Data Science, Demography and Social Media

Challenges and Opportunities

Emilio Zagheni

Department of Sociology and eScience Institute
University of Washington, Seattle

February 2, 2017

Today's seminar

1. How data science and social media are transforming demography
2. How demographic thinking helps us make sense of messy and biased data
3. How misuse of new tools and data may lead to dangerous outcomes

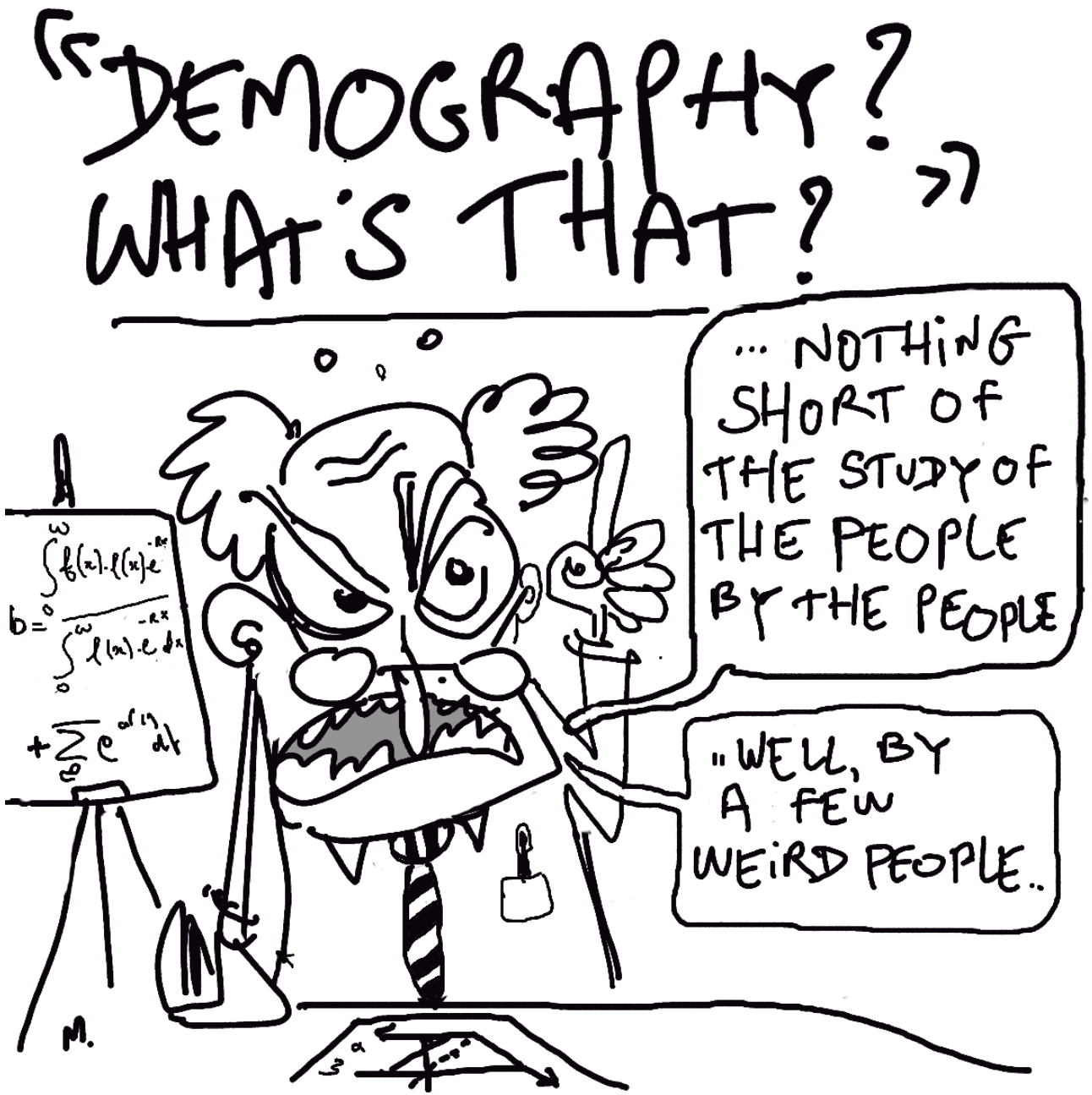
Outline

a. Background on 'digital demography'

b. An example from my own research: Estimating migration using Facebook advertisement data

c. Potential misuse of online advertising platforms

d. Making sense of messy data: an example using Twitter data



credit: Emmanuel Letouzé

Demography is the study of populations (including non-human populations). It deals with processes related to mortality, fertility, migration. It attempts to explain the causes and consequences of population dynamics.

"*Demography is the quintessential quantitative social science. It bears something of the same relationship to other social sciences that physics bears to other natural sciences*"

- Ken Wachter, Essential Demographic Methods

"*Demography is the quintessential quantitative social science. It bears something of the same relationship to other social sciences that physics bears to other natural sciences*"

- Ken Wachter, Essential Demographic Methods

$\Rightarrow$ Demography is a discipline that plays a central role in the social sciences

"*Biodemography fundamentally deepens our understanding of the underlying evolutionary drivers of demographic patterns across the tree of life*"

- Jim Vaupel

"*Biodemography fundamentally deepens our understanding of the underlying evolutionary drivers of demographic patterns across the tree of life*"

- Jim Vaupel

$\Rightarrow$ Demography as an engine of innovation in the biological sciences

One of Demography's many traits

Demography is (or aspires to be) a driver of innovation for *all* sciences and is energized by exchange of ideas with other disciplines

Parallels with Data Science

~~Demography~~ Data Science is (or aspires to be) a driver of innovation for *all* sciences and is energized by exchange of ideas with other disciplines

What has made Demography successful?

- **The intrinsic nature of the discipline**:
  - It deals with quantities that are relatively easy to measure
  - The object of analysis is suitable for mathematical modeling
  - Everything is a population
- **Extrinsic factors:**
  - Data availability (often collected from authorities for a number of purpose)
  - The questions asked have policy relevance
  - Major demographic issues faced by societies (e.g. population growth, population aging)

What is the next frontier in Demography? What are the challenges ahead?

“Digital Demography”

- The Web, social media and smartphones have had a sudden and unprecedented impact on our lives and have given researchers new data to study demographic behavior.

"Digital Demography"

- The Web, social media and smartphones have had a sudden and unprecedented impact on our lives and have given researchers new data to study demographic behavior.

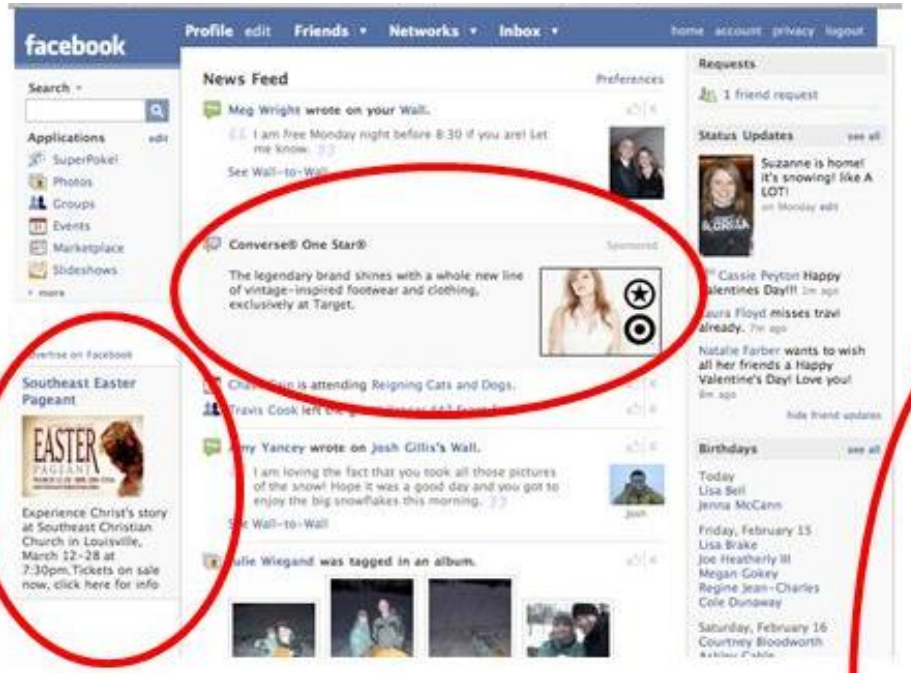
- 'Digital demography' is about:
  1. Studying the implications of the digital revolution on demographic behavior
  2. Using new data sources to better understand demographic processes

Using Facebook Advertisement Data to Estimate Migration

Joint work with Ingmar Weber (QCRI) and Krishna Gummadi (MPI)

What ads looked like in the 1930s...

Today: Online (targeted) advertising



Reach the right people.
Instead of creating an advertisement and hoping that it reaches the right customers, you can create a Facebook Social Ad and target it precisely to the audience you choose. The ads can also be shown to users whose friends have recently engaged with your Facebook Page or engaged with your website through Facebook Beacon. Social Ads are more likely to influence users when they appear next to a story about a friend's interaction with your business.

Advertisers define who they want to reach based on factors like interests, age, location and more.

We show their ads to the people most likely to be interested in their products, services and causes.

When an advertiser wants to reach...

Nearby cyclists

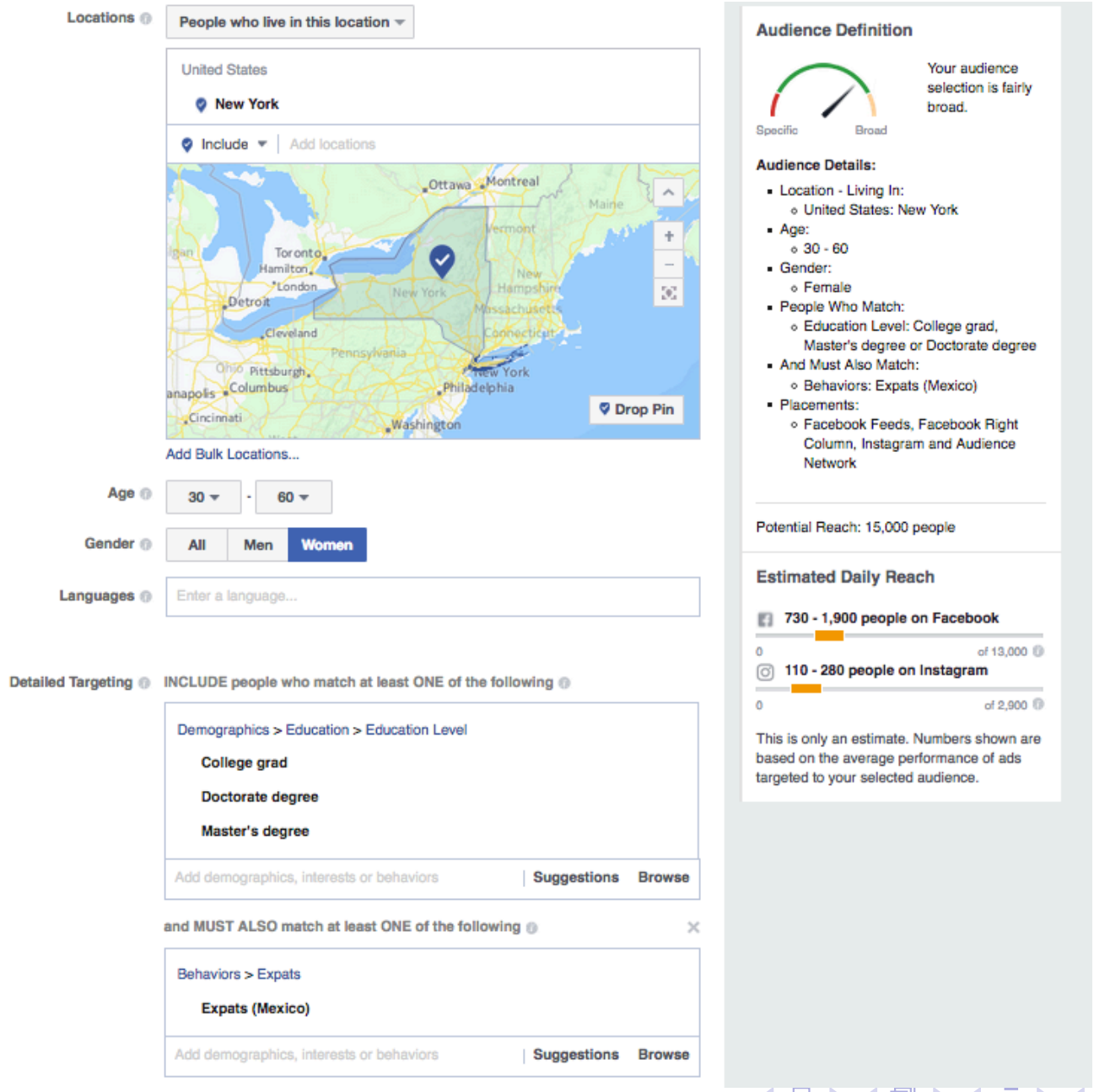
- Between **18 - 35** years old
- **Female**
- Within **20 miles** of my store
- Interested in **bicycling**
- **Mobile** users

We show their ad to people like...

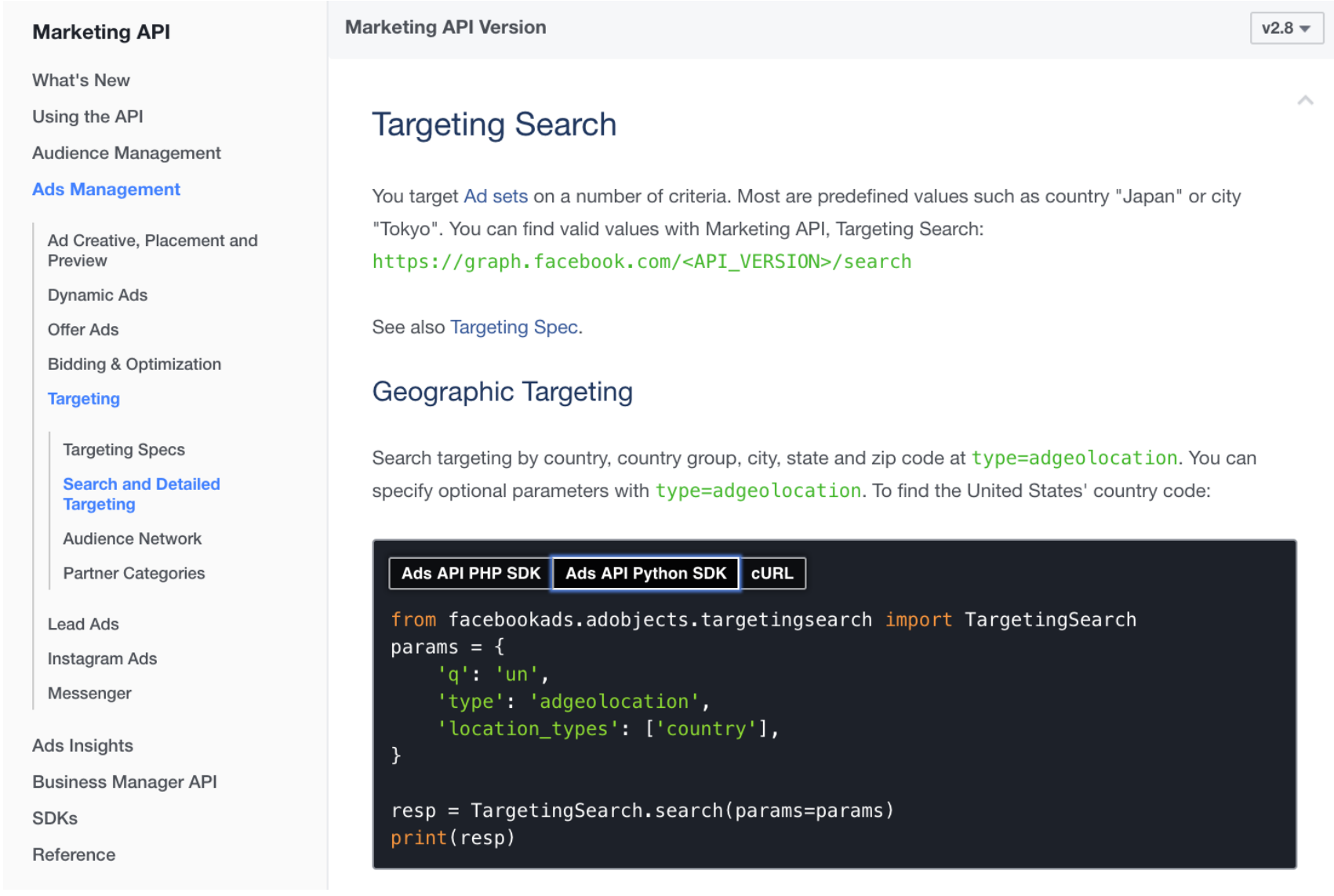
Elena

- **30** years old
- **Female**
- **Menlo Park, CA**
- Interested in **bicycling** movies, cooking
- **iPhone user** car shopper, gamer

Targeting a demographic group on Facebook



You can access the data in a programmatic way



Marketing API

- What's New
- Using the API
- Audience Management
- Ads Management
  - Ad Creative, Placement and Preview
  - Dynamic Ads
  - Offer Ads
  - Bidding & Optimization
  - Targeting
    - Targeting Specs
    - Search and Detailed Targeting
    - Audience Network
    - Partner Categories
  - Lead Ads
  - Instagram Ads
  - Messenger
- Ads Insights
- Business Manager API
- SDKs
- Reference

Marketing API Version v2.8

Targeting Search

You target Ad sets on a number of criteria. Most are predefined values such as country "Japan" or city "Tokyo". You can find valid values with Marketing API, Targeting Search:
`https://graph.facebook.com/<API_VERSION>/search`

See also Targeting Spec.

Geographic Targeting

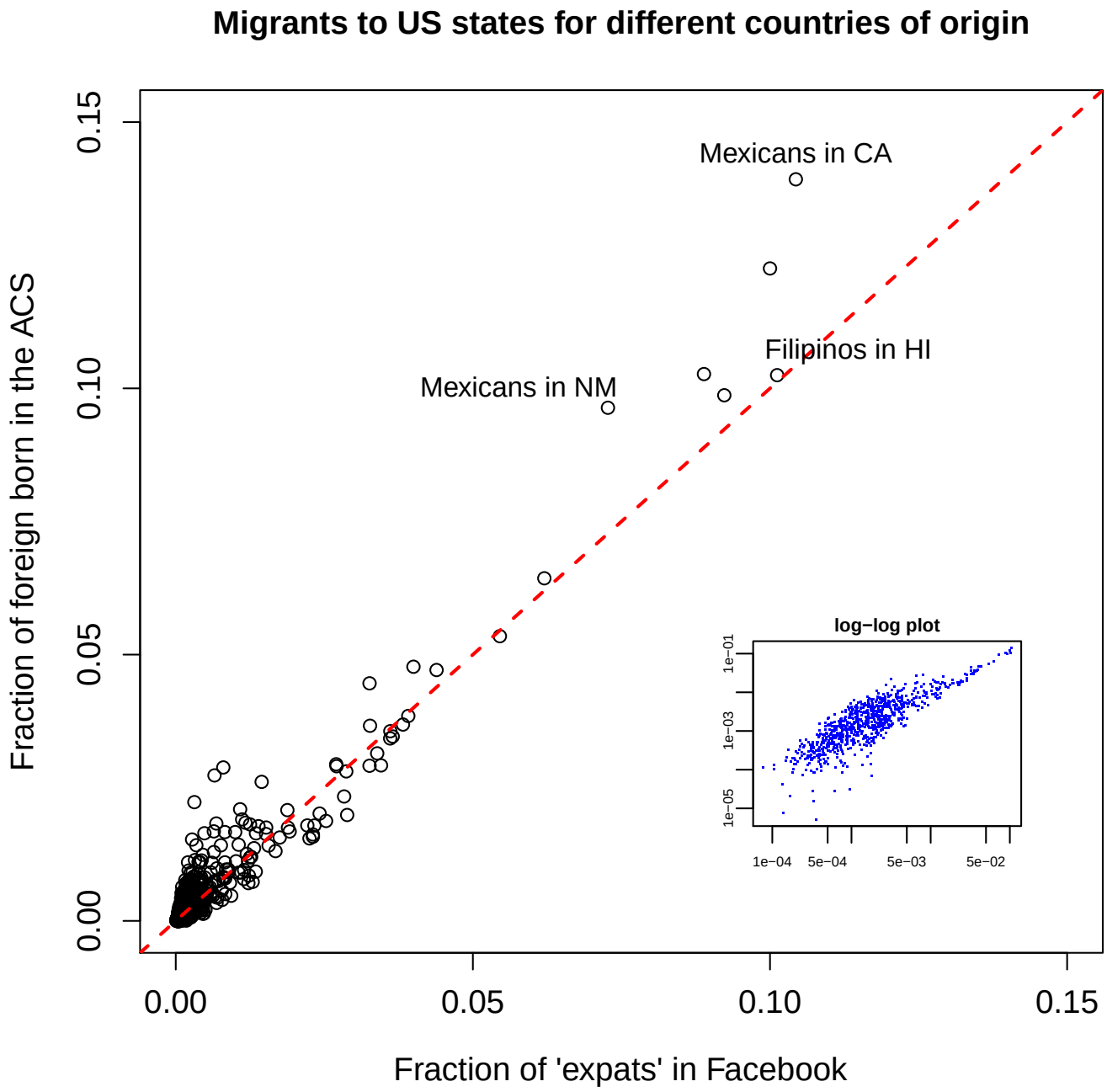
Search targeting by country, country group, city, state and zip code at `type=adgeolocation`. You can specify optional parameters with `type=adgeolocation`. To find the United States' country code:

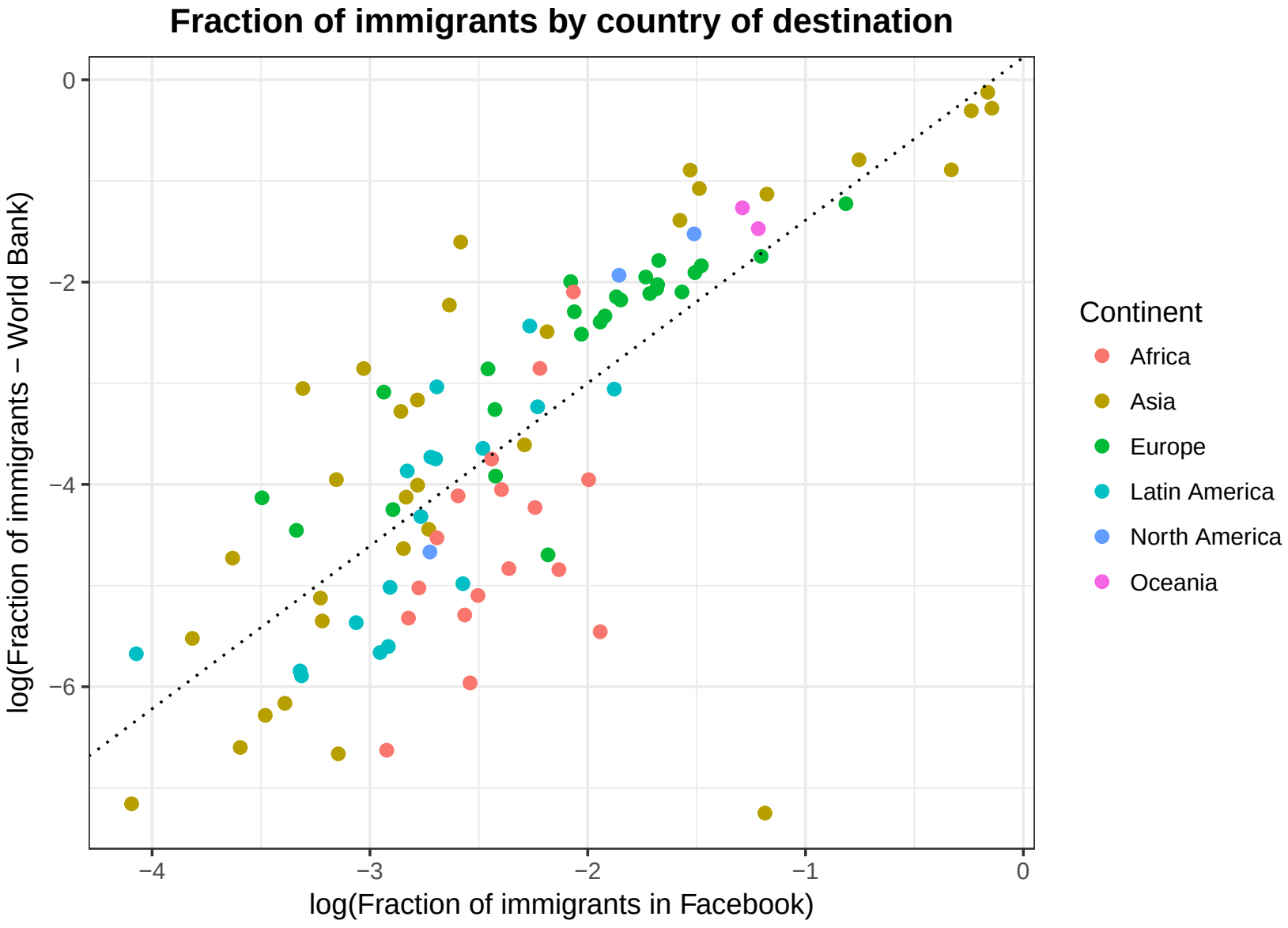
Ads API PHP SDK | Ads API Python SDK | cURL

```
from facebookads.adobjects.targetingsearch import TargetingSearch
params = {
    'q': 'un',
    'type': 'adgeolocation',
    'location_types': ['country'],
}

resp = TargetingSearch.search(params=params)
print(resp)
```

Leveraging Facebook to study Migration


Migrants to US states for different countries of origin
Mexicans in CA
Filipinos in HI
Mexicans in NM
log−log plot
1e−01
1e−03
1e−05
1e−04
5e−04
5e−03
5e−02
0.00
0.05
0.10
0.15
Fraction of foreign born in the ACS
Fraction of 'expats' in Facebook


Fraction of immigrants by country of destination
log(Fraction of immigrants – World Bank)
0
−2
−4
−6
−4
−3
−2
−1
0
log(Fraction of immigrants in Facebook)
Continent
Africa
Asia
Europe
Latin America
North America
Oceania

A tool potentially useful for demographic and survey research, but that could also be misused...

Machine Bias

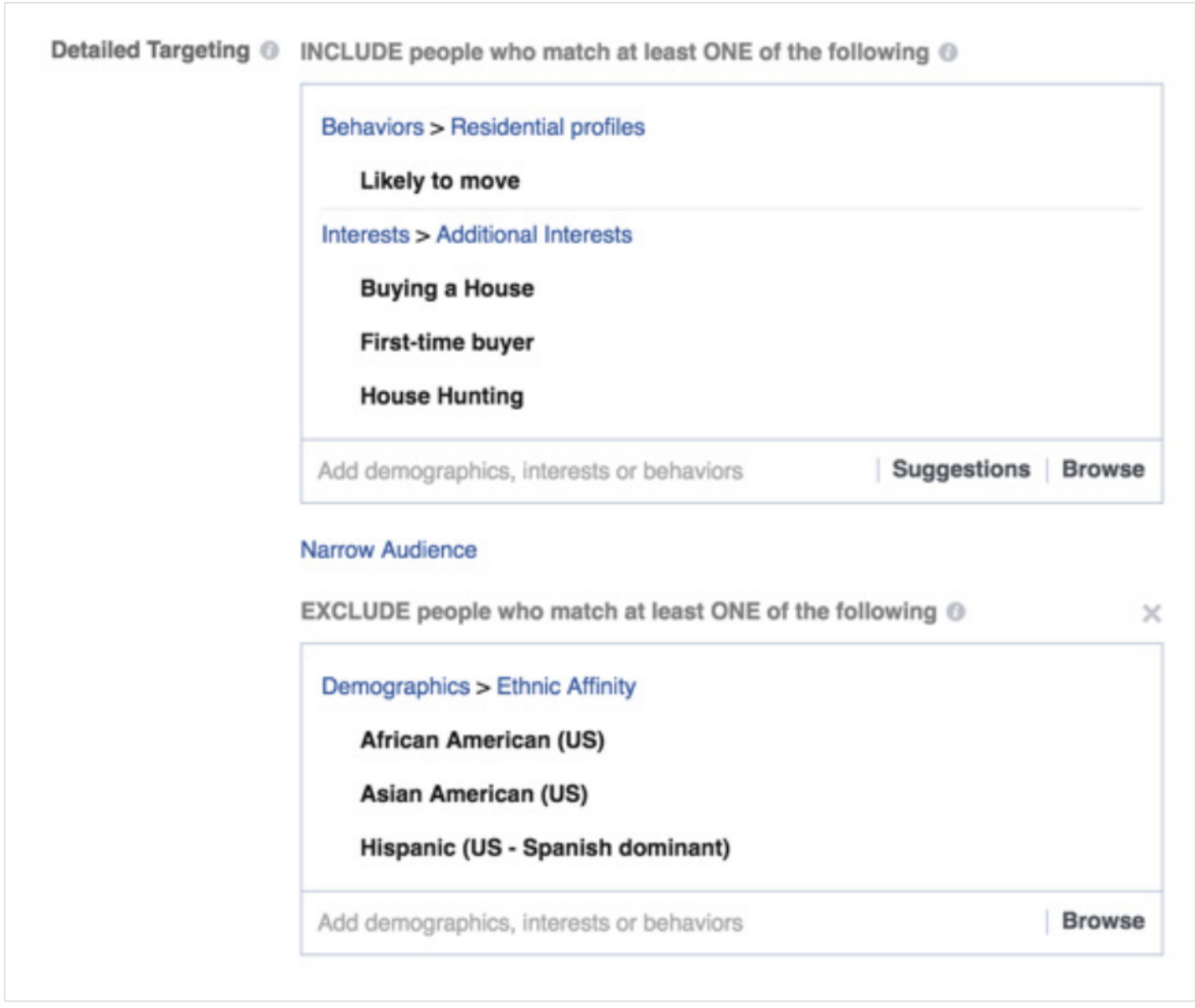
Facebook Lets Advertisers Exclude Users by Race

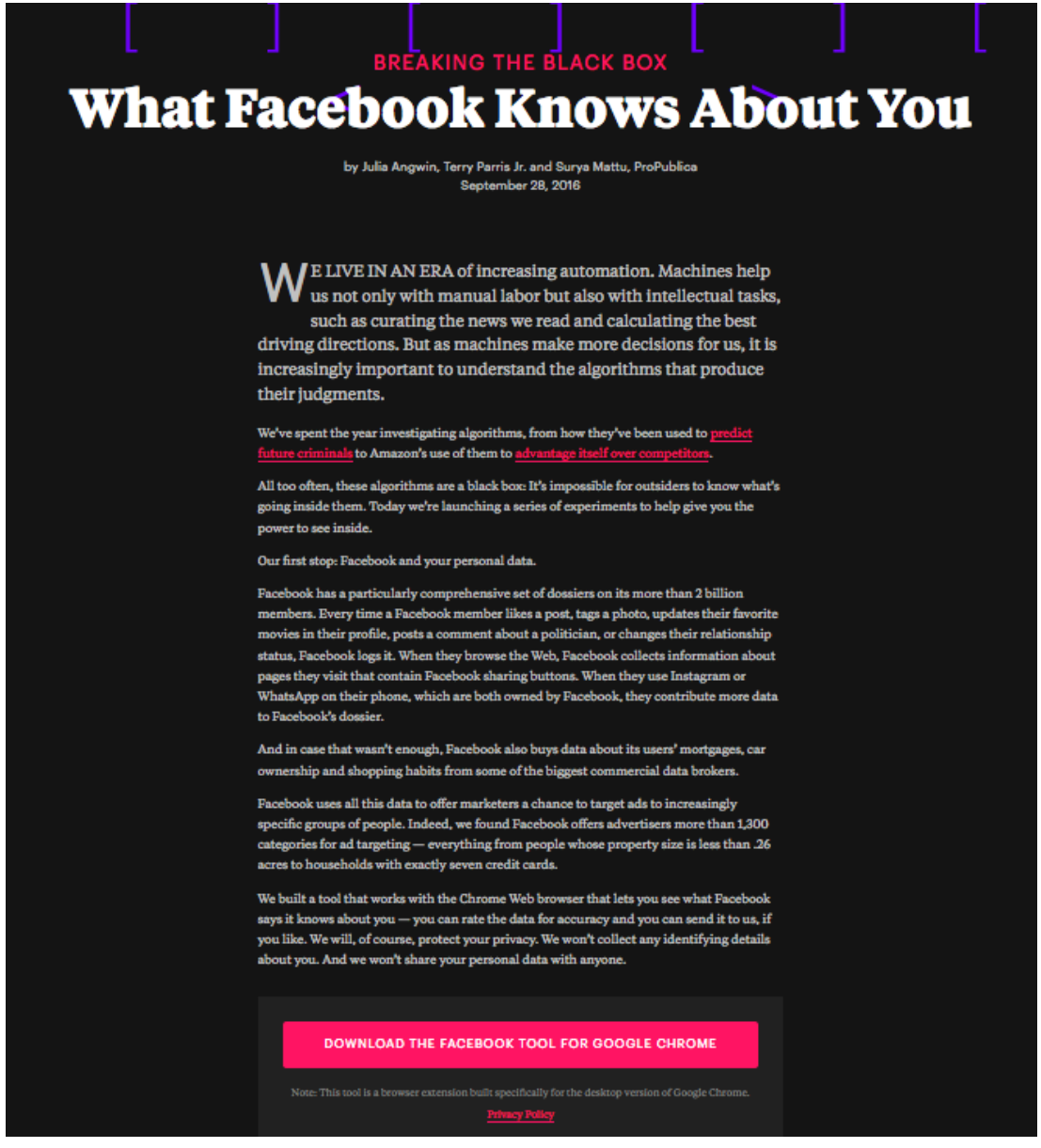
Facebook's system allows advertisers to exclude black, Hispanic, and other "ethnic affinities" from seeing ads.

by Julia Angwin and Terry Parris Jr.
ProPublica, Oct. 28, 2016, 7 a.m.

632 Comments | Print

Here is a screenshot of an ad we purchased in Facebook's housing categories via the company's advertising portal:





BREAKING THE BLACK BOX

What Facebook Knows About You

by Julia Angwin, Terry Parris Jr. and Surya Mattu, ProPublica
September 28, 2016

WE LIVE IN AN ERA of increasing automation. Machines help us not only with manual labor but also with intellectual tasks, such as curating the news we read and calculating the best driving directions. But as machines make more decisions for us, it is increasingly important to understand the algorithms that produce their judgments.

We've spent the year investigating algorithms, from how they've been used to predict future criminals to Amazon's use of them to advantage itself over competitors.

All too often, these algorithms are a black box: It's impossible for outsiders to know what's going inside them. Today we're launching a series of experiments to help give you the power to see inside.

Our first stop: Facebook and your personal data.

Facebook has a particularly comprehensive set of dossiers on its more than 2 billion members. Every time a Facebook member likes a post, tags a photo, updates their favorite movies in their profile, posts a comment about a politician, or changes their relationship status, Facebook logs it. When they browse the Web, Facebook collects information about pages they visit that contain Facebook sharing buttons. When they use Instagram or WhatsApp on their phone, which are both owned by Facebook, they contribute more data to Facebook's dossier.

And in case that wasn't enough, Facebook also buys data about its users' mortgages, car ownership and shopping habits from some of the biggest commercial data brokers.

Facebook uses all this data to offer marketers a chance to target ads to increasingly specific groups of people. Indeed, we found Facebook offers advertisers more than 1,300 categories for ad targeting — everything from people whose property size is less than .26 acres to households with exactly seven credit cards.

We built a tool that works with the Chrome Web browser that lets you see what Facebook says it knows about you — you can rate the data for accuracy and you can send it to us, if you like. We will, of course, protect your privacy. We won't collect any identifying details about you. And we won't share your personal data with anyone.

DOWNLOAD THE FACEBOOK TOOL FOR GOOGLE CHROME

Note: This tool is a browser extension built specifically for the desktop version of Google Chrome.

Privacy Policy

https://www.propublica.org/article/breaking-the-black-box-what-facebook-knows-about-you

Making sense of noisy and messy data

Can you recognize this city?

Does this look a bit more familiar?

The original picture

Can we infer the height of the Space Needle from one of the images?

Can we infer the height of the Space Needle from one of the images?

- No distortions $\Rightarrow$ Compare with buildings around it

Can we infer the height of the Space Needle from one of the images?

- No distortions $\Rightarrow$ Compare with buildings around it
- Distortions consistent across the image $\Rightarrow$ you can still compare with buildings nearby

Can we infer the height of the Space Needle from one of the images?

- No distortions $\Rightarrow$ Compare with buildings around it
- Distortions consistent across the image $\Rightarrow$ you can still compare with buildings nearby
- No clear pattern in distortions $\Rightarrow$ develop a statistical model to understand patterns

Social Media offer a “distorted” image of the real world

Social Media offer a "distorted" image of the real world

- We want to know the true rates for the underlying population

Social Media offer a "distorted" image of the real world

- We want to know the true rates for the underlying population

$\Rightarrow$ Combining different sources of information is key to extracting value from potentially biased data

...Social media data were produced and collected for reasons other than population studies

There is a lot of useful information in big social data, but we need to work hard to interpret the new data sources

Example: Inferring Migration/Mobility patterns from Twitter Data

Zagheni, Garimella, Weber and State 2014

Geo-located Twitter data

Geo-located Twitter data

- We collected a large sample of geo-located Twitter tweets (with geographic coordinates) for the period 2011-2013 in OECD countries

Geo-located Twitter data

- We collected a large sample of geo-located Twitter tweets (with geographic coordinates) for the period 2011-2013 in OECD countries
- We evaluated short-term mobility (periods of 4 months)

Geo-located Twitter data

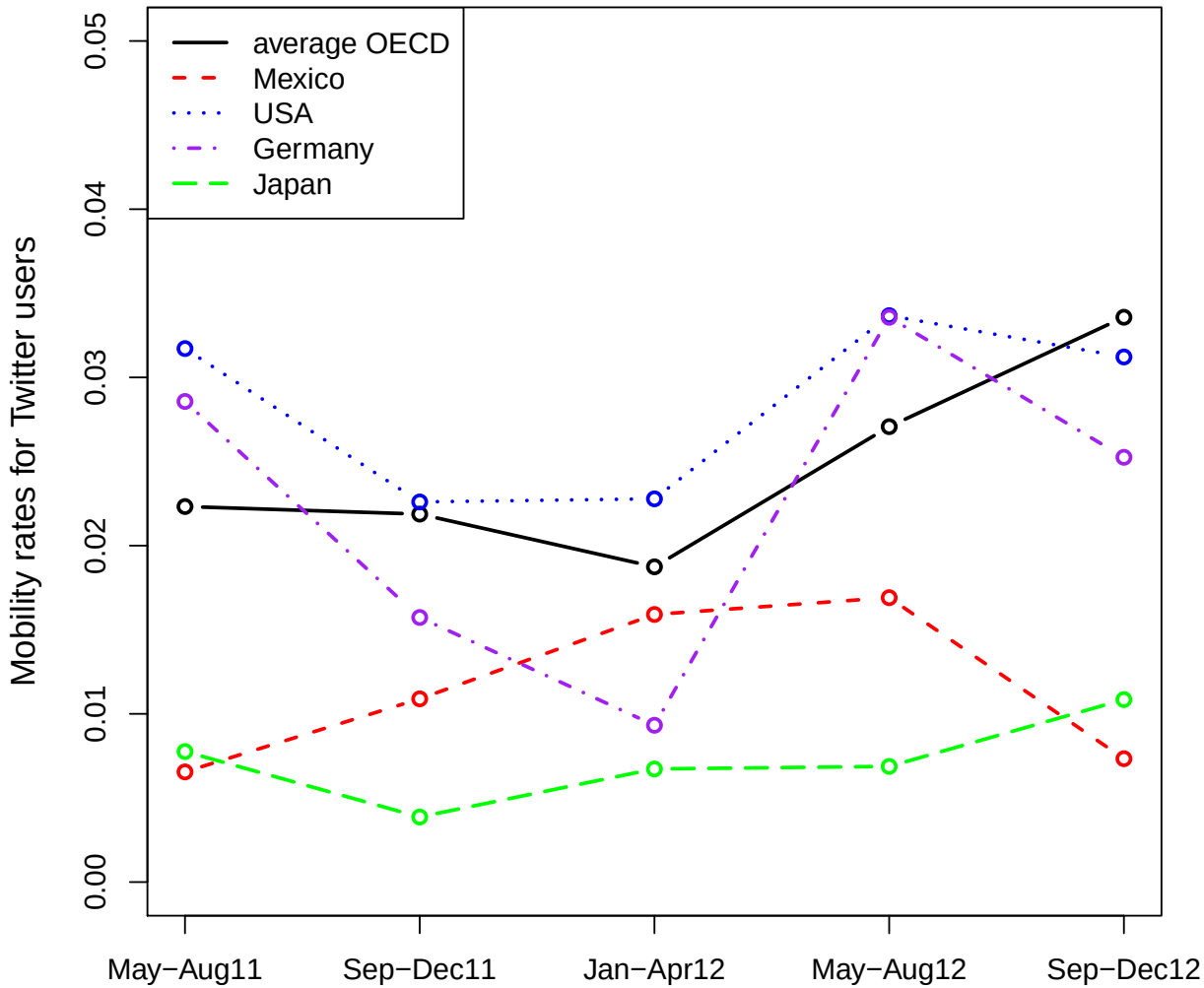
- We collected a large sample of geo-located Twitter tweets (with geographic coordinates) for the period 2011-2013 in OECD countries
- We evaluated short-term mobility (periods of 4 months)
- No official statistics to calibrate a model

Geo-located Twitter data

- We collected a large sample of geo-located Twitter tweets (with geographic coordinates) for the period 2011-2013 in OECD countries
- We evaluated short-term mobility (periods of 4 months)
- No official statistics to calibrate a model

$\Rightarrow$ We proposed a difference-in-differences approach to estimate trends

Geographic mobility for Twitter users



Source: Zagheni, Garimella, Weber and State, WWW'14

Assumptions when ‘ground truth’ data do not exist

Assumptions when 'ground truth' data do not exist

Consider the following situation:

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

and

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

Additive bias different across regions, but constant (or changes by the same amount across regions) over short periods of time

Assume that we knew the 'true' rates (x) for France and Spain

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

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

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

Assume that we knew the 'true' rates (x) for France and Spain

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

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

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

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

Assume that we knew the 'true' rates (x) for France and Spain

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

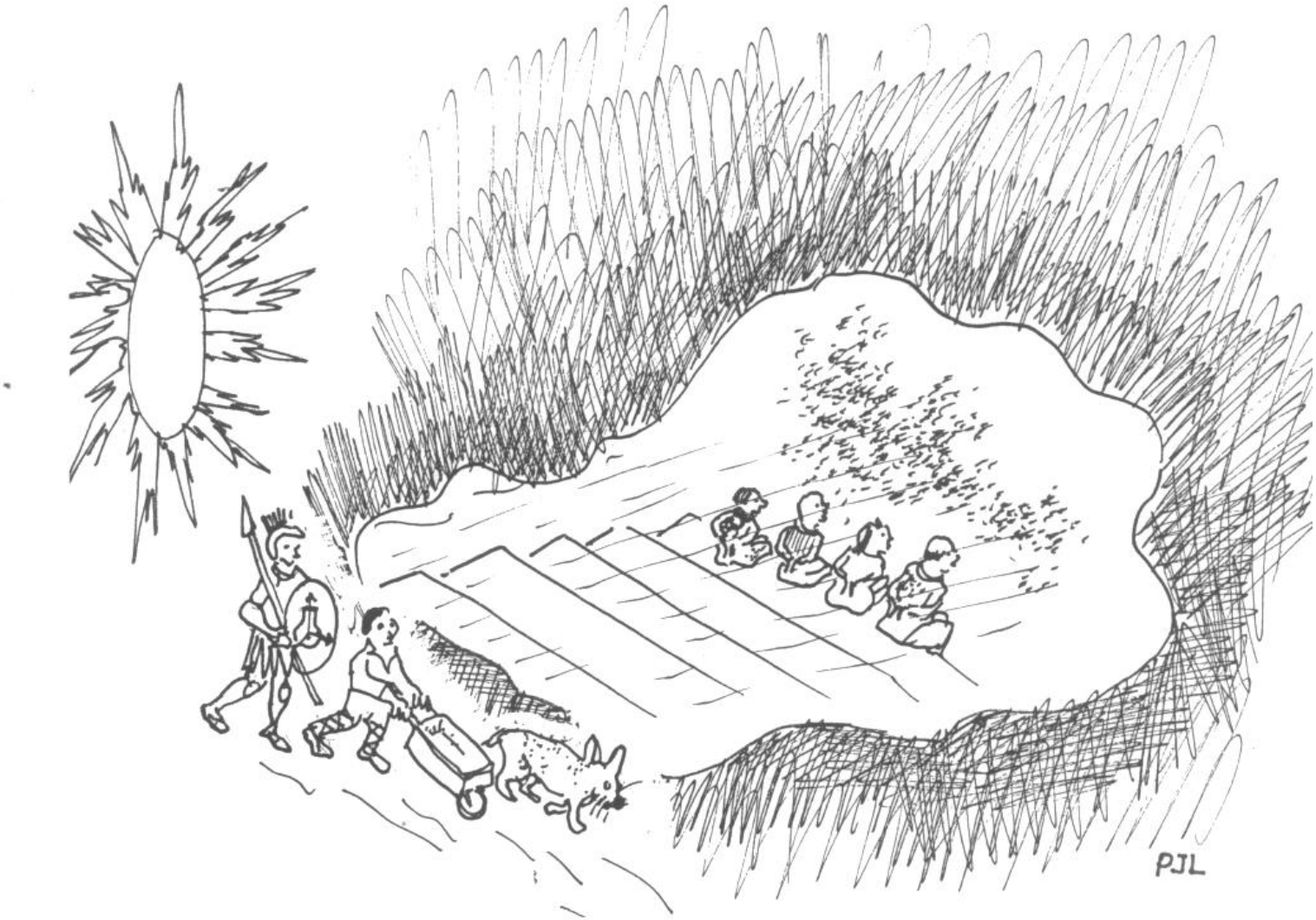
Let's define δ^{t+1} as the differential in the variation of these quantities of interest between time t and $(t+1)$

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

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

$$= 0.2 - 0.1 = 0.1$$

Plato's allegory of the Cave



Plato's Allegory of the Cave

All we see is a distorted image (y) of the 'true' 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}?

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

All we see is a distorted image (y) of the 'true' 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}?

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

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

All we see is a distorted image (y) of the 'true' 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

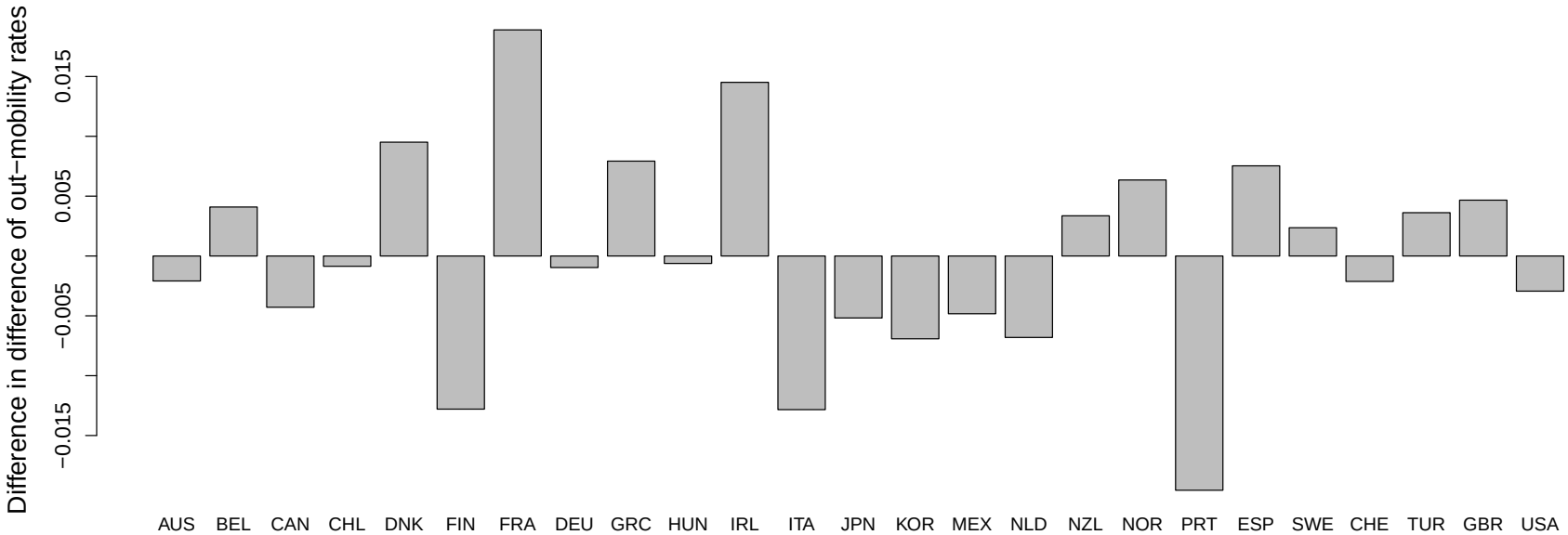
$$\delta^{t+1} = (y_i^{t+1} - y_z^{t+1}) - (y_i^t - y_z^t)$$

After substituting:

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

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

Twitter example



Source: Zagheni, Garimella, Weber and State, WWW'14

Remarks

If the bias is expected to be multiplicative:

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

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

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

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:

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

Tip of the iceberg

Tip of the iceberg

- Digital demography is potentially relevant for every area of population studies. Examples include
  1. How is online dating affecting household formation?
  2. How do people behave in the online marriage market? And how do they react to demographic imbalances and shocks?
  3. How are new technologies affecting intergenerational relationships?
  4. How does online exposure to peers affect health and fertility behavior?
  5. What do Google searches reveal about fertility or abortions?

Underlying themes

Underlying themes

1. How to access and make sense of new, messy and biased data sources?
2. To what extent traditional research design can be re-purposed to address new challenges?
3. What is the impact of new data and new tools on our society?

$\Rightarrow$ SOC 401: "Data Science and Population Processes", is offered in the Fall

CENTER FOR STUDIES IN DEMOGRAPHY & ECOLOGY
UNIVERSITY of WASHINGTON

RESEARCH / TRAINING / SERVICES / COMPUTING / PEOPLE / NEWS & EVENTS / ABOUT US

Center for Studies in Demography and Ecology

CSDE is a community of faculty and students associated to advance population science through research and training
As a federally funded research center with over 70 years of experience, the CSDE community of scholars develops new
demographic measures and methods, advances knowledge about population dynamics, generates new data and
evidence to support population science, and trains the next generation of demographers.

www.csde.washington.edu