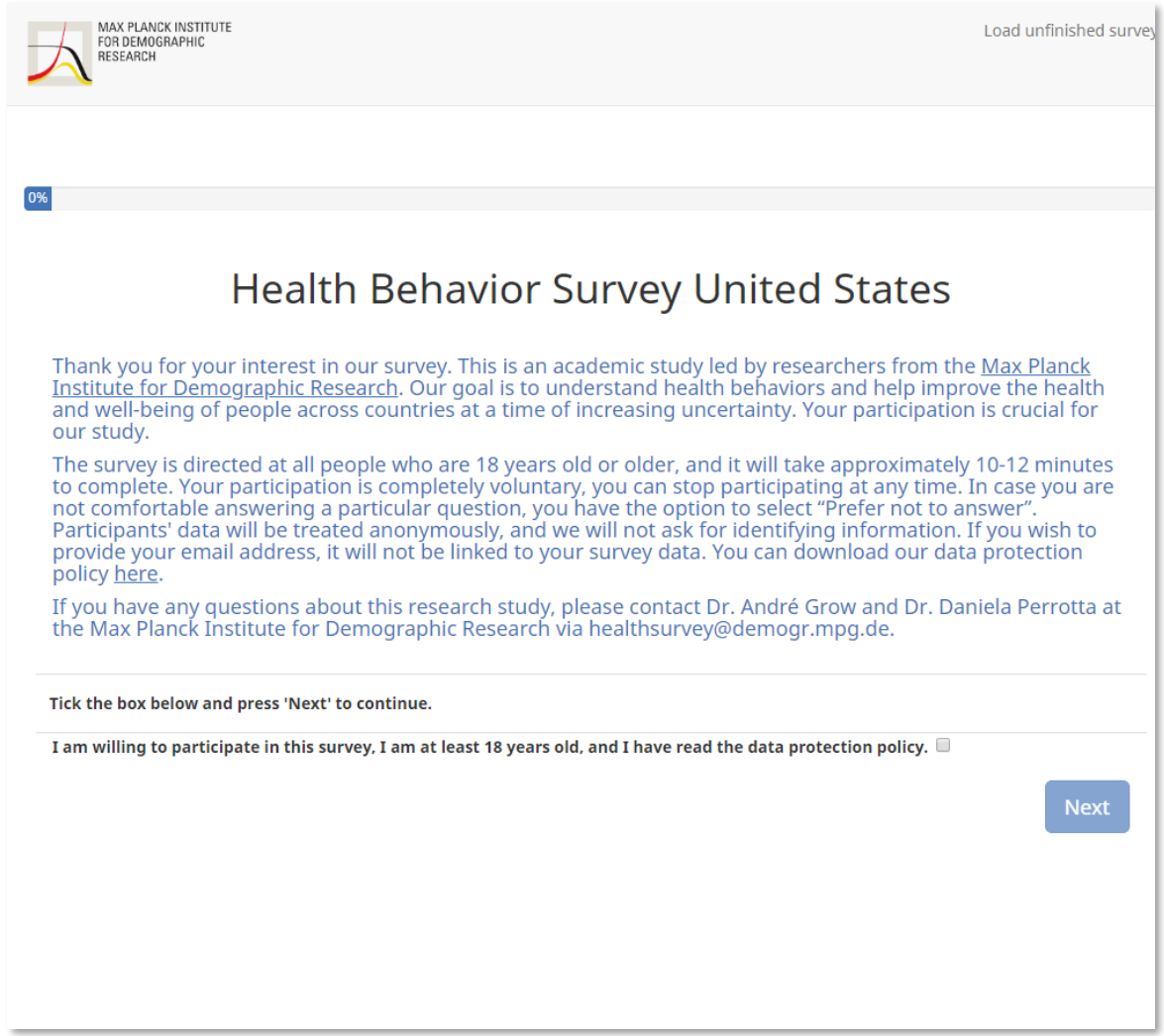


How many users who used to live in the UK, now live in New York City?

www.facebook.com/business

FACEBOOK ADS AS A RECRUITMENT TOOL FOR SURVEY PARTICIPANTS





WARM UP

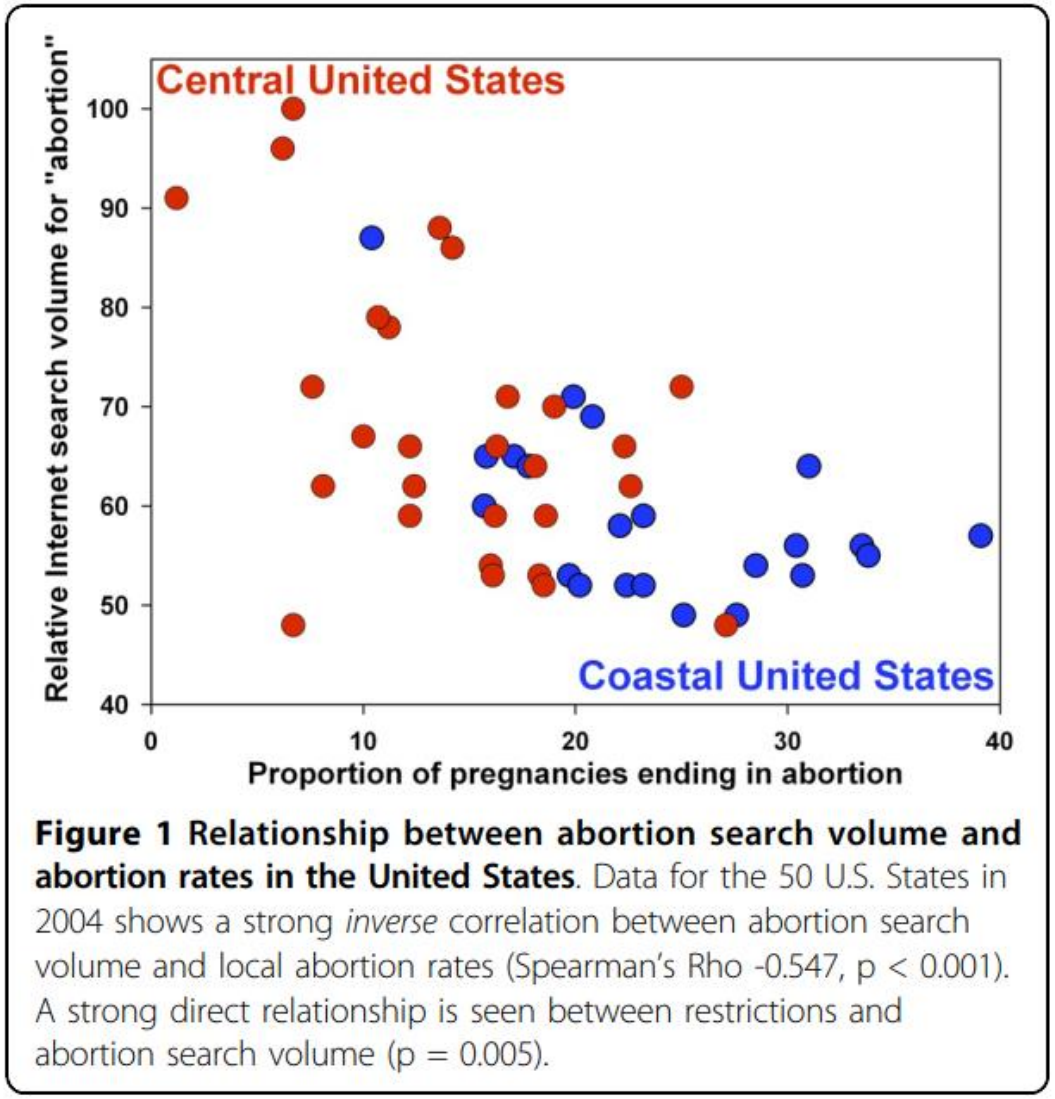


Figure 1 Relationship between abortion search volume and abortion rates in the United States. Data for the 50 U.S. States in 2004 shows a strong *inverse* correlation between abortion search volume and local abortion rates (Spearman's Rho -0.547, $p < 0.001$). A strong direct relationship is seen between restrictions and abortion search volume ($p = 0.005$).

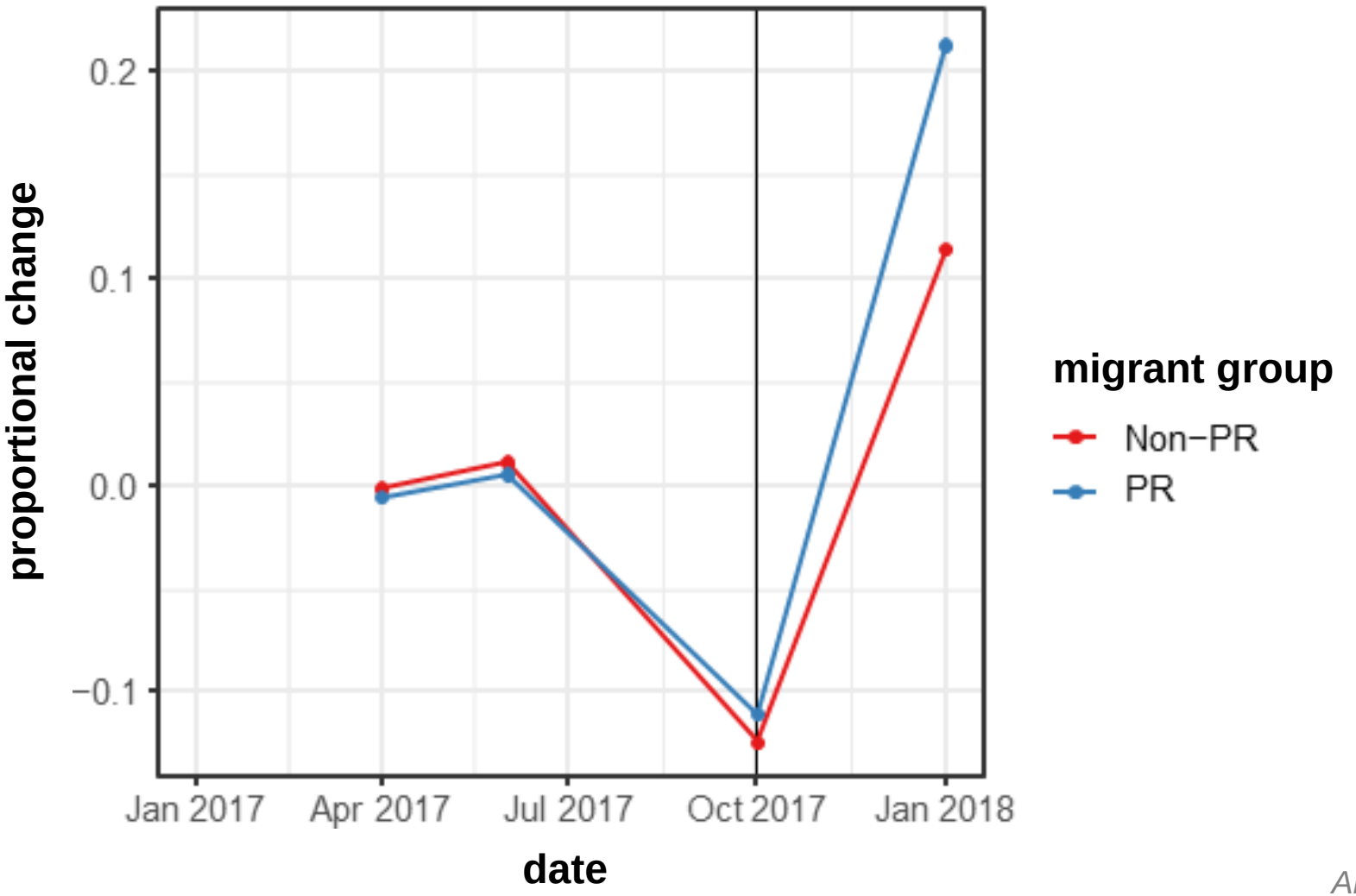
Reis and Brownstein (2010)

TODAY'S OUTLINE

- **TRENDS OVER TIME AND FILTERING OUT BIAS**
- **SURVEYS AND NON-REPRESENTATIVE SAMPLES**
- **CALIBRATION AND WEIGHTING**

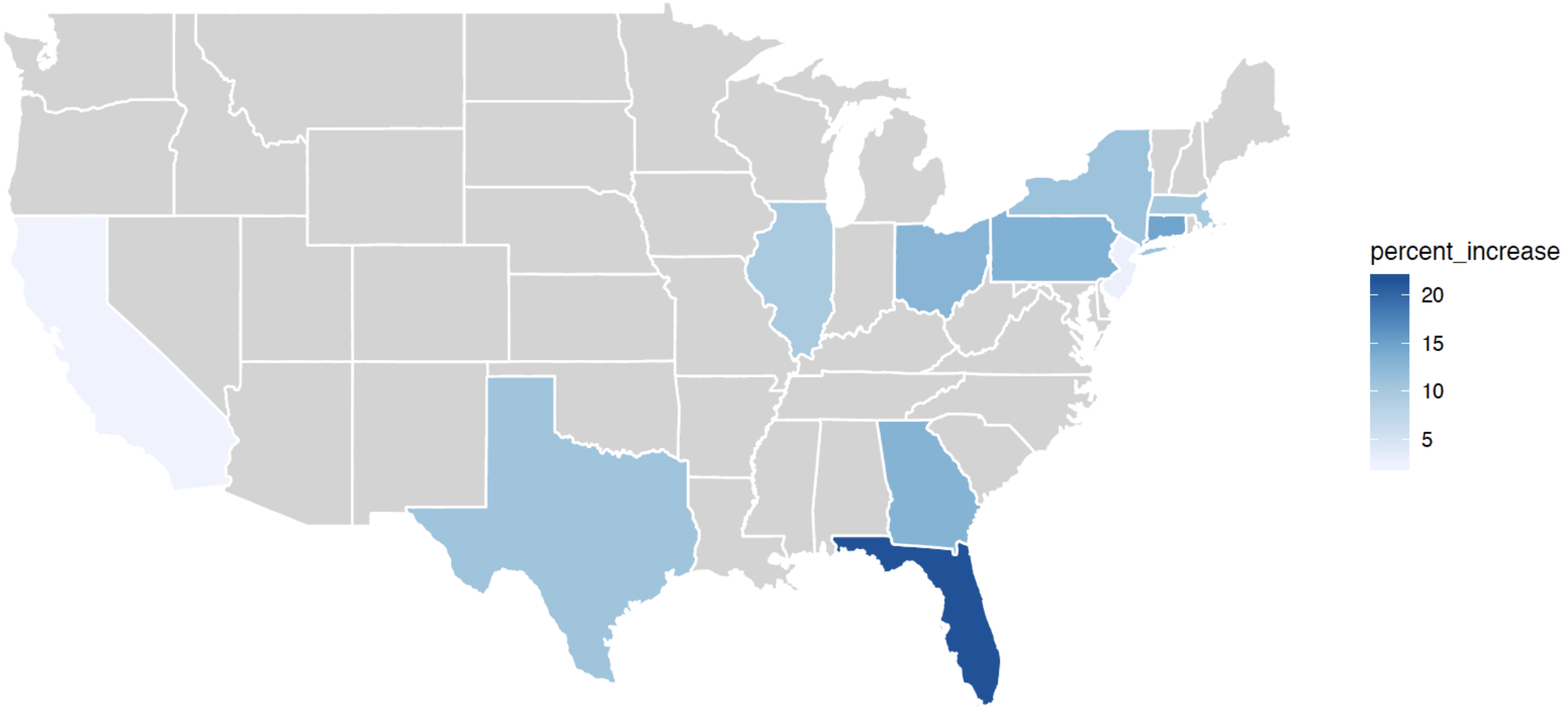
MIGRATION AND MOBILITY AFTER HURRICANE MARIA MADE LANDFALL IN PUERTO RICO (SEPT. 2017)

FACEBOOK DATA AND A DIFF-IN-DIFF APPROACH



Alexander, Polimis and Zagheni (2019)
Population and Development Review

PERCENT INCREASE IN PUERTO RICANS FROM OCT. 2017 TO JAN. 2018

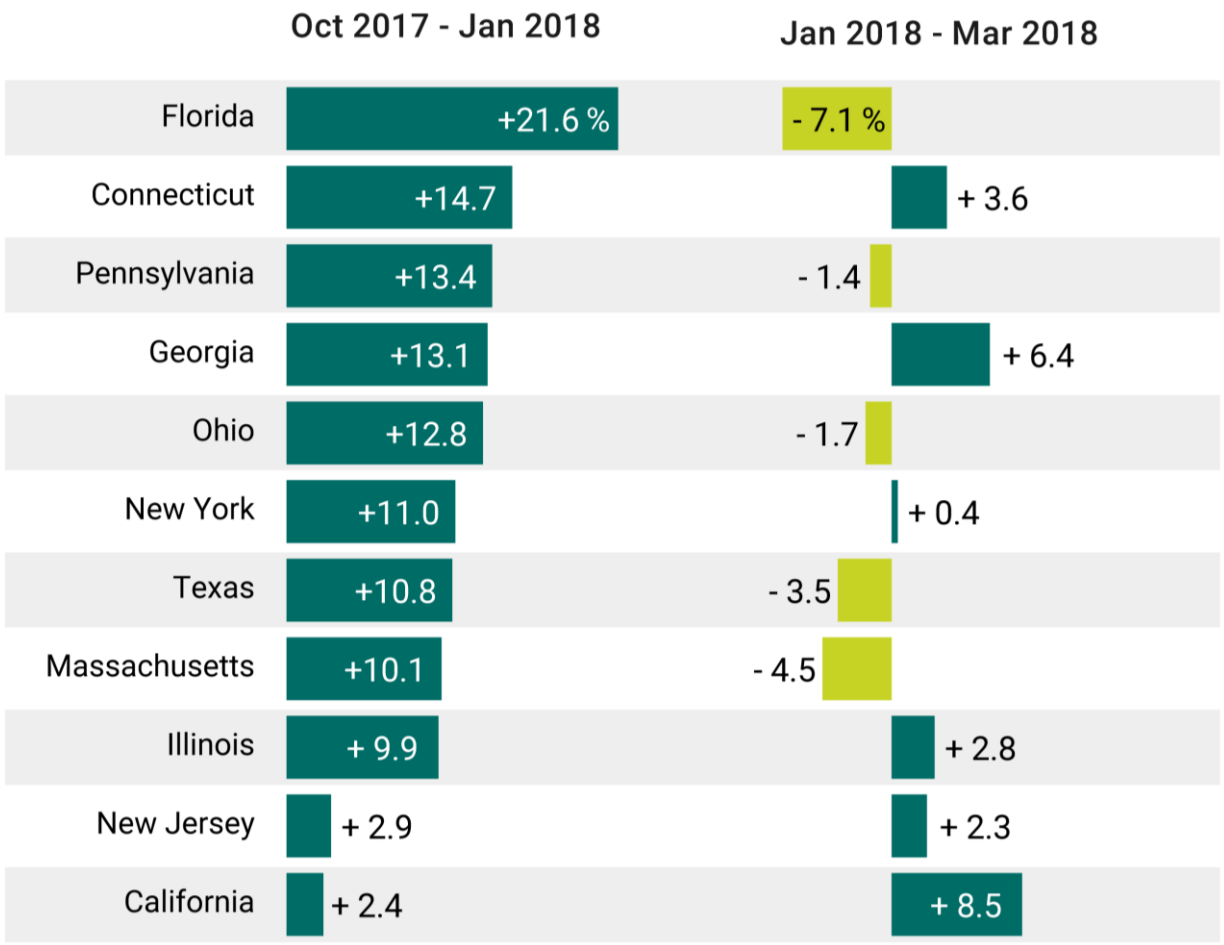


Alexander, Polimis and Zagheni (2019)
Population and Development Review

RETURN MIGRATION OR ONWARD RELOCATION

Migration after Hurricane Maria

Change of Puerto Rican migrant population in the US* (as percentages)



* For the nine US states with the largest populations of Puerto Rican migrants

Alexander, Polimis and Zagheni (2019)
Population and Development Review

CONSIDER THIS MODEL FOR DATA GENERATION

$$\underbrace{y_i^t}_{\substack{\text{Observation from}\\ \text{social media}\\ \text{for location i}}} = \underbrace{n}_{\substack{\text{bias for}\\ \text{location i}}} + \underbrace{x_i^t}_{\substack{\text{“true” rate}\\ \text{for location i}}}$$

and

$$\underbrace{y_z^t}_{\substack{\text{Observation from}\\ \text{social media}\\ \text{for location z}}} = \underbrace{m}_{\substack{\text{bias for}\\ \text{location z}}} + \underbrace{x_z^t}_{\substack{\text{“true” rate}\\ \text{for location z}}}$$

ASSUME THAT WE KNEW THE TRUE RATES (X) FOR TWO COUNTRIES: FRANCE (FR) AND SPAIN (SP)

$x_{FR}^{t+1} = 0.7$	$x_{SP}^{t+1} = 0.5$
$x_{FR}^{t} = 0.5$	$x_{SP}^{t} = 0.4$

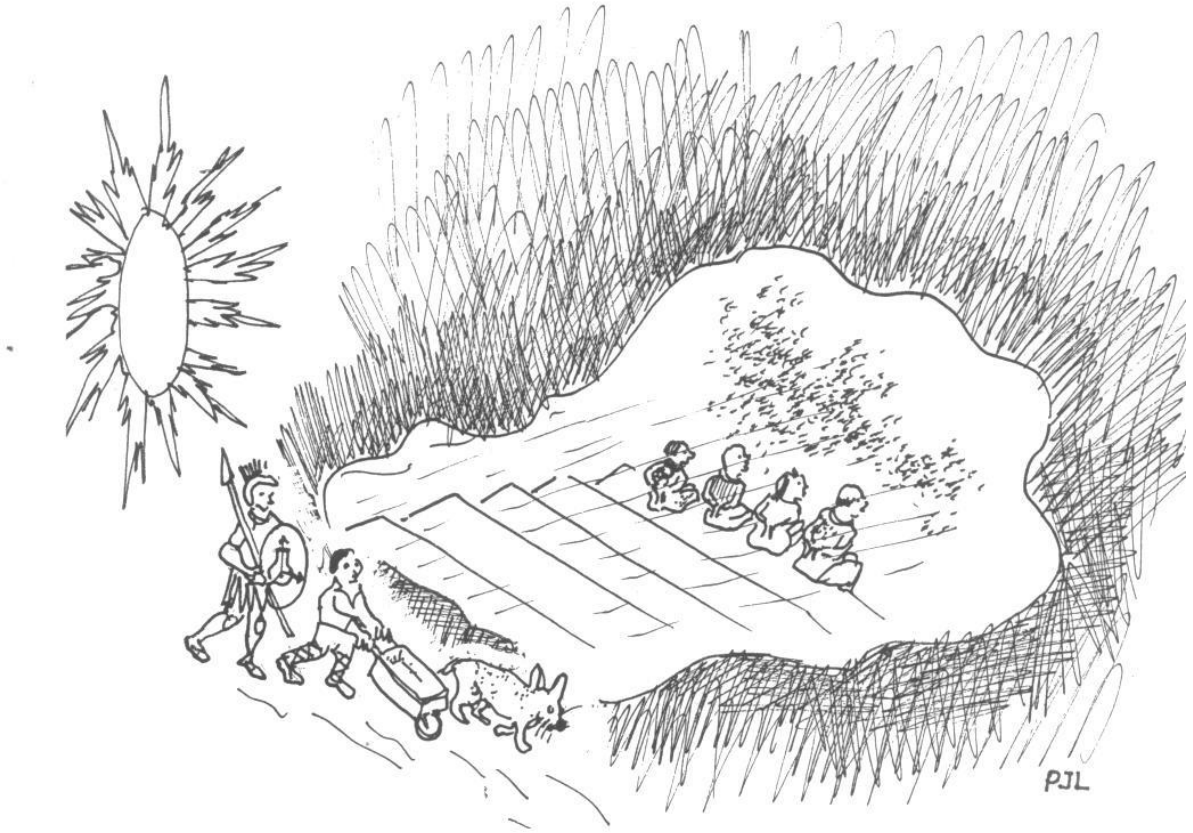
Let's define δ^{t+1} as the differential in the variation of these quantities of interest between time t and $(t+1)$

$$\delta^{t+1} = \underbrace{(x_{FR}^{t+1} - x_{FR}^{t}) - (x_{SP}^{t+1} - x_{SP}^{t})}_{\text{difference in the increments}} =?$$

$$\delta^{t+1} = (0.7 - 0.5) - (0.5 - 0.4) =$$

$$= 0.2 - 0.1 = 0.1$$

PLATO'S ALLEGORY OF THE CAVE



Plato's Allegory of the Cave
(The Republic)

ALL WE SEE IS A DISTORTED IMAGE (Y) OF THE 'TRUE' UNDERLYING RATES (X)

$y_{FR}^{t+1} = 0.2 + 0.7$	$y_{SP}^{t+1} = 0.1 + 0.5$
$y_{FR}^{t} = 0.2 + 0.5$	$y_{SP}^{t} = 0.1 + 0.4$

What is δ^{t+1}?

$$\delta^{t+1} = \underbrace{(y_{FR}^{t+1} - y_{FR}^{t}) - (y_{SP}^{t+1} - y_{SP}^{t})}_{\text{difference in the increments}} =?$$

$$\delta^{t+1} = (0.9 - 0.7) - (0.6 - 0.5) =$$

$$= 0.2 - 0.1 = 0.1$$

Same as before...

DIFFERENCE IN DIFFERENCES ESTIMATOR

- To the extent that the bias is additive and, within each country, is constant over short periods of time, DiD estimates from social media data:

$$\delta^{t+1} = (y_i^{t+1} - y_z^{t+1}) - (y_i^t - y_z^t)$$

are good estimates of the underlying differential:

$$\delta^{t+1} = \underbrace{(x_i^{t+1} - x_i^t) - (x_z^{t+1} - x_z^t)}_{\text{difference in the increments}}$$

- Additive values of the bias (m and n) cancel out

SOME REMARKS

If the bias is expected to be multiplicative:

$$\underbrace{y_i^t}_{\substack{\text{Observation from} \\ \text{social media} \\ \text{for location i}}} = \underbrace{n}_{\substack{\text{bias for} \\ \text{location i}}} \times \underbrace{x_i^t}_{\substack{\text{“true” rate} \\ \text{for location i}}}$$

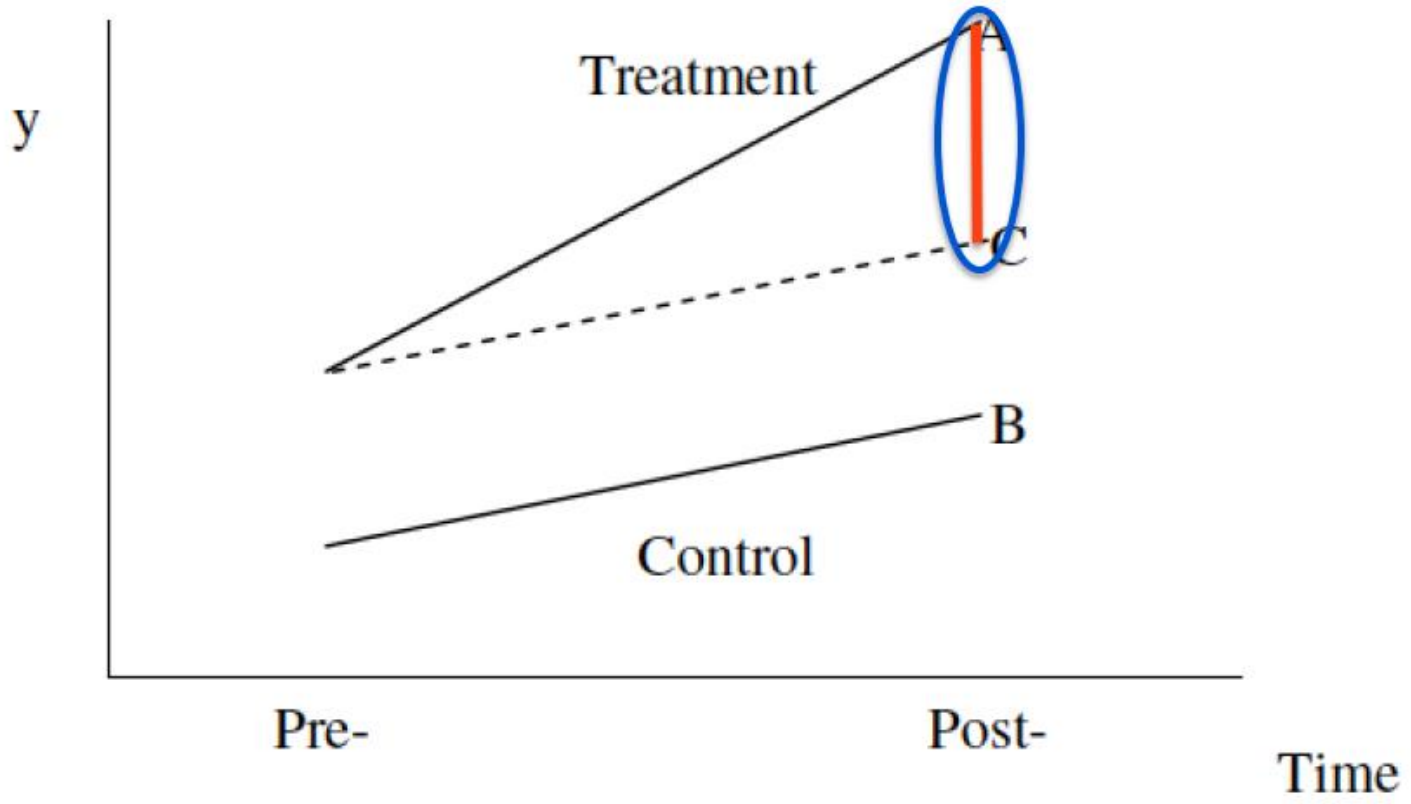
Use a logarithmic transformation

$$\log(y_i^t) = \log(n) + \log(x_i^t)$$

Then use the difference-in-differences estimator on the logs:

$$\delta^{t+1} = [\log(y_i^{t+1}) - \log(y_z^{t+1})] - [log(y_i^t) - log(y_z^t)]$$

DIFFERENCE IN DIFFERENCES – IN REGRESSION FORM



$$\hat{\delta} = (\bar{Y}_{treat,post} - \bar{Y}_{control,post}) - (\bar{Y}_{treat,pre} - \bar{Y}_{control,pre})$$

DIFFERENCE IN DIFFERENCES – IN REGRESSION FORM

$$y_{it} = \beta_0 + \beta_1 I(treat_{it}) + \beta_2 I(post_{it}) + \beta_3 I(treat_{it}) I(post_{it}) + \epsilon_{it}$$

where:
$I(treat_{it}) = 1$ if treatment; 0 otherwise
$I(post_{it}) = 1$ if post; 0 otherwise

β_3 is the estimate of δ, the difference-in-difference estimator

Why?

DIFFERENCE IN DIFFERENCES – IN REGRESSION FORM

$$y_{it} = \beta_0 + \beta_1 I(treat_{it}) + \beta_2 I(post_{it}) + \beta_3 I(treat_{it}) I(post_{it}) + \epsilon_{it}$$

- $E[y_{it}|I(treat_{it}) = 1, I(post_{it}) = 1] = \beta_0 + \beta_1 + \beta_2 + \beta_3$
- $E[y_{it}|I(treat_{it}) = 1, I(post_{it}) = 0] = \beta_0 + \beta_1$
- $E[y_{it}|I(treat_{it}) = 0, I(post_{it}) = 1] = \beta_0 + \beta_2$
- $E[y_{it}|I(treat_{it}) = 0, I(post_{it}) = 0] = \beta_0$

DIFFERENCE IN DIFFERENCES – IN REGRESSION FORM

$$y_{it} = \beta_0 + \beta_1 I(treat_{it}) + \beta_2 I(post_{it}) + \beta_3 I(treat_{it}) I(post_{it}) + \epsilon_{it}$$

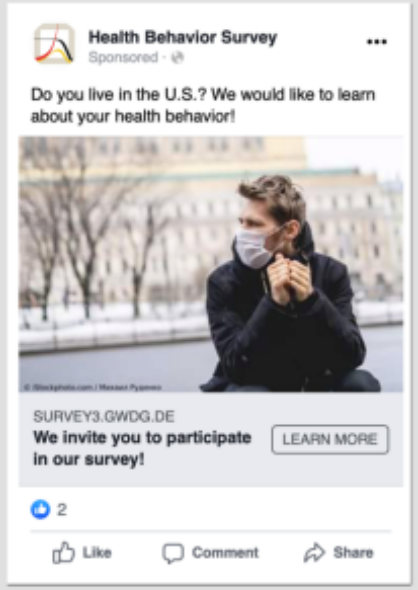
	Post	Pre	Difference
Treatment	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_1$	$\beta_2 + \beta_3$
Control	$\beta_0 + \beta_2$	β_0	β_2
Difference	$\beta_1 + \beta_3$	β_1	β_3

TODAY'S OUTLINE

- TRENDS OVER TIME AND FILTERING OUT BIAS
- SURVEYS AND NON-REPRESENTATIVE SAMPLES
- CALIBRATION AND WEIGHTING

THE COVID-19 HEALTH BEHAVIOR SURVEY (CHBS)

RESPONDENTS RECRUITMENT VIA FACEBOOK ADS



Example of Facebook ad in the US.

COVID-19 HEALTH BEHAVIOR SURVEY (CHBS)

- **Socio-demographic characteristics**
- **Health indicators** (e.g. chronic disease, symptoms, health seeking behavior)
- **Behaviors and attitudes** (e.g. threat perception, preventive behaviours, disruptions to daily routine)
- **Social contacts** (at home, at school, at work, in the general community)

POST-STRATIFICATION WEIGHTING

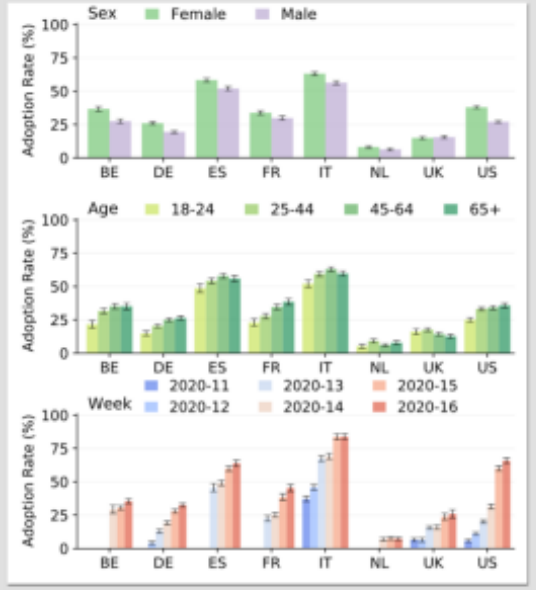
on central demographic variables

Compute weights by age, sex, and region of residence

POPULATION DATA

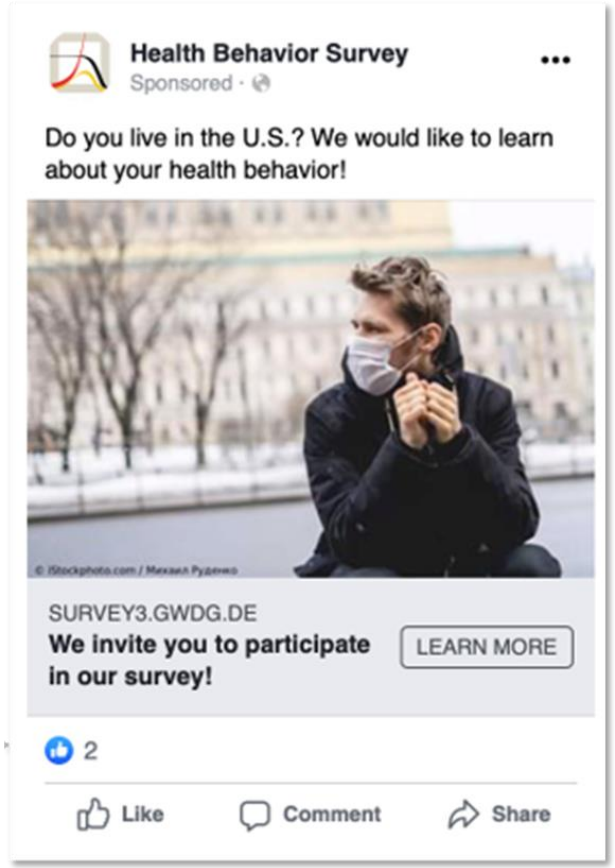
from Eurostat and US census

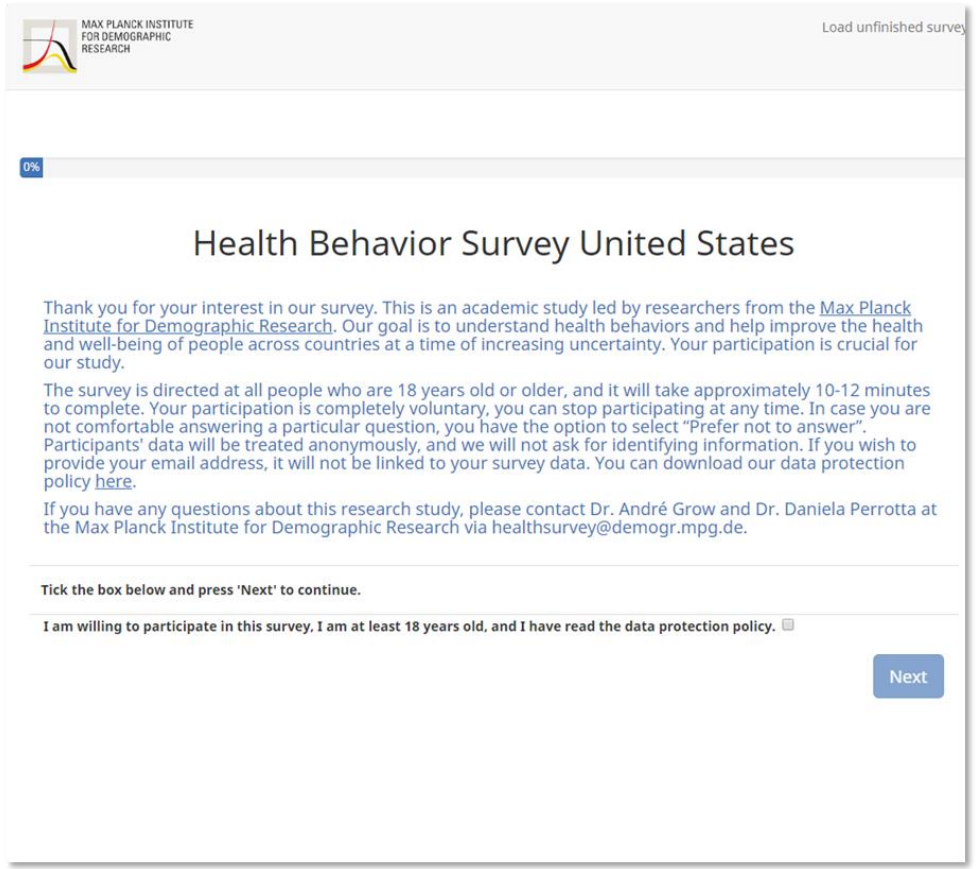
ESTIMATES OF BEHAVIORAL CHANGES ACROSS COUNTRIES



Example of wearing face masks.

Grow, Perrotta, Del Fava, Cimentada, Rampazzo, Gil-Clavel and Zagheni (2020)
Journal of Medical Internet Research





Coverage: 8 Countries (BE, DE, ES, FR, IT, NL, UK, US)

Field period: Continuously running between March 13 and August 12, 2020

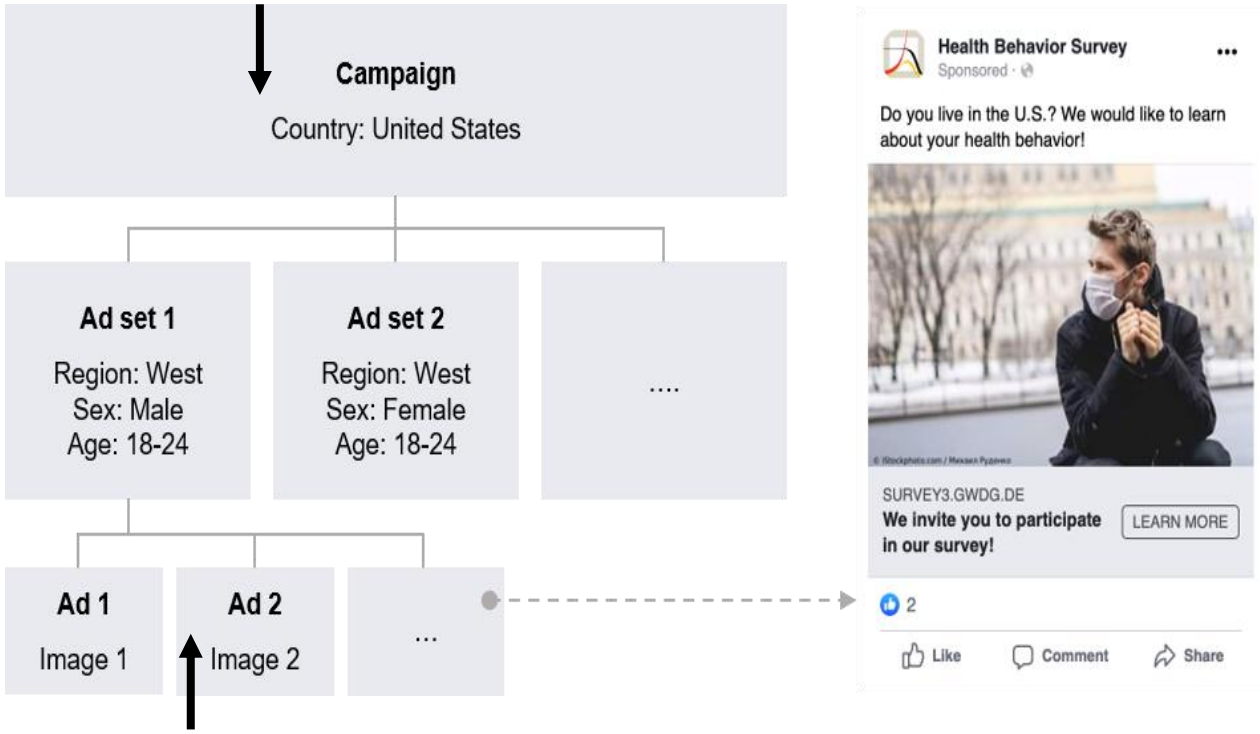
Topics: Socio-demographics, attitudes, behavior, health, social contacts

Total N: 144,034

THE ADVERTISING CAMPAIGNS

- **One ad set/stratum per demographic group:**
  - sex (m/f)
  - age (18-24, 25-44, 44-64, 65+)
  - Region of residence (NUTS1/US Census regions)

One campaign per country

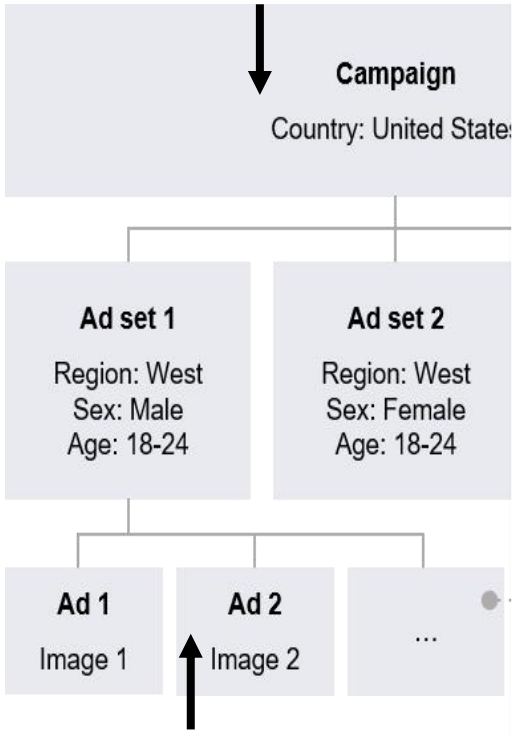


Six different ad images within each ad set

THE ADVERTISING CAMPAIGNS

- **One ad set/stratum per demographic group:**
  - sex (m/f)
  - age (18-24, 25-44, 44-64, 65+)
  - Region of residence (NUTS1/US Census regions)

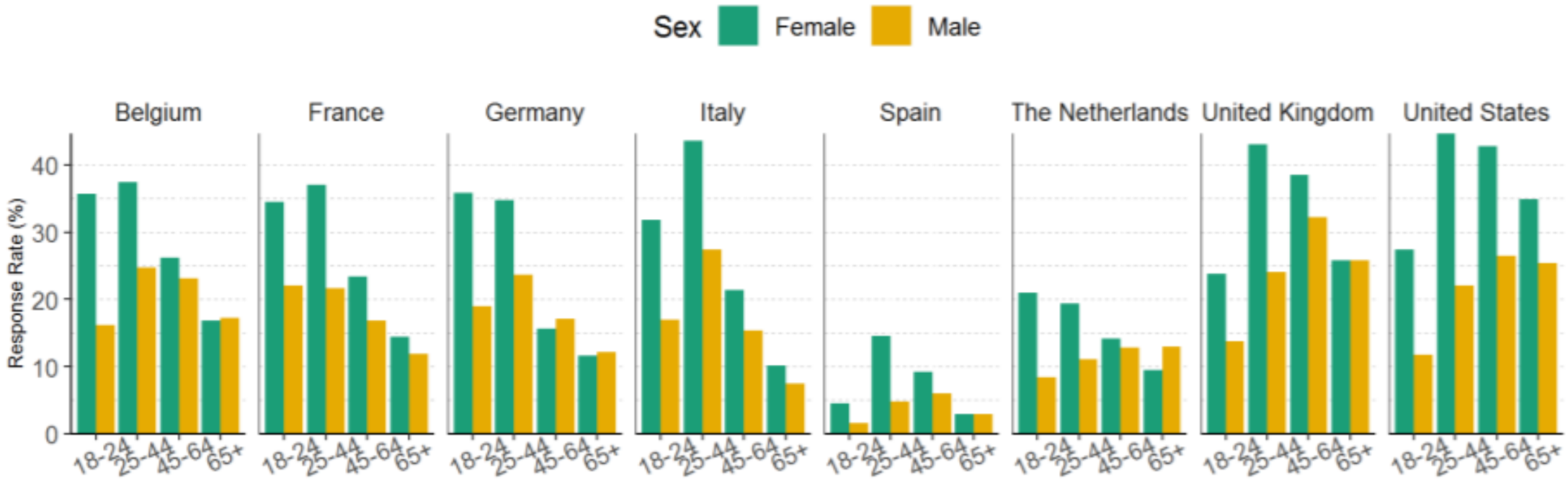
One campaign per country



Six different ad images within each ad set

RESPONSE RATE

- **Response rate** (% Facebook users who completed the questionnaire after clicking on the ad) generally reasonable with some variability
- **Observation 1:** Spain
- **Observation 2:** Women > Men



D. Perrotta, A. Grow, F. Rampazzo, J. Cimentada, E. Del Fava, S. Gil-Clavel & E. Zagheni. 2020. "Behaviors and Attitudes in Response to the COVID-19 Pandemic: Insights from a Cross-National Facebook Survey". *medRxiv*, doi: https://doi.org/10.1101/2020.05.09.20096388.

POST-STRATIFICATION WEIGHTING

After the data had been collected, adjustments were produced to obtain a closer approximation to a representative sample

Weight assigned to respondents in stratum *i* in the survey

$$w_i = \frac{p_i}{\hat{p}_i}$$

Proportion of individuals who belong to stratum *i* in the population

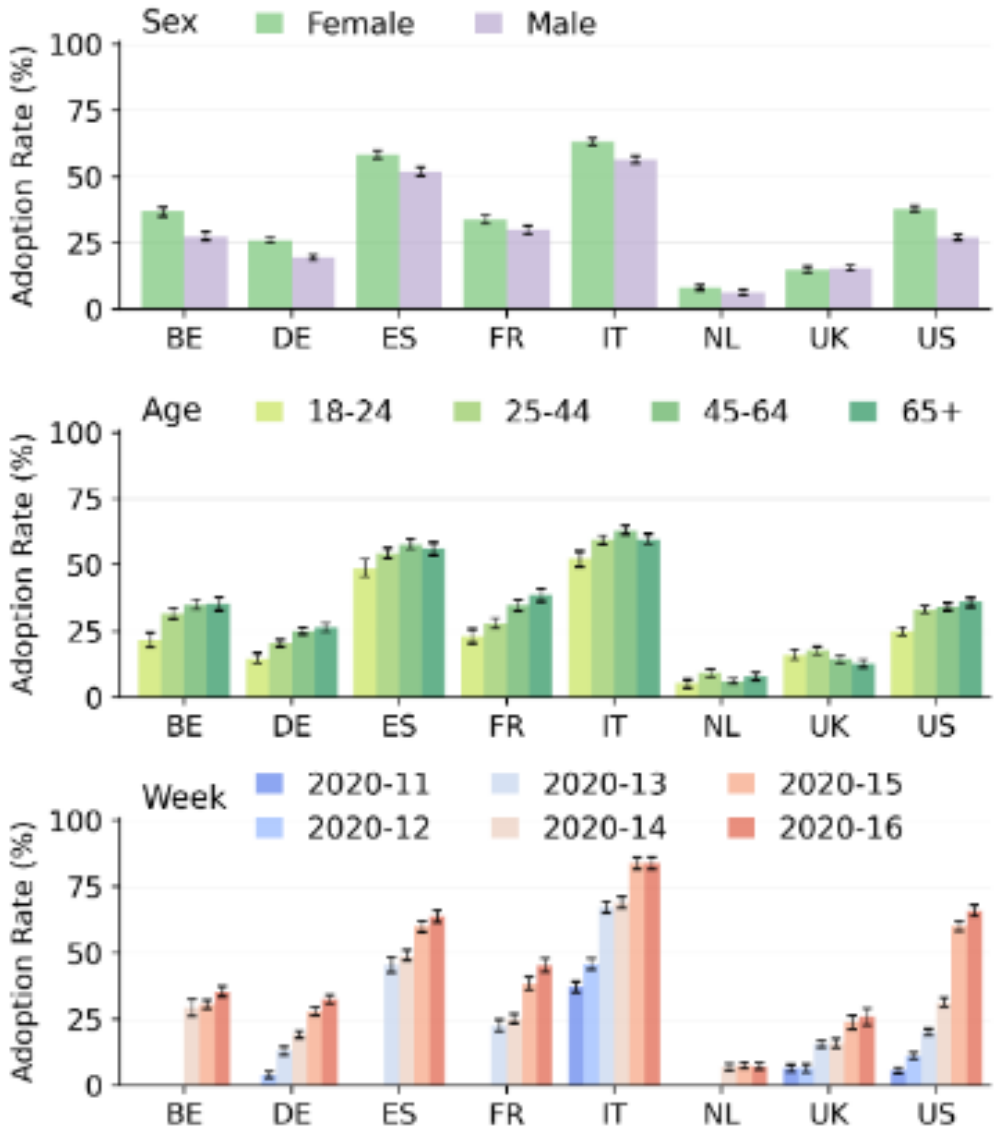
Proportion of respondents who belong to stratum *i* in the survey

			Facebook		Eurostat		
Region	Sex	Age	$\hat{N}_i$	$\hat{p}_i$	N_i	p_i	w_i
Central	Female	18-24	199	0.027	522,730	0.010	0.363
Central	Female	25-44	384	0.053	1,453,810	0.028	0.523
Central	Female	45-64	265	0.037	1,876,591	0.036	0.979
Central	Female	65+	104	0.014	1,605,018	0.031	2.133
Central	Male	18-24	119	0.016	572,911	0.011	0.665
Central	Male	25-44	214	0.030	1,454,223	0.028	0.939
Central	Male	45-64	176	0.024	1,758,096	0.034	1.380
Central	Male	65+	78	0.011	1,221,951	0.023	2.165

Table S1. Example of post-stratification reweigh using the central region of Italy.

D. Perrotta, A. Grow, F. Rampazzo, J. Cimentada, E. Del Fava, S. Gil-Clavel & E. Zagheni. 2020. "Behaviors and Attitudes in Response to the COVID-19 Pandemic: Insights from a Cross-National Facebook Survey". *medRxiv*, doi: https://doi.org/10.1101/2020.05.09.20096388.

ADOPTION RATE OF WEARING A FACE MASK



Perrotta, Grow, Rampazzo, Cimentada, Del Fava, Gil-Clavel and Zagheni (2020)

INITIAL VALIDATION OF FACEBOOK DATA

User's actual characteristics

Facebook's classification

FRACTION OF ALL CLASSIFICATIONS THAT ARE CORRECT

	sex	age	region
Belgium	0.98	0.96	0.91
France	0.98	0.93	0.95
Germany	0.98	0.95	0.97
Italy	0.99	0.95	0.97
Netherlands	0.98	0.96	0.98
Spain	0.99	0.94	0.97
United Kingdom	0.99	0.94	0.93
United States	0.99	0.94	0.98

Belgium and the United Kingdom somewhat stand out

Across countries, respondents were most likely to be correctly classified based on sex, and a little less likely based on age and region

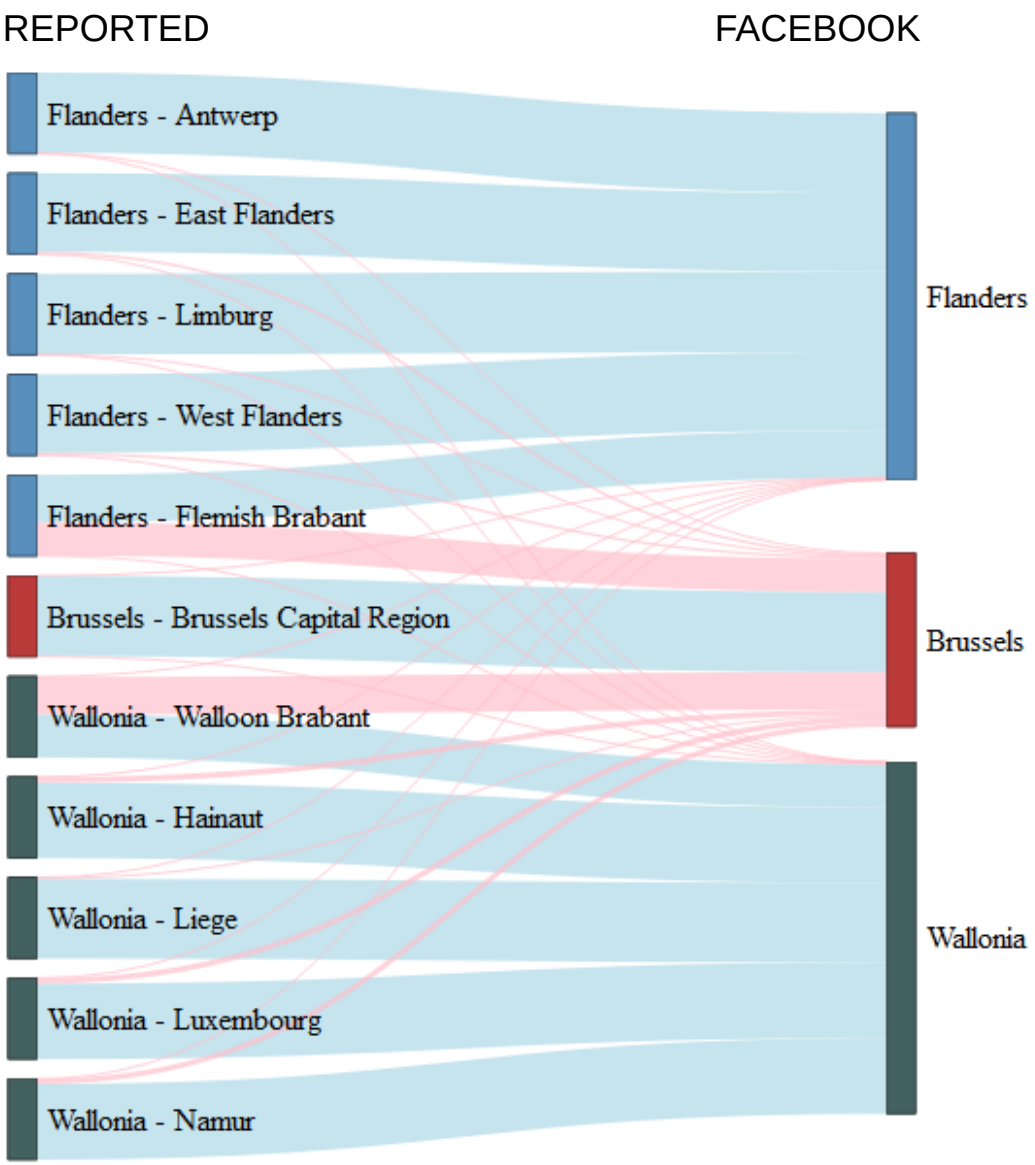
FRACTION OF ALL CLASSIFICATIONS AS A SPECIFIC REGION THAT ARE CORRECT

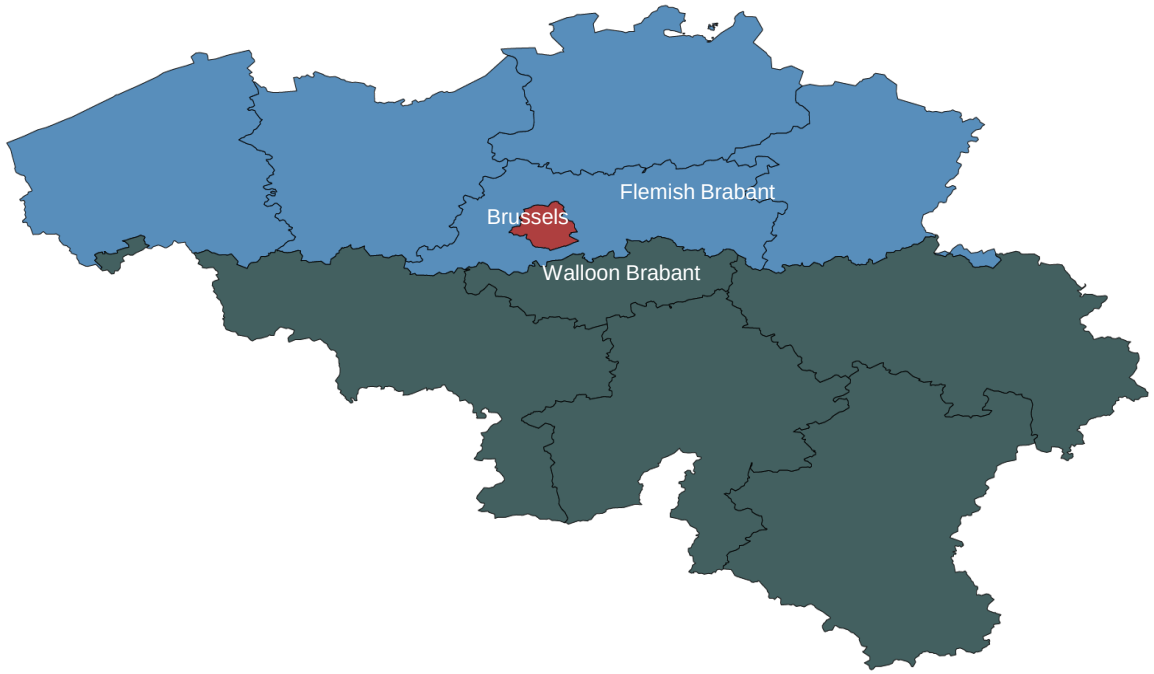
Country	Region	Precision
Belgium	Brussels	0.664
	Flanders	0.992
	Wallonia	0.992
Germany	East	0.955
	North	0.976
	South	0.981
	West	0.978
United Kingdom	England	0.992
	London	0.678
	Northern Ireland	0.993
	Scotland	0.991
	Wales	0.908
United States	Midwest	0.986
	Northeast	0.982
	South	0.979
	West	0.985

In most countries, the fraction was >.9 across regions

Belgium and the United Kingdom deviate from this

A CLOSER LOOK AT REGIONS IN BELGIUM

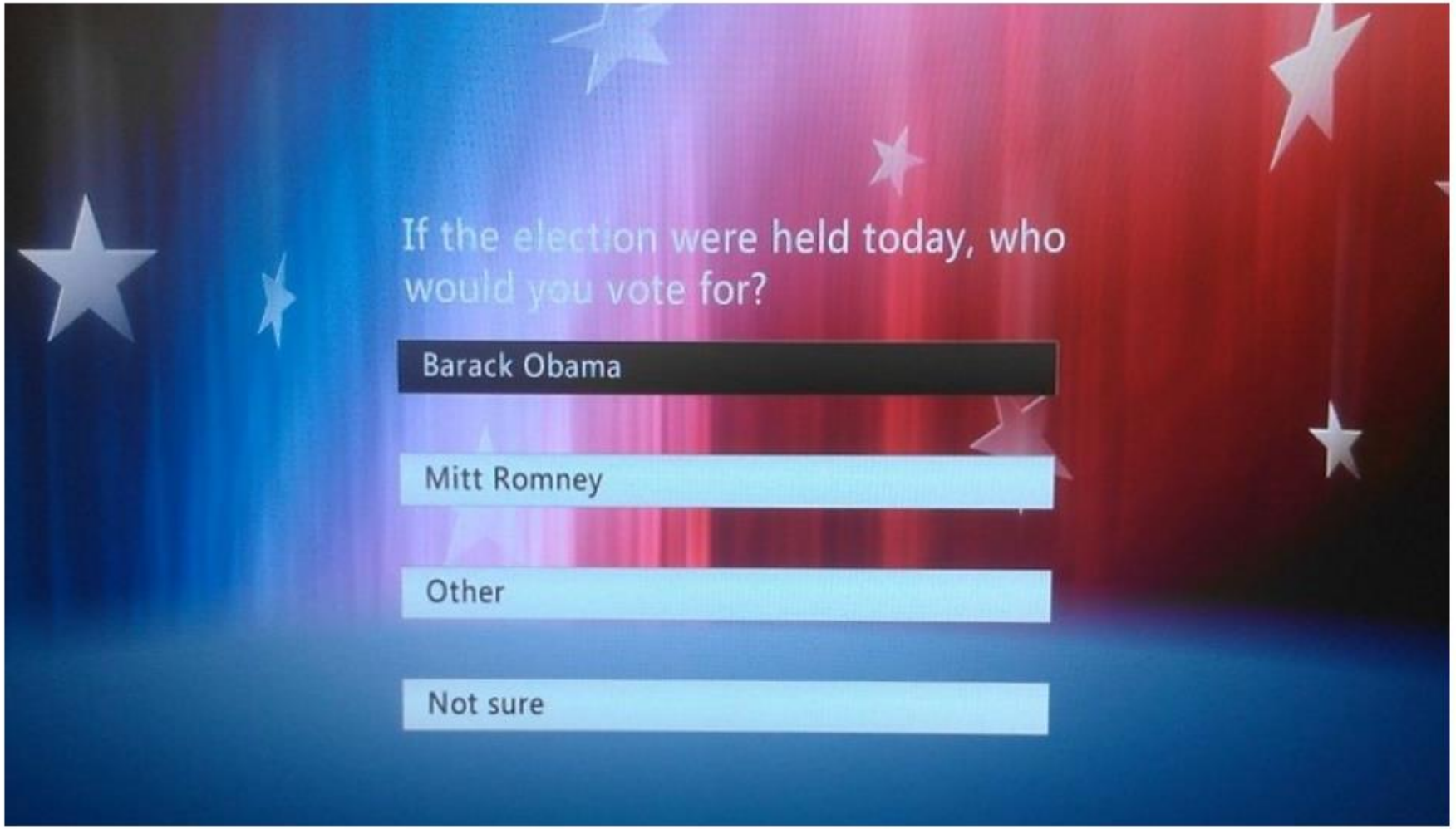




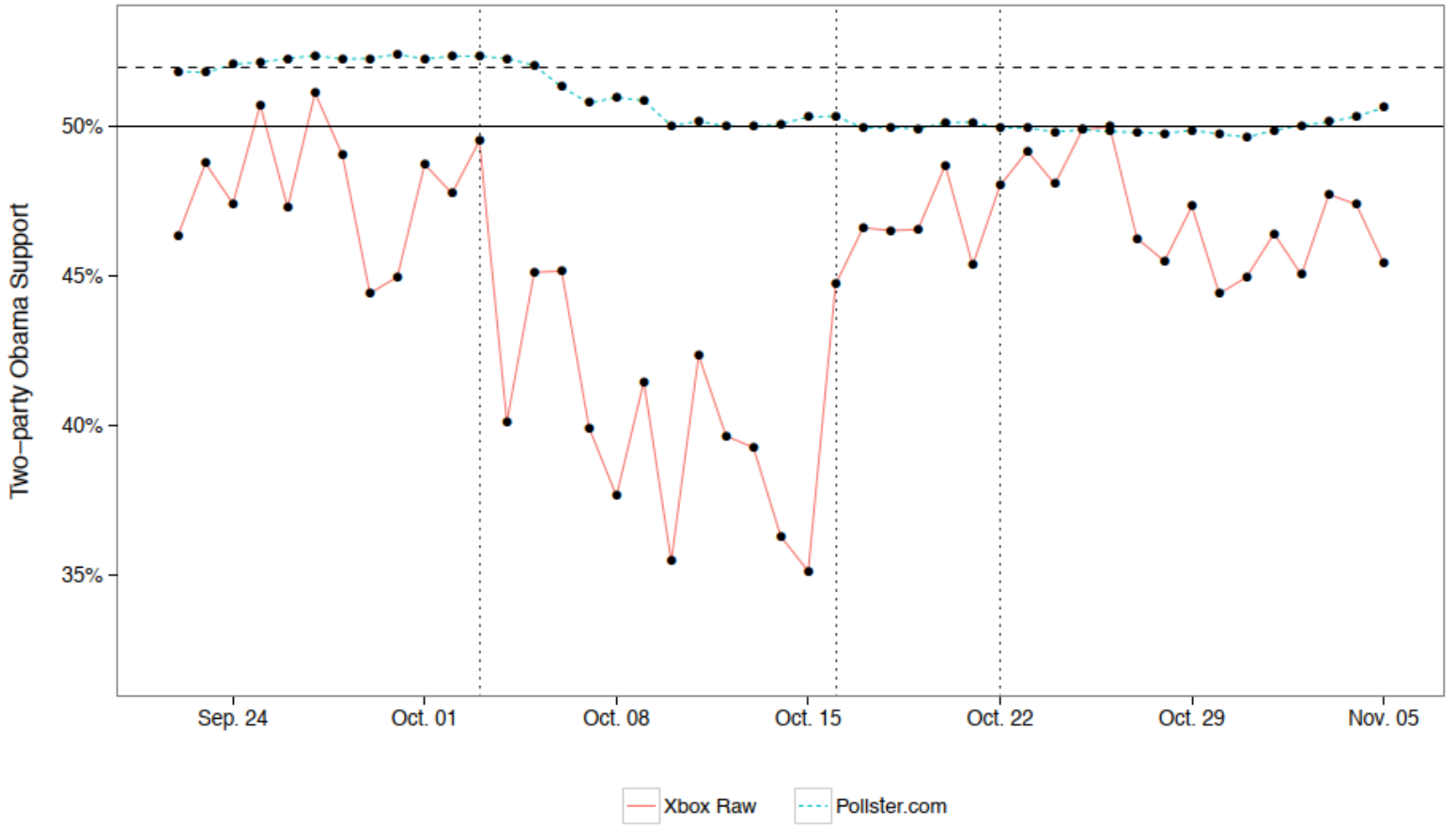
Misclassifications for Brussels mostly concern people who live in adjacent regions

MORE NON-REPRESENTATIVE SAMPLES: A SURVEY OF XBOX USERS

MORE NON-REPRESENTATIVE SAMPLES: A SURVEY OF XBOX USERS



RAW DATA FROM THE XBOX SURVEY



Wang, Rothschild, Goel and Gelman (2014)

APPROACH

Post-stratification:

- Partition the population into cells, j (e.g., based on combinations of various demographic attributes)
- Use the sample to estimate the response variable, $\hat{y}_j$, within each cell
- Aggregate the cell-level estimates, by weighting them for their relative proportion in the population:

$$\hat{y}^{PS} = \frac{\sum_{j=1}^{J} N_j \hat{y}_j}{\sum_{j=1}^{J} N_j}$$

WHAT IS THE PROBLEM?

WHAT IS THE PROBLEM?

- When considering all possible combination of sex, age, race, education, state, etc., the number of cells becomes very big
- Some cells would be zeroes, or would have high stochasticity

- The proposed solution: a multilevel regression model
- The model is used to estimate the expected value in each cell

LOGISTIC MODEL FOR BINARY RESPONSE (E.G., VOTER VS NON-VOTER)

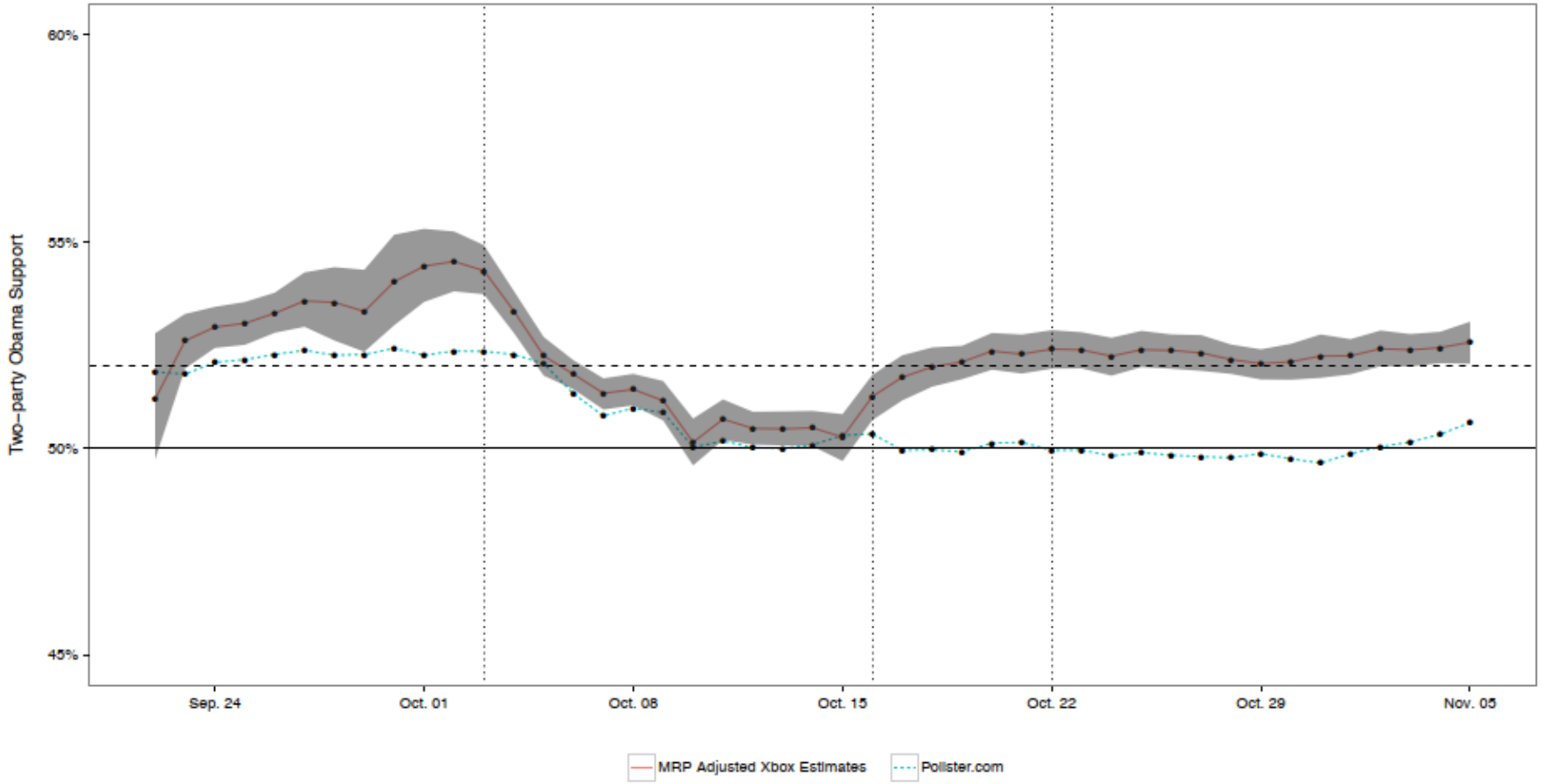
$$\Pr(Y_i \in \{\text{Obama, Romney}\}) =$$
$$\text{logit}^{-1}\big(\alpha_0 + \alpha_1(\text{state last vote share})$$
$$+ a^{\text{state}}_{j[i]} + a^{\text{edu}}_{j[i]} + a^{\text{sex}}_{j[i]} + a^{\text{age}}_{j[i]} + a^{\text{race}}_{j[i]} + a^{\text{party ID}}_{j[i]} + b^{\text{ideology}}_{j[i]} + b^{\text{last vote}}_{j[i]}\big)$$

LOGISTIC MODEL FOR BINARY RESPONSE (E.G., OBAMA CONDITIONAL ON BEING A VOTER)

$$\Pr(Y_i = \text{Obama} \,|\, Y_i \in \{\text{Obama, Romney}\}) =$$
$$\text{logit}^{-1}\big(\beta_0 + \beta_1(\text{state last vote share})$$
$$+ b^{\text{state}}_{j[i]} + b^{\text{edu}}_{j[i]} + b^{\text{sex}}_{j[i]} + b^{\text{age}}_{j[i]} + b^{\text{race}}_{j[i]} + b^{\text{party ID}}_{j[i]} + b^{\text{ideology}}_{j[i]} + b^{\text{last vote}}_{j[i]}\big)$$

$$b^{\text{var}}_{j[i]} \sim N(0, \eta^2_{\text{var}}),$$
$$\eta^2_{\text{var}} \sim \text{inv-}\chi^2(\mu, \eta^2_0).$$

RESULTS FROM THE XBOX SURVEY AFTER POST-STRATIFICATION ADJUSTMENTS



Wang, Rothschild, Goel and Gelman (2014)

TODAY'S OUTLINE

- TRENDS OVER TIME AND FILTERING OUT BIAS
- SURVEYS AND NON-REPRESENTATIVE SAMPLES
- CALIBRATION AND WEIGHTING

THE PROBLEM

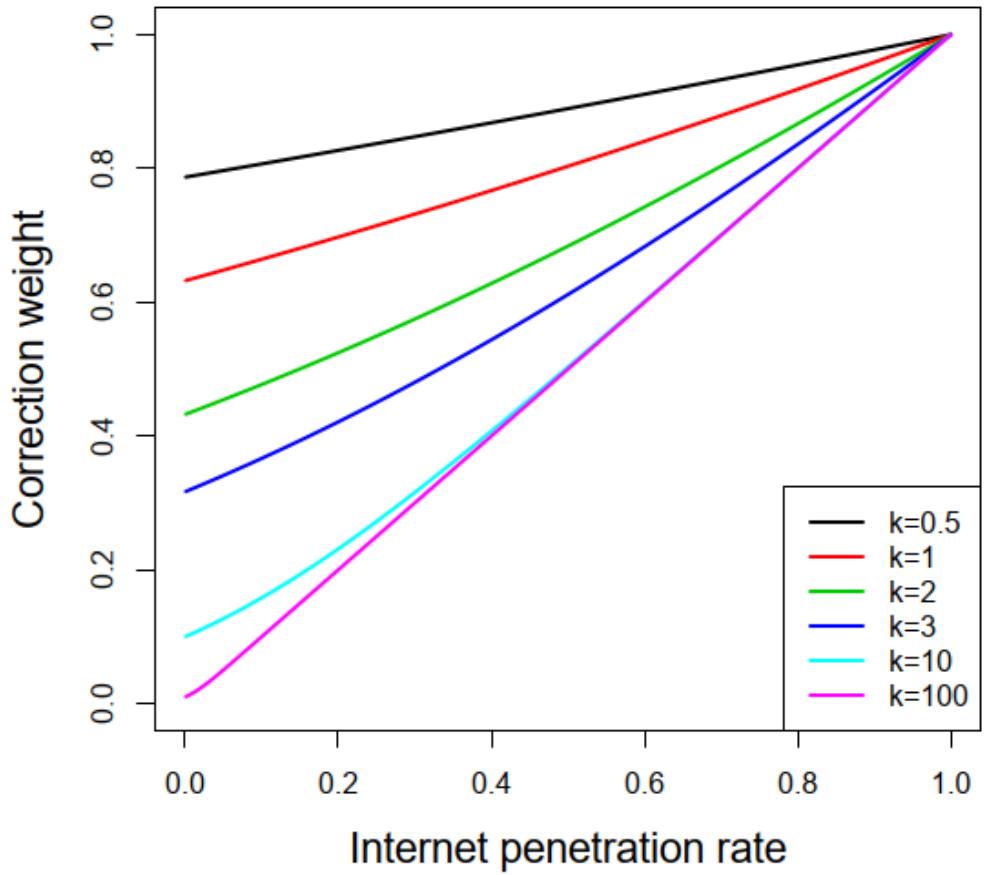
- We have some statistics considered as “silver standard” for Europe
- We have digital trace data on migration flows by age and sex for most countries, but with biases
- We want to be able to say something about migration flows for other regions of the world

Source: Zagheni and Weber (2012)

THE APPROACH – WHAT ARE THE ASSUMPTIONS?

Define a function (of a parameter k) that corrects for over representation, given age- and sex-specific Internet penetration rates.

$$CF = \frac{pen_{gac}(e^{-k} - 1)}{(e^{-k \times pen_{gac}} - 1)}$$



Source: Zagheni and Weber (2012)

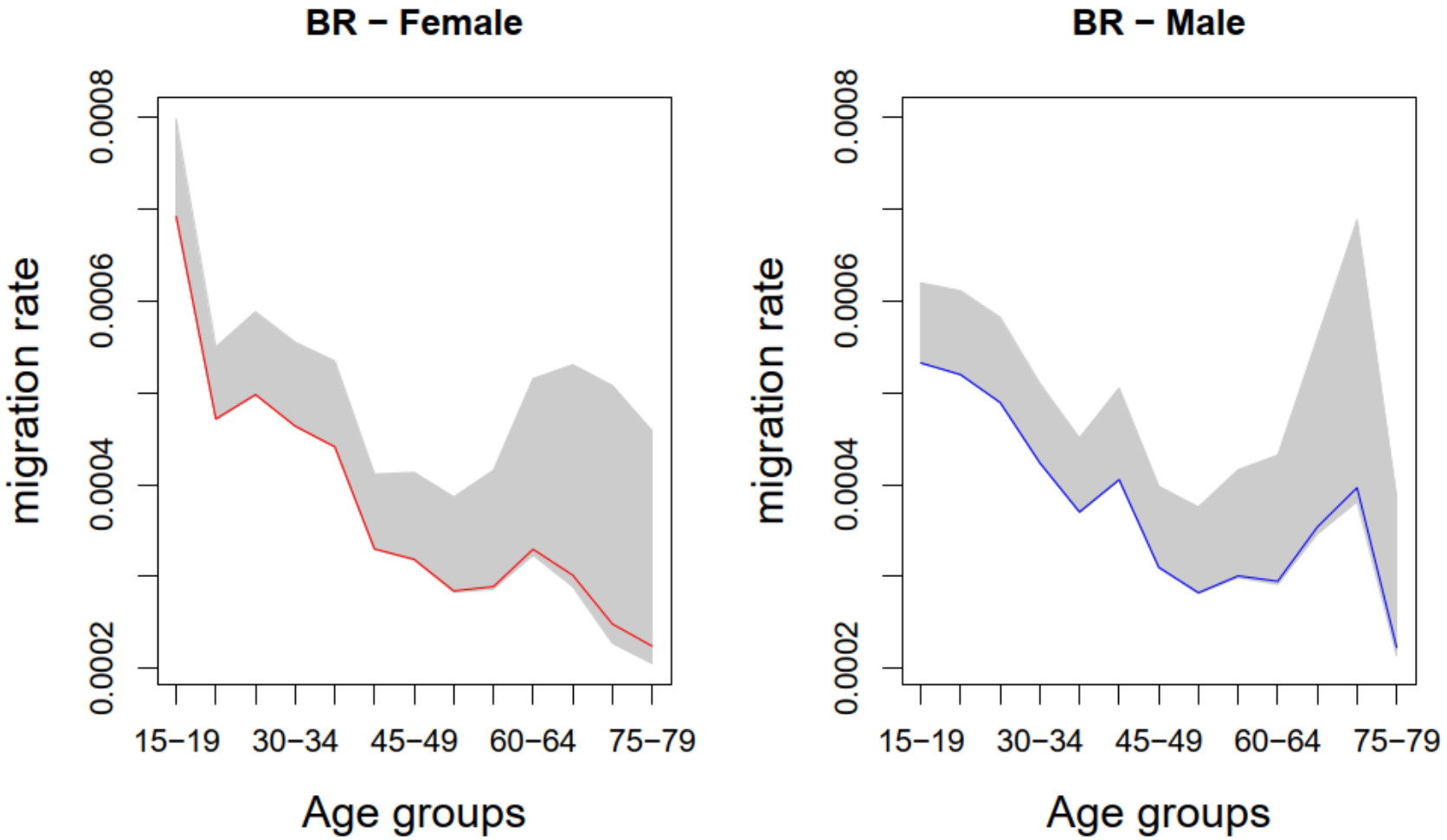
KEY STEPS

- Estimate the parameters that make the corrected estimates from digital trace data as close as possible to the “official” or “best quality” migration data available (e.g., for Europe)

- Use the estimated parameters and their level of uncertainty to extrapolate to other countries/regions

Source: Zagheni and Weber (2012)

EXAMPLE FOR BRAZIL



Source: Zagheni and Weber (2012)

POST-STRATIFICATION WEIGHTS + ADJUSTMENTS

LET'S EXPLORE IT TOGETHER

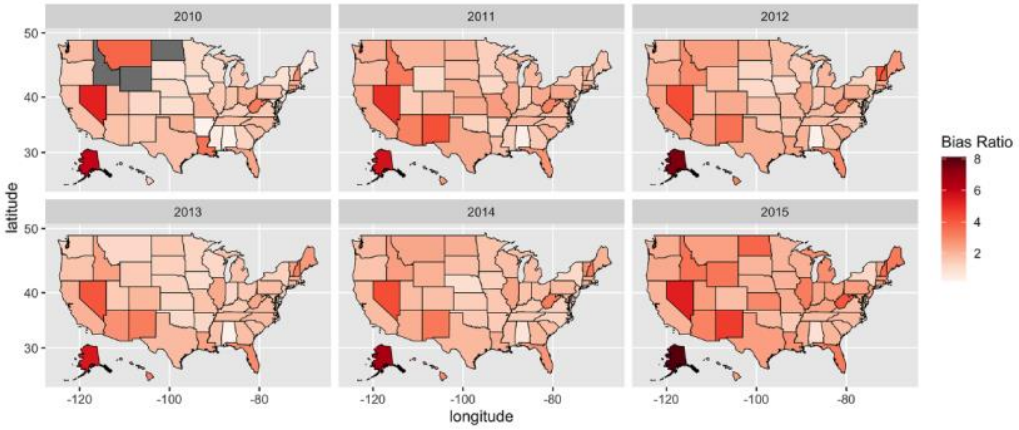
$$\text{Income}_o = \frac{GNIpc_o}{\max\left(GNIpc_o\right)}$$

$$W_{o,t} = \frac{1}{\text{Income}_o \times \frac{FBUsers_{o,t}}{Population_{o,t}} + (1 - \text{Income}_o) \times r_t}$$

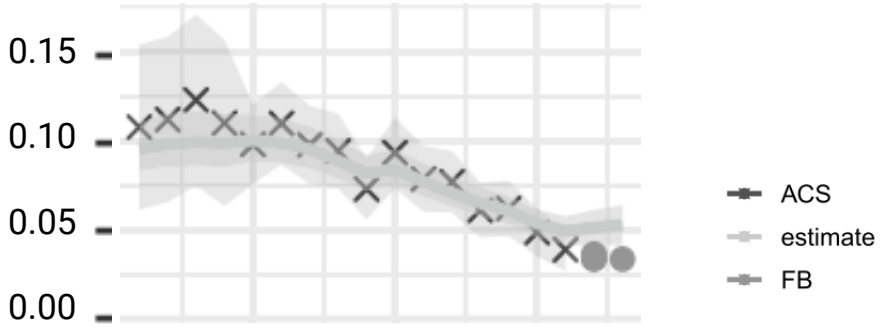
Source: Chi, Abel et al. (2025) PNAS

MORE ADVANCED STATISTICAL METHODS TO LEVERAGE TIMELY SOCIAL MEDIA DATA FOR PROVISIONAL ESTIMATES

Bias ratios by state



Fraction of Mexican men, 25-29, in Georgia US



Original Article

Sociological Methods & Research
1–39
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00491241221140144
journals.sagepub.com/home/smr
SAGE

Modeling the Bias of Digital Data: An Approach to Combining Digital With Official Statistics to Estimate and Predict Migration Trends

Yuan Hsiao[1], Lee Fiorio[2], Jonath
and Emilio Zagheni[4]

Abstract

Obtaining reliable and timely estimates of migrati
cing the migration theory and guiding policy dec
lenge. Digital data provide granular information
not draw from representative samples of the popu

Population Research and Policy Review (2022) 41:1–28
https://doi.org/10.1007/s11113-020-09599-3

ORIGINAL RESEARCH

Check for updates

Combining Social Media and Survey Data to Nowcast Migrant Stocks in the United States

Monica Alexander[1] · Kivan Polimis[2] · Emilio Zagheni[3]

Received: 11 April 2020 / Accepted: 20 July 2020 / Published online: 18 August 2020
© The Author(s) 2020

Demography (2021) 58(6):2193–2218
DOI 10.1215/00703370-9578562 © 2021 The Authors
Published online: 9 November 2021
This is an open access article distributed under the terms of a Creative Commons license (CC BY-NC-ND 4.0).

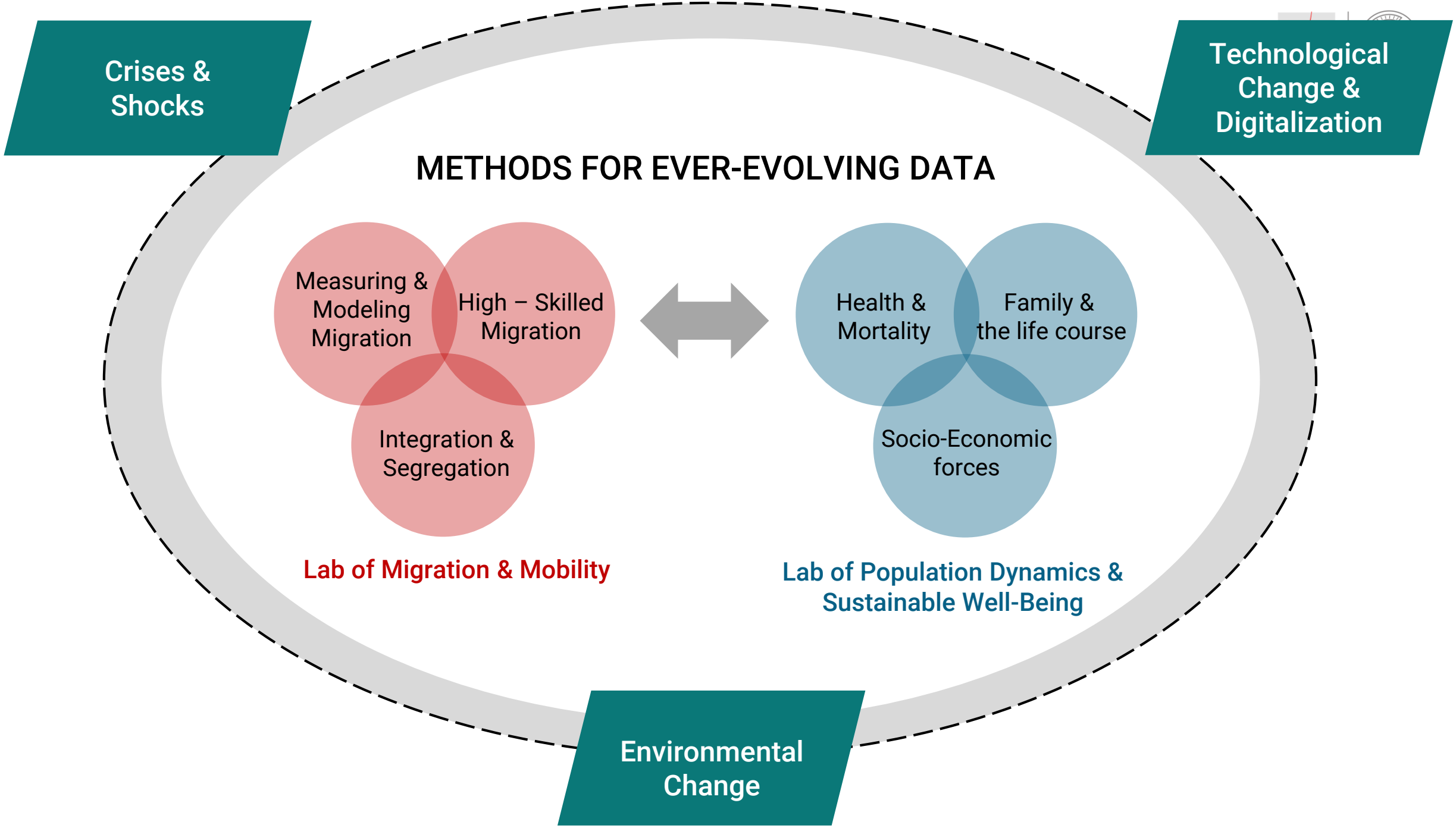
A Framework for Estimating Migrant Stocks Using Digital Traces and Survey Data: An Application in the United Kingdom

Francesco Rampazzo, Jakub Bijak, Agnese Vitali, Ingmar Weber, and Emilio Zagheni

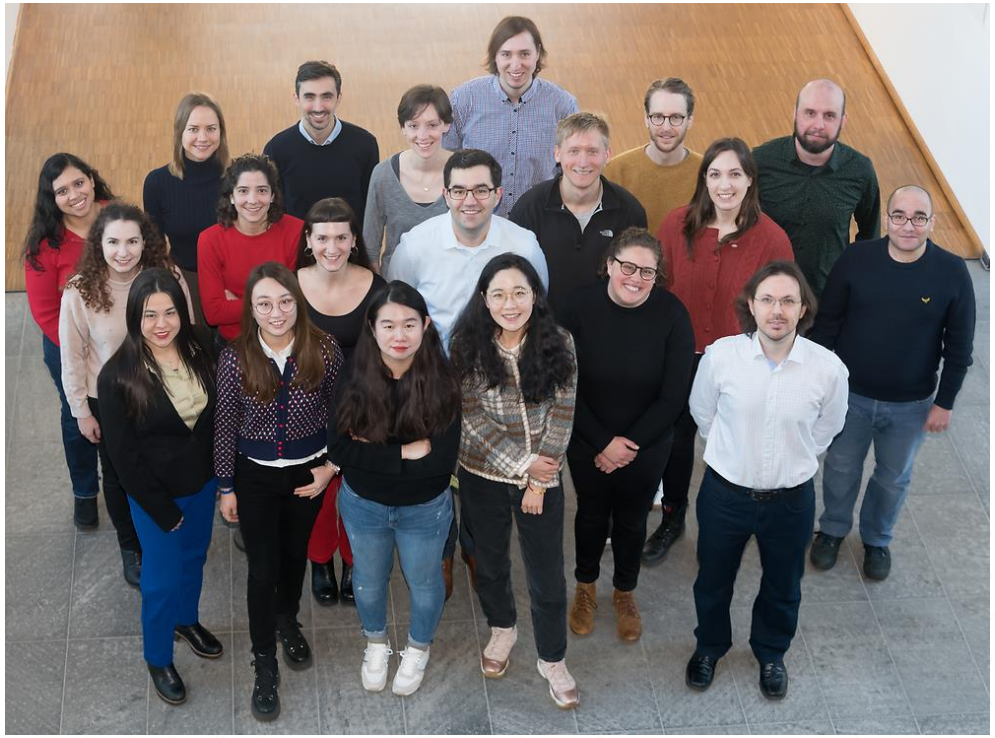
ABSTRACT An accurate estimation of international migration is hampered by a lack of timely and comprehensive data, and by the use of different definitions and measures of migration in different countries. In an effort to address this situation, we complement traditional data sources for the United Kingdom with social media data: our aim is to understand whether information from digital traces can help measure international migration. The Bayesian framework proposed is used to combine data from the Labour Force Survey (LFS) and the Facebook Advertising Platform to study the number of European migrants in the United Kingdom, with the aim of producing more accurate estimates of the numbers of European migrants. The overarching model is divided into a Theory-Based Model of migration and a Measurement Error Model. We review the

TODAY'S OUTLINE

- **TRENDS OVER TIME AND FILTERING OUT BIAS**
- **SURVEYS AND NON-REPRESENTATIVE SAMPLES**
- **CALIBRATION AND WEIGHTING**


Crises & Shocks
Technological Change & Digitalization
METHODS FOR EVER-EVOLVING DATA
Measuring & Modeling Migration
High – Skilled Migration
Integration & Segregation
Health & Mortality
Family & the life course
Socio-Economic forces
Lab of Migration & Mobility
Lab of Population Dynamics & Sustainable Well-Being
Environmental Change

ACKNOWLEDGMENTS & FURTHER DISCUSSION



DIGITAL AND COMPUTATIONAL DEMOGRAPHY @MPIDR

THE COMMUNITY AT THE MPI FOR DEMOGRAPHIC RESEARCH