UNIVERSITÄT BONN

Transdisciplinary Research Area
Innovation and Technology for
Sustainable Futures (TRA 6)

zef
Center for
Development Research
University of Bonn

WELCOME TO THE TRA 6 LECTURE SERIES INNOVATION PATHWAYS TO SUSTAINABILITY / ZEF COLLOQUIUM SERIES

DIGITAL DATA FOR MIGRATION RESEARCH

INGMAR WEBER, QATAR COMPUTING RESEARCH INSTITUTE (QCRI), HAMAD BIN KHALIFA UNIVERSITY

MODERATORS:
JAN BÖRNER, TRA 6-SPEAKER,
RIBANA ROSCHER, FACULTY OF AGRICULTURE,
UNIVERSITY OF BONN

Digital Data for Migration Research

Ingmar Weber

March 18, 2021

@ingmarweber

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Amazing Collaborators! (Alphab.)

Natalie Adler, Musa Al-Asad, Jakub Bijak, Carlos Callejo, Masoomali Fatehkia, Manuel Garcia, **Kiran Garimella**, Krishna Gummadi, Alfredo Morales, Fabrizio Natale, Joao Palotti, Francesco Rampazzo, Marzia Rango, Vedran Sekara, Elise Sonne, Spyridon Spyratos, Bogdan State, Michele Vespe, Jeffrey Villaveces, Agnese Vitali, **Emilio Zagheni**, …

How can we use non-traditional data for complementing and improving existing migration statistics?

More than numbers

How migration data can deliver real-life benefits for migrants and governments

IOM's Glo
Data Ana
Gl

ıpany
ERNMENT

Hope:
Better data => better outcomes

Development

- Creating transparency on remittance costs could **increase the money sent back to low- and middle-income countries by $15 billion–$20 billion.**
- Data-driven programs can help **promote the targeted return of 20% of high-skilled professionals working abroad.**

Journey

- Data-driven interventions can **double the number of human-trafficking cases identified. This could enable support to around 150,000 victims.**

Arrival

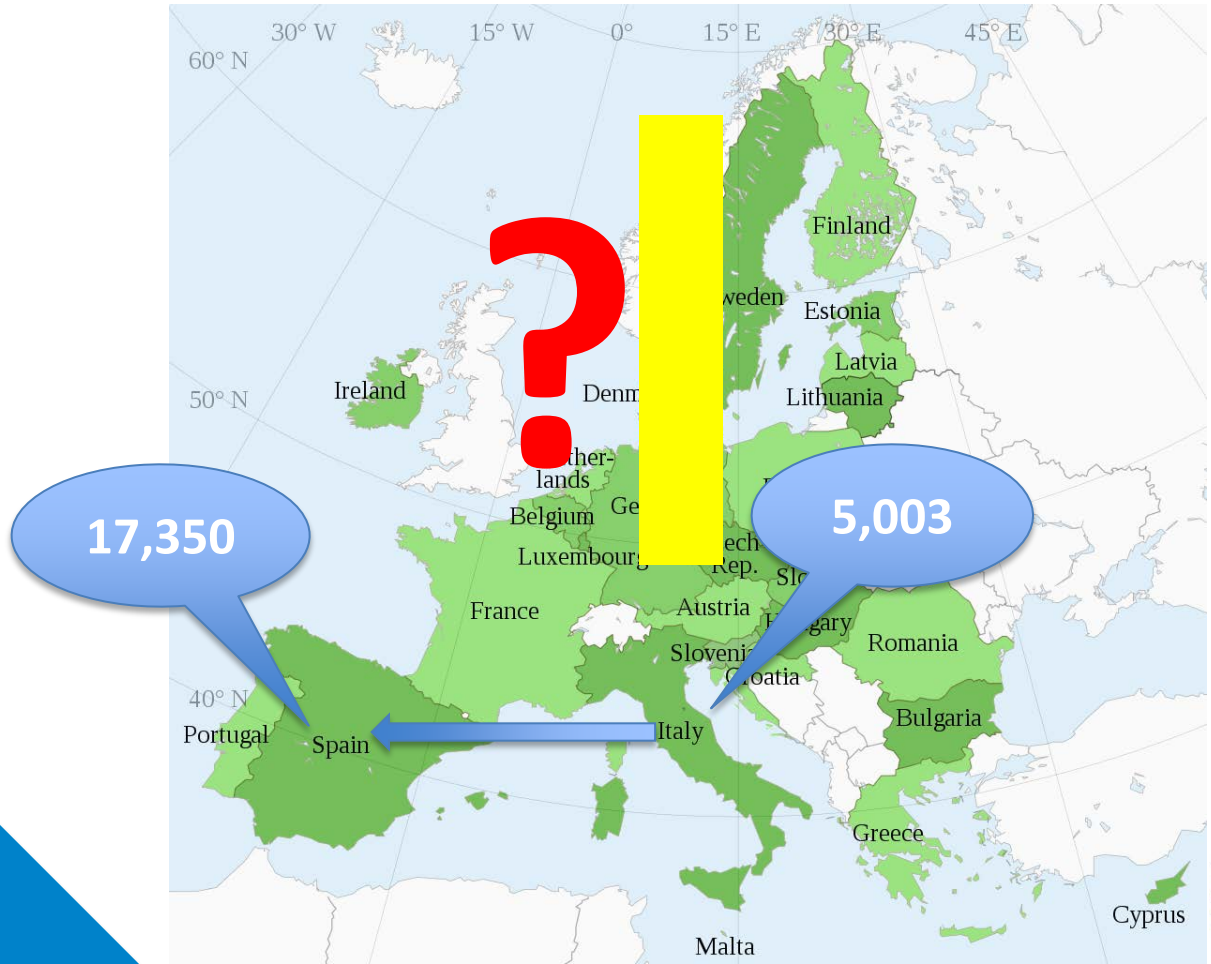
- Data can **enable a significant reduction in the number of duplicate registrations,** using centralized databases and modern technologies (eg, top-in-class fingerprint and facial-recognition systems have an accuracy of more than 99%).

Return and reintegration

- Using data to tailor assisted voluntary-return programs to migrants' needs could **increase the share of voluntary returns. A 10 percentage-point increase would have saved EU governments up to €350 million for the years 2008–16.**

What’s wrong with current data?

Question: how many people moved from Italy to Spain in 2015?



QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

The Guardian International edition

News | Opinion | Sport | Culture | Lifestyle

World ▸ Europe US Americas Asia Australia Middle East Africa Inequality Global development

Migration

This article is more than 1 year old

EU migration to UK underestimated by ONS, analysts say

Aug. 21, 2019

In a humiliating admission by the Office for National Statistics (ONS), its migration statistics quarterly report – which the UK government use to inform policy – has been reclassified as “experimental”.

The move comes after statisticians discovered earlier estimates of net migration between 2009 and 2016 from the EU8 countries - Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Slovakia, Slovenia – were too low.

https://www.theguardian.com/world/2019/aug/21/eu-migration-to-uk-underestimated-by-ons-analysts-say

Even for migration among rich countries data can be bad.

Even worse during humanitarian crises.

Talk Outline

- New Data, New Concepts
  - Google+: from migration corridors to migration clusters
  - Twitter: effect of definitions on results
- Knowing the Present
  - Yahoo IP addresses: measuring recent migration flows
  - Facebook: how many Venezuelan FB users are in Colombia
  - Satellite imagery: counting cars from space
- Predicting the Future
  - Google Trends: migrants searching for a better life
- The Importance of Goats
  - Beyond big data: context, context, context

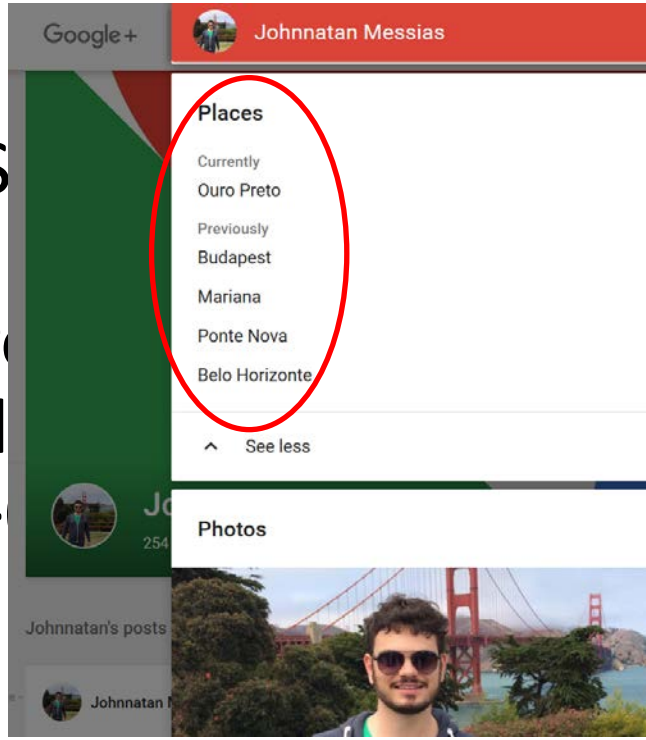
Exploring new concepts and definitions

NEW DATA, NEW CONCEPTS

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Beyond Origin-Destination Migration Analysis

I'm a German citizen living in Qatar. S
from Germany to Qatar?
Yes, according to Qatari border contr
But: Germany (78->99), United Kingd
Germany (03->07), Switzerland (07->
Spain (09->12), Qatar (12->now)



Use the "places lived" on Google+
Done in 2012, no "currently", just set of places
Get tuples of co-lived countries
Obtained for 22.6M users

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Are there "higher order" patterns of migration?

Scenario 1	Countries Lived In				Bilateral Flows
	A	B	C	D	
M1	x	x	x		(A,B), (A,C), (A,D), (B,C), (B,D), (C,D)
M2	x			x	
M3		x		x	
M4			x	x	

different

same

(A,B), (A,C), (A,D), (B,C), (B,D), (C,D)

Identifying “unexpected” clusters

- Given *large* users counts for all of (A,B), (A,C), and (B,C) we would expect a *large* count for (A,B,C)
- Given *small* users counts for all of (A,B), (A,C), and (B,C) we would expect a *small* count for (A,B,C)

Formalized expectation:

$$|ABC| \sim \min(|AB|,|AC|,|BC|) * \mathrm{avg}(|AB|,|AC|,|BC|)$$

Spearman rank correlation .75 with |ABC|

Which Triples Occur More/Less Than Expected

Description	IG Rank	IG Value	χ^2 Rank	χ^2 Value
Min Distance	1	0.231	1	984.742
Max Distance	2	0.180	3	767.547
Common Region	3	0.178	2	780.458
Avg Distance	4	0.173	4	745.858
Max GDP	5	0.102	5	474.392
Avg GDP	6	0.089	6	408.225
Min GDP	7	0.070	7	312.460
Common Civ.	8	0.033	8	147.838
Common Visa	9	0.017	9	80.004
Com. Col. Link	10	0.0001	10	0.679

Most discriminative features for 3-class distinction

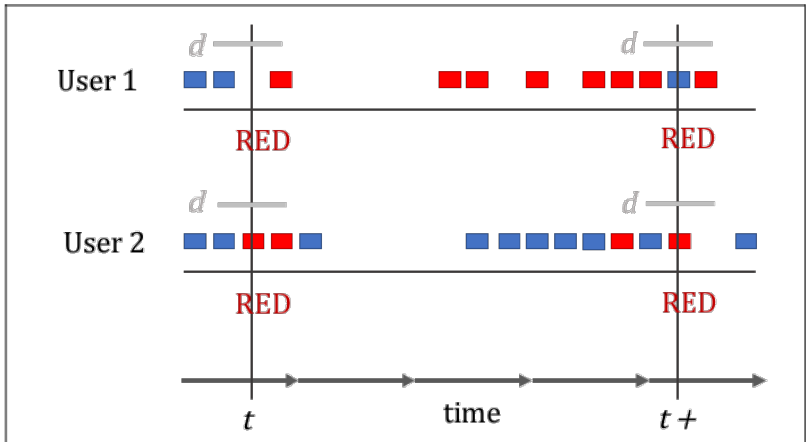
More than expected:

- (Spain, France, Italy)
- (UAE, India, Singapore)

Less than expected:

- (Brazil, Mexico, USA)
- (Canada, China, UK)

Duration and Interval



Your main residence for past X months?

duration

[redacted]

@[redacted]

Hier @ArminLaschet am Samstag habe ich Geburtstag und ich hätte gerne als Geschenk eine Impfung.

Danke

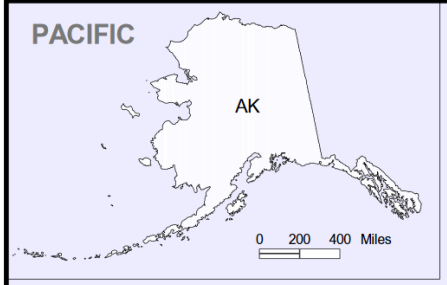
Translate Tweet

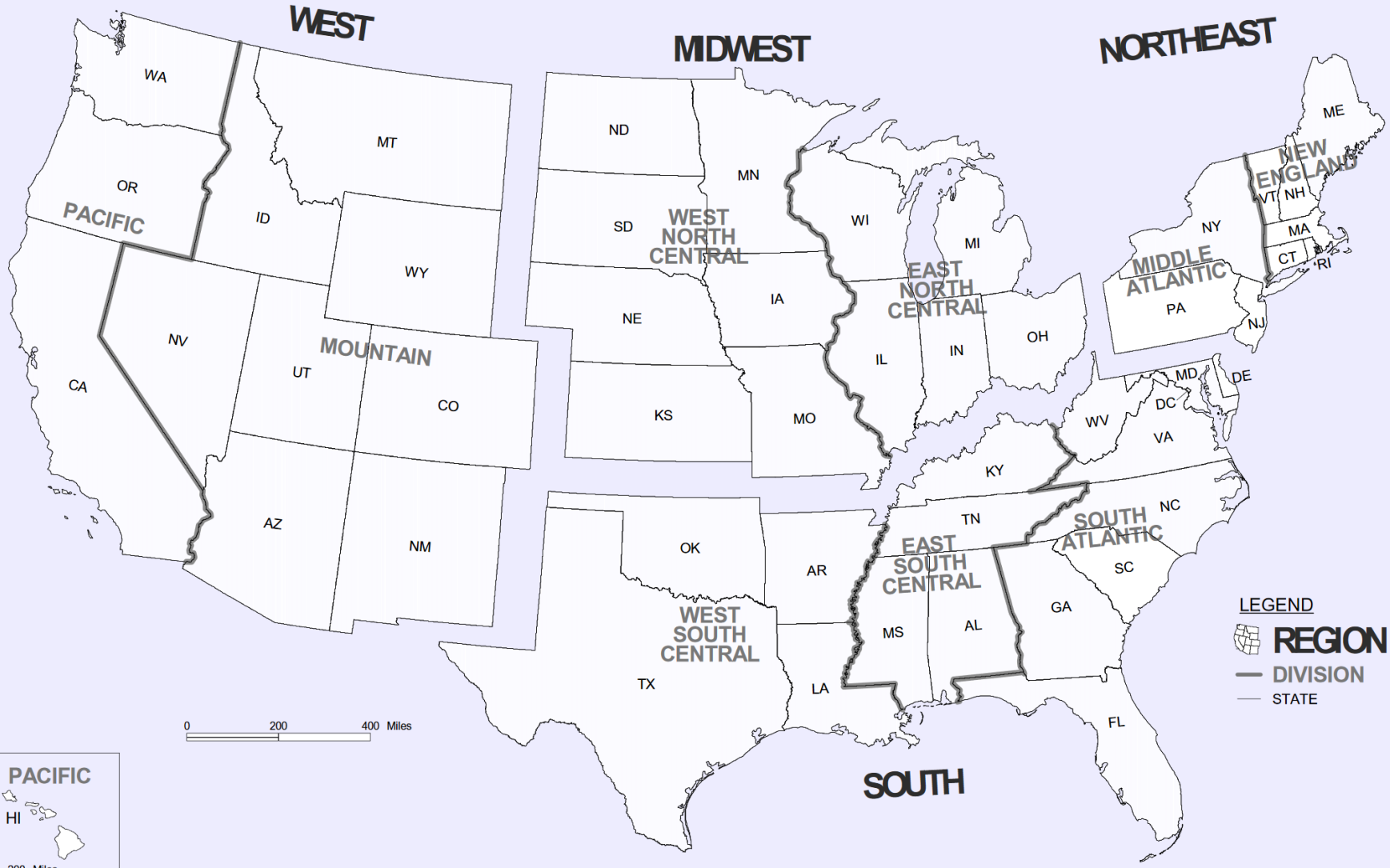
6:03 PM · Mar 16, 2021 from Bonn, Germany · Twitter for Android

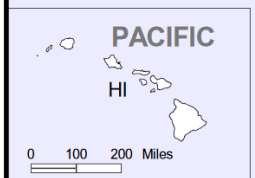
- Collected user tweet histories for ~62k users with at least some tweets in the US
- 15.3 million geo-tagged tweets
- Mean of 250 tweets per user
- Migration between 9 US census divisions

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Census Regions and Divisions of the United States



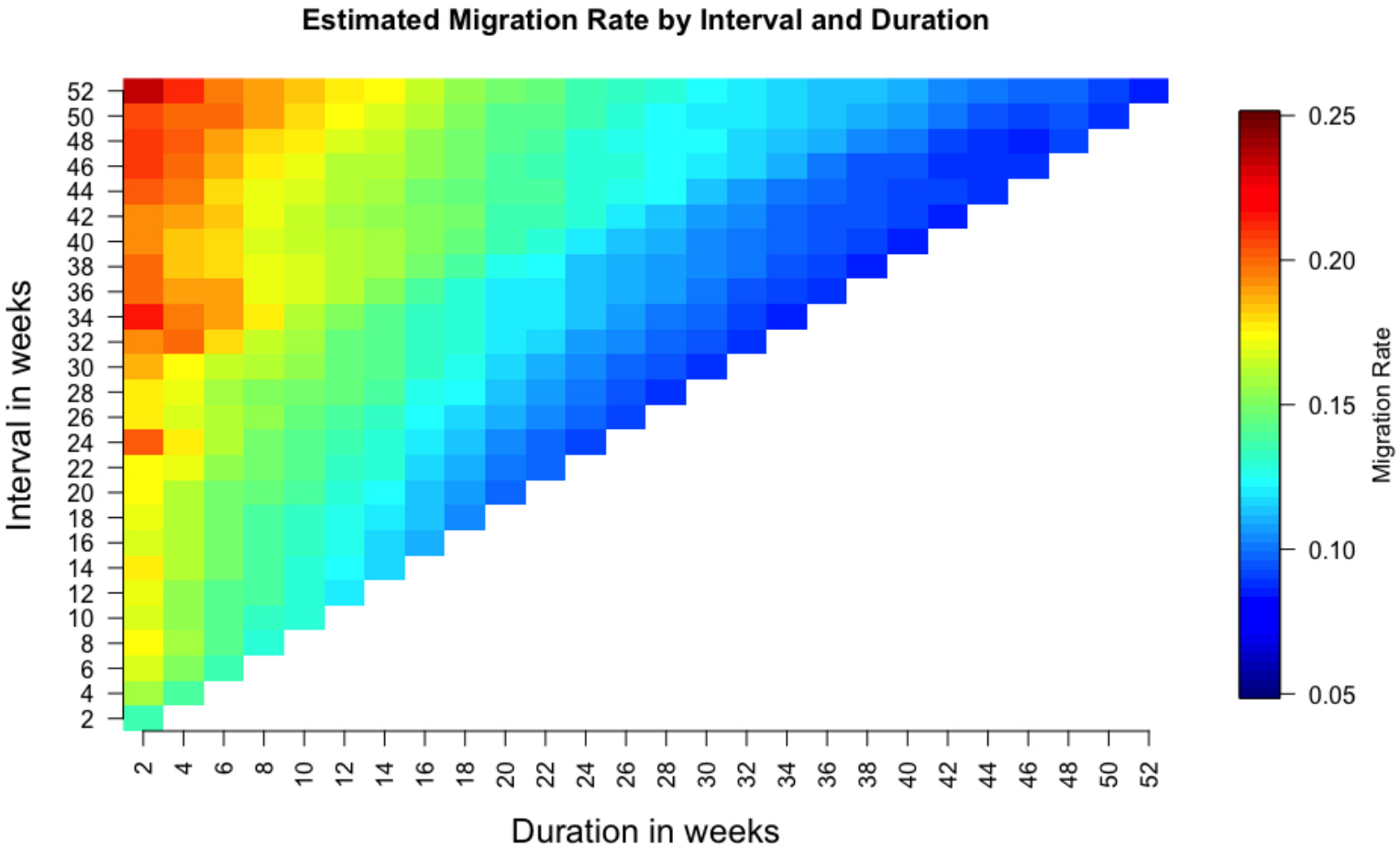




U.S. Department of Commerce Economics and Statistics Administration U.S. Census Bureau

Prepared by the Geography Division

The Duration-Interval Interplay



Plot of estimated migration rate as a function of interval and duration length. Rates were estimated fixing July 1st 2012 as the starting point.

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

What’s happening?

KNOWING THE PRESENT

YAHOO!®

How many people emigrated from the US in 2020?

Emigration from the United States

From Wikipedia, the free encyclopedia

Emigration from the United States is the process where individuals from the United States move to live in other countries, creating an **American diaspora** (**overseas Americans**). The process is the reverse of the immigration to the United States. The United States does not keep track of emigration, and counts of Americans abroad are thus only available based on statistics kept by the destination countries.

American diaspora

Total population
9,000,000[1] (2016, est.)

Contents [hide]

IP Address => Approximate Geolocation

GeoIP2 City Results

IP Address	Country Code	Location	Network	Postal Code	Approximate Coordinates*	Accuracy Radius (km)	ISP	Organization
78.100.40.90	QA	Doha, Baladiyat ad Dawhah, Qatar, Asia	78.100.40.0/24		25.2925, 51.5321	1000 ~10km	Ooredoo	Ooredoo

https://www.maxmind.com/en/geoip-demo

- Any online service you frequently use knows your coarse-grained mobility pattern
- We used anonymized data from Yahoo

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Data Collection

- Large sample of anonymized Yahoo email meta data, including self-reported birth year and gender
- Sent email between September 2009 and June 2011, at least once a month
- 43 million users, half from the US
- Migration: different modal country for [Sep 2009, Jun 2010] and [Jul 2010, Jun 2011]
- Also obtained internet penetration for (country, age, gender) group
- And migration data for European countries from Eurostat (for calibration)

Internet => Young & Educated => More Mobile

- Expect a particular type of selection bias:
- *Highly mobile people are early adopters for internet (and email) use*
- Introduce an ad-hoc correction factor (CF)

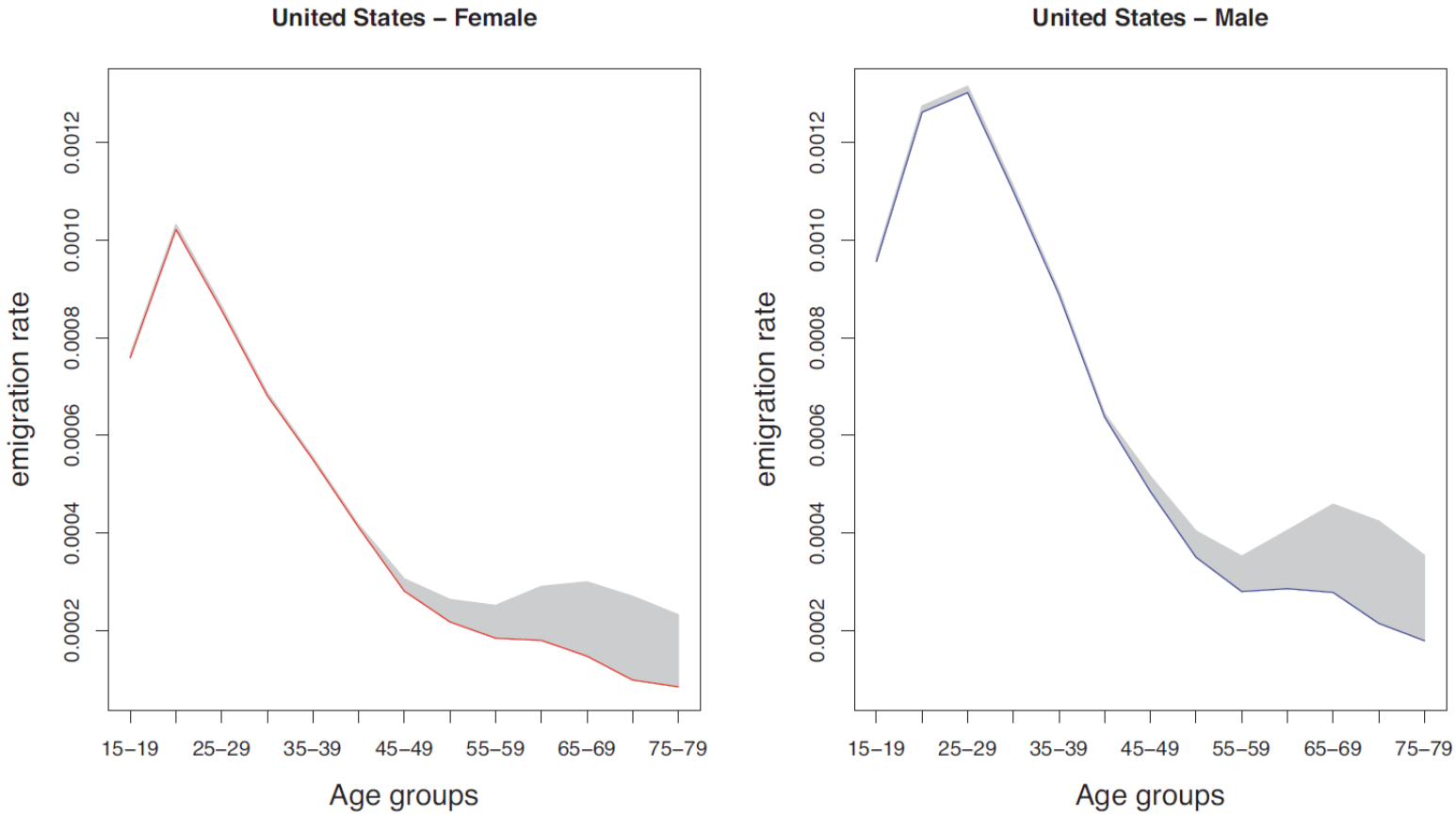
$$CF = \frac{p_{gac}(e^{-k} - 1)}{(e^{-kp_{gac}} - 1)}$$

100% internet penetration => *CF* = 1.0

Low internet penetration => *CF* < 1.0

Tune *k* using data from Europe

Results for the United States



Red line: after applying correction factor. Top of gray area: estimates from raw data.

for Business

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

How many Venezuelans are living in Colombia?

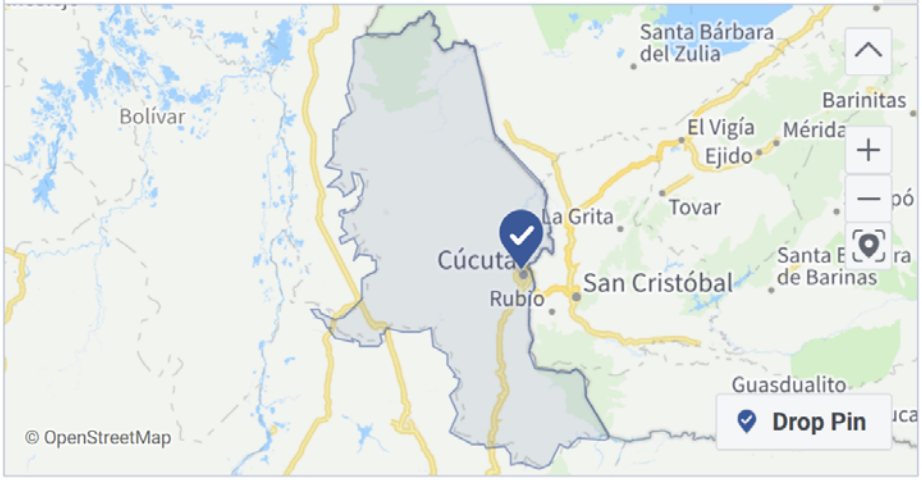
Ads Manager

Search Busi

Locations — People who live in this location

Colombia

Norte de Santander



Add locations in bulk

Age — 13 - 65+

Gender — All | Men | Women

Detailed targeting — Include people who match

Behaviours > Ex-pats

Lived in Venezuela (formerly Expats – Venezuela)

Audience size

Your audience selection is fairly broad.

Potential reach: 98,000 people

Estimated daily results

Reach

9.5K-27K

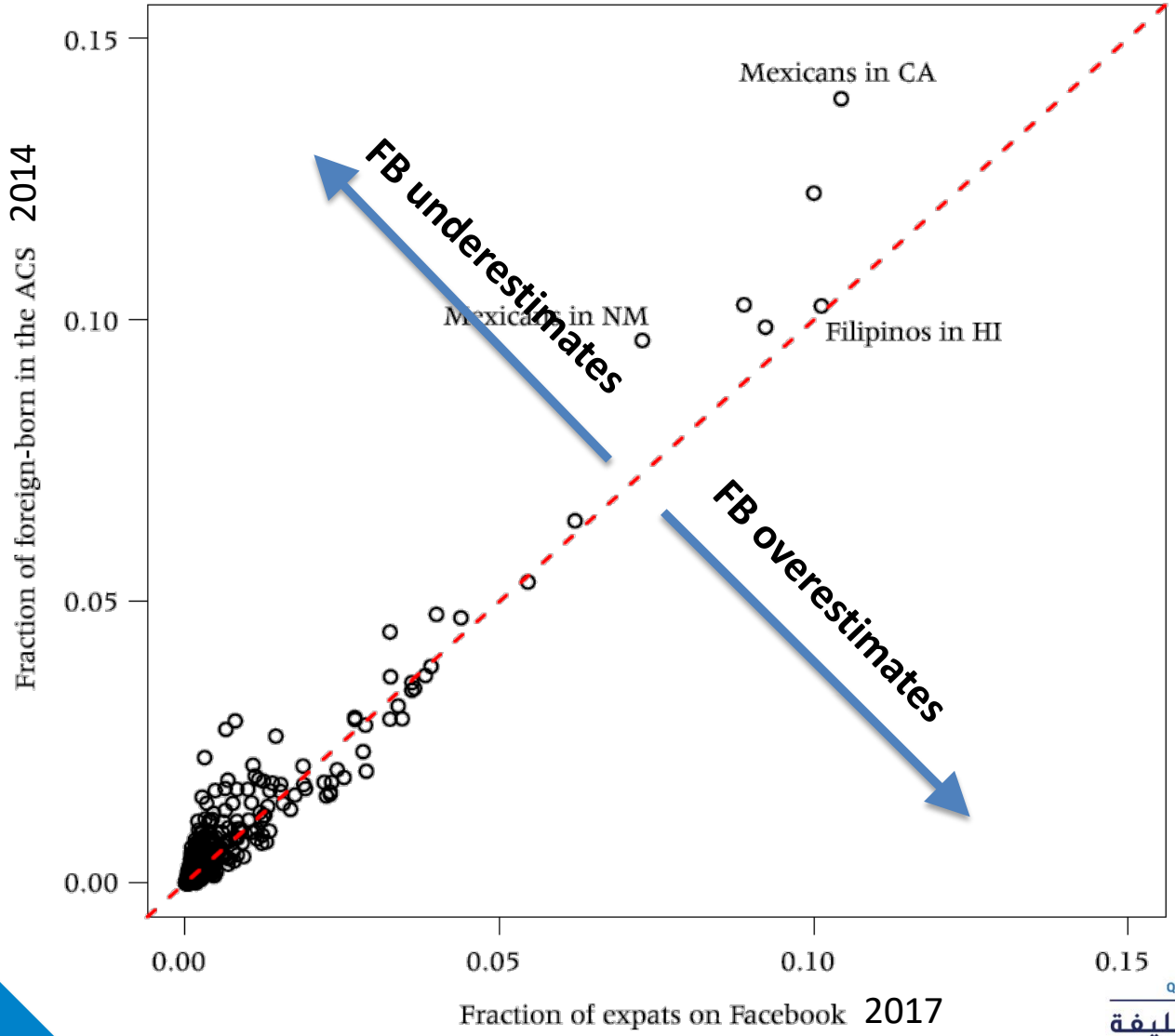
Link clicks

71-206

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 guarantee results.

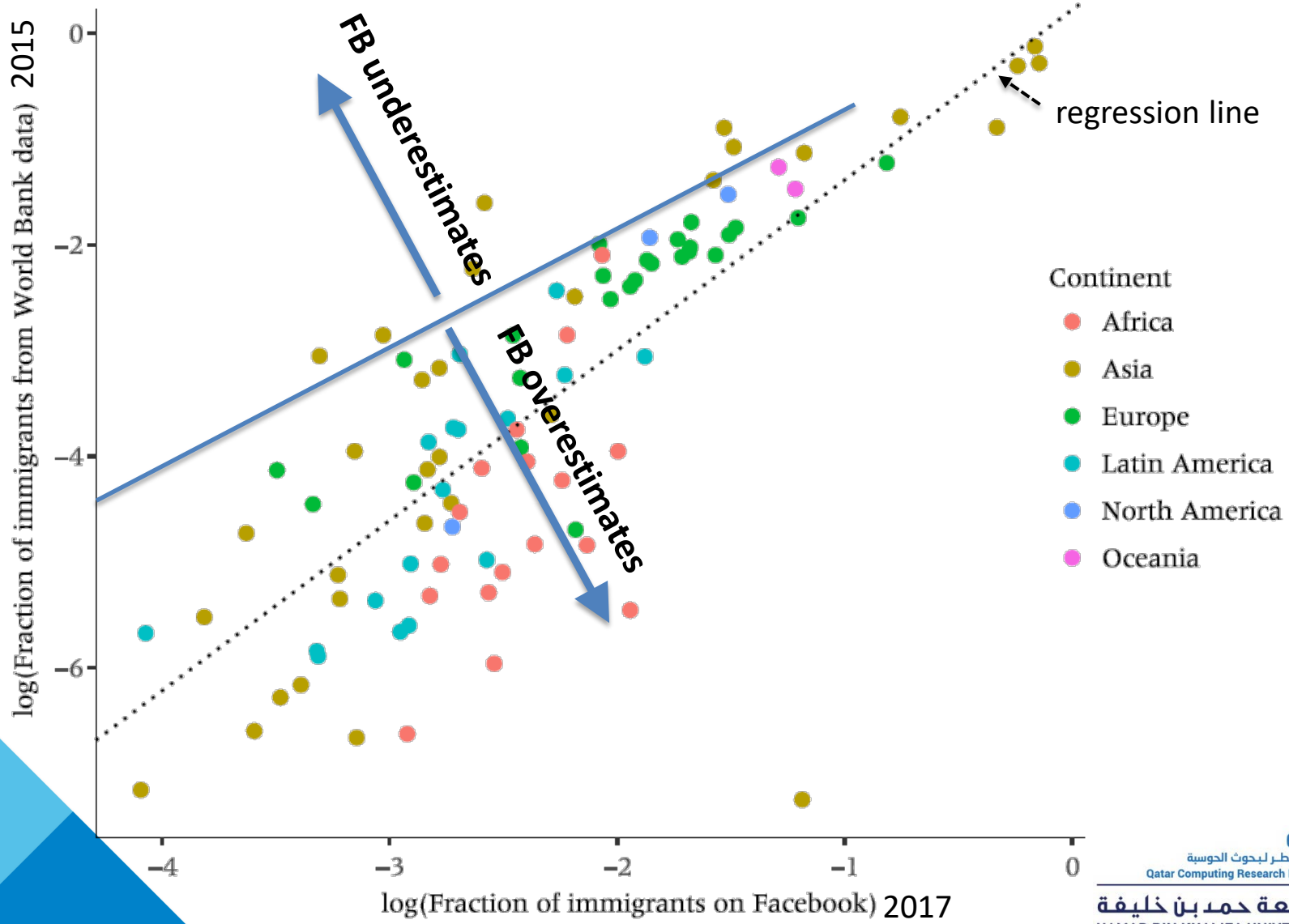
Were these estimates helpful?

Expats Across US States



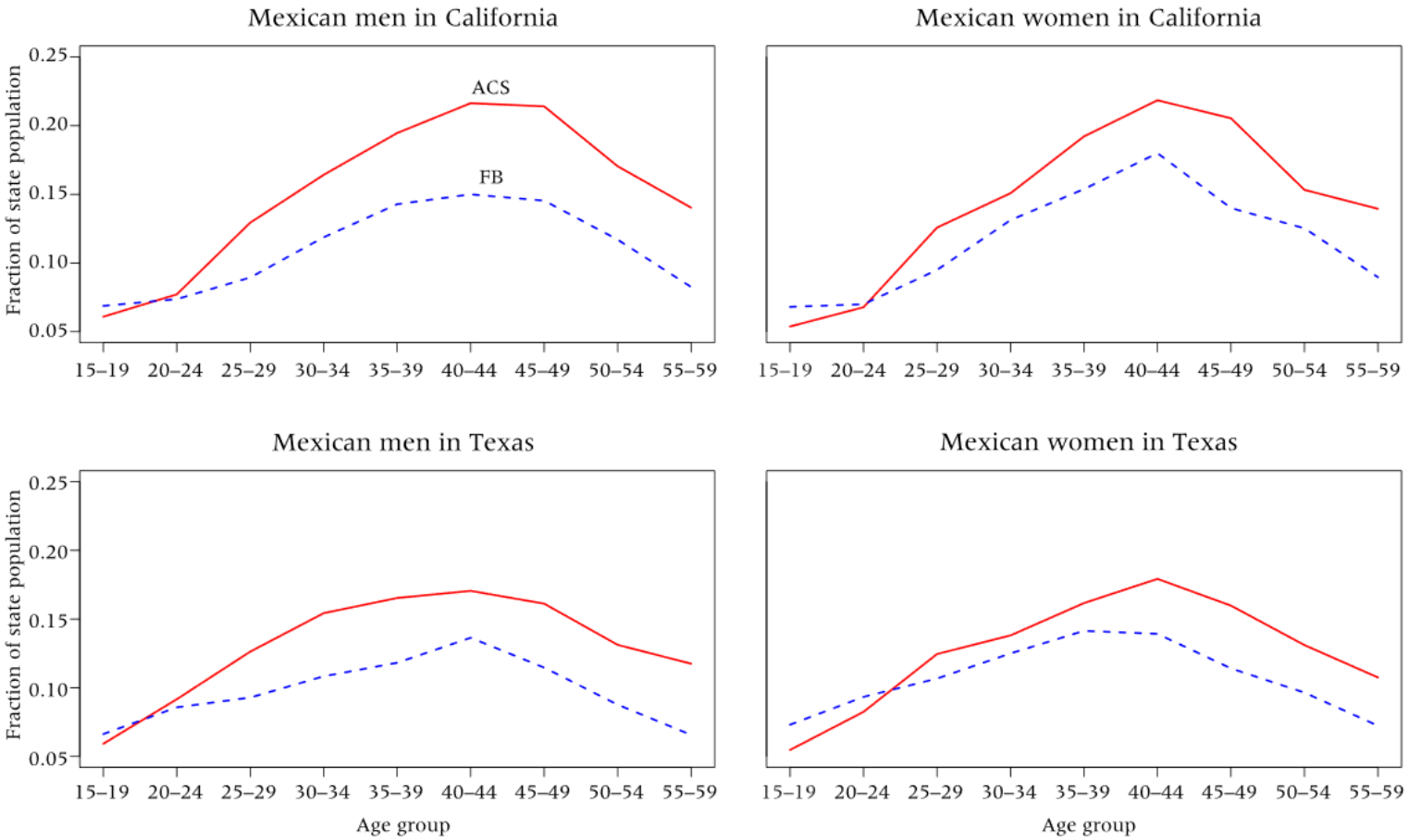
QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Expats Across Countries



QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Age-Specific Selection Biases



SOURCES: American Community Survey (ACS 2014); Facebook Adverts Manager.

Bias Reduction via Model-Fitting

z = age-gender group

$$log(\text{ACS foreign-born pop}_{ij}^{z})$$

i = country of birth

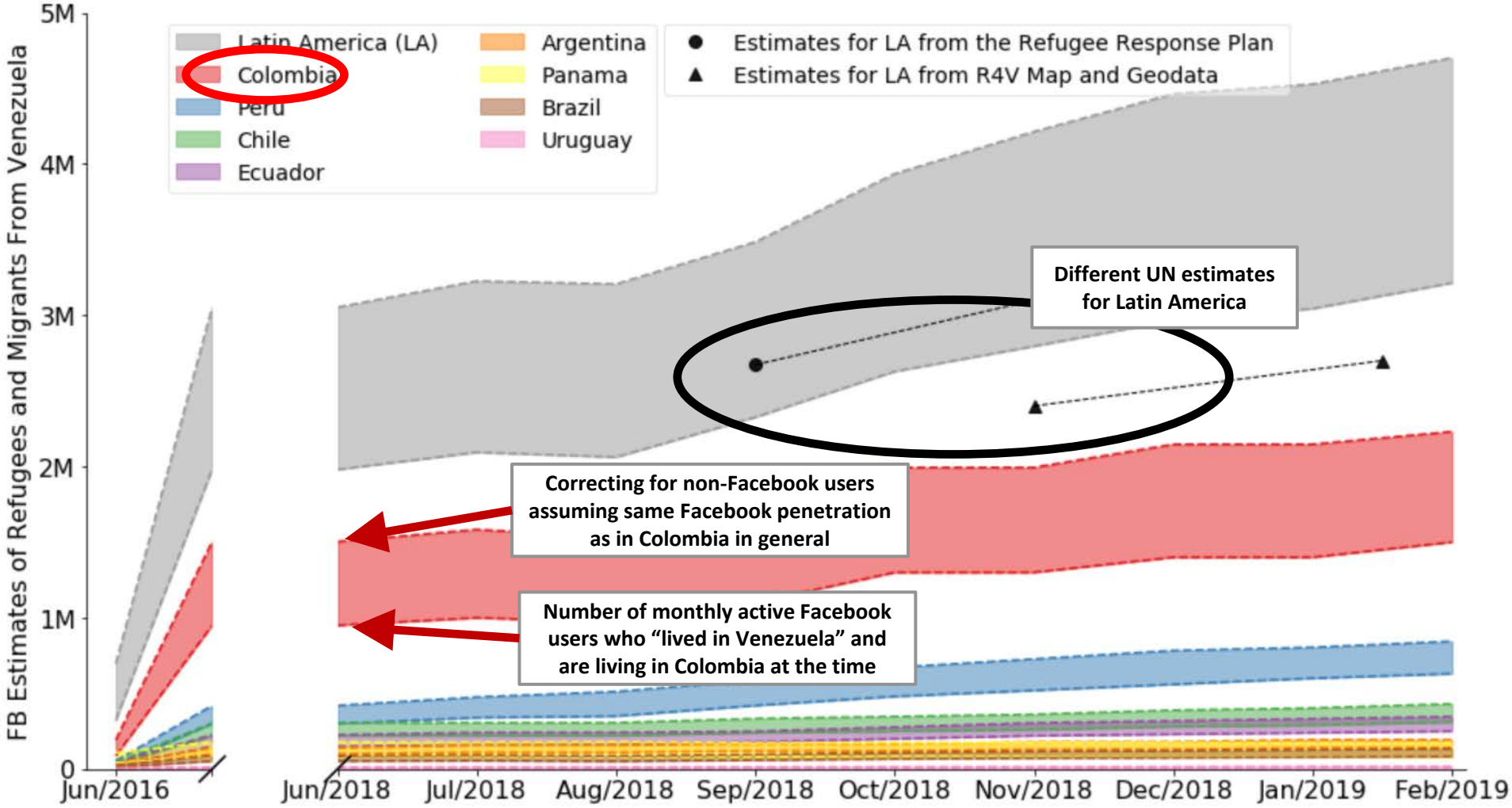
j = US state of residence

Mean out-of-sample absolute percentage error 37%, down from 56% without origin-age bias correction

Adjusted R^2 = .70

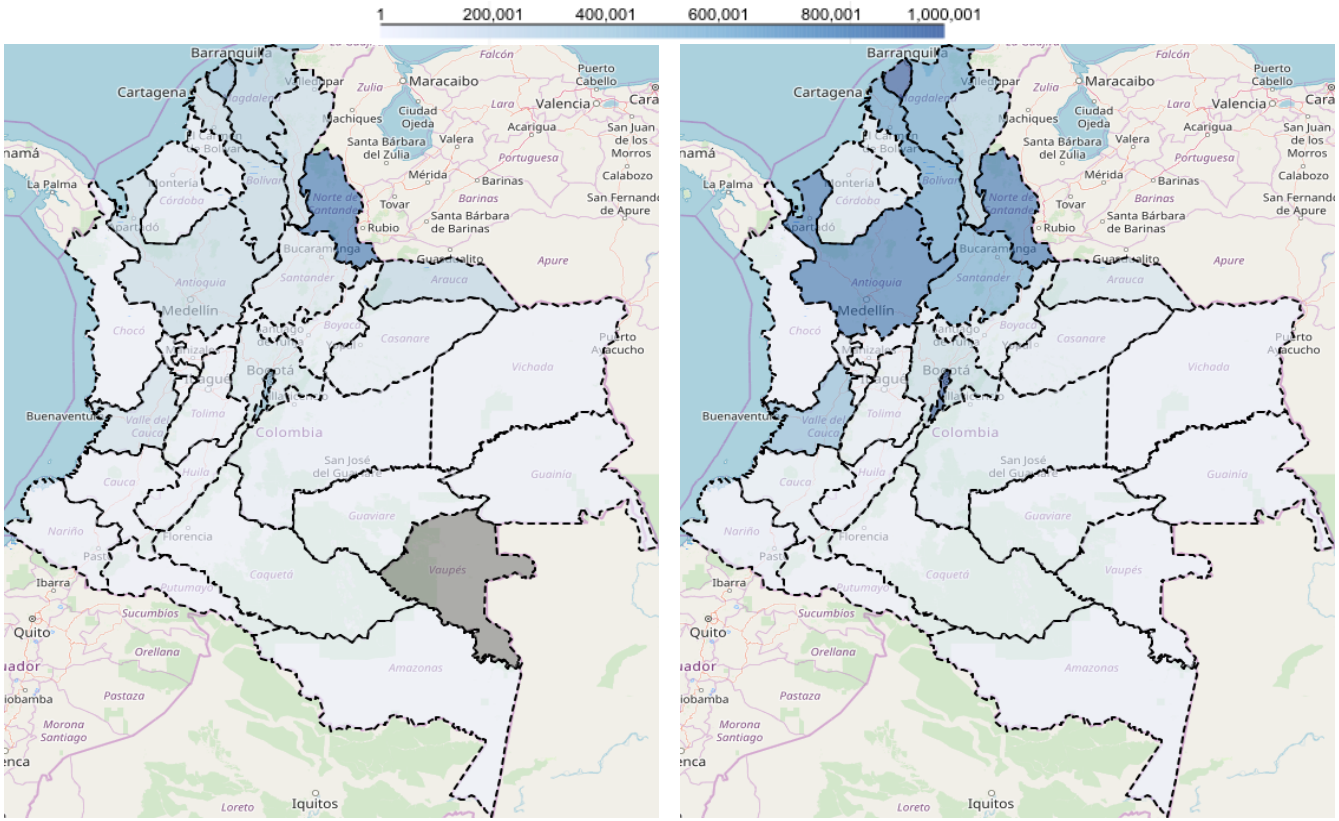
Does not use GDP, language, internet penetration, …

Venezuelan Exodus



QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Validation w/ (Few) Available Data



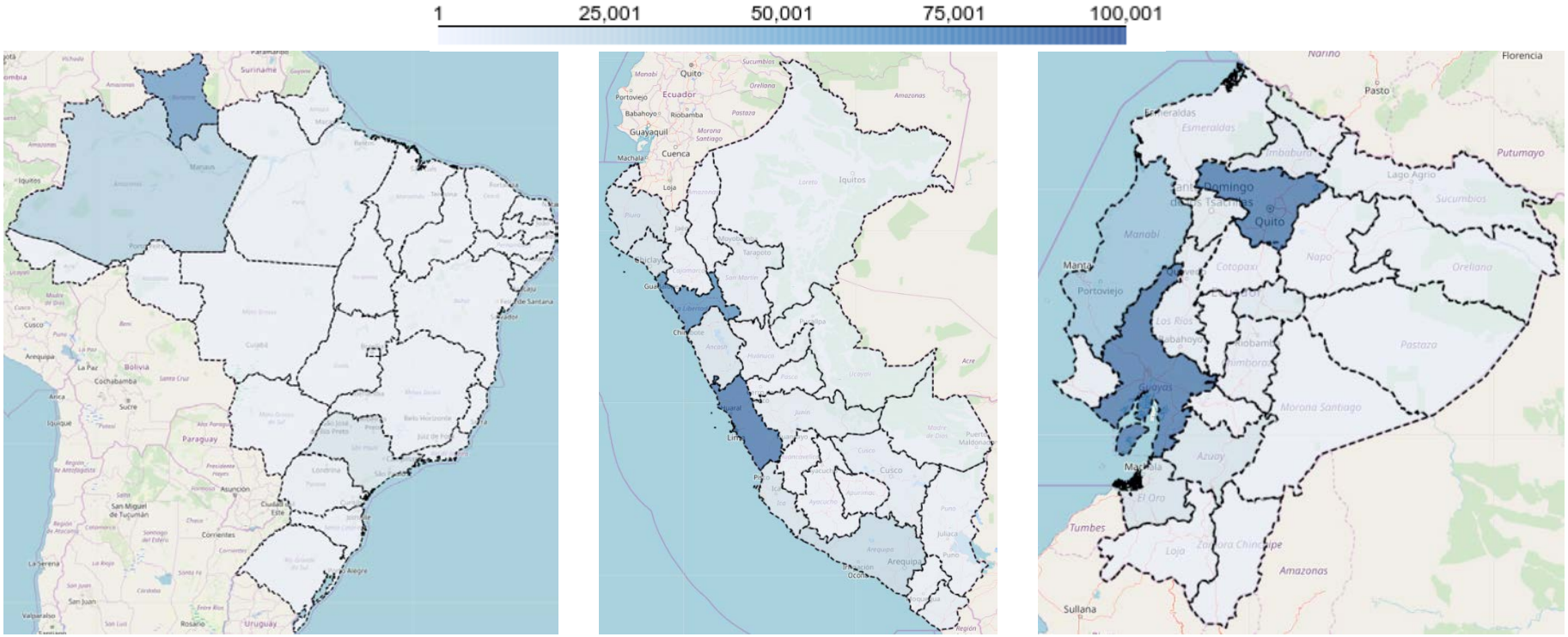
Registro de Administrativo de Migrantes Venezolanos (RAMV) - Jun, 2018

Facebook - Jun, 2018

Kendall's τ = .71 (n=31)

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Previously Unavailable Estimates



Brazil - Facebook. Feb 2019 Peru - Facebook. Feb 2019 Ecuador - Facebook. Feb 2019

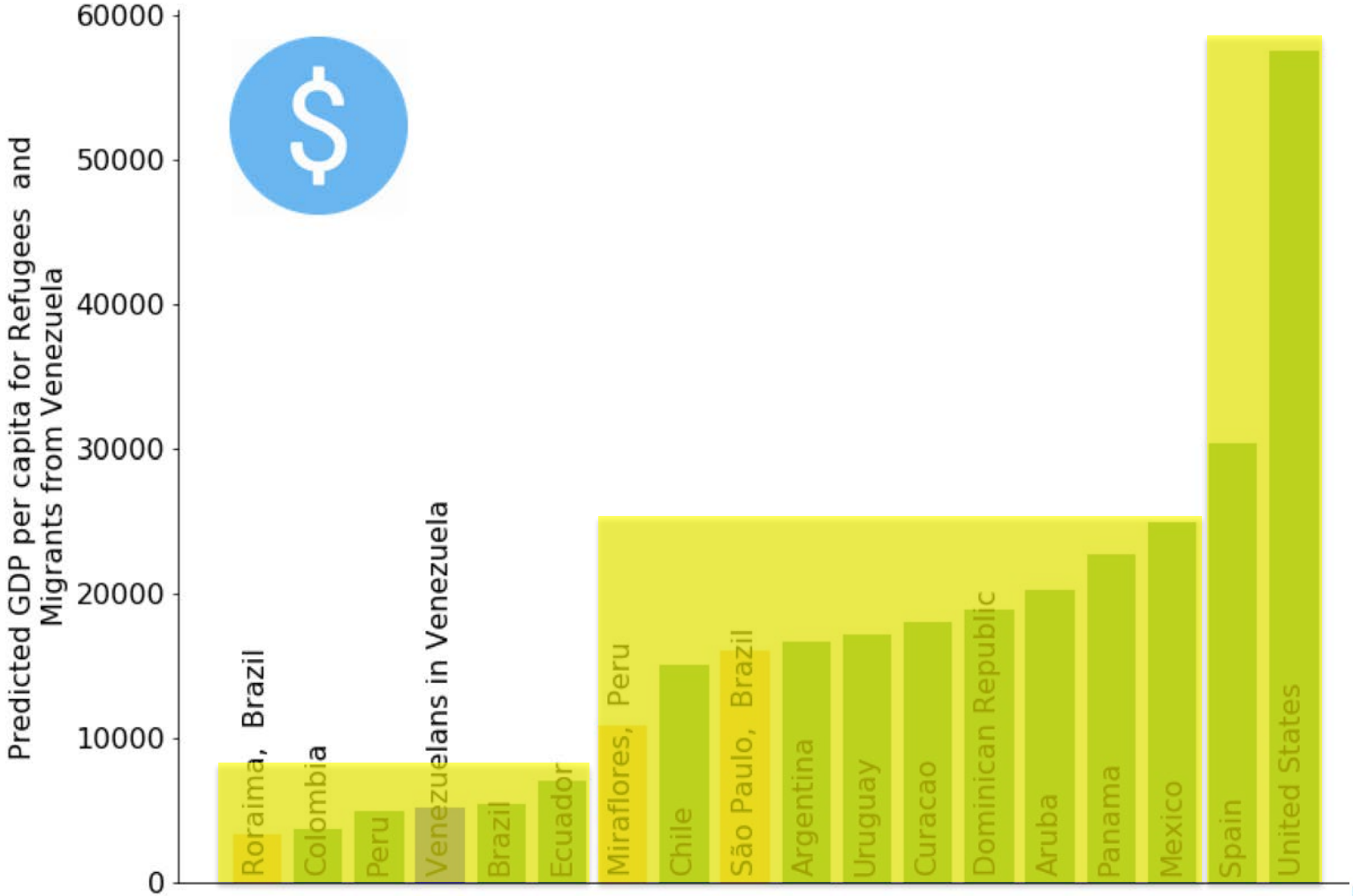
Predicted Income Based on OS

We fit a regression model to predict country-level GDP across host countries (n=15) using the percentage of FB users using iOS devices (adj. R^2 = .87).

=> “Given that X% of FB users have an iOS device, I would expect a per-capita GDP of Y USD per year.”

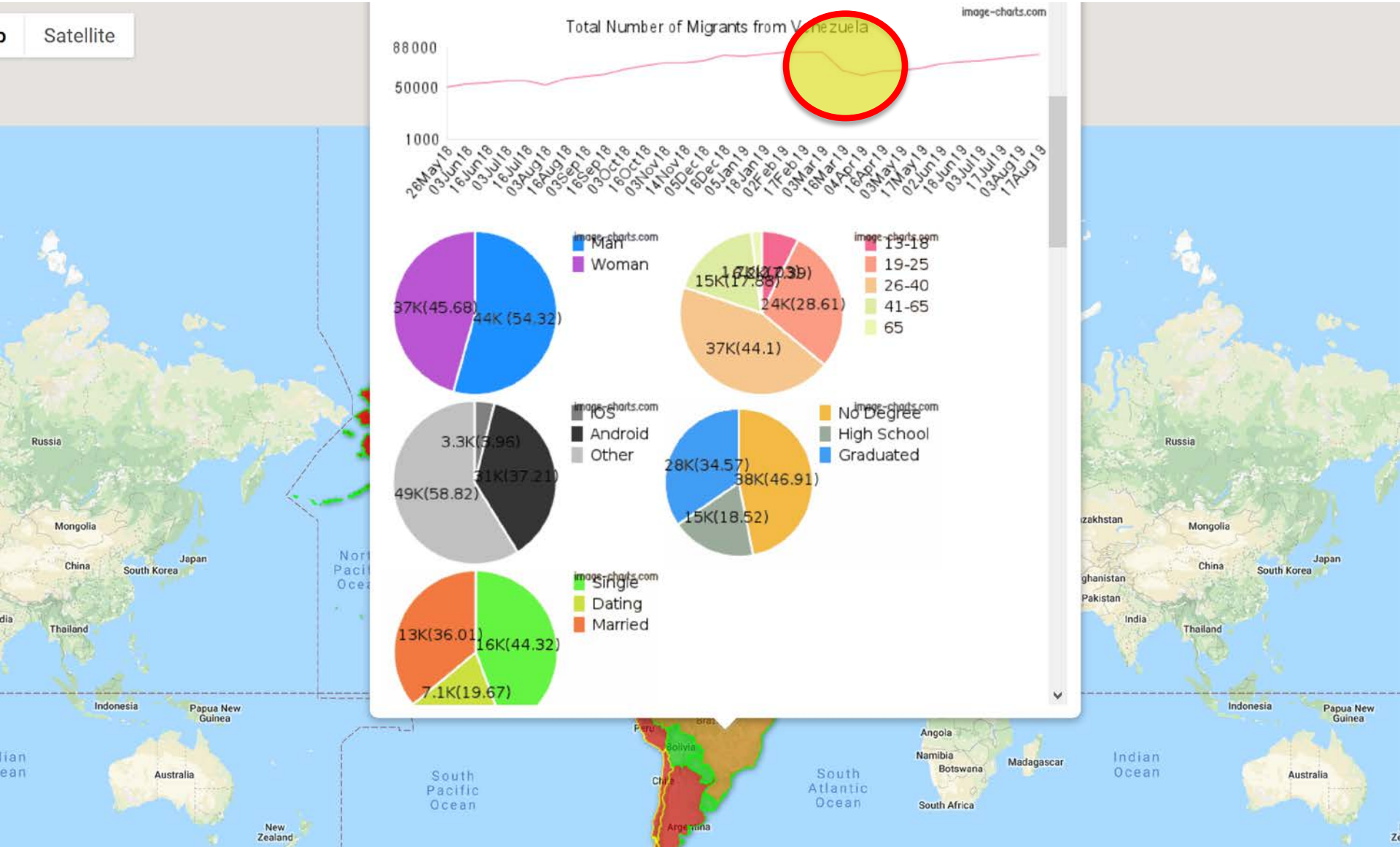
Apply this model to the Venezuelan migrants and refugees in different host countries.

Predicted Income Based on OS



معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Changes to Facebook's Backend



Operational Impact

USAID DEL PUEBLO DE LOS ESTADOS UNIDOS DE AMÉRICA | iMMAP

Detección de Usuarios venezolanos conectados en Facebook

Municipal | Departamental | Regional | Perú | Ecuador | Tendencia

IMMAP localiza a los migrantes venezolanos en América Latina a través del uso de el api de Facebook advertising data mostrando las conexiones de usuarios que antes vivian en Venezuela y ahora viven en el extranjero.

Total usuarios conectados	Dato Oficial Migración Colombia
1,845,200	**1,729,537**
Usuarios conectados Febrero 15	Venezolanos en Colombia

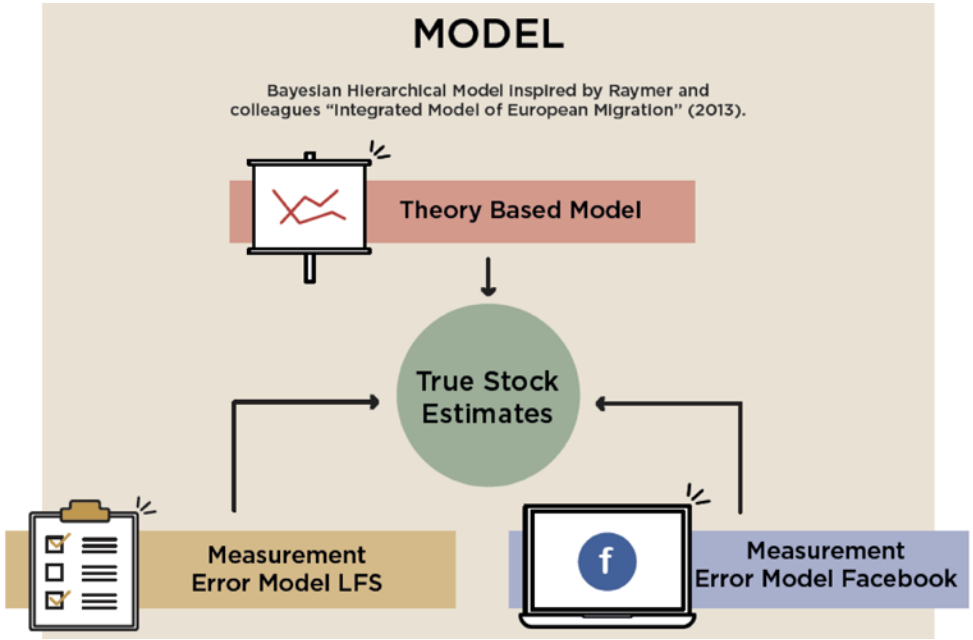
- Los datos son recolectados cada 15 días, mediante el API de mercadeo de Facebook. Los estimados mostrados estan basados en los filtros utilizados y en el comportamiento de los usuarios en los últimos 30 días. Obtén mas información sobre el alcance potencial
- Los estimados presentados no están diseñados para coincidir con censos u otras fuentes oficiales. Facebook no provee censos digitales o conteos de migrantes y/o refugiados. Estos estimados deben ser vistos como una señal par ser utilizada en triangulación.
- Facebook solo provee la definición del comportamiento seleccionado (Expatriados Venezuela). No provee datos estadísticos ni históricos
- El comportamiento depende de la información proporcionada por el usuario en Facebook, su ciudad actual y ciudad de origen y la estructura de la red de amigos (por ejemplo, tener al menos dos amigos de Facebook en el país de origen y dos amigos de Facebook en el país de destino). Leer más: Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants, ZAGHENI, Emilio. WEBER, Ingmar. GUMMADI, Krishna

Importante: En ningún momento durante la recolección de datos se accede a información privada de usuarios de Facebook.
Los datos no representan conteo de migrantes o refugiados.

USAID DEL PUEBLO DE LOS ESTADOS UNIDOS DE AMÉRICA | iMMAP

Fuente de datos: Facebook API Marketing.
Calculos: IMMAP contacto@immap.org

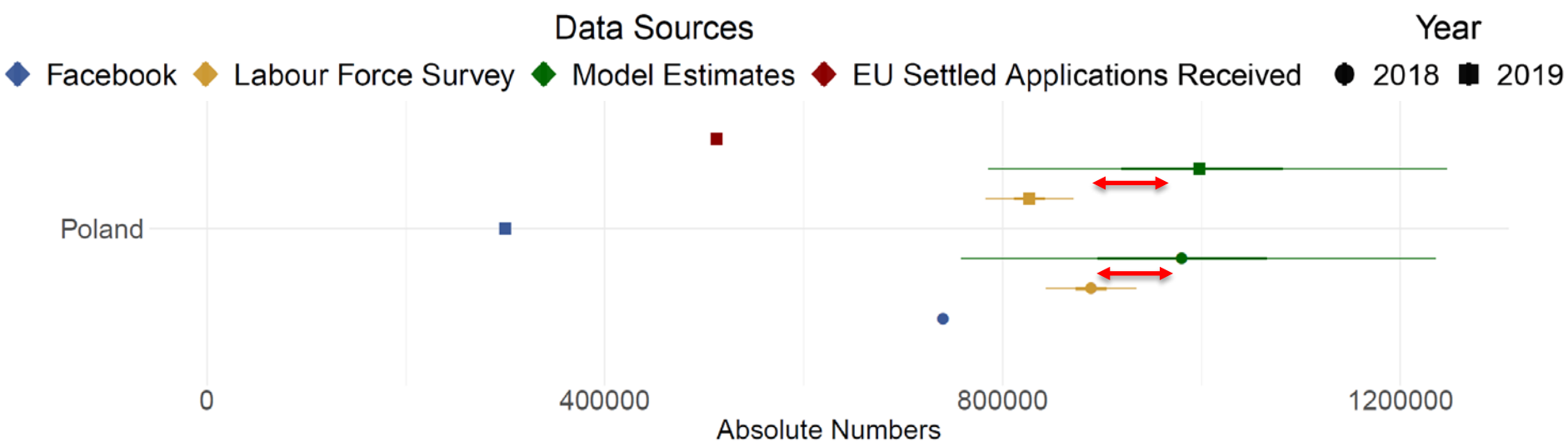
Combining Different Data Sources

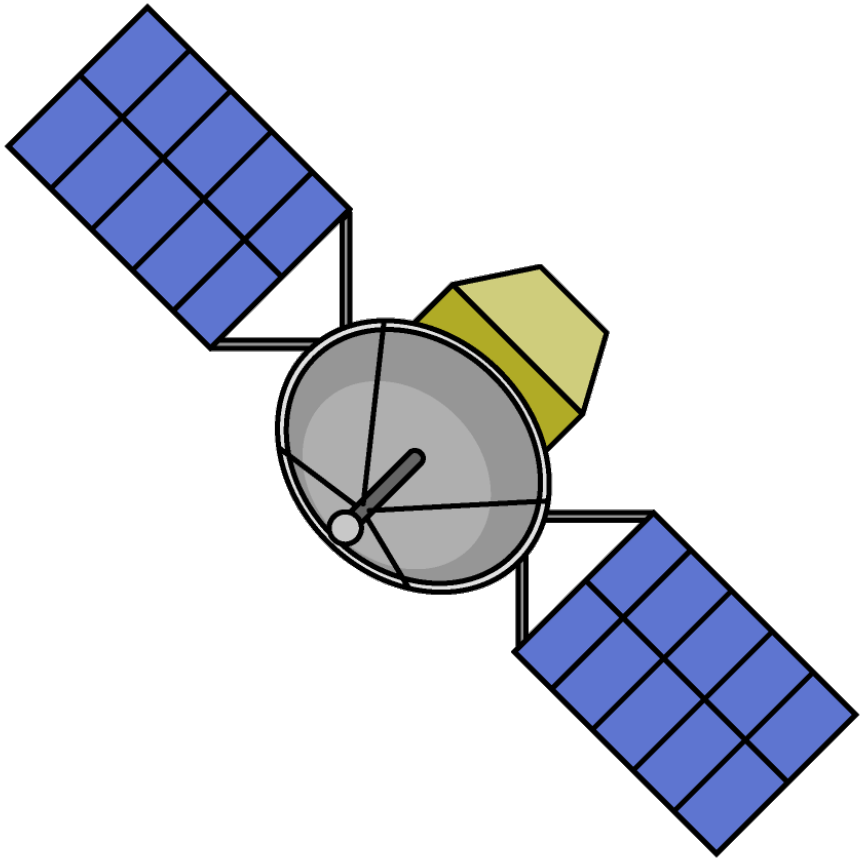


Give up on "gold standard" data

Make certain model and bias assumptions

Estimate which true data would be most likely to satisfy those assumptions and generate the observed data





Are traces of migration visible from space?

Can we see this from space?

https://www.nytimes.com/interactive/2020/02/14/world/middleeast/syria-idlib-displaced.html

Does Movement Happen by Vehicle?

"Transport: most IDPs made the first move in vehicles while others walked. For vehicles, sometimes they had their own motorcycles or cars, or family members had space in their vehicles."

No Where Safe to Go: A Study of Internally Displaced Population Movement for the Affected Population in Syria
NGO Forum - Humanitarian Trends Analysis Unit, December 2015

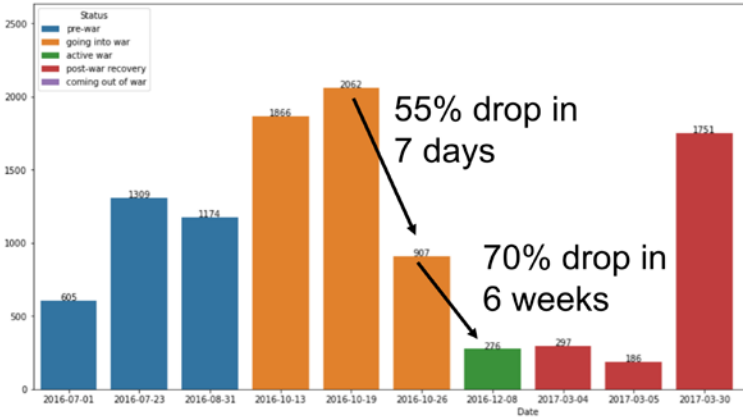
Very different from the "Caminantes" in Venezuela/Colombia


9/2016
2019
Before fighting
Al Bab
Image © 2021 CNES / Airbus
Imagery Date: 9/13/2016 36°22'23.36" N 37°30'54.07" E elev 0 ft


2/2017
2019
During fighting
Al Bab
Image © 2021 CNES / Airbus
Imagery Date: 2/16/2017 36°22'23.33" N 37°30'54.13" E elev


11/2017
2019
After fighting
Al Bab
Image © 2021 Maxar Technologies
Imagery Date: 11/9/2017 36°22'19.87" N 37°30'58.11" E elev 0 ft

Automated Car Detection



Lots of biases that need fixing
- nadir angle, urban vs. rural, day-of-week
Build a meta model on top
- Count of cars -> displacement estimate

Uses model from "Measuring Human and Economic Activity from Satellite Imagery to Support City-Scale Decision-Making during COVID-19 Pandemic", IEEE TBDATA

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Are people getting ready to leave their homes?

PREDICTING THE FUTURE

Google Trends

Are people searching on Google for a better life?

Google Trends

Journal of Development Economics 142 (2020) 102347

Contents lists available at ScienceDirect

Journal of Development Economics

journal homepage: www.elsevier.com/locate/devec

Searching for a better life: Predicting international migration with online search keywords☆

Marcus H. Böhme[a,1], André Gröger[b,c,*], Tobias Stöhr[d,e]

[a] German Federal Ministry of Finance, Germany
[b] Universitat Autònoma de Barcelona (UAB), Spain
[c] Barcelona Graduate School of Economics (BGSE), Spain
[d] Kiel Institute for the World Economy (IfW), Germany
[e] IZA, Germany

ARTICLE INFO

JEL classification:
F22
C82
J61

ABSTRACT

Migration data remains scarce, particularly in the context of developing countries. We demonstrate how geo-referenced online search data can be used to measure migration intentions in origin countries and to predict bilateral migration flows. Our approach provides strong additional predictive power for international migration flows when compared to reference models from the migration and trade literature. We provide evidence, based

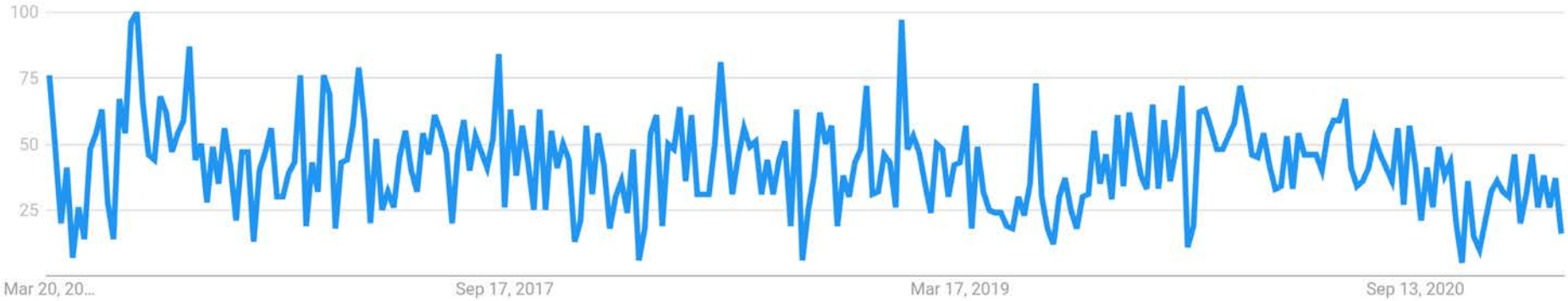
Google Trends | Explore

jobs germany
Search term

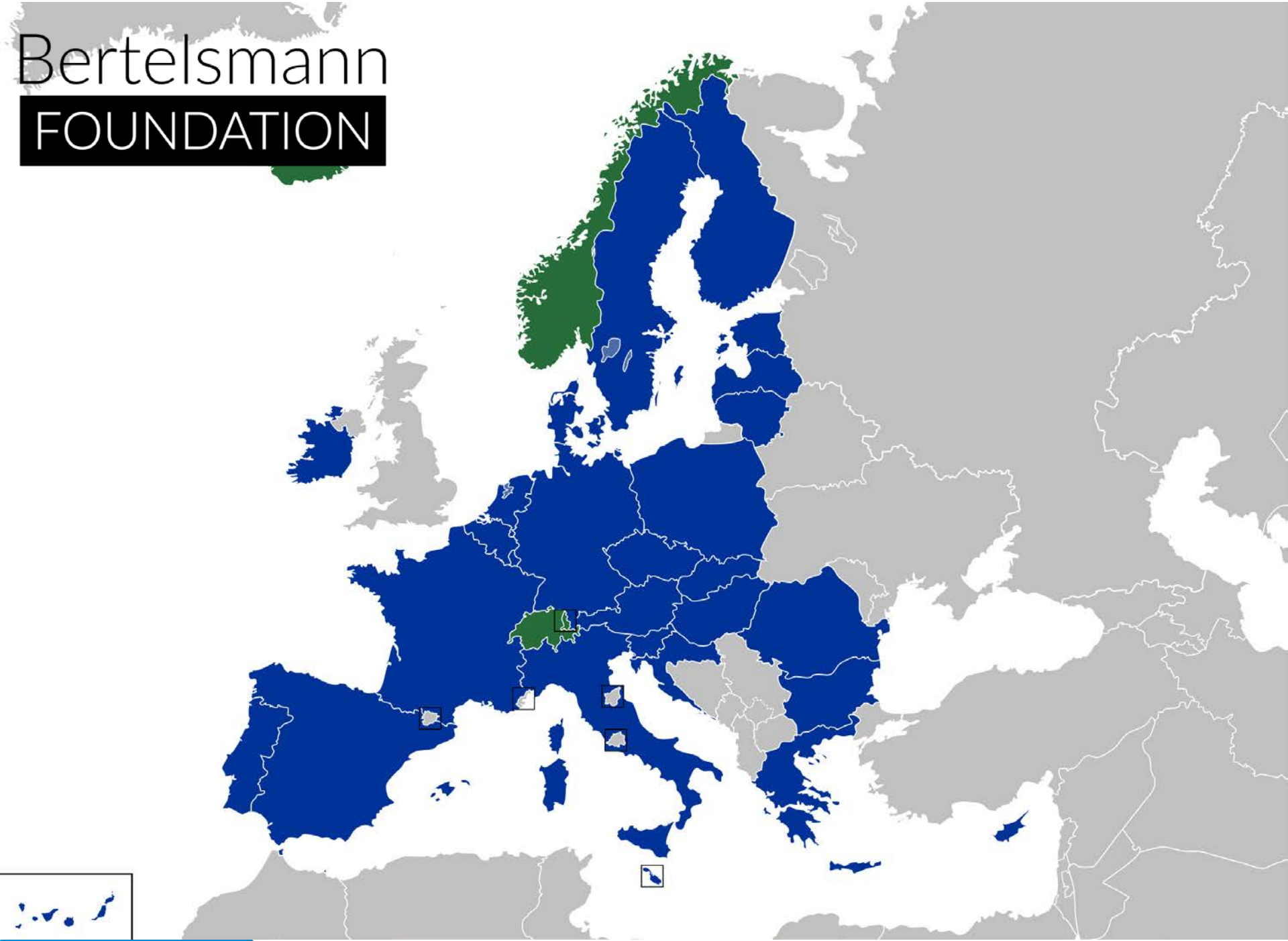
+ Compare

United Kingdom | Past 5 years | All categories | Web Search

Interest over time



Bertelsmann
FOUNDATION

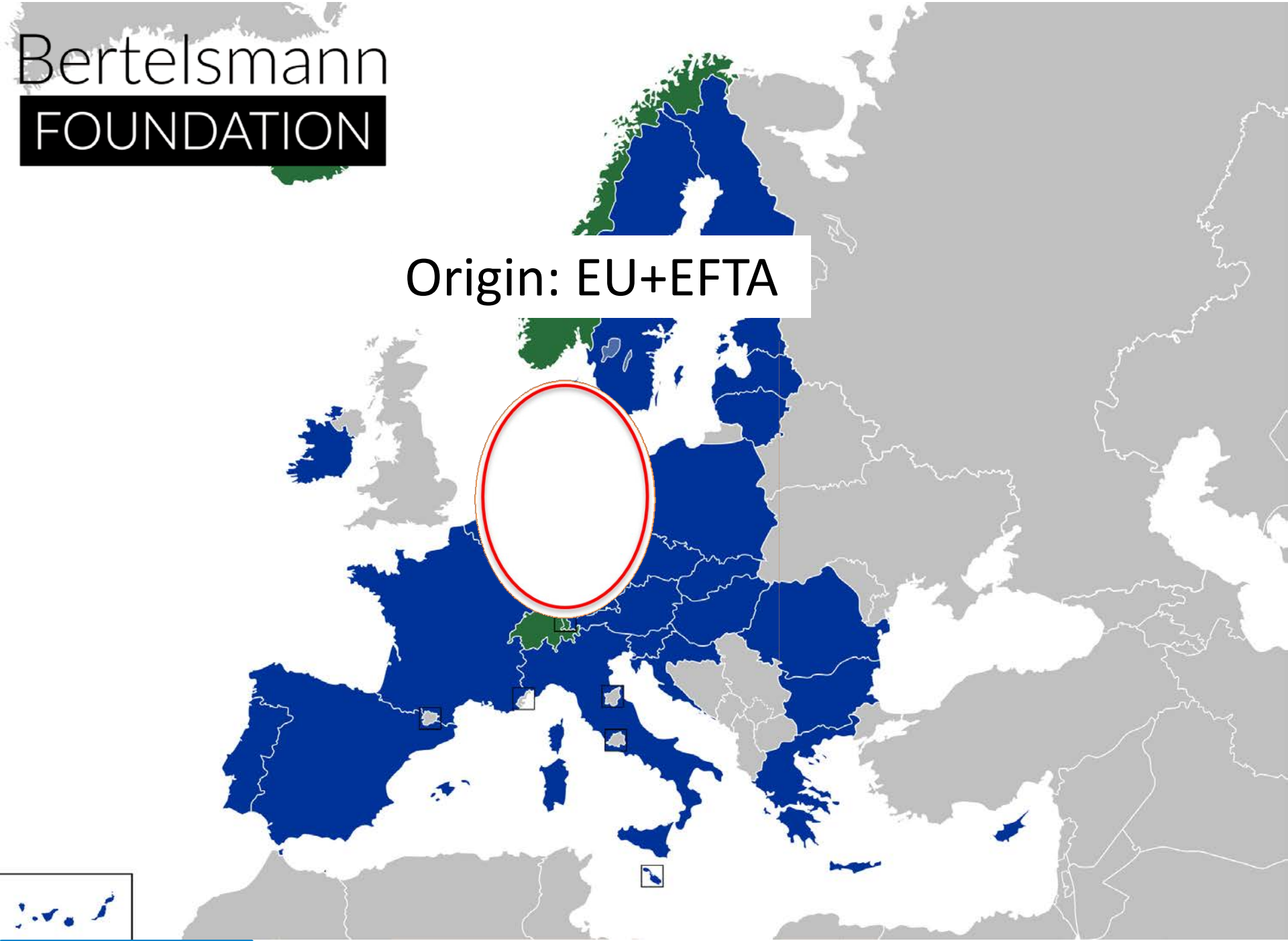


Bertelsmann
FOUNDATION

Destination: Germany

Bertelsmann
FOUNDATION

Origin: EU+EFTA



Predicting EU-Germany Migration

More limited and controlled setup than related projects

Online survey to help guide lag analysis:
"Approximately how much time has elapsed between the beginning of your planning for your move to Germany and your arrival in Germany?"

Additional data:
* Data for registration of EU nationals in Germany ("ground truth")
* Gallup World Poll, including questions on emigration intent
* Employment data for EU countries
* Data from the Goethe Institute on interest in German language courses

Currently: trying to address sparsity for smaller countries

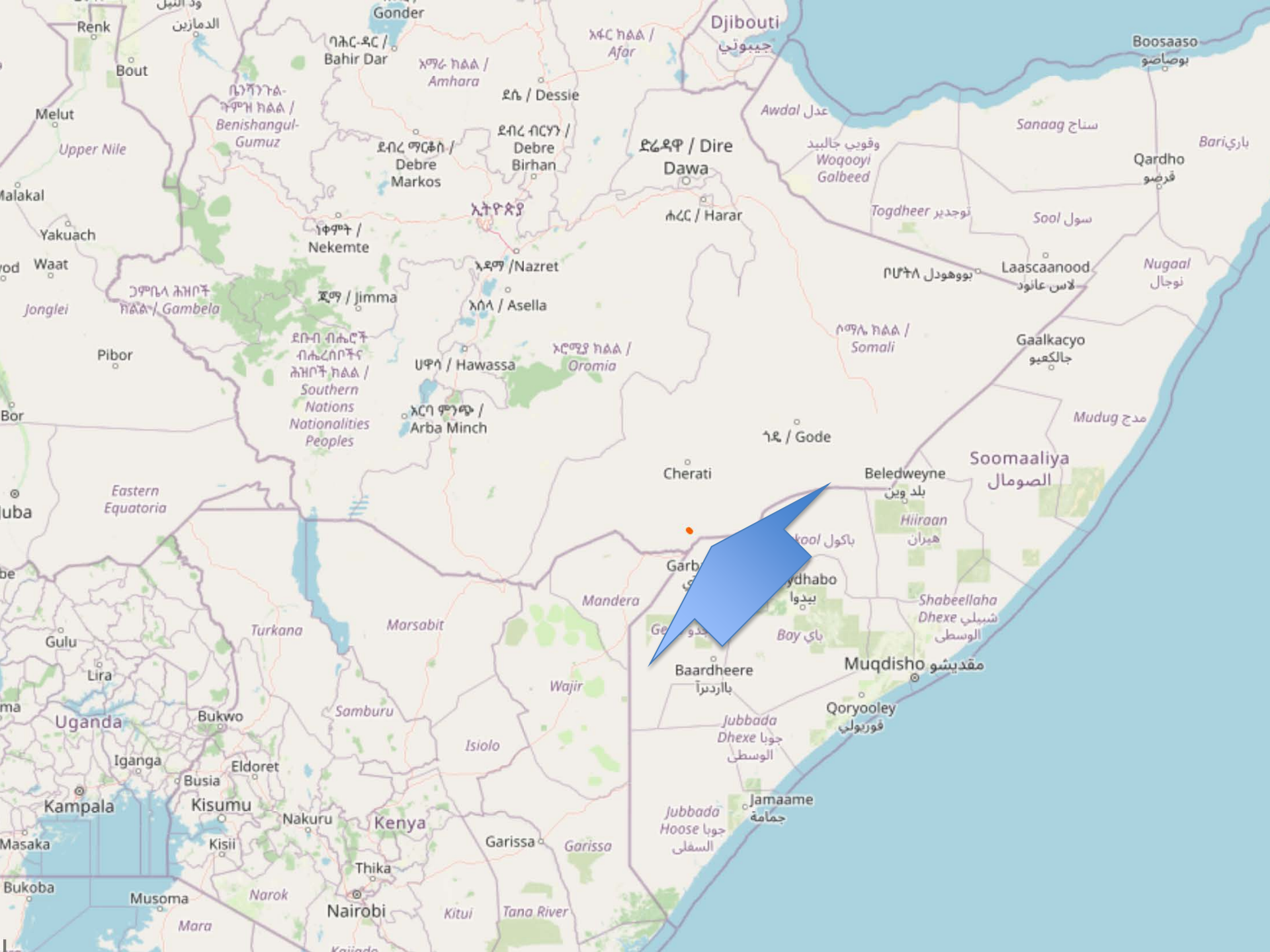
Ongoing …

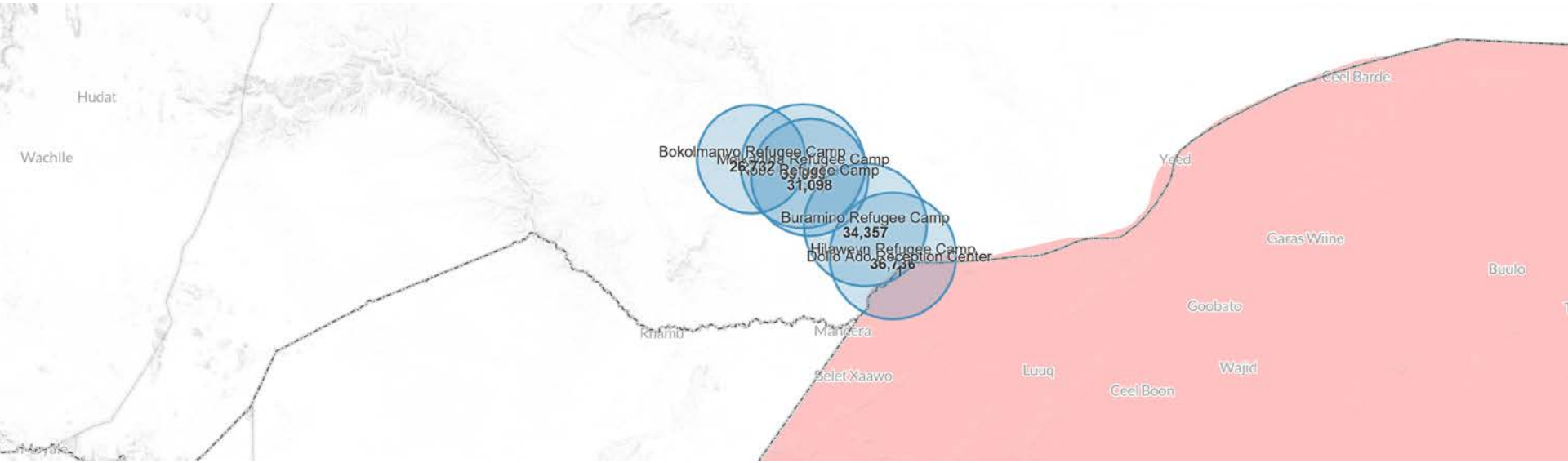
Context, context, context

THE IMPORTANCE OF GOATS

Project Jetson

Welcome to the Project Jetson website.
Jetson is a machine learning-based experiment that provides predictions on the movement(s) of displaced people. This experimental project combines data science, statistical processes, design-thinking techniques, and qualitative research methods. Jetson actively seeks new data sources, new narratives, and new collaborations in order to keep iterating, and improving. It has further underlined the importance of partnership, of collaboration, and of transparency.


Renk
Bout
Melut
Upper Nile
Malakal
Yakuach
Waat
Jonglei
Pibor
Bor
Juba
Eastern Equatoria
Gonder
ባሕር-ዳር / Bahir Dar
አማራ ክልል / Amhara
ደሴ / Dessie
ደብረ ማርቆስ / Debre Markos
ደብረ ብርሃን / Debre Birhan
ኢትዮጵያ
ነቀምት / Nekemte
ጂማ / Jimma
ጋምቤላ ሕዝቦች ክልል / Gambela
ደቡብ ብሔሮች ብሔረሰቦችና ሕዝቦች ክልል / Southern Nations Nationalities Peoples
ሀዋሳ / Hawassa
አርባ ምንጭ / Arba Minch
ናዝሬት /Nazret
አሰላ / Asella
ኦሮሚያ ክልል / Oromia
ሐረር / Harar
ድሬዳዋ / Dire Dawa
Djibouti
ጂቡቲ
አፋር ክልል / Afar
Benishangul-Gumuz
Awdal عدل
Woqooyi Galbeed
Togdheer توجدير
Sool سول
Sanaag سناج
Bari باري
Qardho قردضو
Boosaaso بوصاصو
Laascaanood لاس عانود
Nugaal نوجال
Somali
Gaalkacyo جالكعيو
Mudug مدج
Gode
Cherati
Beledweyne بلد وين
Soomaaliya الصومال
Hiiraan هيران
Shabeellaha Dhexe شبيلي الوسطى
Mandera
Garb
Bay باي
Baardheere باردير
Muqdisho مقديشو
Qoryooley قوريولي
Jubbada Dhexe جوبا الوسطى
Jubbada Hoose جوبا السفلى
Jamaame جمامة
Turkana
Marsabit
Wajir
Samburu
Isiolo
Garissa
Kitui
Tana River
Narok
Mara
Uganda
Gulu
Lira
Bukwo
Iganga
Busia
Eldoret
Kampala
Kisumu
Nakuru
Kenya
Kisii
Masaka
Bukoba
Musoma
Thika
Nairobi


Hudat
Wachlle
Bokolmanyo Refugee Camp
Refugee Camp
31,098
Buramino Refugee Camp
34,357
Hilawevn Refugee Camp
Dollo Ado Reception Center
Ceel Barde
Yeed
Garas Wiine
Buulo
Goobato
Luuq
Wajid
Ceel Boon
Belet Xaawo
Mandera

Somali Goats

By Cpl. Paula M. Fitzgerald, USMC - United States Department of Defense, Public Domain, https://commons.wikimedia.org/w/index.php?curid=28141538

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Rebeca Moreno Jiménez

“It’s all about context,” she says. “Silicon Valley data scientists are far away from the reality and the ethical problems data [present] in an unequal world. Trying to explain the complexity of this world with a few models and some datasets are too limited in [the humanitarian] sector. [The humanitarian sector] is about humans, and sometimes irrational and rational human choices. ”

https://medium.com/unhcr-innovation-service/a-goat-story-3ed6bdd2b237

Hooked? => Big Data for Migration Alliance

BD4M - BIG DATA FOR MIGRATION ALLIANCE

EVENTS RESOURCES JOIN

WEBINAR SERIES: Harnessing Data Innovation for Migration Policy

Key Actors

Marzia Rango

International Organization for Migration | GMDAC

Michele Vespe

European Commission | KCMD

Stefaan Verhulst

The Govlab

Thanks!

Data literacy of stakeholders?

Fake accounts?

Biggest risks?

Who's left behind?

Will FB disappear?

Privacy?

Questions?

Cui bono?

Ease of data collection?

Relation to Big Tech?

Integration processes?

Power asymmetries?

Other digital data sources?