3 - Digital trace data

Diego Alburez-Gutierrez
MPIDR
European Doctoral School of Demography 2019-20

01/04/2020

MAX-PLANCK-INSTITUT FÜR DEMOGRAFISCHE FORSCHUNG
MAX PLANCK INSTITUTE FOR DEMOGRAPHIC RESEARCH

Agenda

1. Q&A
2. Introduction to digital trace and marketing data
3. Example 1: Migration
4. Example 2: Internet users
5. Discussion

Q&A

- Questions about on `Exercise 1` from the final assignment
- Issues with Familinx data
- Other?

Digital traces are incidental to our online presence

- Digital breadcrumbs are unavoidable
- Pre-GDPR, largely unchecked
- Marketing-led
- Not collected for social-scientific research

Some data sources

1. Marketing platforms
   - Facebook/Instagram/WhatsApp API
   - Linkedin API

Some data sources

1. Marketing platforms
   - Facebook/Instagram/WhatsApp API
   - Linkedin API
2. Online platforms and communication
   - Twitter (API)
   - Google Trends
   - Email, IP address, mobile phones

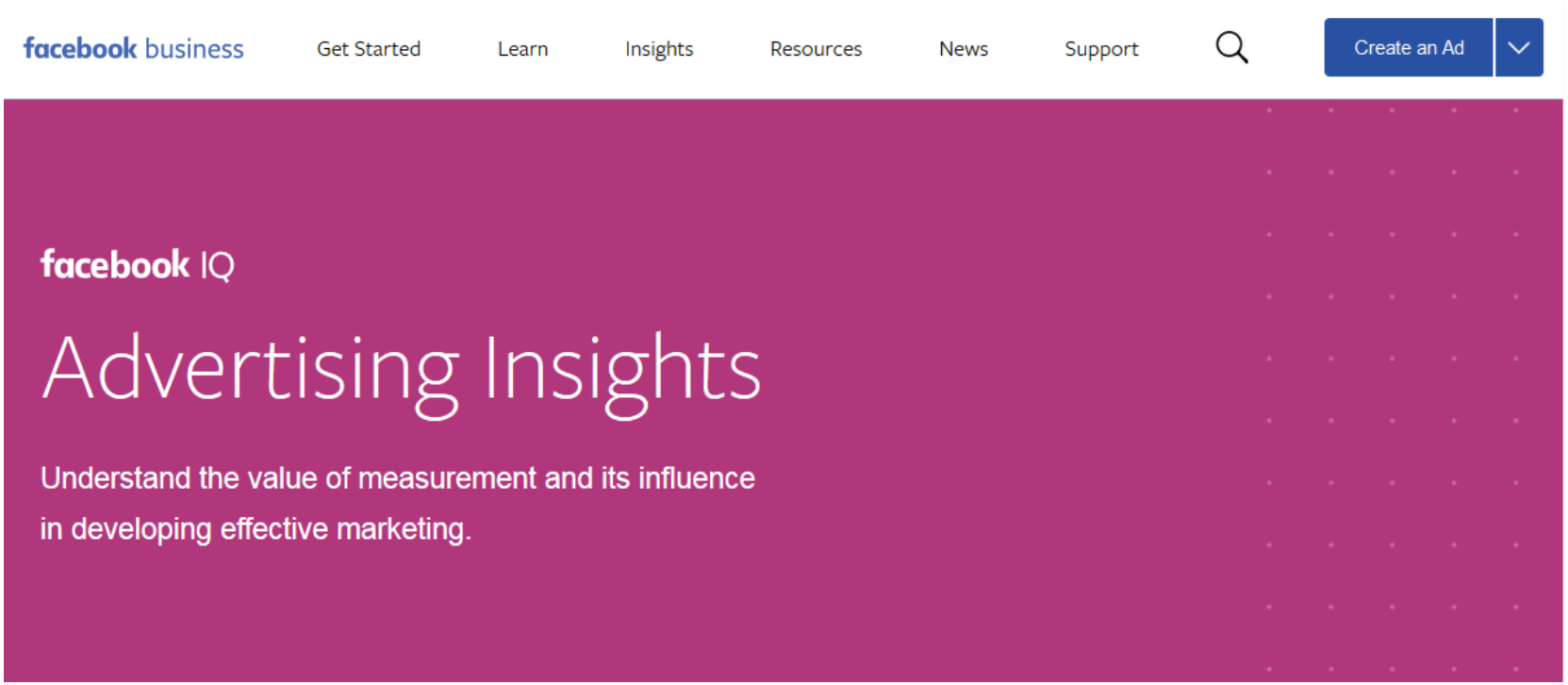
Some data sources

1. Marketing platforms
   - Facebook/Instagram/WhatsApp API
   - Linkedin API
2. Online platforms and communication
   - Twitter (API)
   - Google Trends
   - Email, IP address, mobile phones
3. Internet of Things
   - Activity trackers and wearable medical devices
   - Wearable sensors (see Cito Cattuto)

Facebook marketing platforms and APIs

- Sofia Gil's tutorial:
  https://github.com/SofiaG1l/Using_Facebook_API
- For python users, Carol Coimbra's:
  https://github.com/carolcoimbra/facebook-ads

Using online marketing tools for demographic research



‘Audience estimates’: FB users in Guatemala

Ads Manager

Search

Diego

Diego Alburez (371284279)

Ad set name 18+

Switch to Quick Creation

Campaign
- Objective

Ad account
- Create new

Ad set
- Page
- Audience
- Placements
- Budget & schedule

Ad
- Identity
- Format
- Media
- Text

Close

Locations

Guatemala

Guatemala City, Guatemala Department

+ 40 km

Include | Type to add more locations | Browse

Cobán, Tapachula, Quetzaltenango (Xela), Guatemala City, Antigua Guatemala, Escuintla, Jutiapa, Atiquizaya, Acajutla, Zacapa, Chiquimula

Drop Pin

Add locations in bulk

Age 18 - 65+

Gender All | Men | Women

Languages Enter a language...

Include people who match

Behaviours > Mobile Device User

Estimate doesn't include Facebook Stories

Because Facebook Stories is a new placement being released gradually, audience and reach estimates aren't currently available. These estimates are based on the other placements that you've selected.

Audience size

Specific — Broad

Your audience is defined.

Potential reach: 3,700,000 people

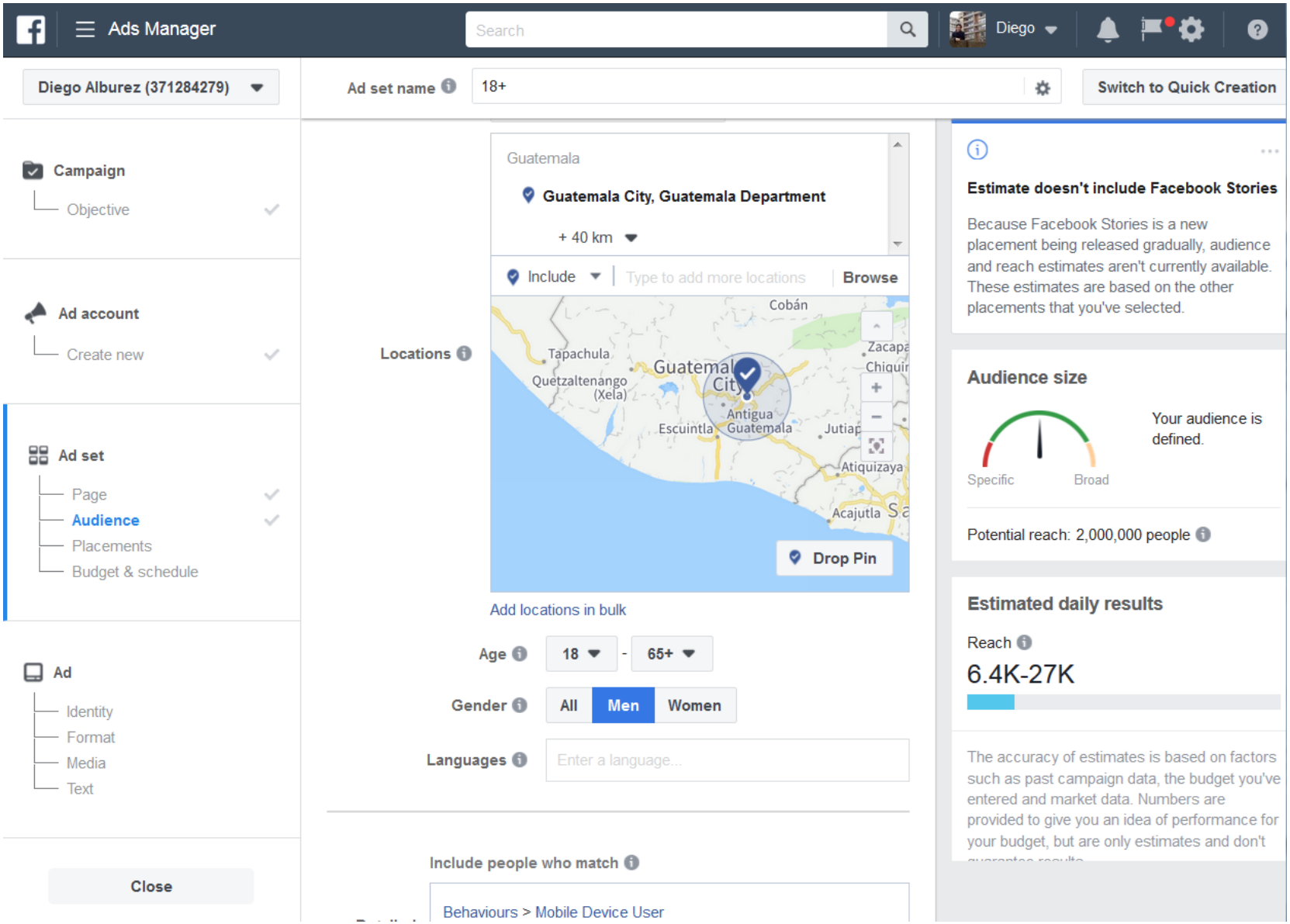
Estimated daily results

Reach

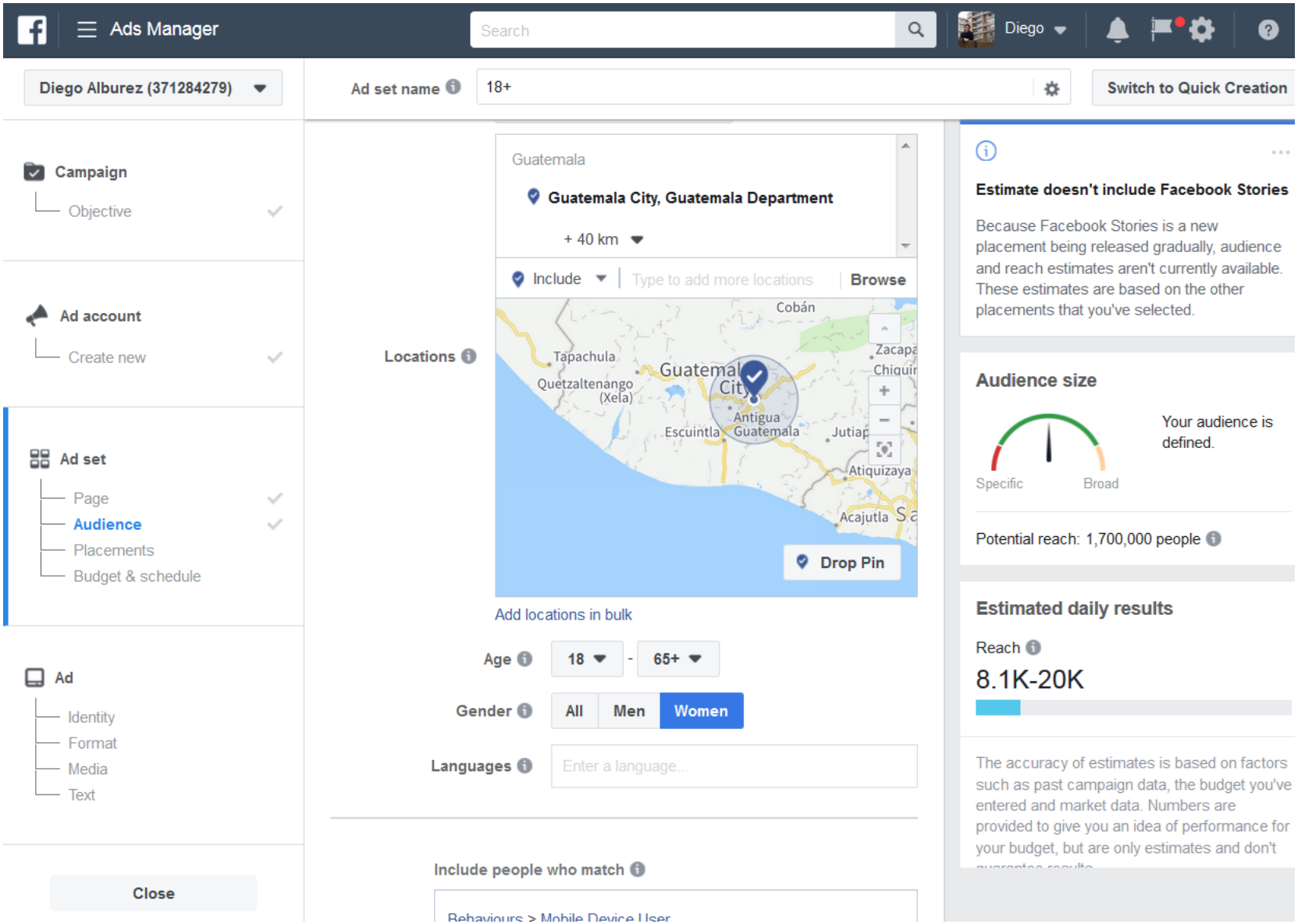
7.5K-22K

The accuracy of estimates is based on factors such as past campaign data, the budget you've entered and market data. Numbers are provided to give you an idea of performance for your budget, but are only estimates and don't

Male FB users, aged 18+ in Guatemala City



Female FB users, aged 18+ in Guatemala City



Question time!

FB audience estimates are used for micro-targeted advertisment.

1. What is this micro-targeting and who uses it?
2. How can it be used for demographic research?

Some magic sampling. . .

```
## [1] "Octavio" "Niall"   "Andres"
```

Some good practices for digital demography

1. Acknowledge non-representativeness
2. Conduct reality checks: compare to IRL data
3. Account for drifting (population, system and behavioural)
4. Remember algorithmic confounding (observing a casino?)
5. Think of ethics, be transparent and upfront

Question time (again)!

We'll review two studies. Identify the

1. **strengths**
2. **weaknesses**

of their reliance on digital trace data.

Some magic sampling. . .

what	who
gender gap	Niall
gender gap	Rustam
migration	Octavio
migration	Andres

Example 1: Migration

Research at a glance

- RQ: Estimate out-migration from Puerto Rico in the months after 2017 Hurricane Maria
- Data: FB advertising platform and American Community Survey (ACS)
- Findings: Flows by age, sex, and US State

Alexander, M., Polimis, K. and Zagheni, E. (2019), The Impact of Hurricane Maria on Out-migration from Puerto Rico: Evidence from Facebook Data. Population and Development Review, 45: 617-630.

Sanity checks

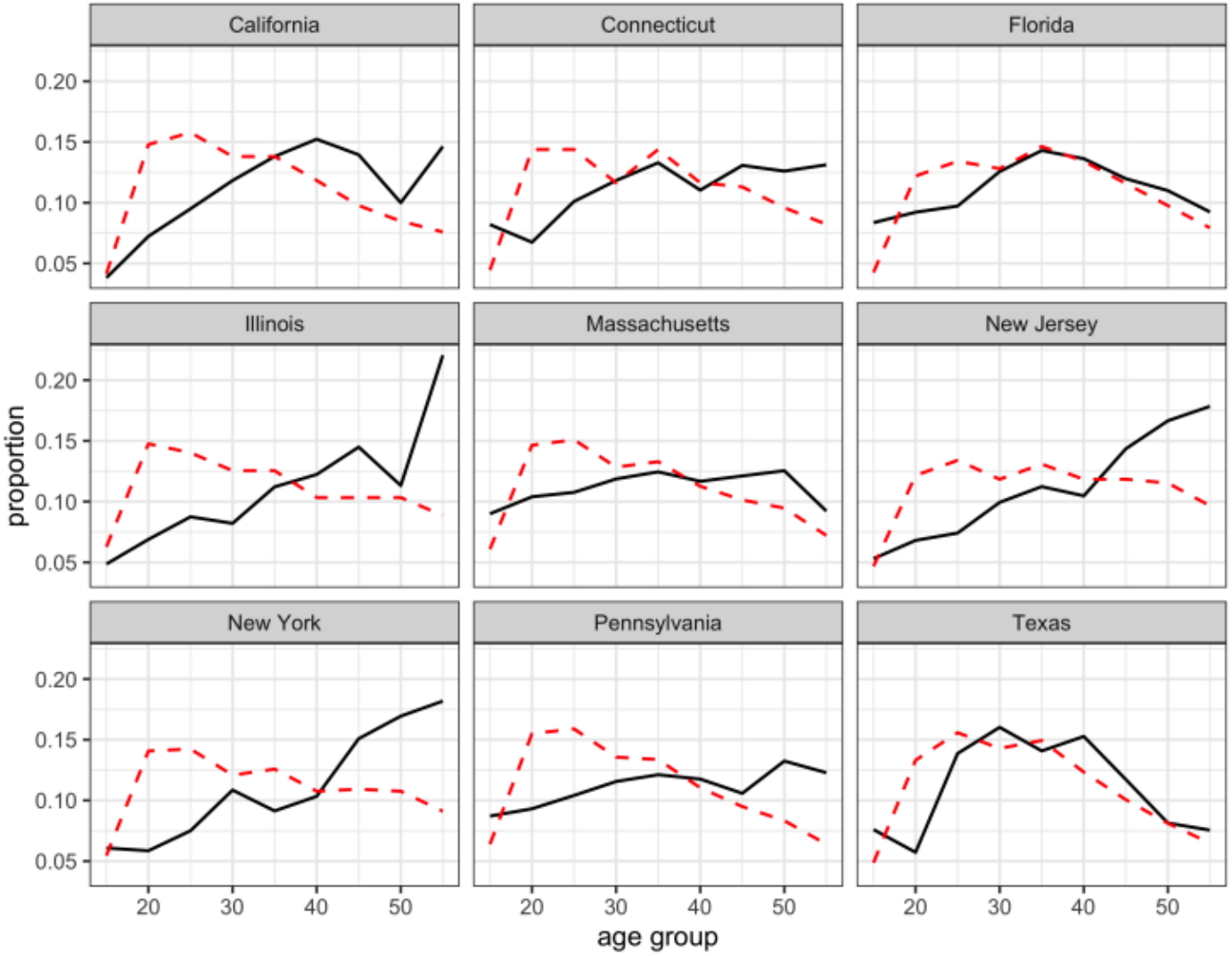


Figure 1: Age distribution of Puerto Rican migrants in FB data (red dashed line) and American Community Survey (black solid line).

Population increase

Table 2: Estimated increase in Puerto Rican migrant stocks from October 2017 to January 2018. The 95% confidence intervals are shown in parentheses.

State (95% CI)	% Increase (95% CI)	Population Increase
Florida	21.6 (20.9, 22.3)	65433 (63342, 67525)
New York	11 (10.3, 11.7)	14477 (13584, 15371)
Pennsylvania	13.4 (12.7, 14.1)	13441 (12700, 14181)
Connecticut	14.7 (12.9, 16.5)	9402 (8244, 10560)
Massachusetts	10.1 (8.82, 11.4)	8957 (7824, 10090)
Texas	10.8 (10.4, 11.2)	5678 (5452, 5904)
Ohio	12.8 (12.2, 13.4)	3274 (3125, 3424)
Illinois	9.9 (9.15, 10.6)	2641 (2441, 2841)
Georgia	13.1 (12.4, 13.8)	2606 (2470, 2742)
New Jersey	2.9 (1.56, 4.24)	2282 (1228, 3336)
California	2.4 (1.86, 2.94)	573 (444, 702)

Alexander, M., Polimis, K. and Zagheni, E. (2019), The Impact of Hurricane Maria on Out-migration from Puerto Rico: Evidence from Facebook Data. Population and Development Review, 45: 617-630.

Percent change by age groups

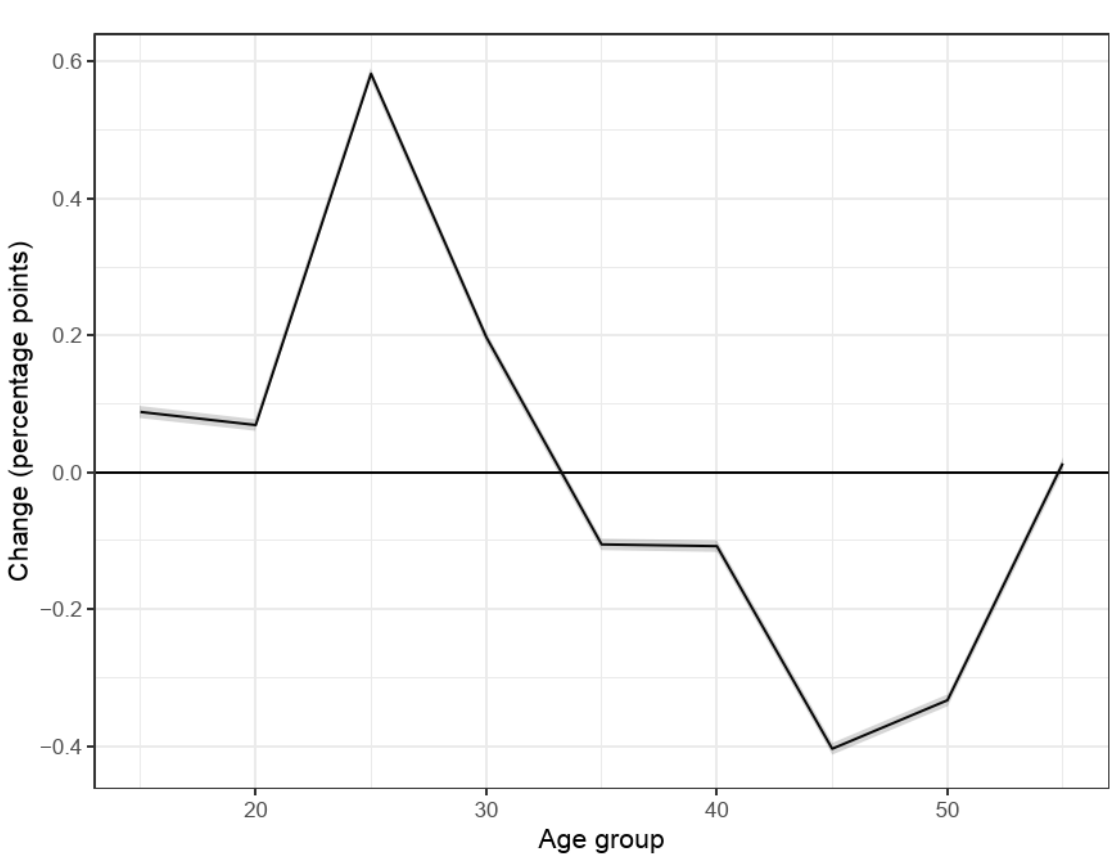


Figure 3: Estimated change in Puerto Rican migrant age distribution from October 2017 to January 2018.

Alexander, M., Polimis, K. and Zagheni, E. (2019), The Impact of Hurricane Maria on Out-migration from Puerto Rico: Evidence from Facebook Data. Population and Development Review, 45: 617-630.

Example 2: Digital use

Summary

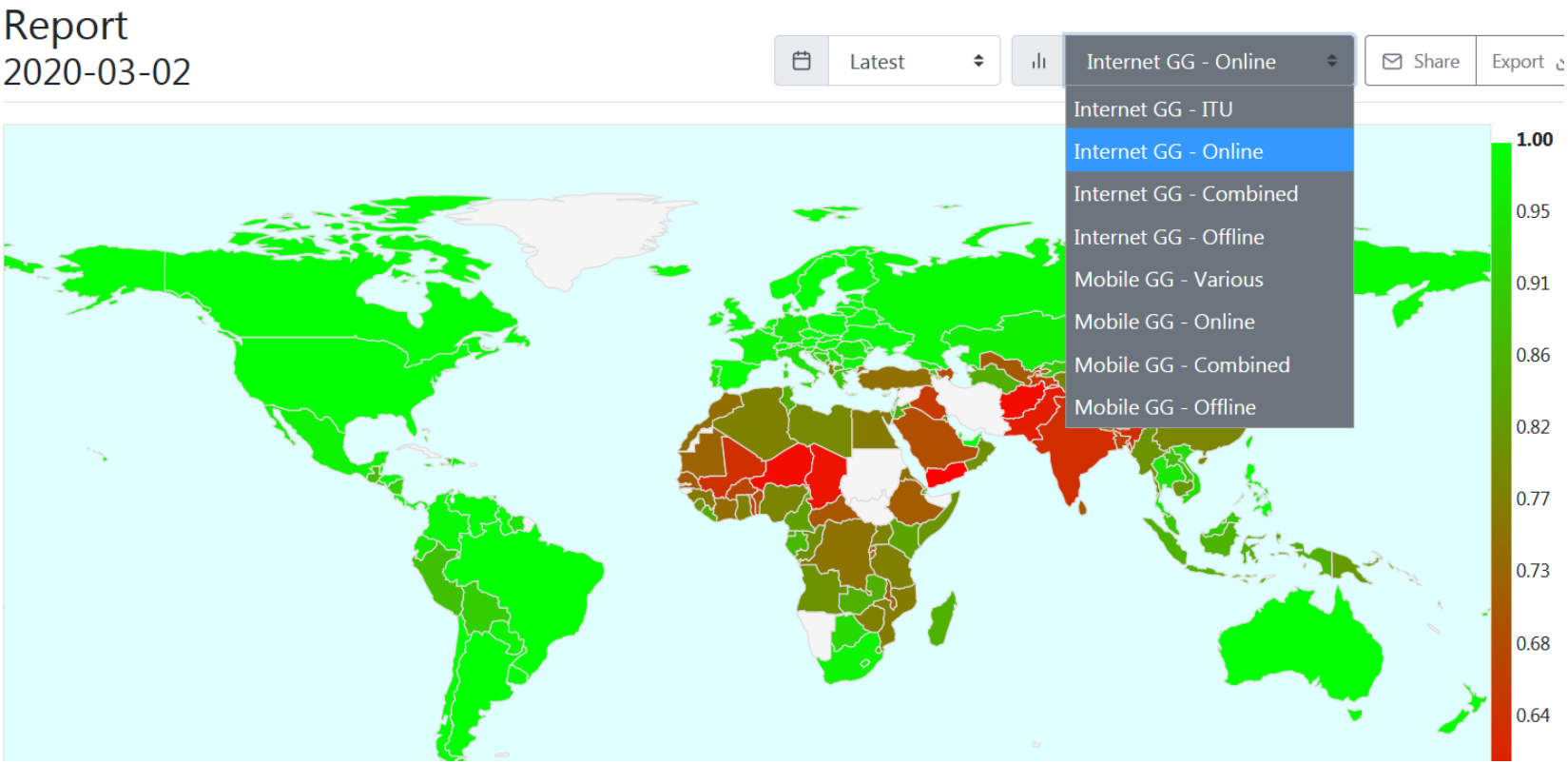
- RQ: Predict internet and mobile phone use gender gaps
- Data: FB advertising platform and indicators from offline sources
- Estimating rates: Facebook Gender Gap Index:

$$\frac{\text{Female to male gender ratio of people with characteristic}}{\text{Female to Male gender ratio of the population}}$$

- Findings:
  - Facebook-based measure performed well compared to ground truth
  - Online+offline measure: best estimates

Fatehkia, M., Kashyap, R., and Weber, I. (2018). Using Facebook ad data to track the global digital gender gap. World Development 107:189–209.

Measuring the gender gap in real-time



https://www.digitalgendergaps.org/data/?report=2020-03-02

Discussion

Question time (refresher)!

We'll review two studies. Identify the

1. **strengths**
2. **weaknesses**

of their reliance on digital trace data.

what	who
gender gap	Niall
gender gap	Rustam
migration	Octavio
migration	Andres

Strengths and weaknesses: Puerto Rico migration

- Pro: Real-time data
- Con: No 'ground-truth' data (?)
- Con: Non-representative sample based on unknown algorithms
- Pro: Difference-in-difference to adjust for bias

Strengths and weaknesses: Digital gender gap

- Pro: Nowcasting
- Pro: Ideal data for the job?
- Pro: 'Ground-truth' data: Internet Gender Gap Index
- Con: Rates are unadjusted - what is the data representative of?

Challenges going ahead

> *Whoever you are... I've always depended on the kindness of strangers.*
>
> — Blanche DuBois, A Streetcar Named Desire

1. Ensuring sustainable data access
2. Addressing systematic bias
3. No information information about algorithms that companies use internally (eg. rounding errors)
4. Privacy and ethical digital research

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. Journal of Information Technology 30(1):75–89.

Make yourself heard!

1. What are the main ethical concerns when using digital trace data?
2. Do all/any apply to digital demographers?
3. How can we minimise risk for users?

Homework

- Start with `Exercise 2`
- Think for tomorrow: How does all of this relate to your interests (if at all)?