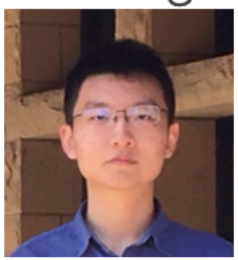
VolkswagenStiftung

Demographic Inference and Representative Population Estimates from Multilingual Social Media Data

Zijian Wang



EECS

Scott Hale

OXFORD INTERNET INSTITUTE

David Adelani

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

Przemyslaw Grabowicz

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

Timo Hartmann

gesis Leibniz-Institut für Sozialwissenschaften

Fabian Flöck

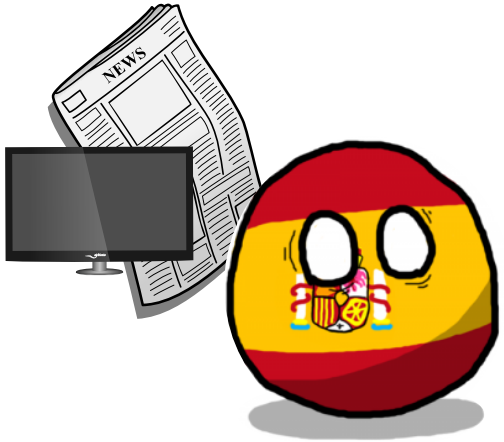
gesis Leibniz-Institut für Sozialwissenschaften

David Jurgens

UMSI

In Proceedings of the Web Conference (WWW), 2019

"What is the most important issue facing the country today?"



Unemployment!

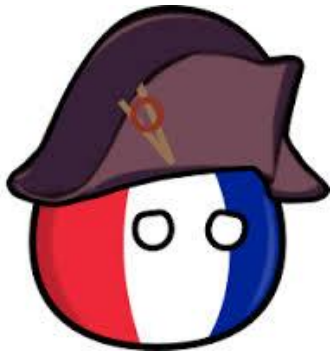
"What is the most important issue facing the country today?"

L'environnement!

Refugee crisis!

"What is the most important issue facing the ~~country~~ today?"

European Union



Surveys often used to measure issue importance

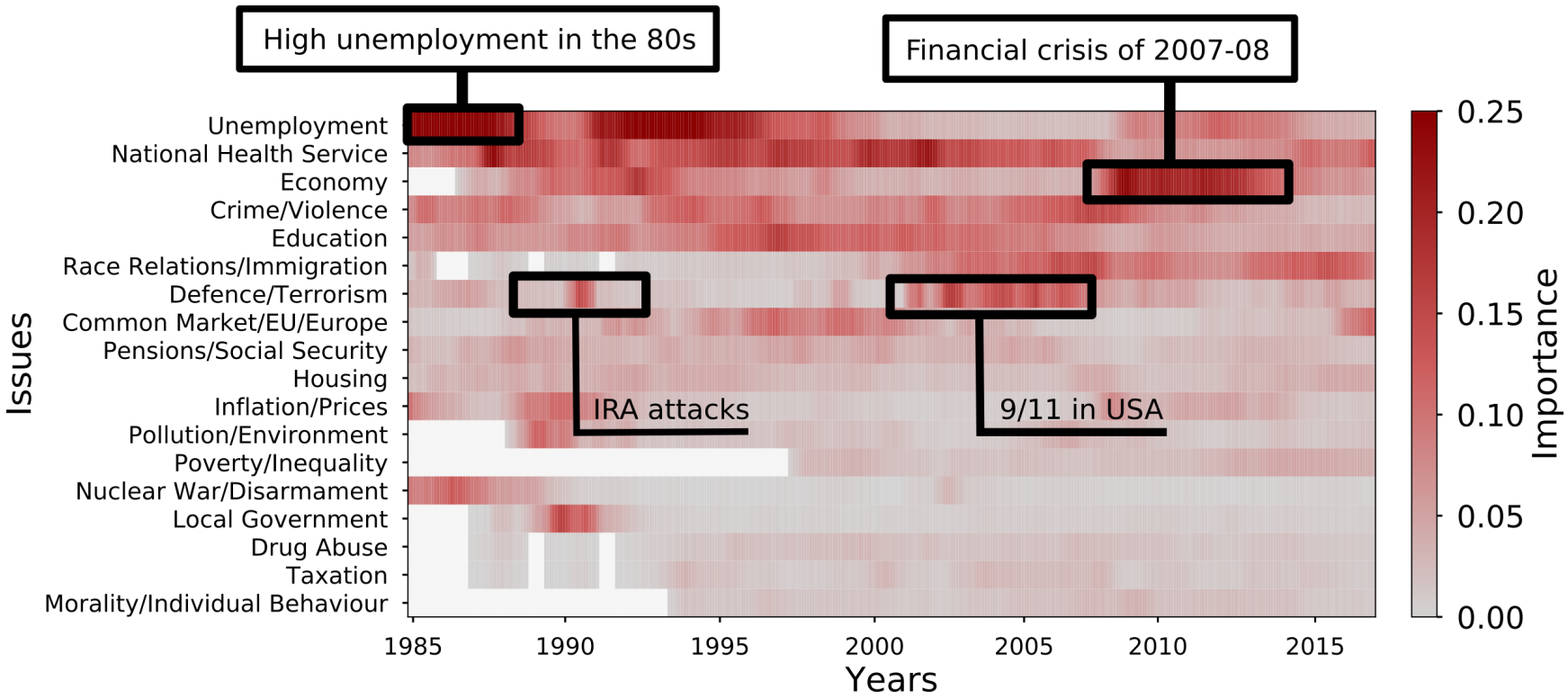


Figure Public attention to policy issues by month from 1985 to 2016 as reported by representative surveys of the UK population collected by Ipsos MORI.

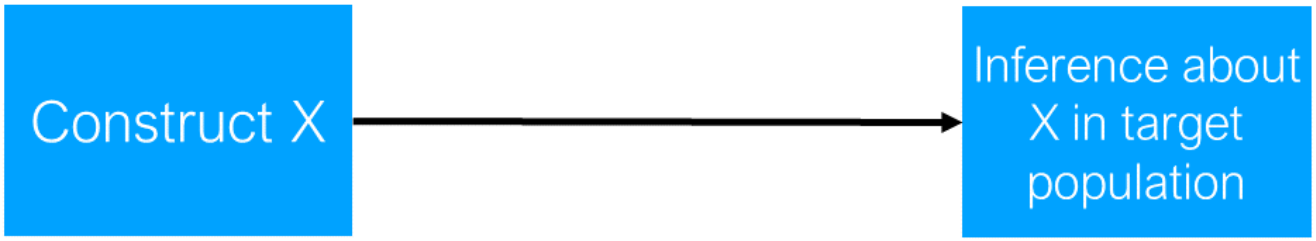
Could social media measure issue importance?

Surveying has been the gold standard, but observation on social media is cheap and easy...

	Survey	Social Media
Purposefully sampled	✓	X
Known demographics	✓	X
Easy to get	X	✓
Cheap	X	✓
Quick reaction to events	X	✓
Available retrospectively	X	✓
Social desirability	✓	✓

Can we use social media data to measure public attention to societal issues?

Typical survey research design



Approval of the head of state

Smoking

Voting intention

Attitude towards immigrants / Racism

Consumer confidence

...

All Canadian Citizens

All Americans over 35

...

Using social media to measure X

Construct X → Inference about X in target population

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). **Detecting influenza epidemics using search engine query data**. Nature, 457(7232), 1012.

Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. **"Predicting elections with twitter: What 140 characters reveal about political sentiment."** Proceedings of ICWSM, 2010

Using social media to measure X

Construct X → Inference about X in target population

CAUTION!

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. **"The parable of Google Flu: traps in big data analysis."** Science 343, no. 6176 (2014): 1203-1205.

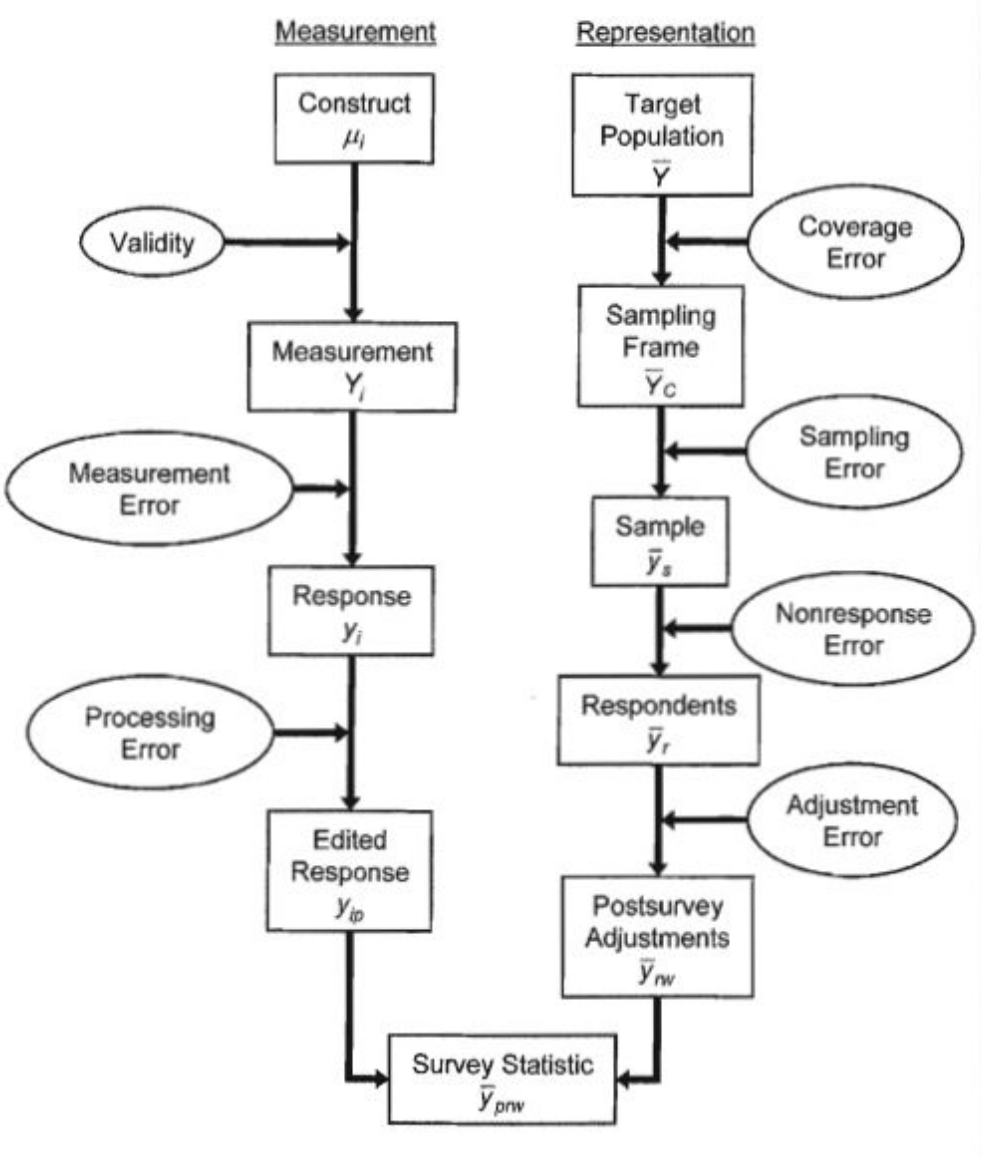
Kim, A., Murphy, J., Richards, A., Hansen, H., Powell, R. and Haney, C., 2013. **Can Tweets Replace Polls? A US Health-Care Reform Case Study.** *Social media, sociality, and survey research*, pp.61-86.

CAUTION!

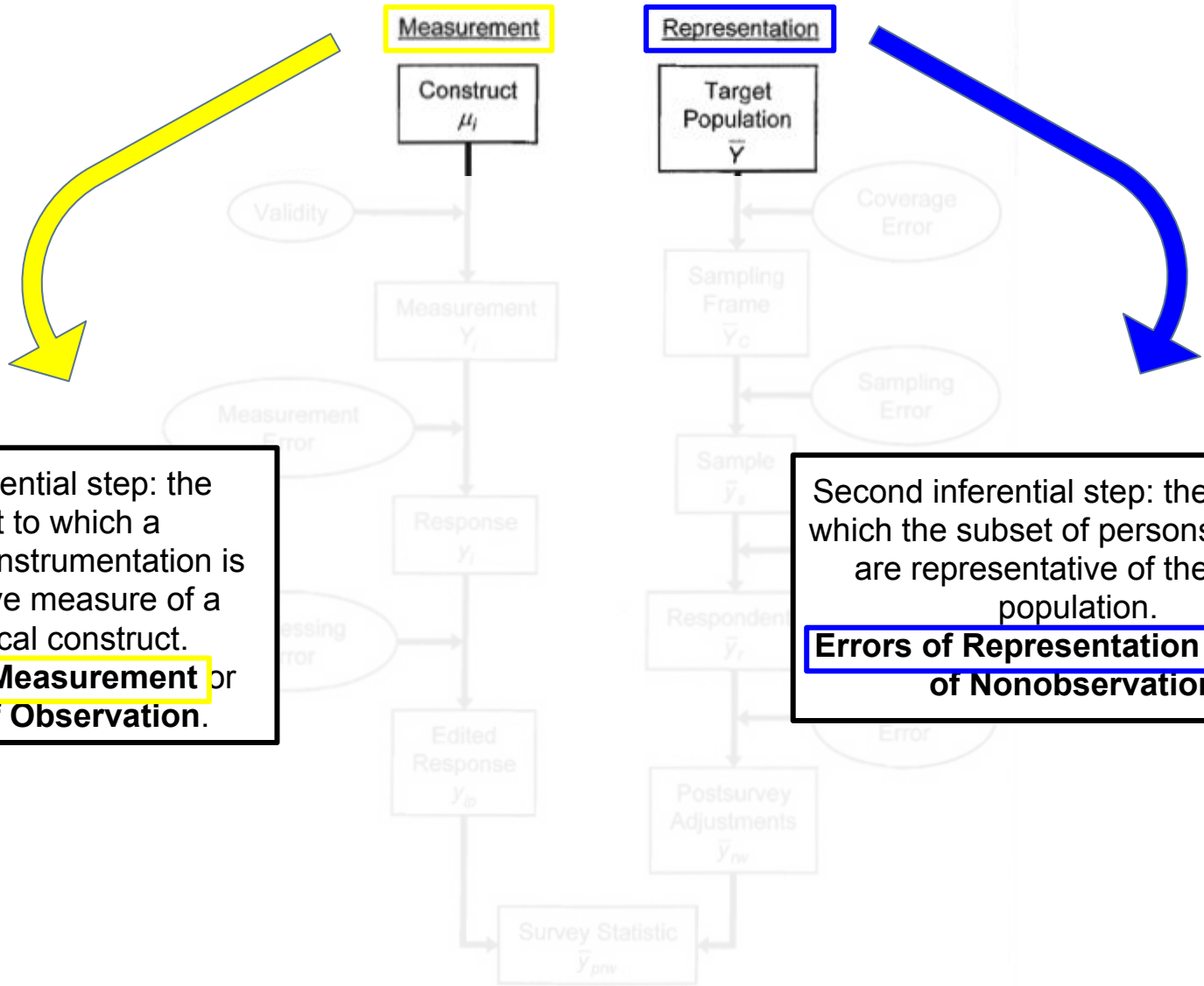
Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. **"Predicting elections with twitter: What 140 characters reveal about political sentiment."** Proceedings of ICWSM, 2010

Jungherr, A., Jürgens, P. and Schoen, H., 2012. **Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im "predicting elections with twitter: What 140 characters reveal about political sentiment"**. *Social science computer review*, *30*(2), pp.229-234.

Gayo-Avello, D., 2012. **"I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"--A Balanced Survey on Election Prediction using Twitter Data**. *arXiv preprint arXiv:1204.6441*.



Total Survey Error Framework: Survey Lifestyle from a Quality perspective. Adapted from Groves et al.(2011) P. 48

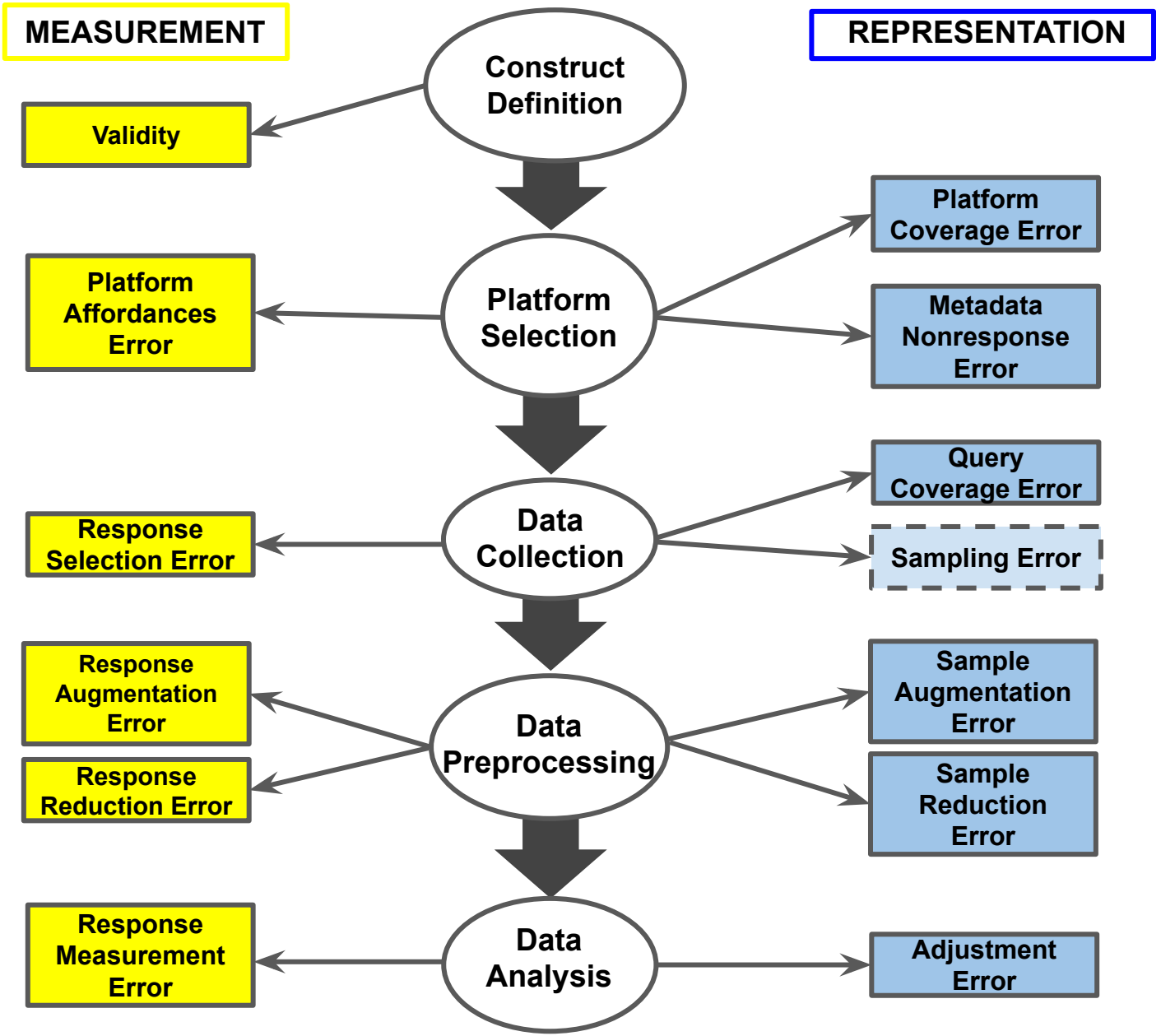


First inferential step: the extent to which a given test/instrumentation is an effective measure of a theoretical construct. **Errors of Measurement** or **errors of Observation**.

Second inferential step: the extent to which the subset of persons sampled are representative of the larger population. **Errors of Representation** or **errors of Nonobservation**.

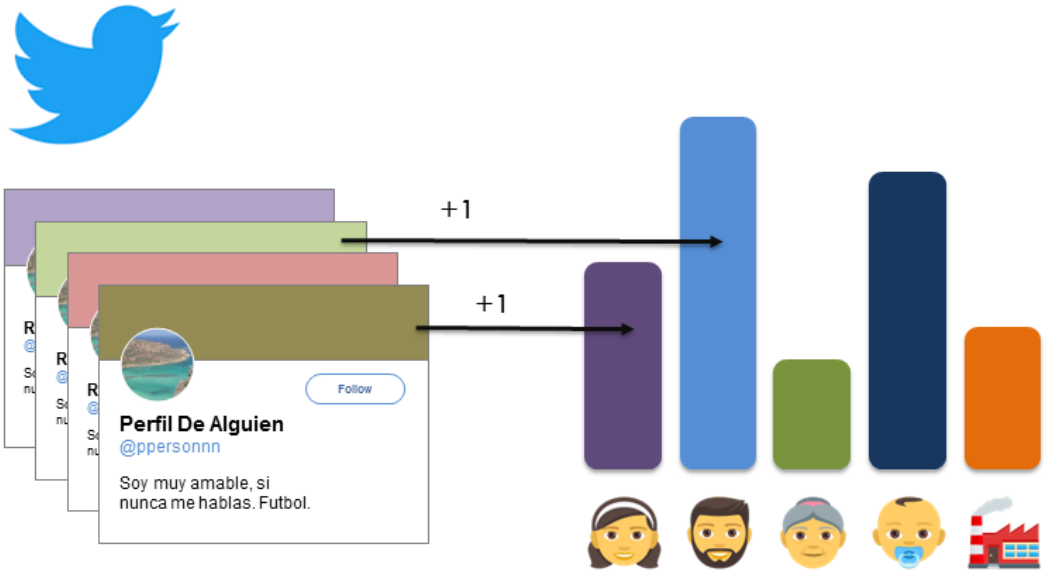
Total Survey Error Framework: Survey Lifestyle from a Quality perspective. Adapted from Groves et al.(2011) P. 48

The Digital Trace Data Error Framework



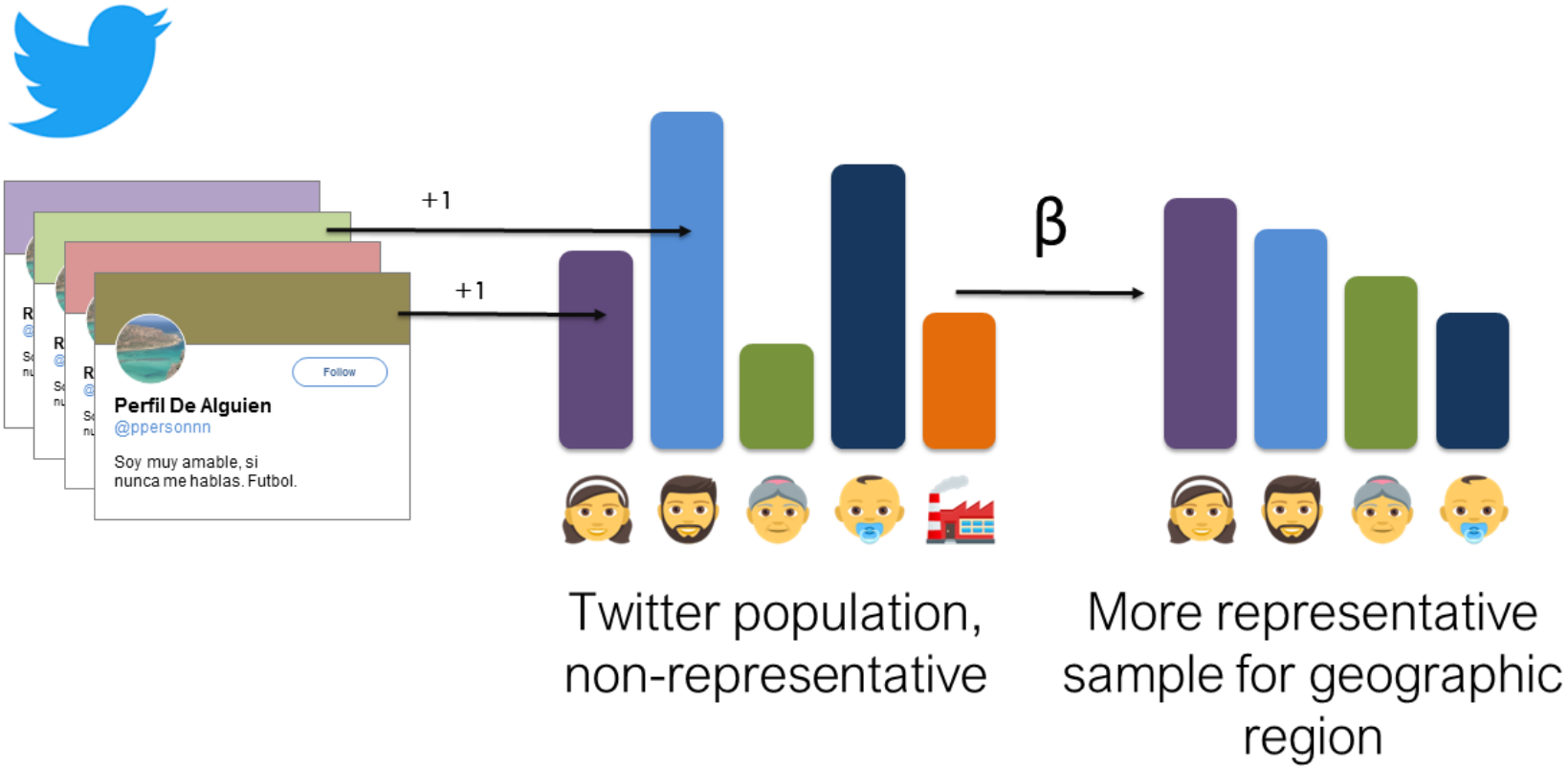
How might debiasing work in practice in social media?

How might debiasing work in practice in social media?

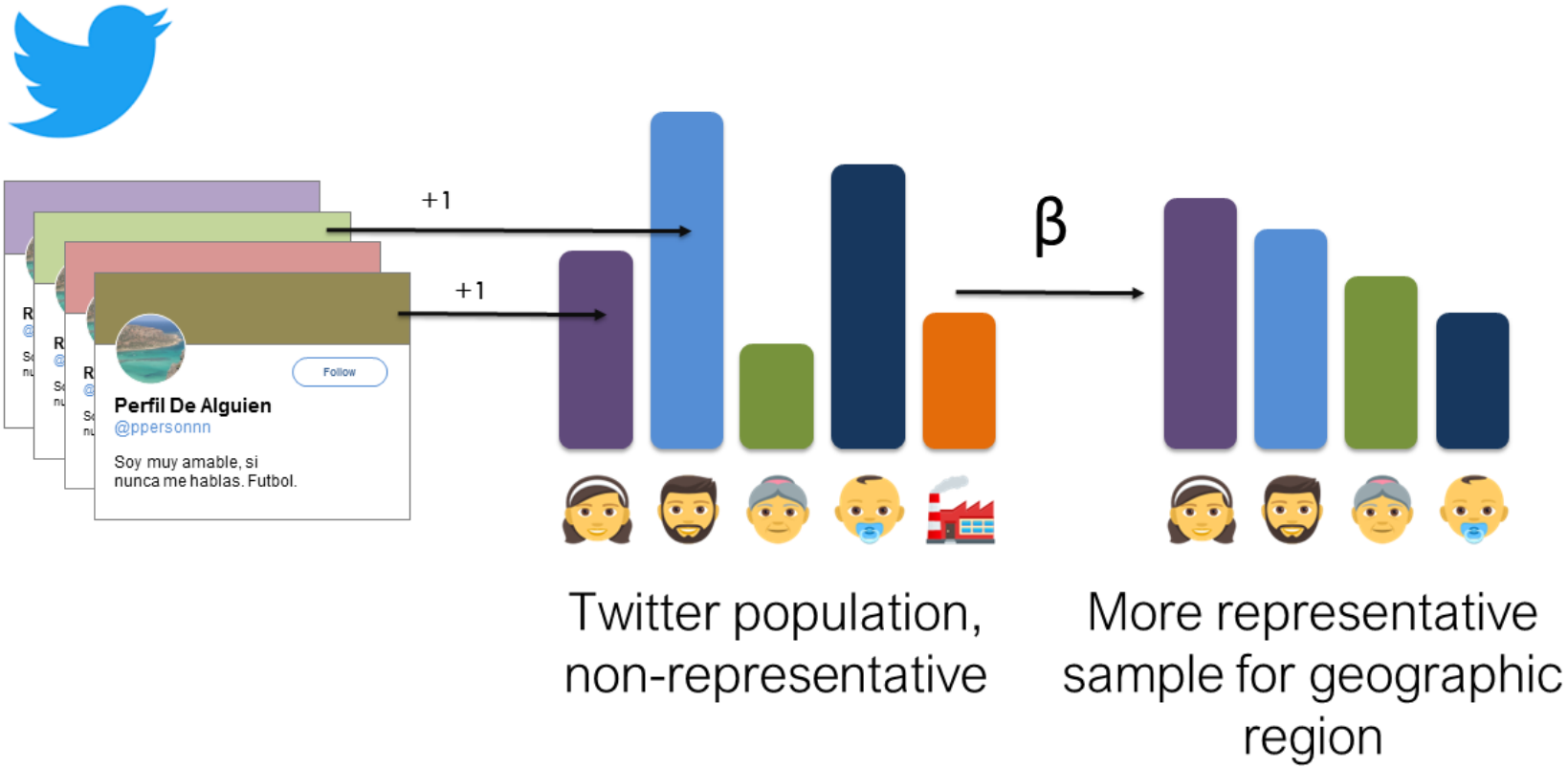


Twitter population,
non-representative

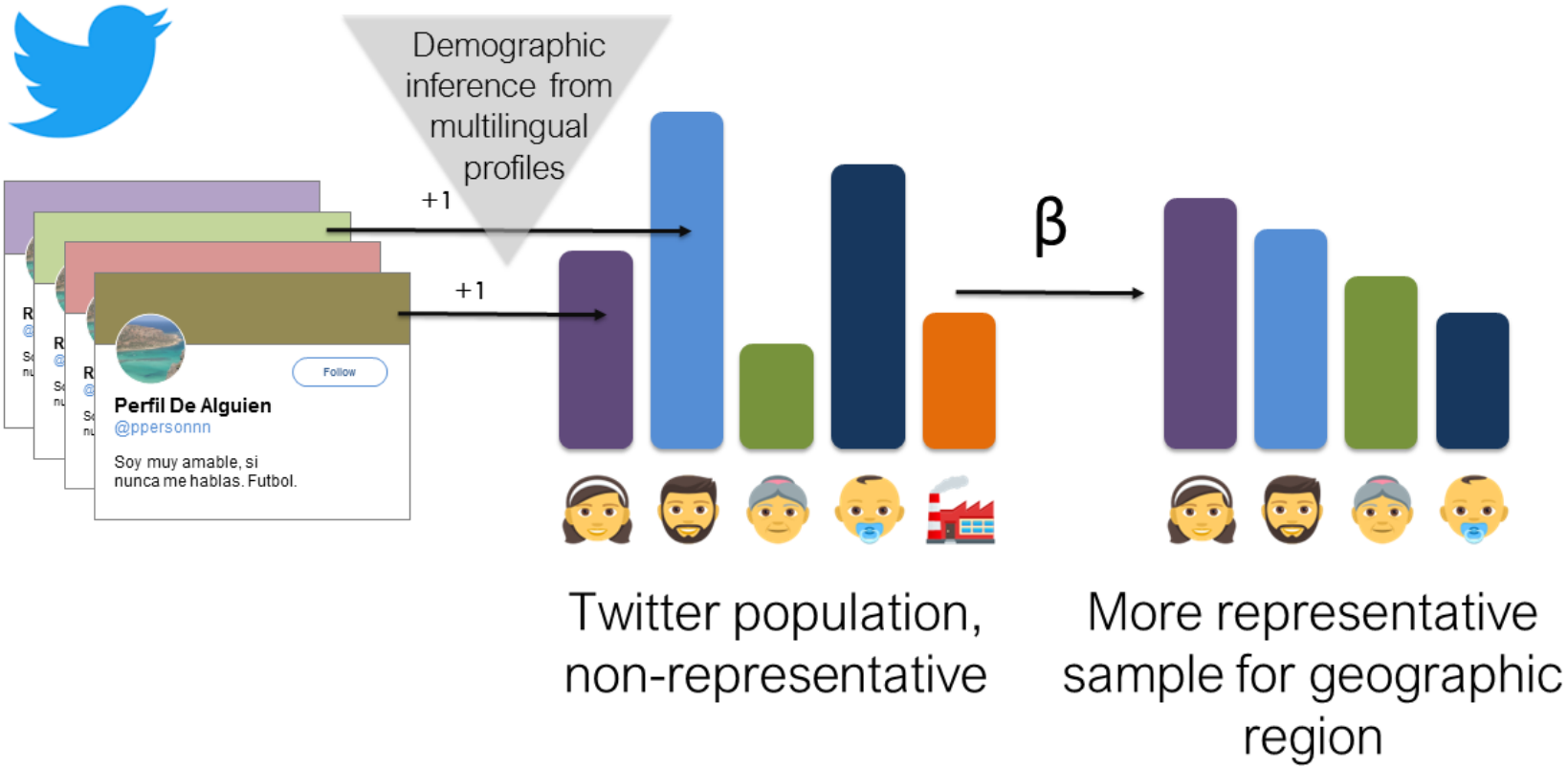
How might debiasing work in practice in social media?



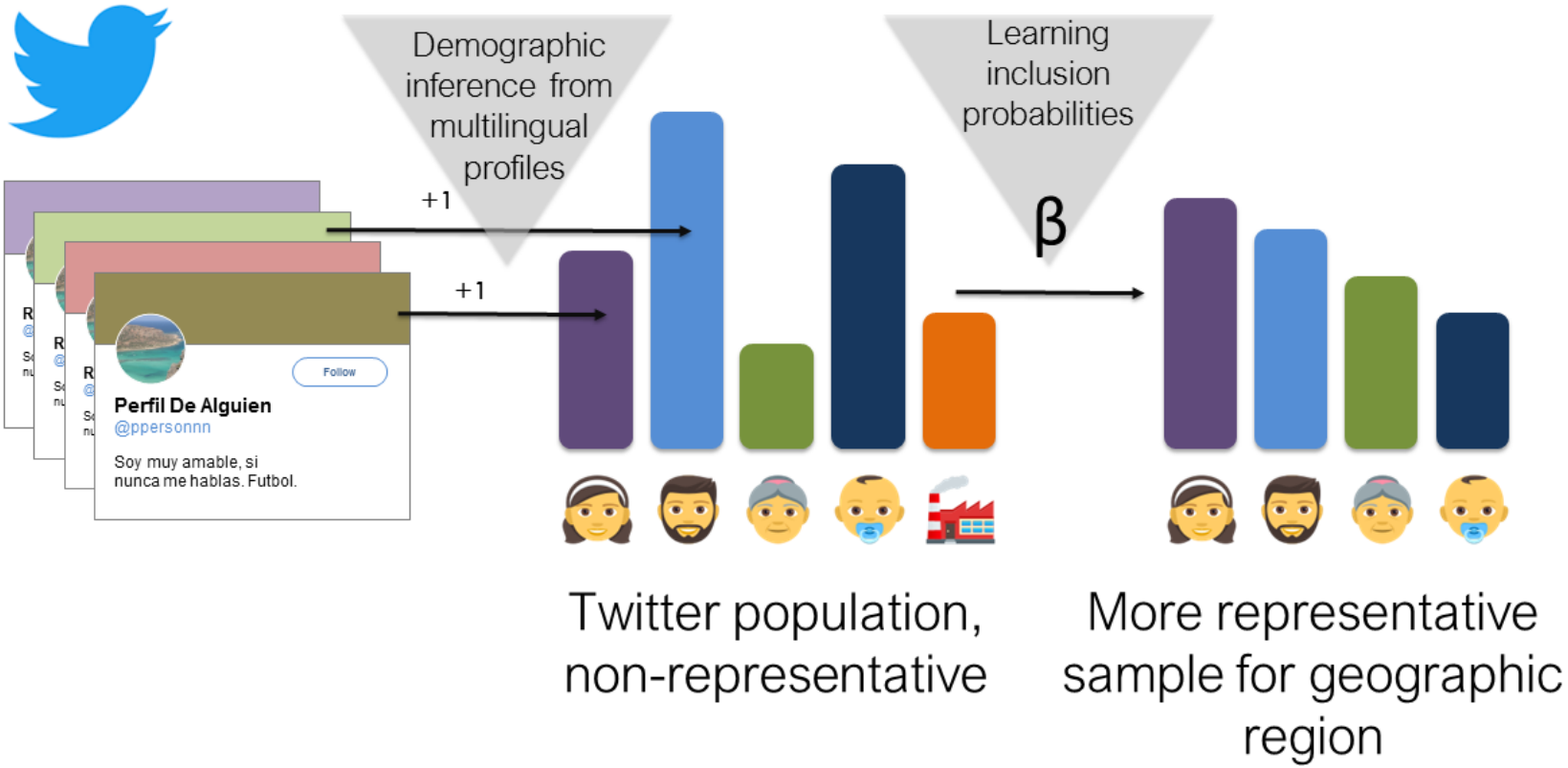
Poststratification requires the ability to first stratify!



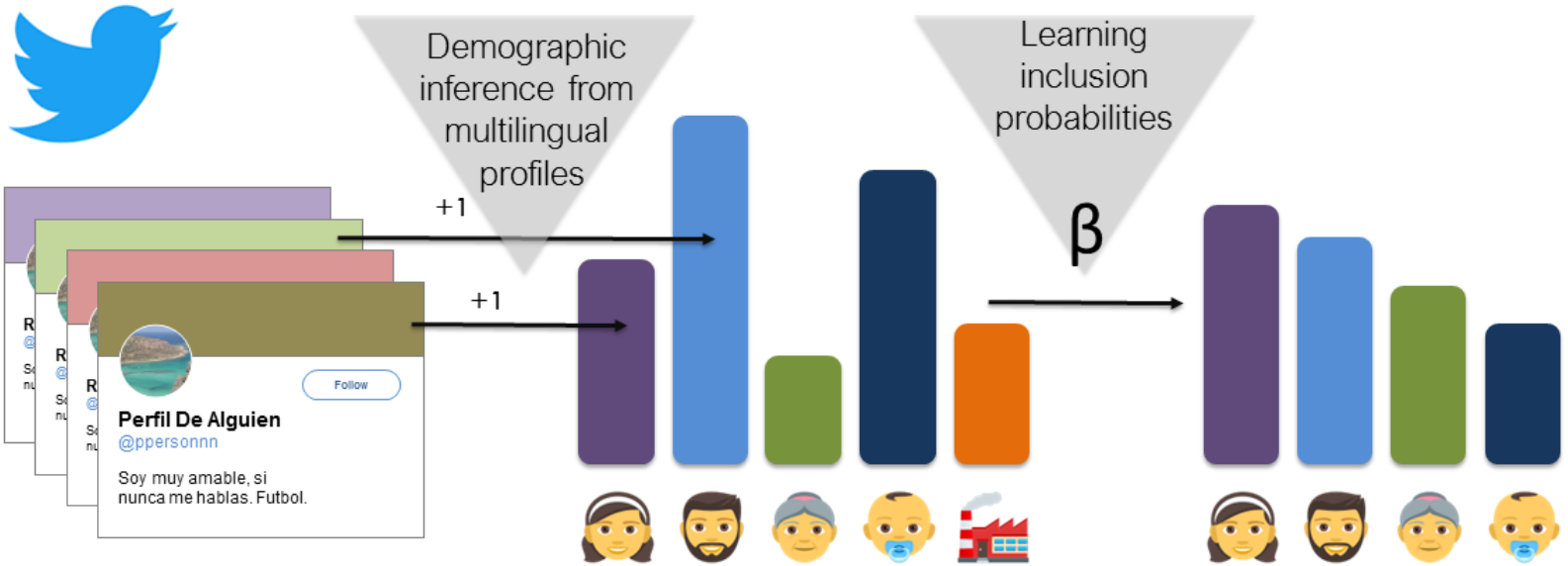
Poststratification requires the ability to first stratify!



Poststratification requires the ability to first stratify!



Poststratification requires the ability to first stratify!



Twitter population, non-representative

More representative sample for geographic region

For 32 languages (EU-based)

binary gender, 4-category age, binary is-org status

Aside: Not all users are human

Which of the following is not from a personal account?

This is it – it's all come down to the last few minutes to change our country #VoteLabour

---

Unemployment sucks but it sucks more when it affects your friendships. We've all been there bit.ly/1lfGDpW

In swine flu season, sanitizer for the hand of God – fluviruses.com/?p=72702 #Flu #viruses #virus #disease

---

Depression is not feeling sad once in a while. It's so much more complicated

---

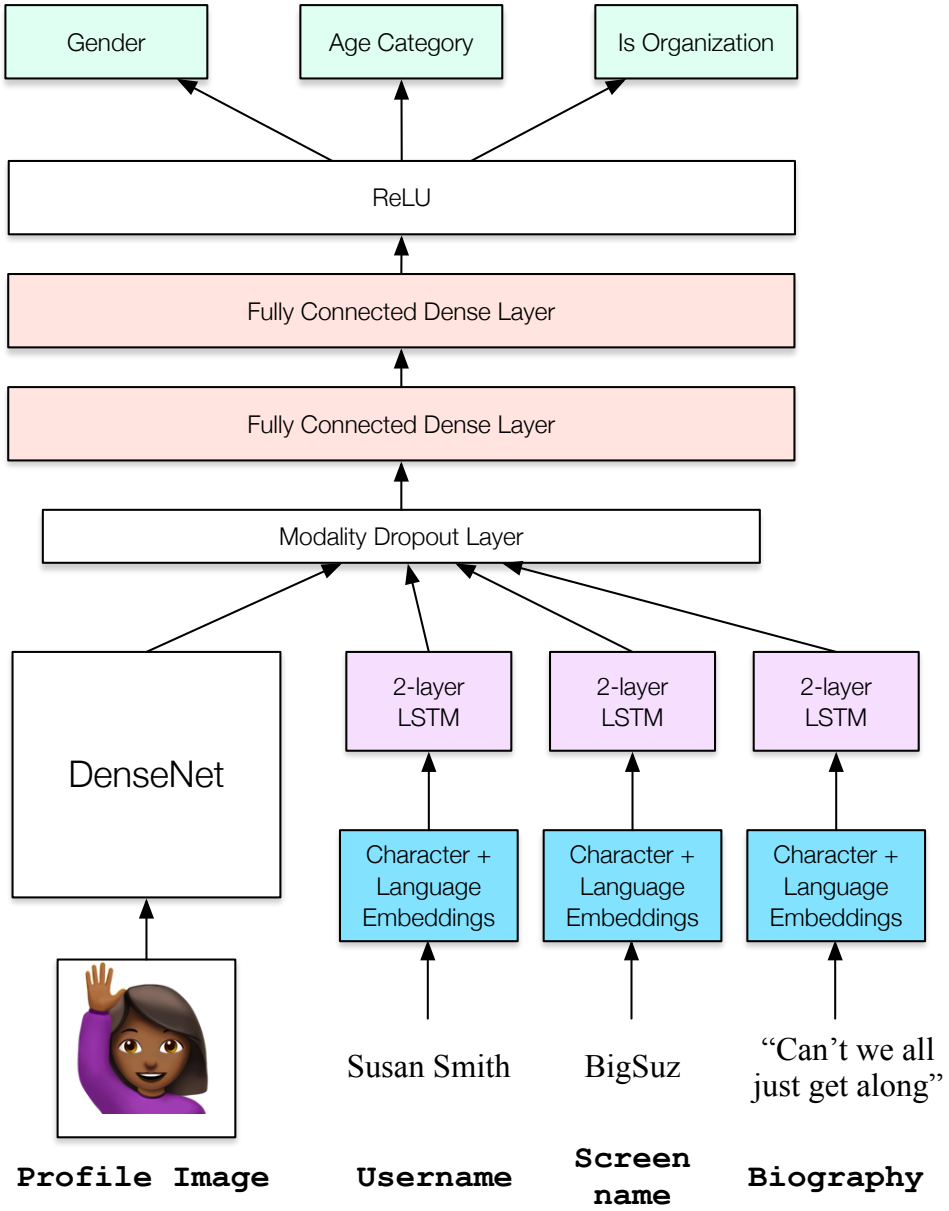
Happy Mother's Day to all the amazing Hockey Moms out there!

Aside: Not all users are human

Nearly 10% of all Twitter accounts belong to organisations. They can play a large rule in agenda setting, but cannot participate in human phenomena.

Organizations are Users Too: Characterizing and Detecting the Presence of Organizations on Twitter. James McCorriston, David Jurgens, and Derek Ruths. Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM). 2015.

Demographic inference (m3)



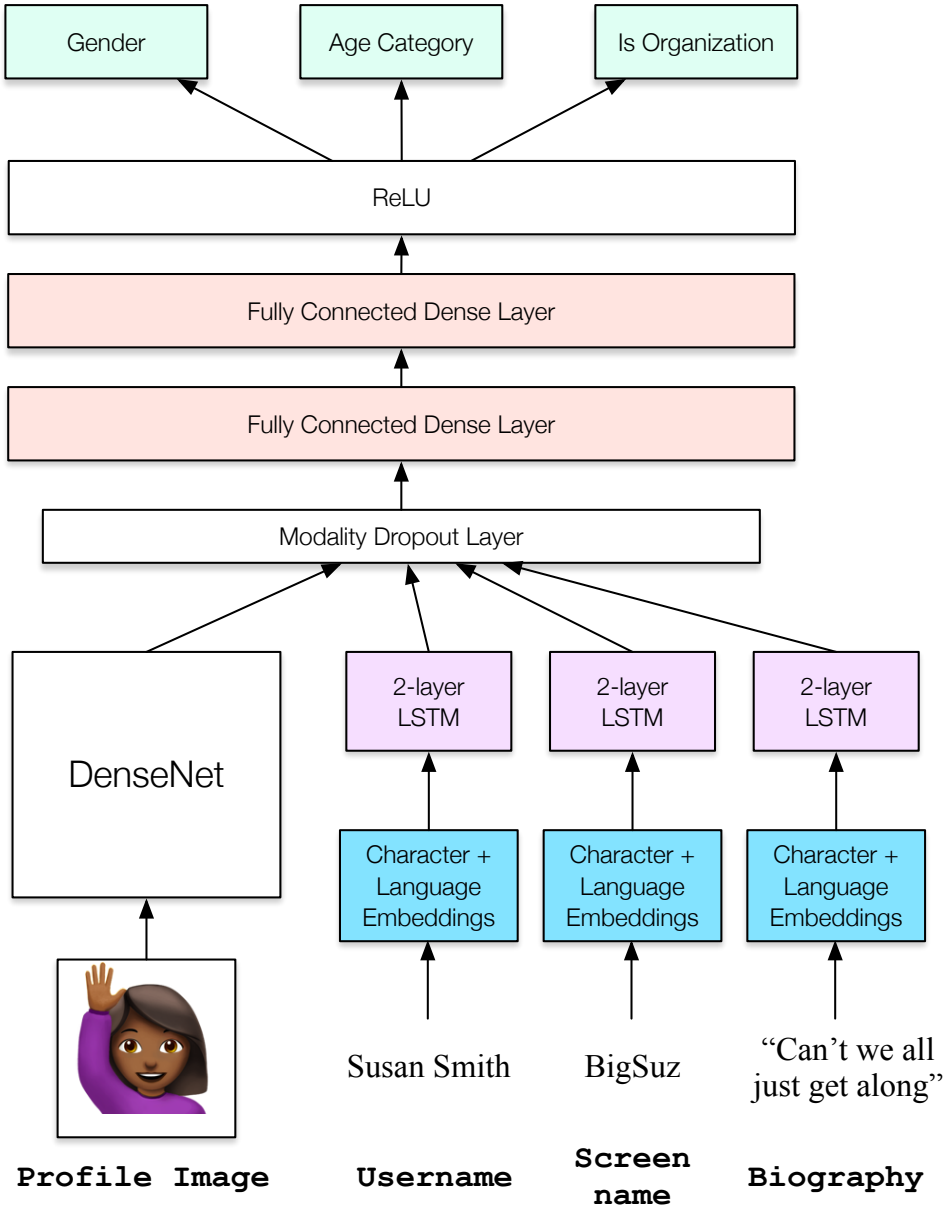
m3 inference:

multimodal Images and text

multiattribute Age, Gender, and is-organization

multilingual For 32 languages spoken in Europe

Demographic inference (m3)



Data:

- Twitter (random sample)
- Twitter (hand-curated lists of organisations)
- IMDB/Wikipedia
- Crowdsourcing (evaluation only)

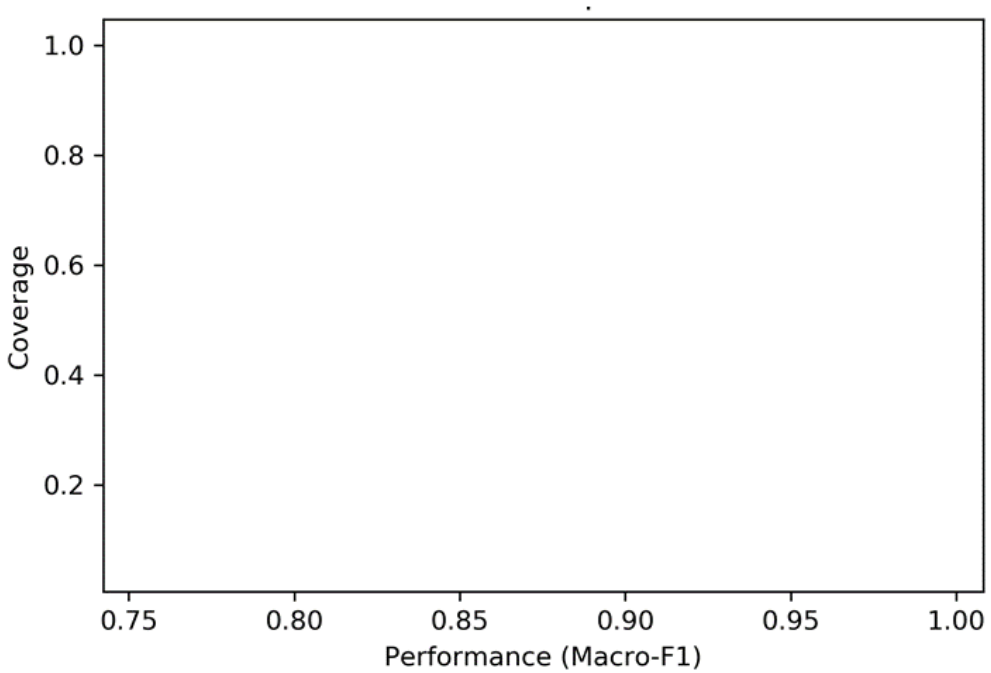
Training and Evaluation Data

- 10% sample of Twitter from 2014 to 2017
  - Heuristically labeled for age/gender and organizations
  - ~40% English
- Headshots of actors in IMDB and biographies in Wikipedia: age and gender
- Cross-language augmentation from EN
- Co-training → expanding data set on large unlabeled twitter corpus

	Gender	Age	Organization
Initial	3.98M	1.20M	59.92K
Total after augment	14.53M	2.61M	23.86M
Held-out Evaluation	0.38M	0.36M	6.91K

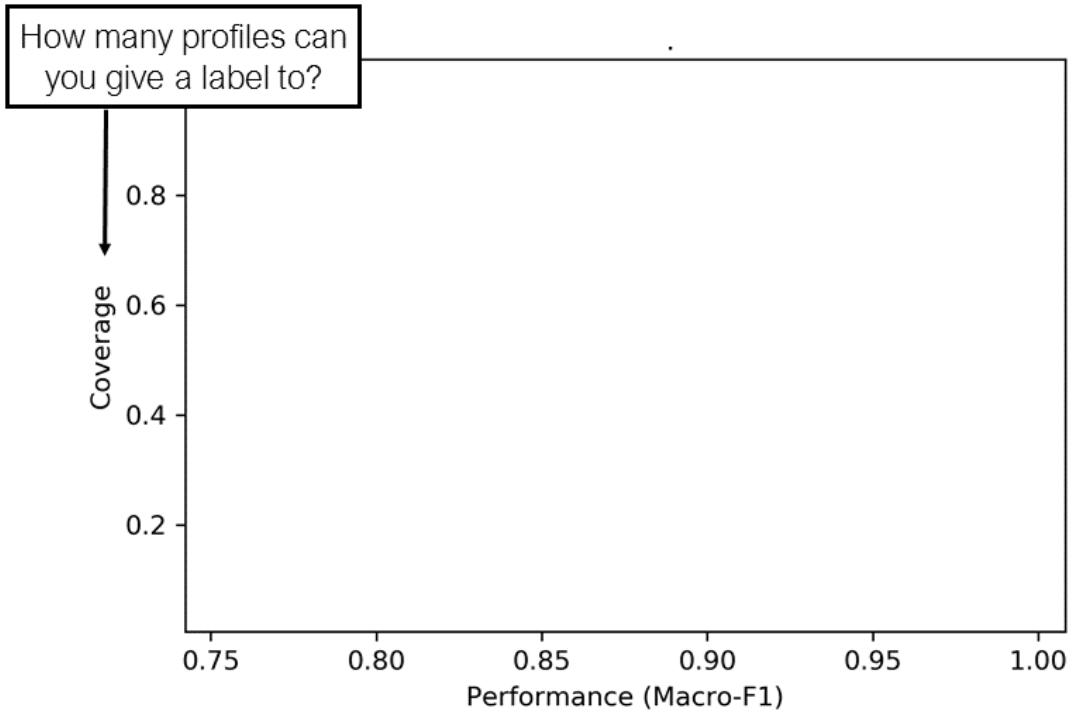
State of the art for gender inference

Gender from Images



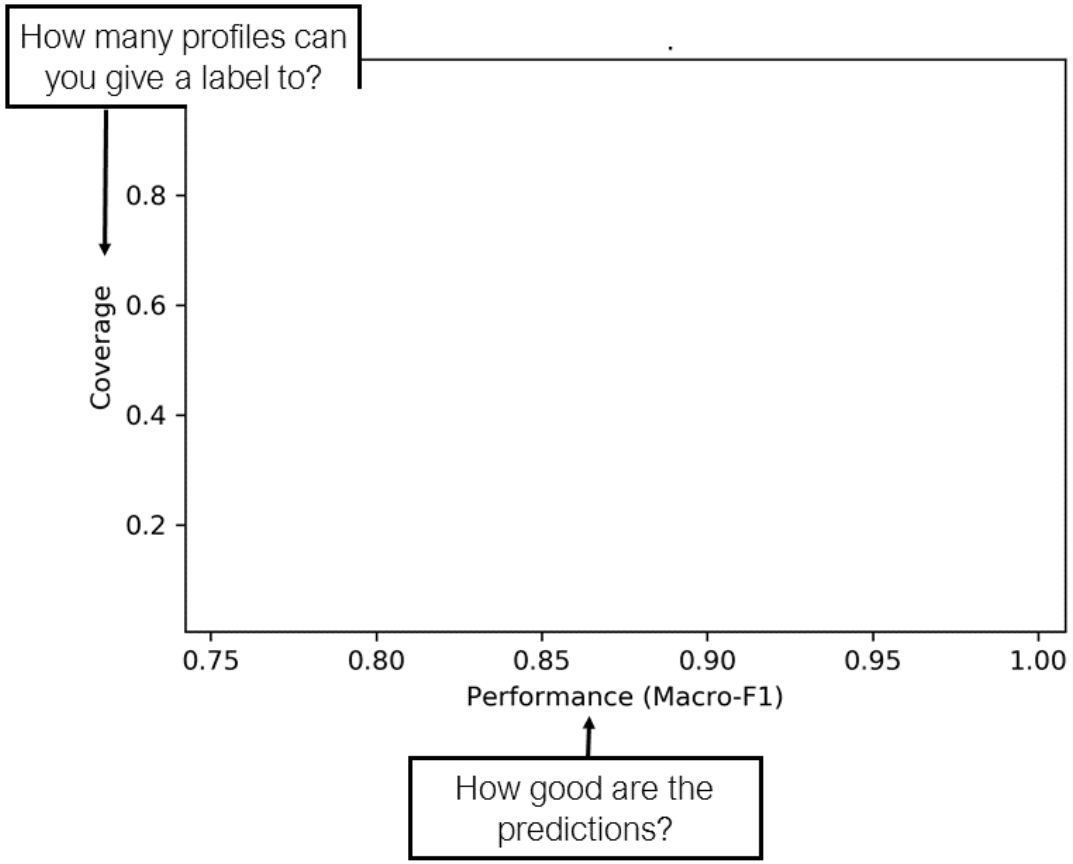
State of the art for gender inference

Gender from Images



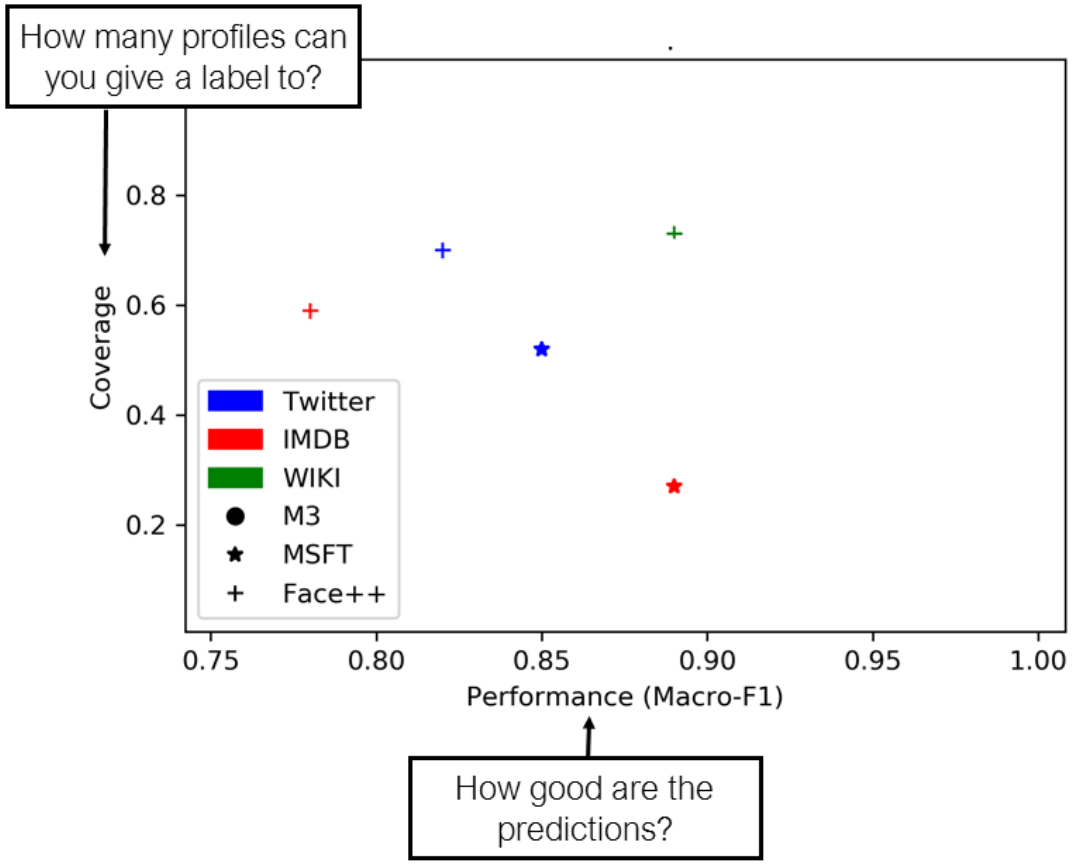
State of the art for gender inference

Gender from Images



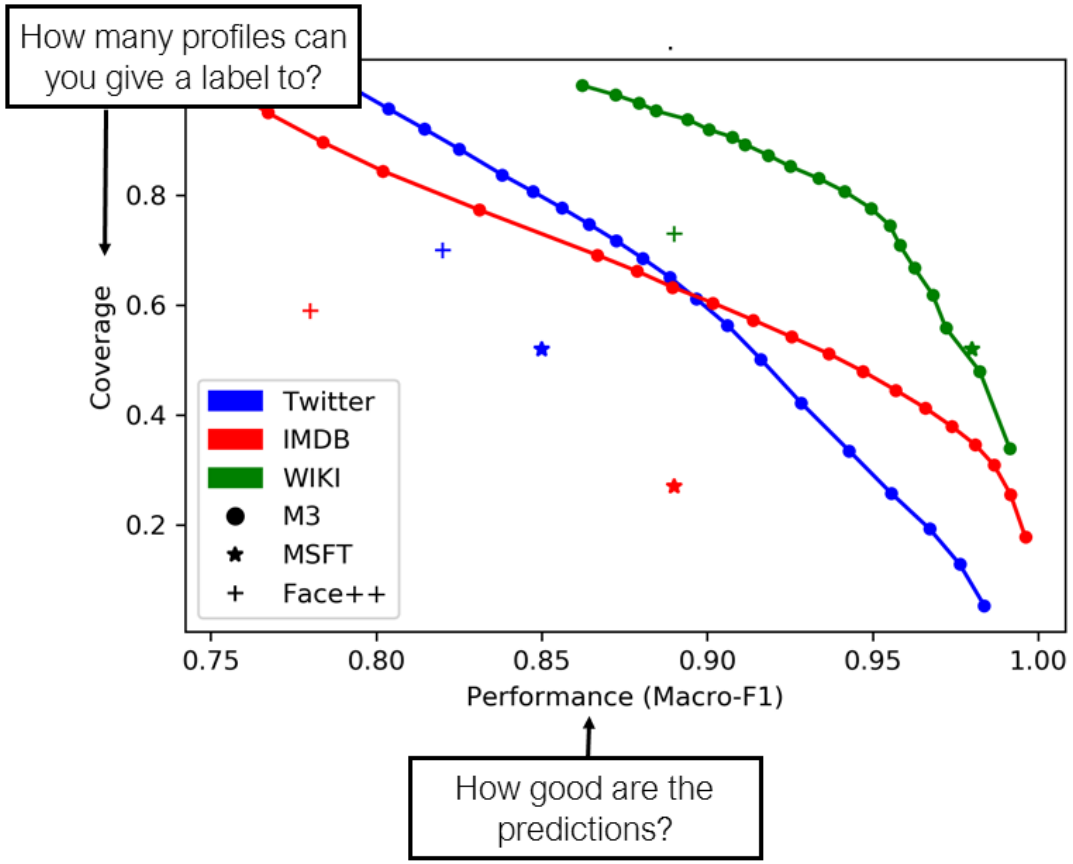
State of the art for gender inference

Gender from Images



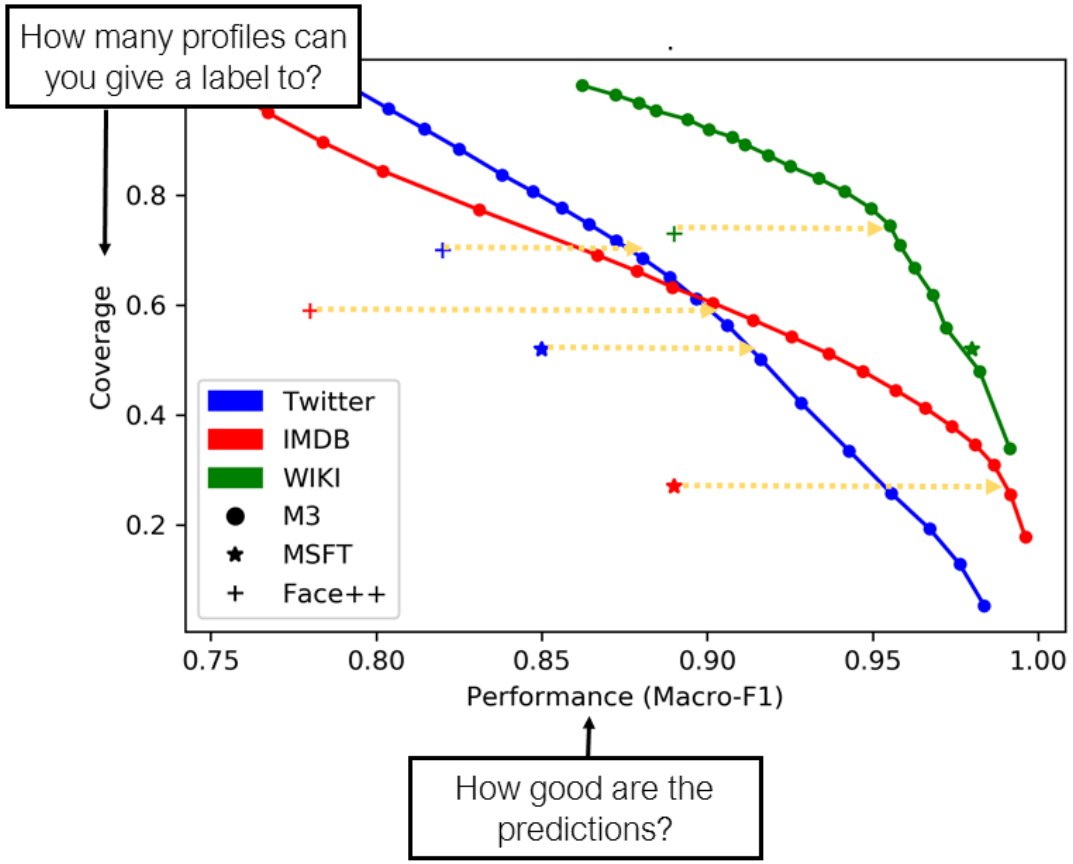
State of the art for gender inference

Gender from Images



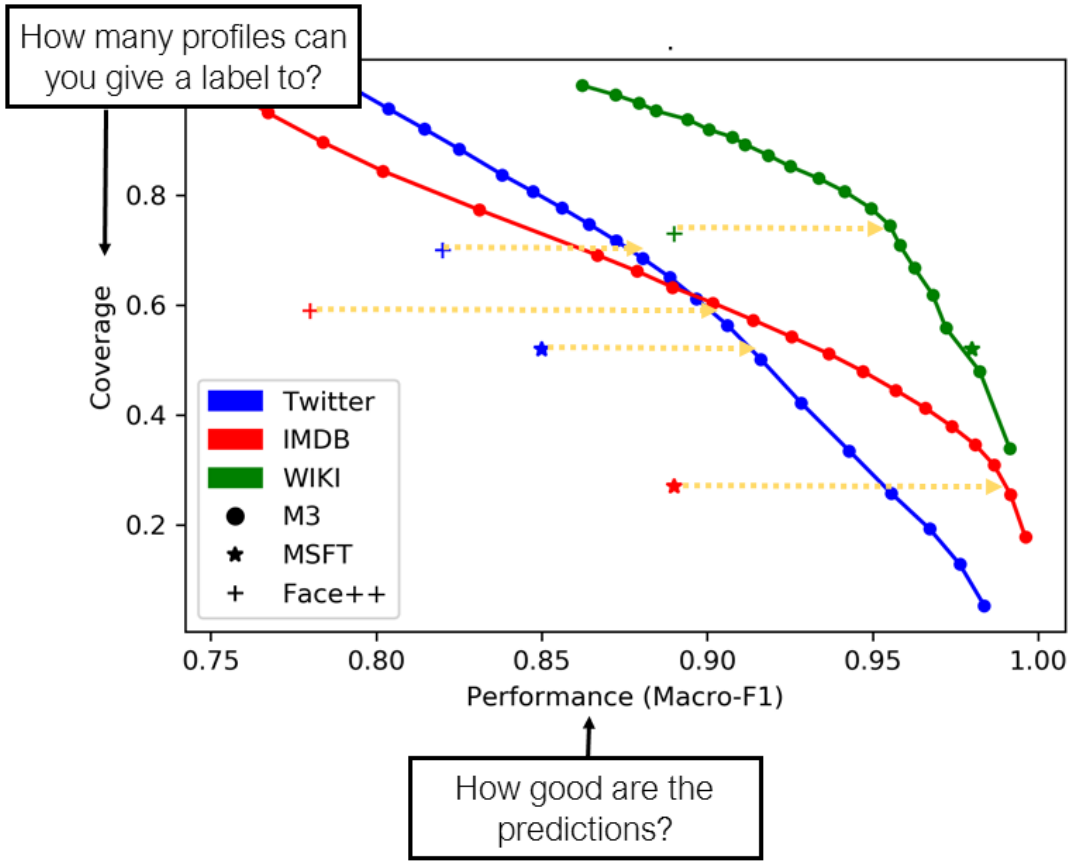
State of the art for gender inference

Gender from Images

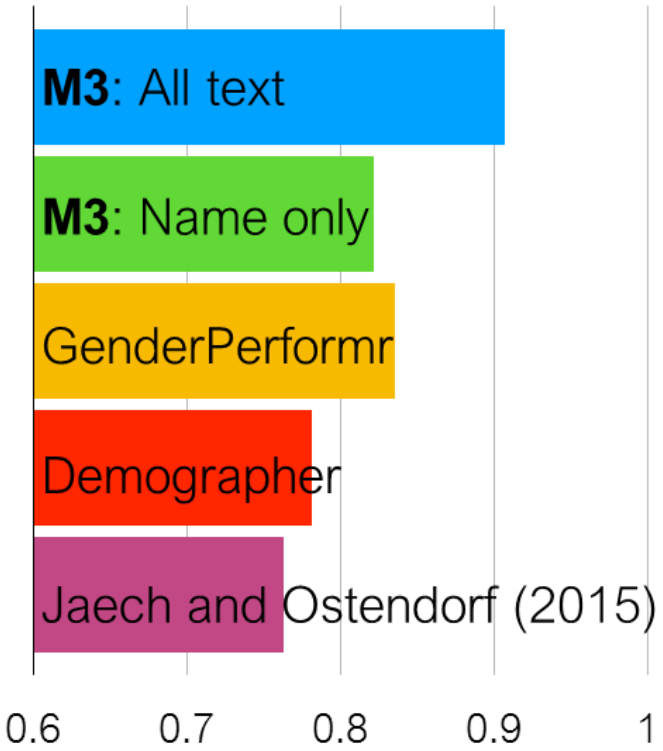


State of the art for gender inference

Gender from Images



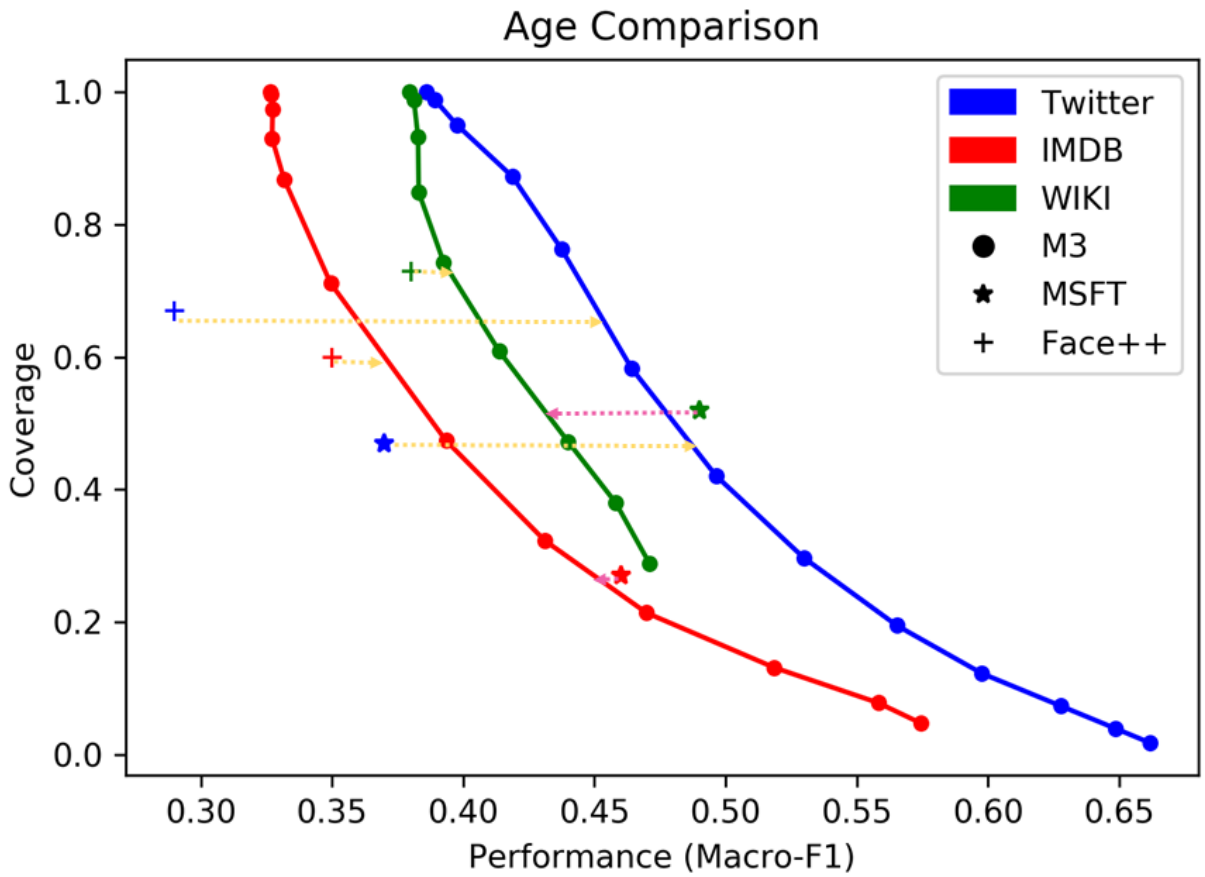
Gender from Text



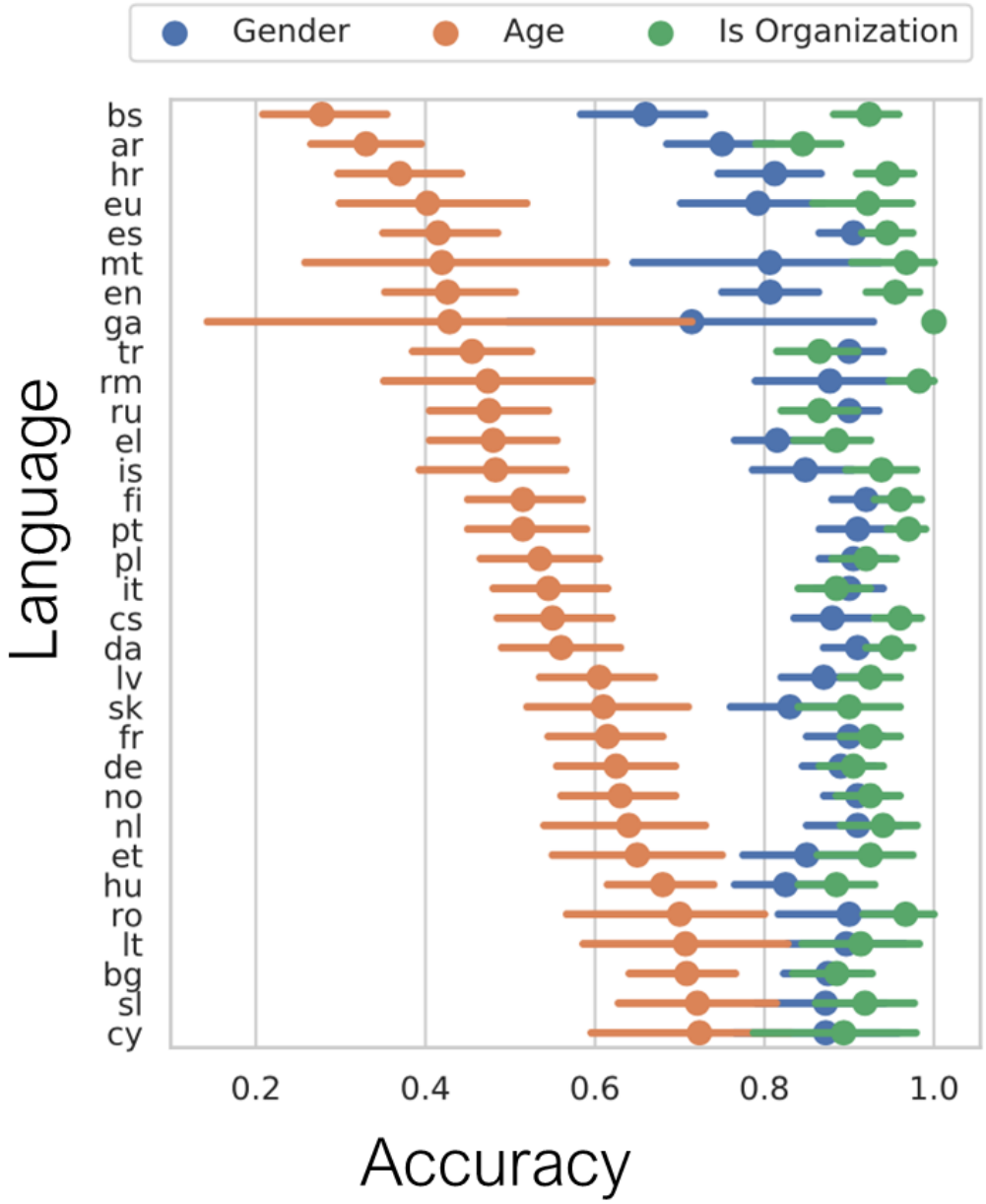
Quick aside: why the increase in coverage with M3?

State of the art for age inference

Age from Image

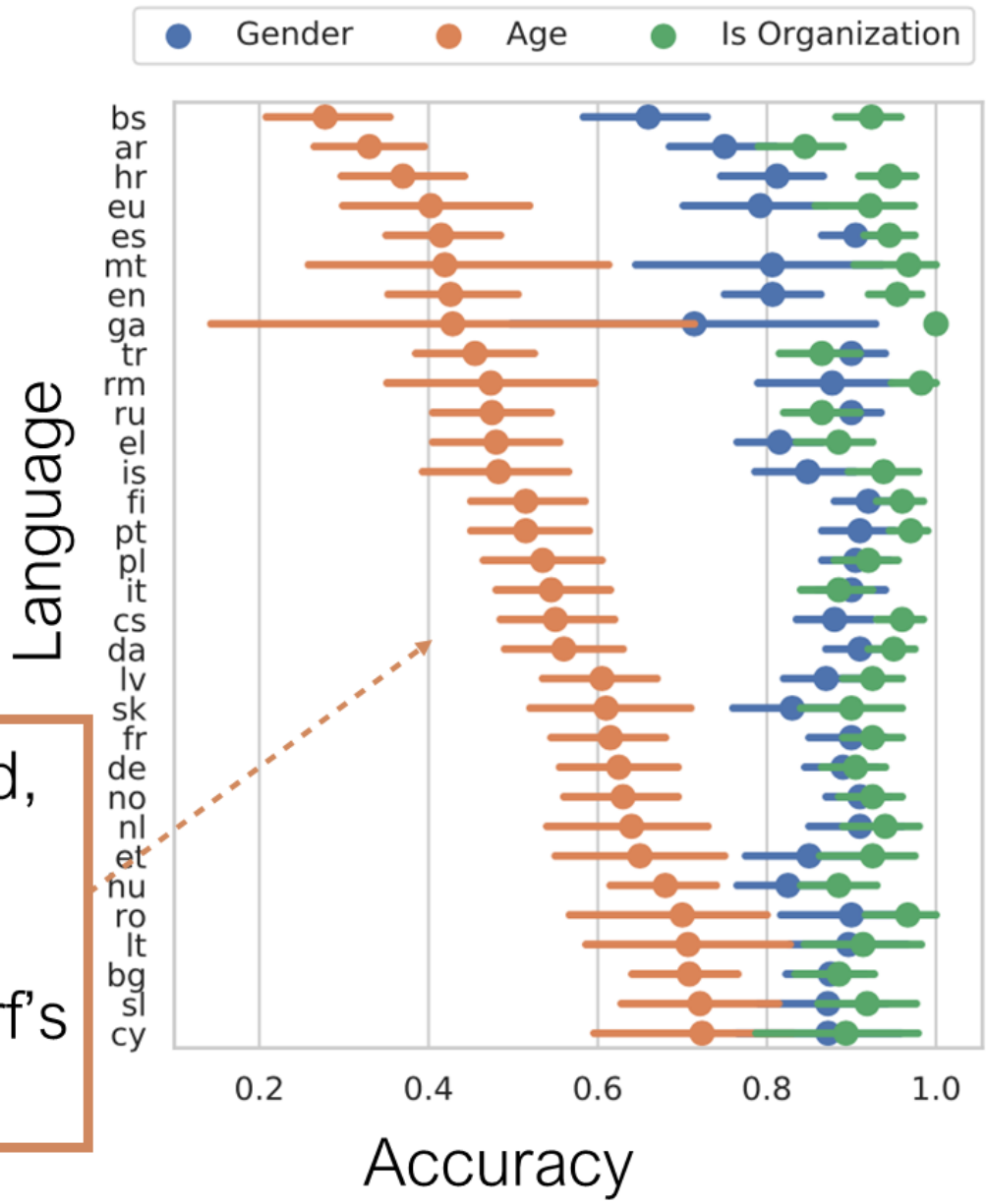


Works pretty well for multiple languages



200 profiles from
32 languages,
3 annotators,
majority voting

Works pretty well for multiple languages

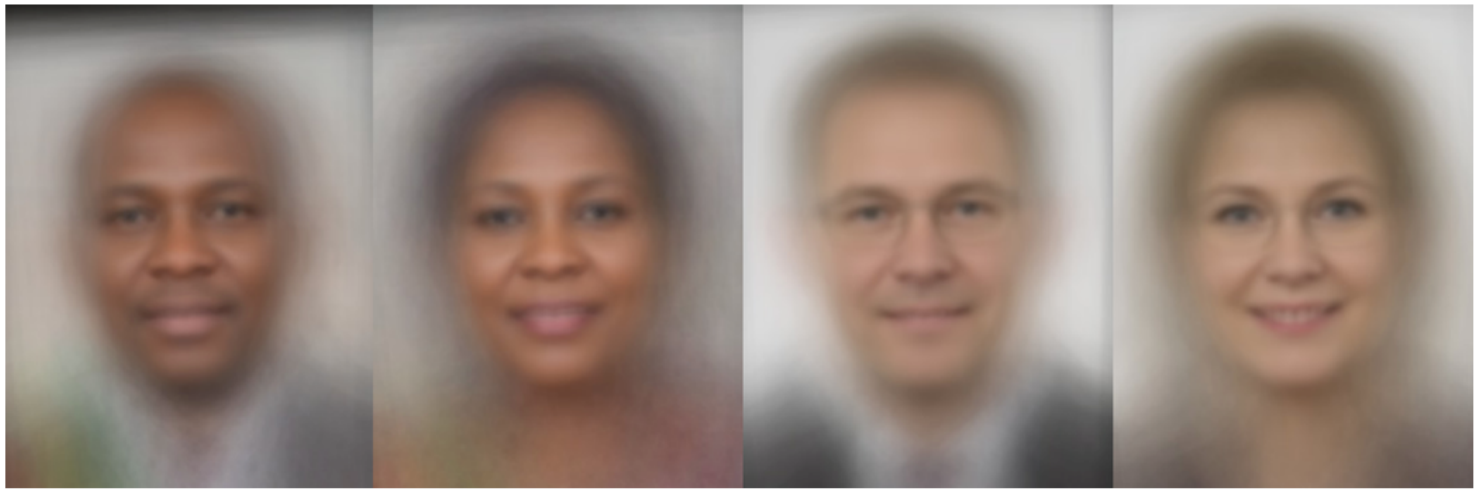


Age is hard, even for humans! (Krippendorf's $\alpha = 0.57$)

200 profiles from 32 languages, 3 annotators, majority voting

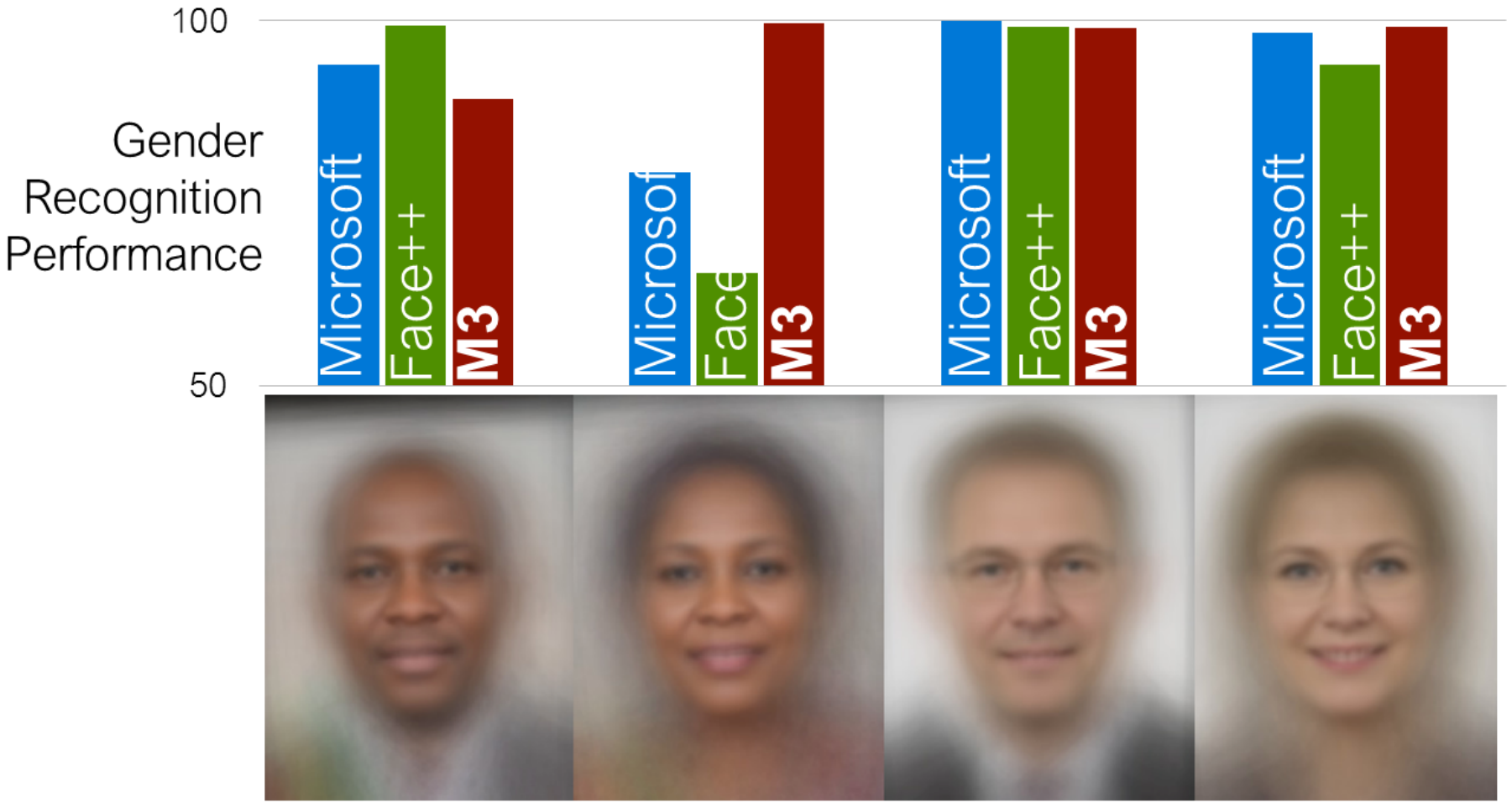
Far less algorithmic bias than current commercial methods

Gender Recognition Performance



data from Buolamwini and Gebru (2018)

Far less algorithmic bias than current commercial methods

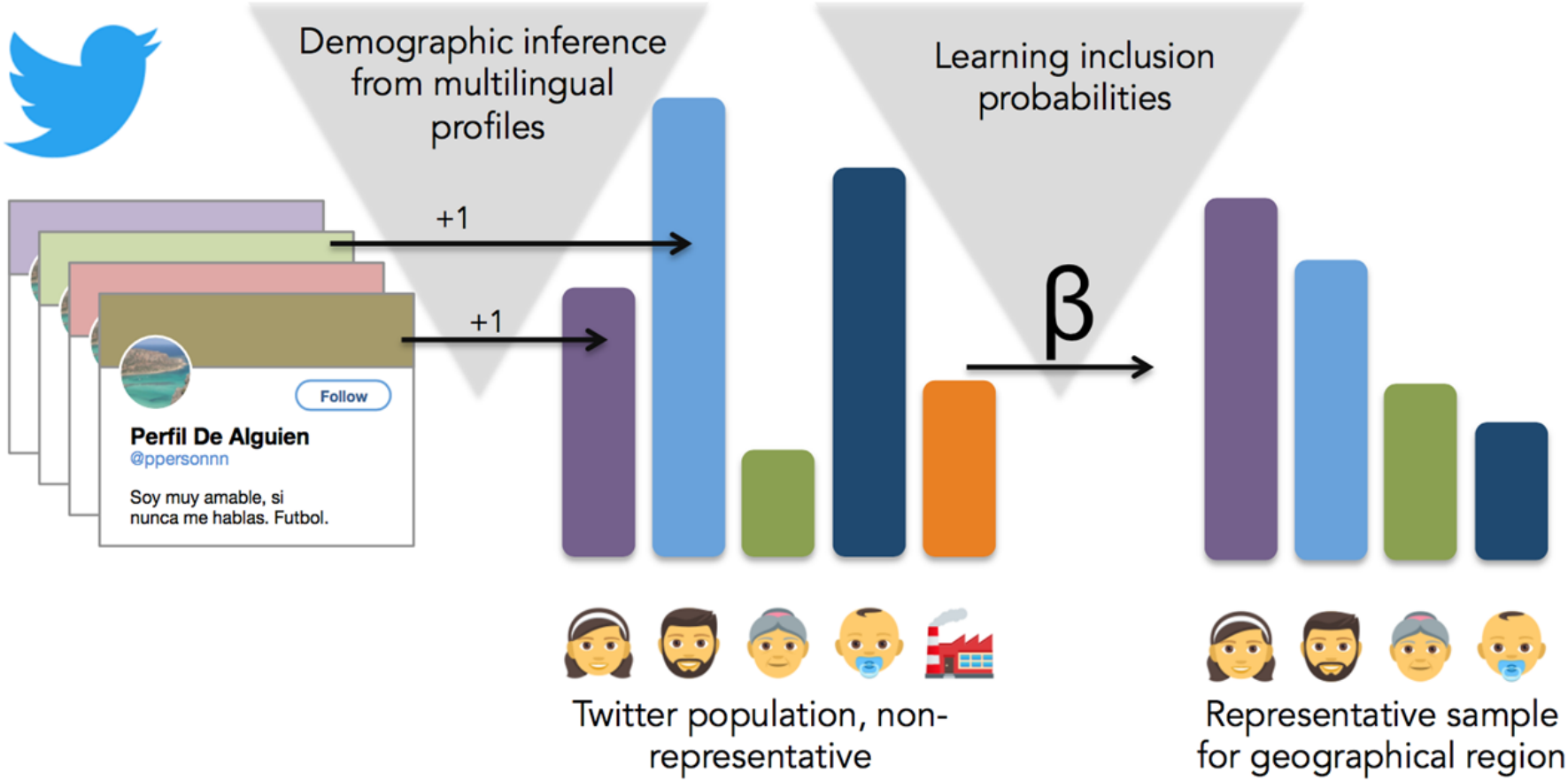


data from Buolamwini and Gebru (2018)

Evaluation summary

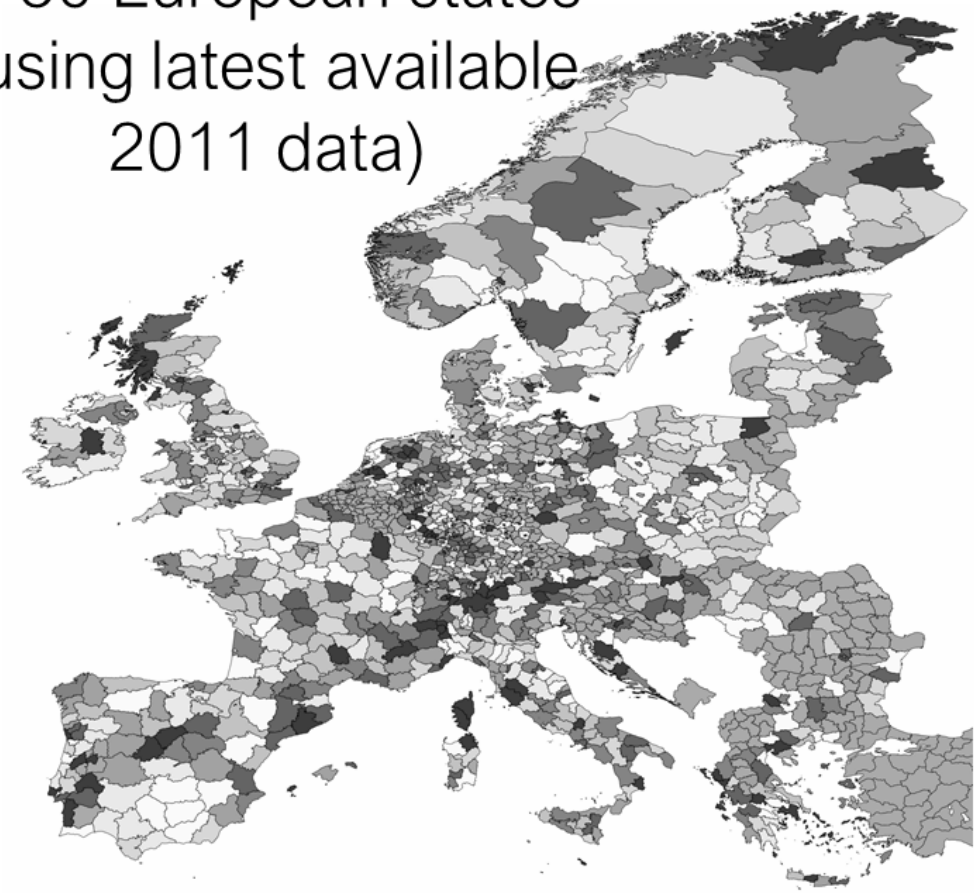
- State-of-the-art results, generally outperforms Face++ and Microsoft Face
- Less biased by skin tones (GenderShades)
- Text-only mode also exceeds comparison systems
- Age is hard, even for humans

Learn the inclusion probabilities for poststratification through population estimation



Population prediction

NUTS3 census regions in 30 European states (using latest available 2011 data)



3,202,964 EU-based Twitter users from September-December 2016

Population prediction

We run simple regression-based approaches using census data (N) and Twitter (M) for NUTS3 level regions across the EU.

NUTS3 regions range in population from 150k to 800k.

We run leave-one-NUTS3-region-out prediction and measure error with *mean absolute percentage error* (MAPE).

An aside: Location

Online social ties can reveal substantial information about a person

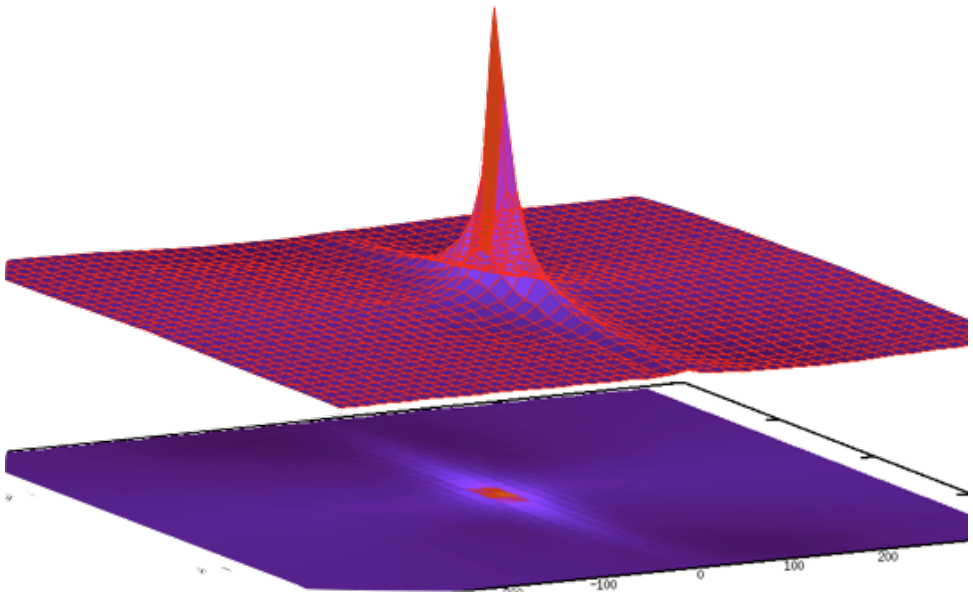
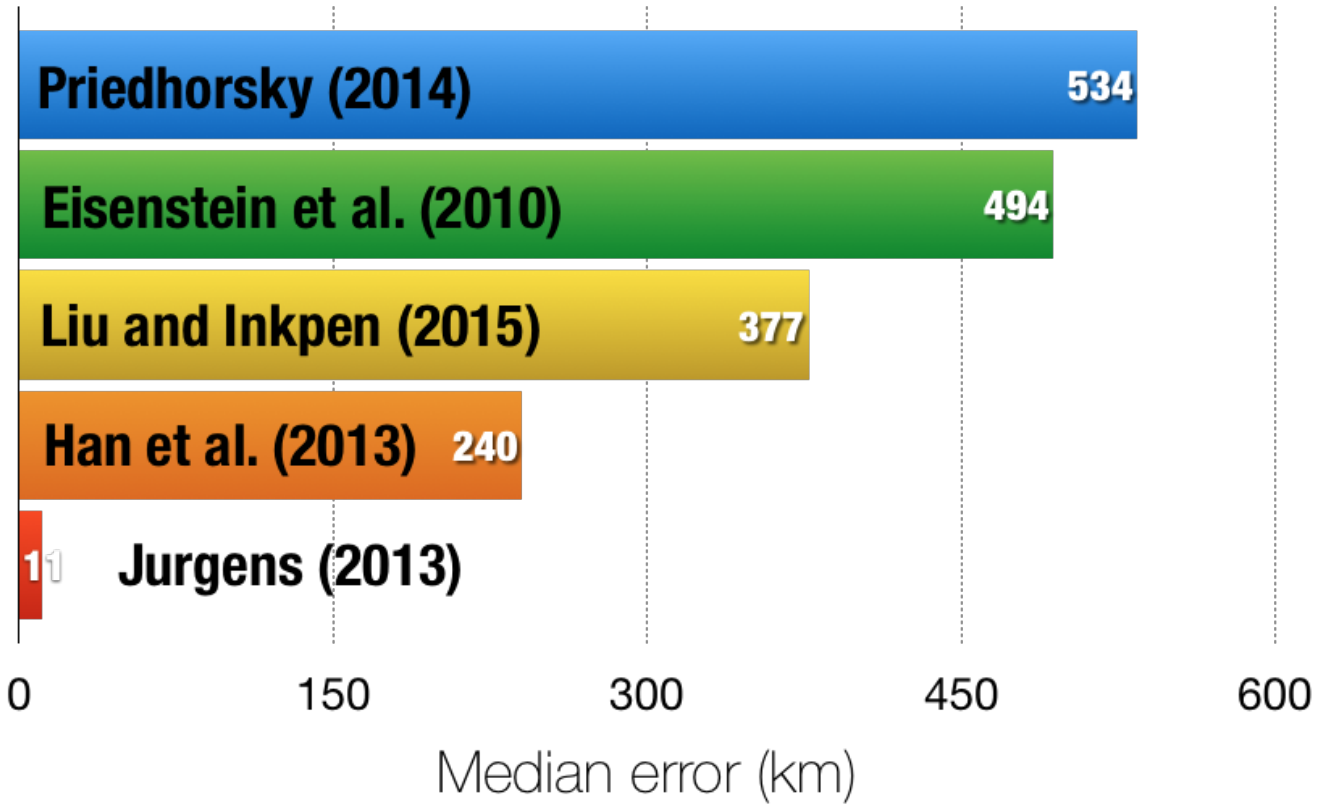


Figure Frequency distribution of where your friends are relative to you

An aside: Location

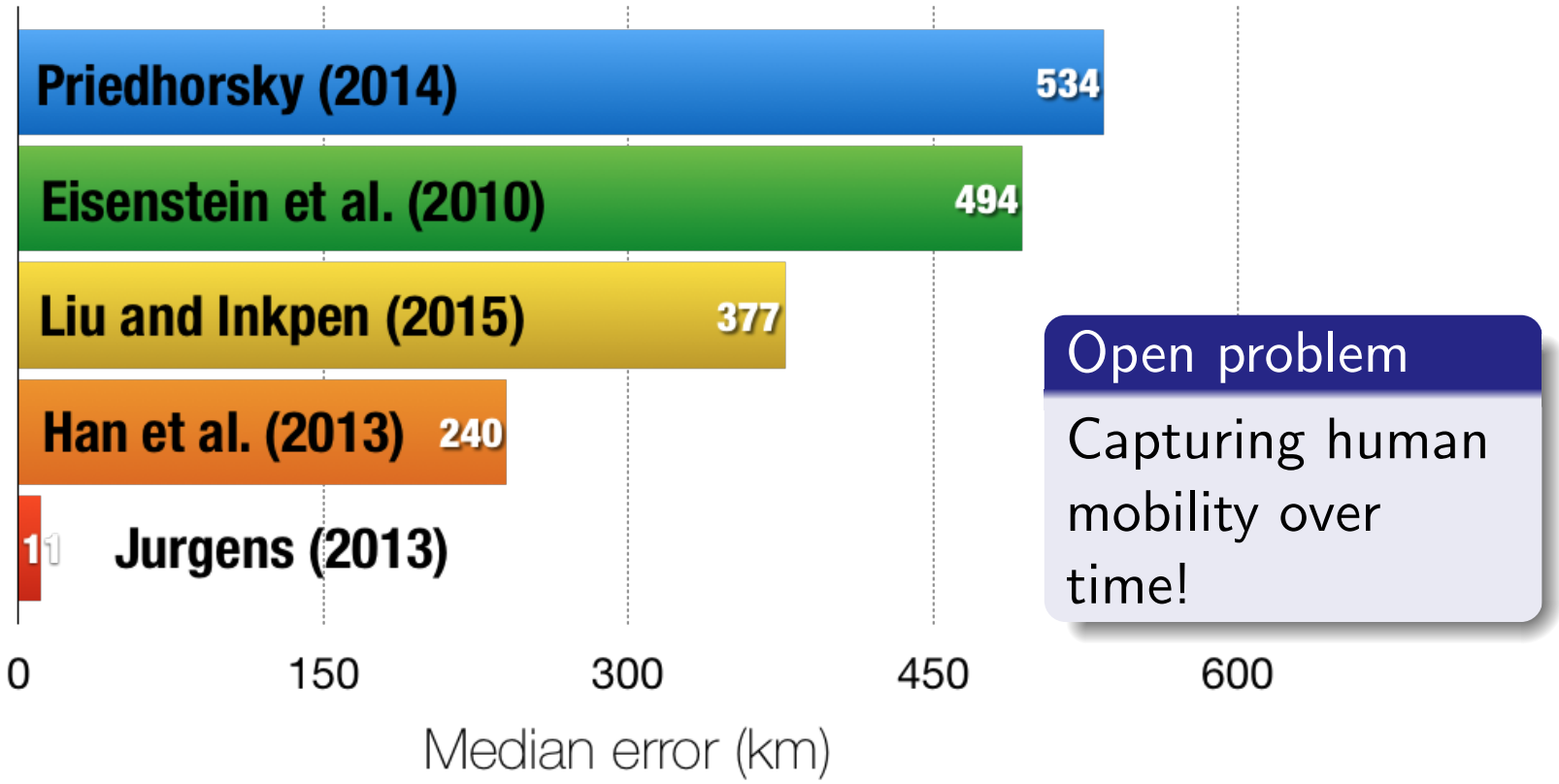
Inferring location on Twitter



That's what friends are for: Inferring location in online communities based on social relationships. David Jurgens. Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM) 2013.

An aside: Location

Inferring location on Twitter



Open problem

Capturing human mobility over time!

That's what friends are for: Inferring location in online communities based on social relationships. David Jurgens. Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM) 2013.

Models

$\mathbf{N \sim M}$ is the base model using only total population from the census (N) and Twitter (M)

$\mathbf{N \sim \sum_g M(g)}$ uses gender marginal counts only (i.e., the total counts of males and females not broken down by ages)

$\mathbf{N \sim \sum_a M(a)}$ uses age marginal counts only

$\mathbf{N \sim \sum_{a,g} M(a,g)}$ uses the joint histograms inferred from Twitter but only the total population values from the census

$\mathbf{\log N(a,g) \sim \log M(a,g) + a + g}$ uses the joint histograms inferred from Twitter and the joint histograms from the census

Results

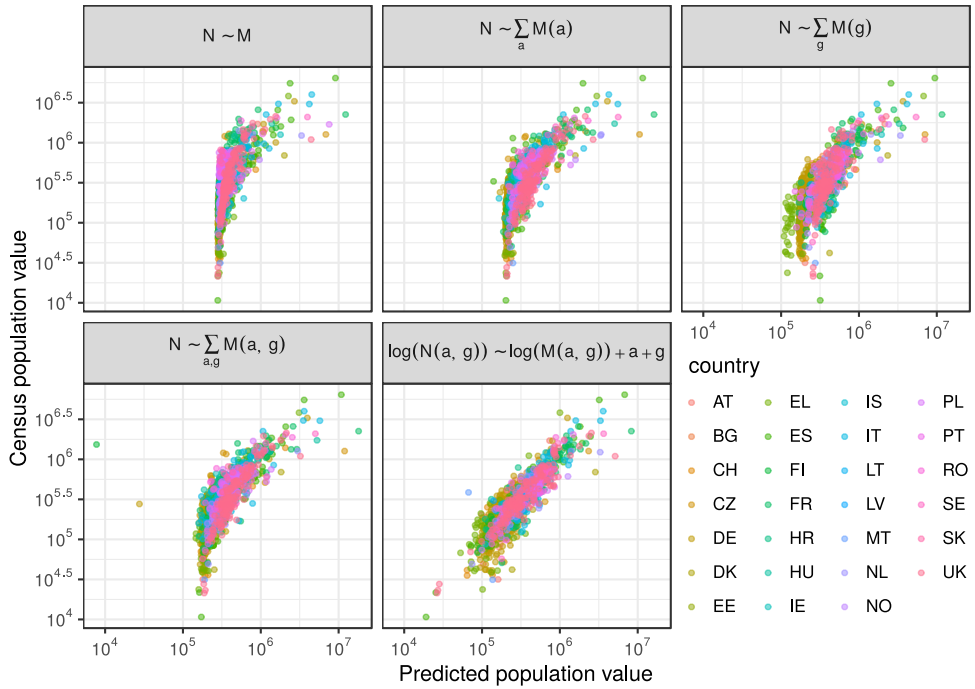
$\mathbf{N \sim M}$ is the base model using only total population from the census (N) and Twitter (M)

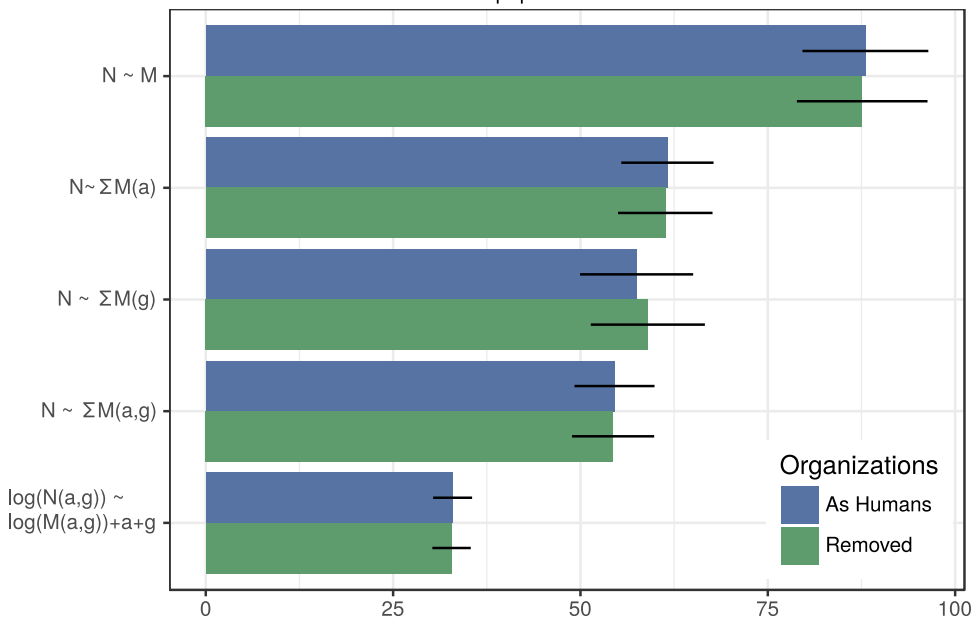
$\mathbf{N \sim \sum_g M(g)}$ uses gender marginal counts only (i.e., the total counts of males and females not broken down by ages)

$\mathbf{N \sim \sum_a M(a)}$ uses age marginal counts only

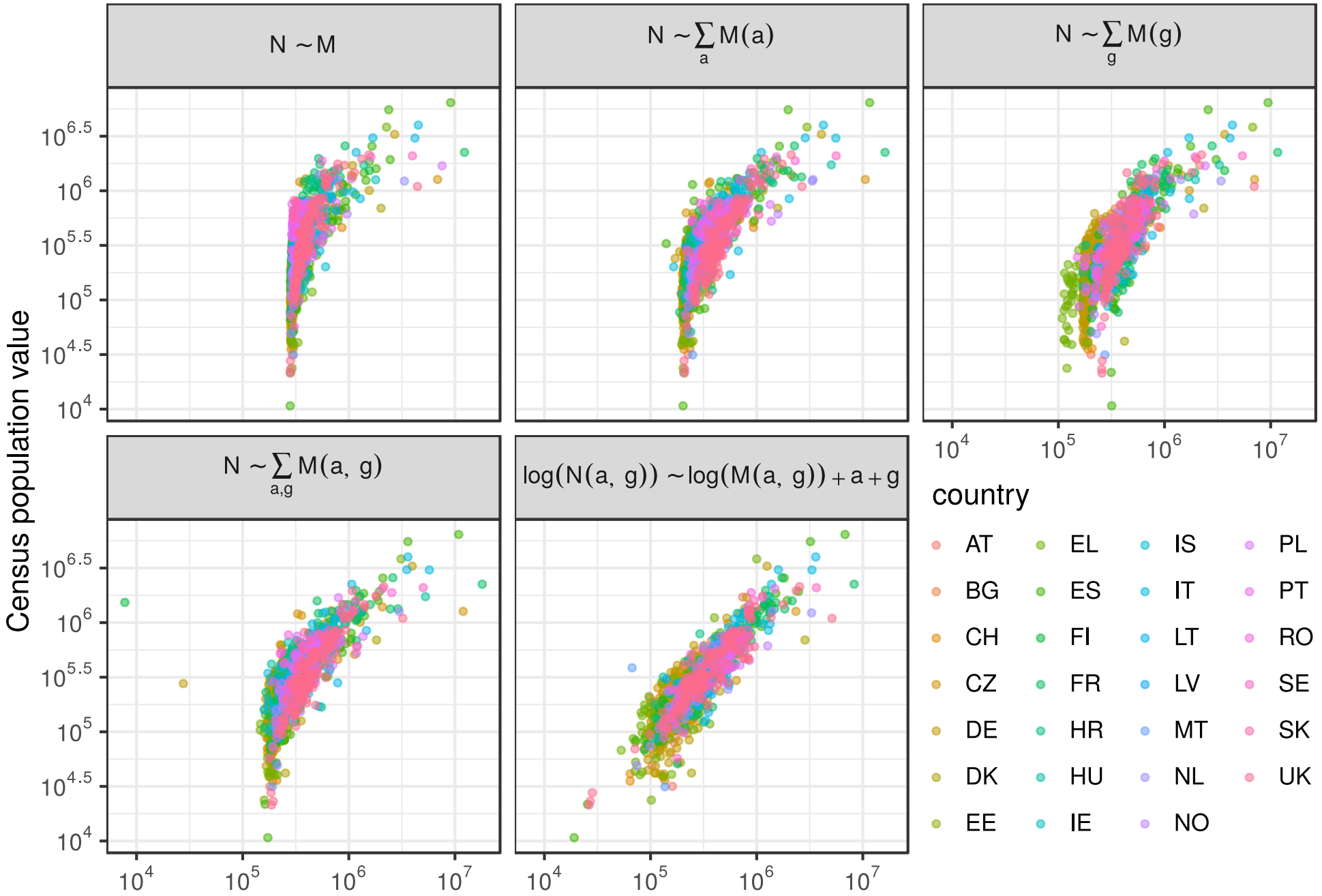
$\mathbf{N \sim \sum_{a,g} M(a,g)}$ uses the joint histograms inferred from Twitter but only the total population values from the census

$\mathbf{\log N(a,g) \sim \log M(a,g) + a + g}$ uses the joint histograms inferred from Twitter and the joint histograms from the census

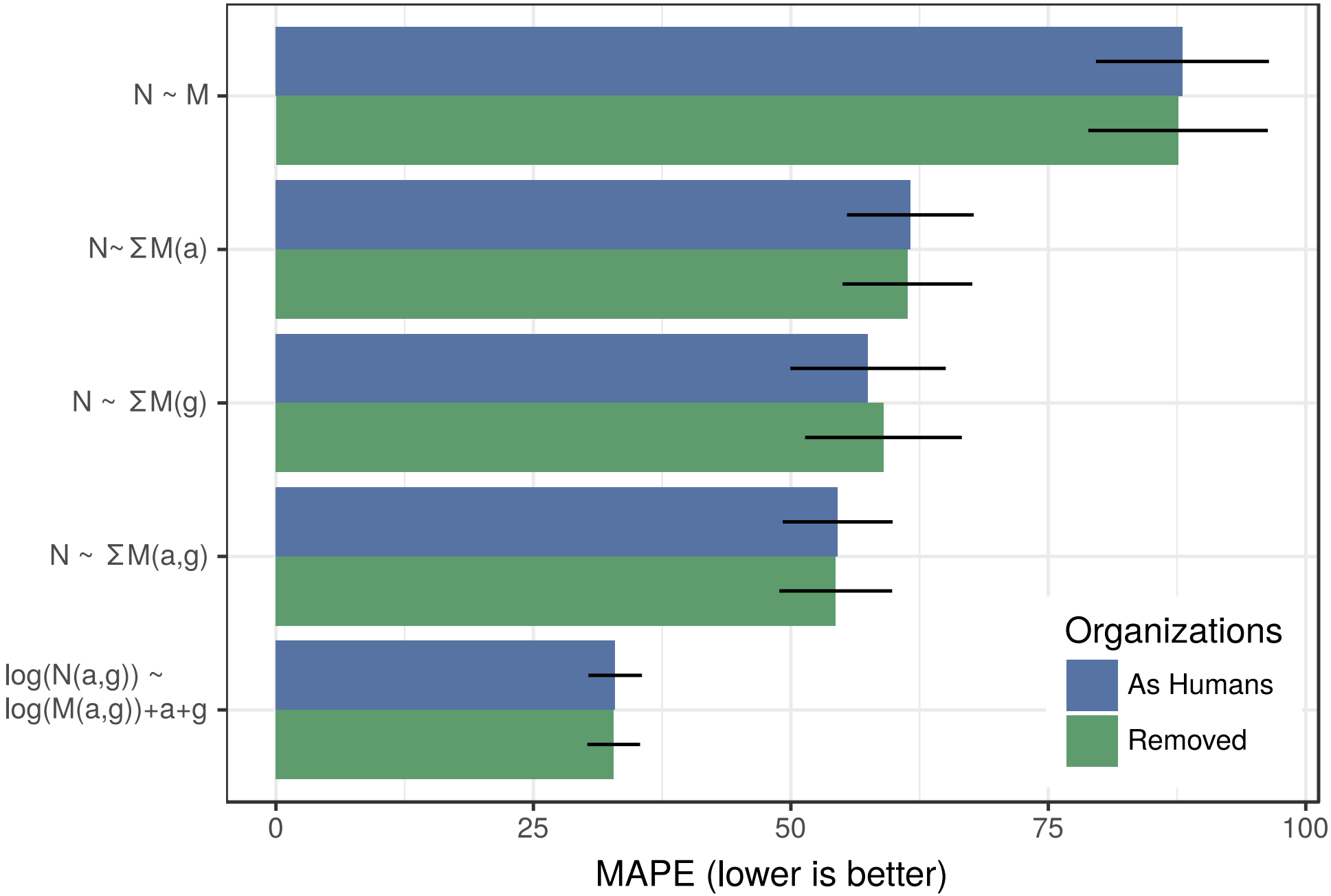




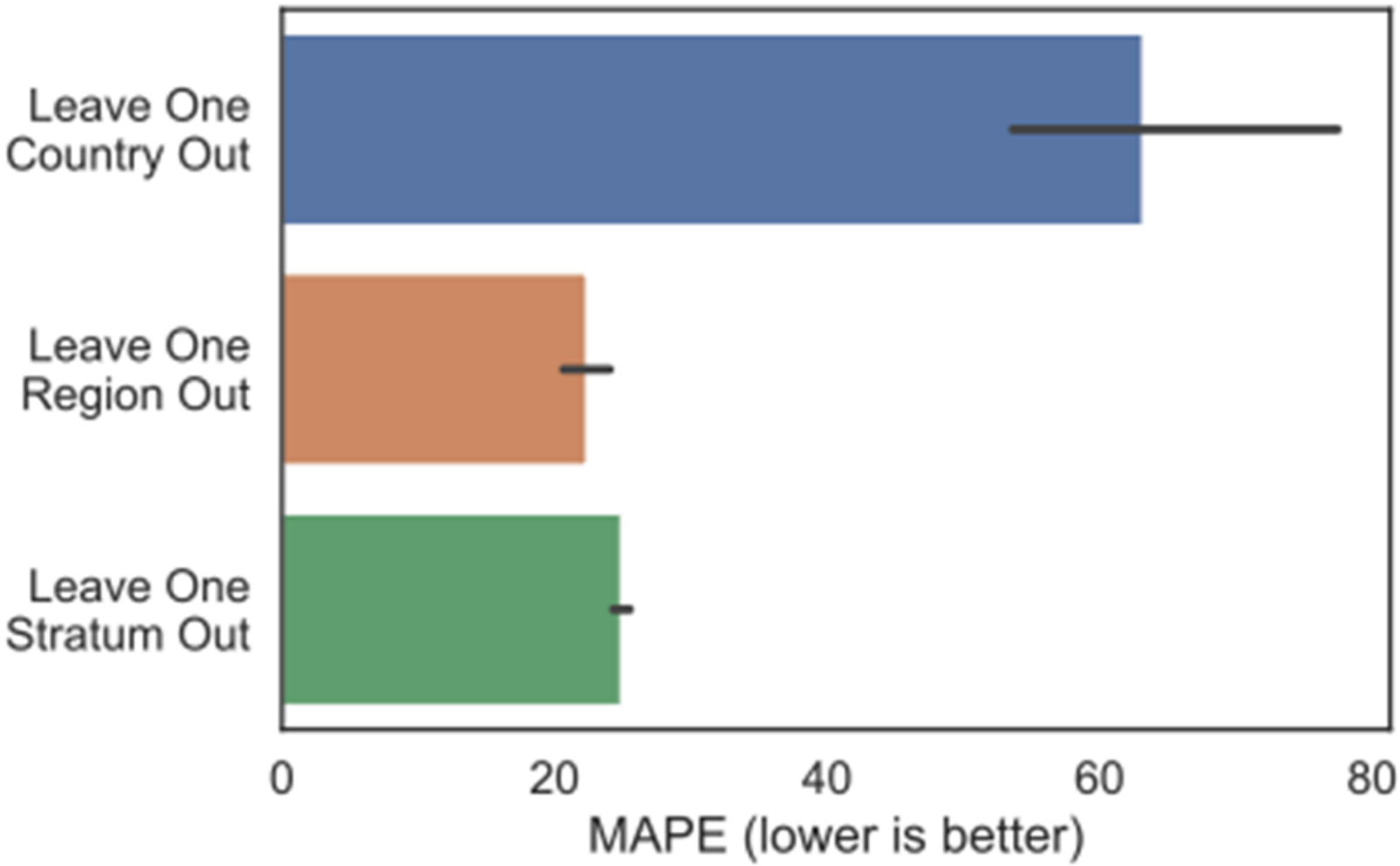
Results


N ~ M
N ~ ∑a M(a)
N ~ ∑g M(g)
N ~ ∑a,g M(a, g)
log(N(a, g)) ~ log(M(a, g)) + a + g
Census population value
country
AT
BE
BG
CH
CZ
DE
DK
EE
EL
ES
FI
FR
HR
HU
IE
IS
IT
LT
LV
MT
NL
NO
PL
PT
RO
SE
SK
UK

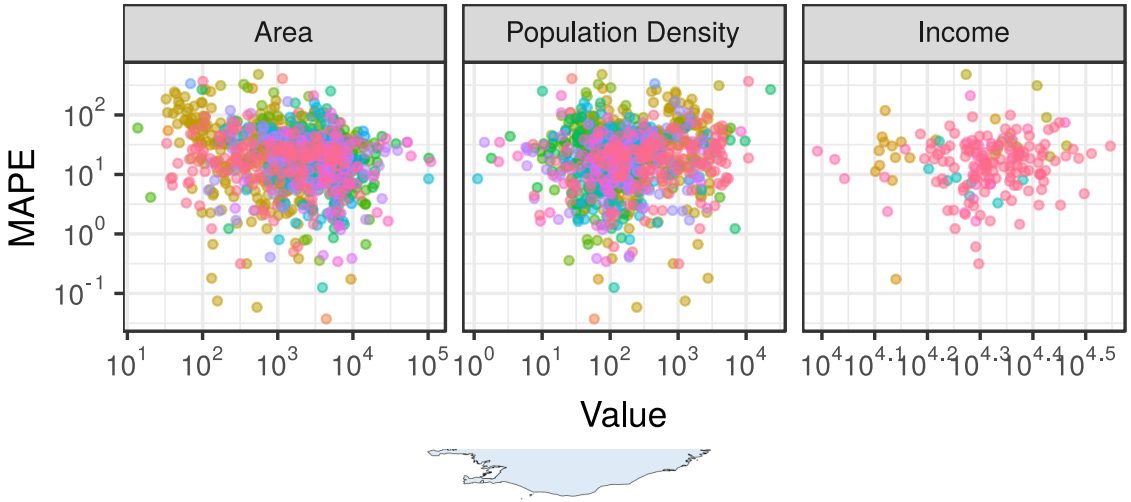
Results

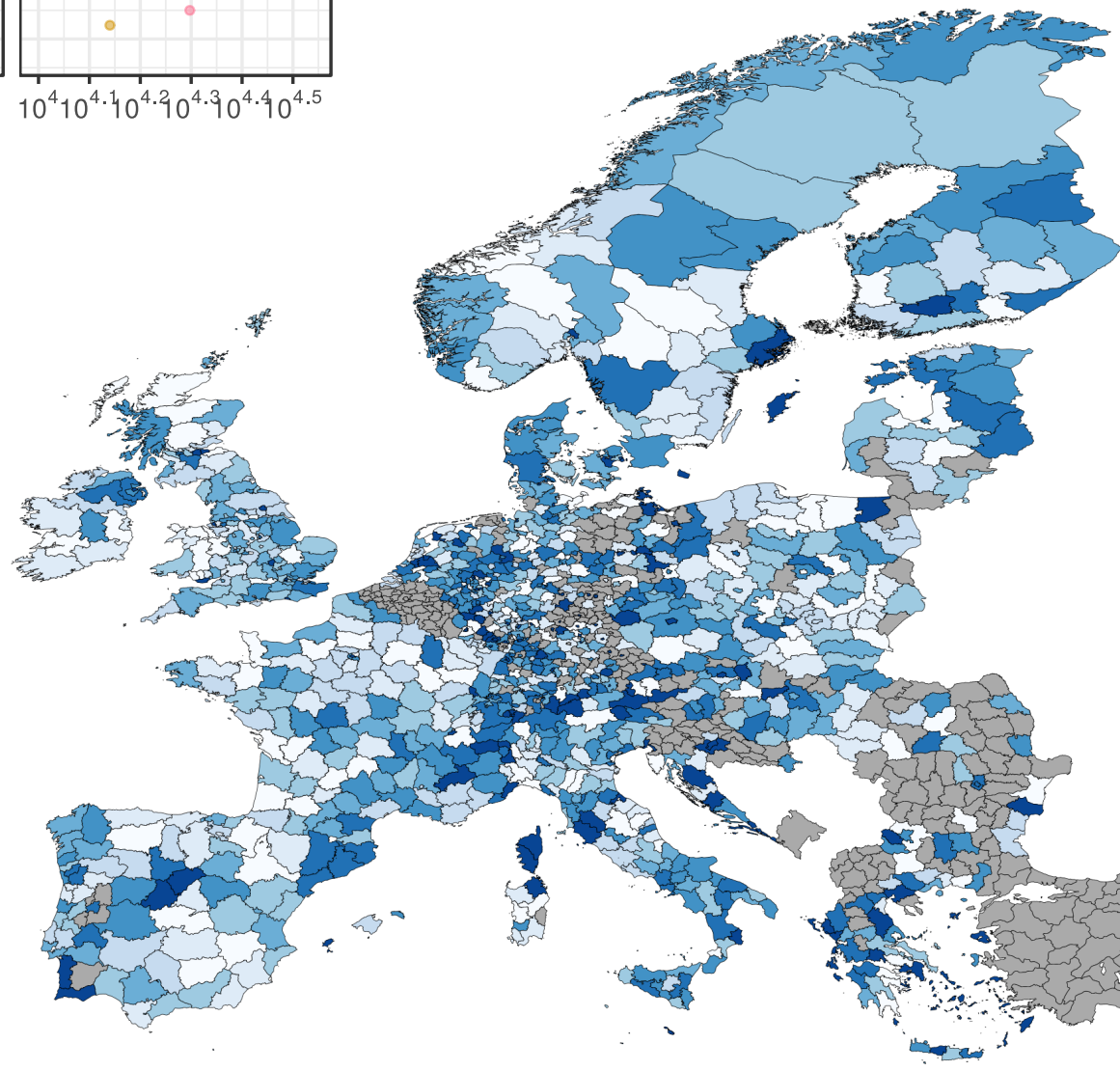

N ~ M
N~ ΣM(a)
N ~ ΣM(g)
N ~ ΣM(a,g)
log(N(a,g)) ~
log(M(a,g))+a+g
0
25
50
75
100
MAPE (lower is better)
Organizations
As Humans
Removed

Inclusion weights generalize to population estimation on held-out data



Sources of error





What have we learnt?

- Possible to infer demographics/org-status fairly reliably in multilingual settings

What have we learnt?

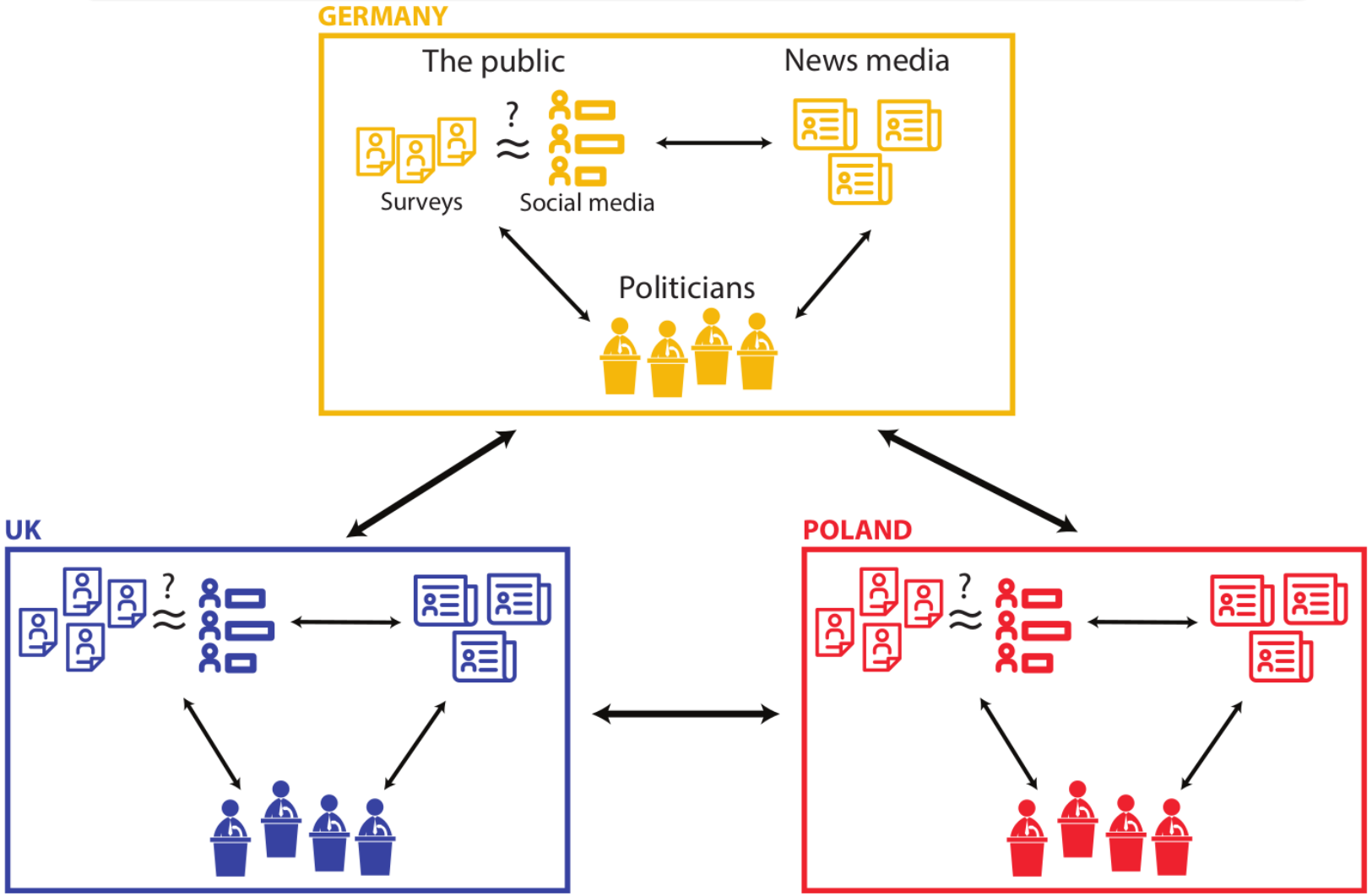
- Possible to infer demographics/org-status fairly reliably in multilingual settings
- Reweighting helps significantly, but...
  - Heterogeneity in social media inclusion probabilities that needs to be accounted for, especially in multicountry setting
  - More analysis needed to understand how reweighting helps with specific dependent variables

What have we learnt?

- Possible to infer demographics/org-status fairly reliably in multilingual settings
- Reweighting helps significantly, but...
  - Heterogeneity in social media inclusion probabilities that needs to be accounted for, especially in multicountry setting
  - More analysis needed to understand how reweighting helps with specific dependent variables
- Open problems
  - Capturing human mobility over time
  - Non-independence of social media observations (e.g., homophily)
  - How to best include probability output of M3 inference into downstream tasks

Back to measuring the attention to societal issues

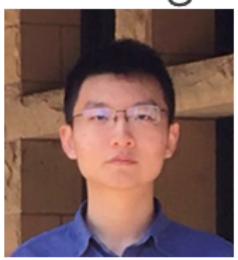
Comparing political, media, and public 'agendas' across languages and countries within the EU, over time.



VolkswagenStiftung

Demographic Inference and Representative Population Estimates from Multilingual Social Media Data

Zijian Wang



EECS

Scott Hale

OXFORD INTERNET INSTITUTE

David Adelani

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

Przemyslaw Grabowicz

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

Timo Hartmann

gesis Leibniz-Institut für Sozialwissenschaften

Fabian Flöck

gesis Leibniz-Institut für Sozialwissenschaften

David Jurgens

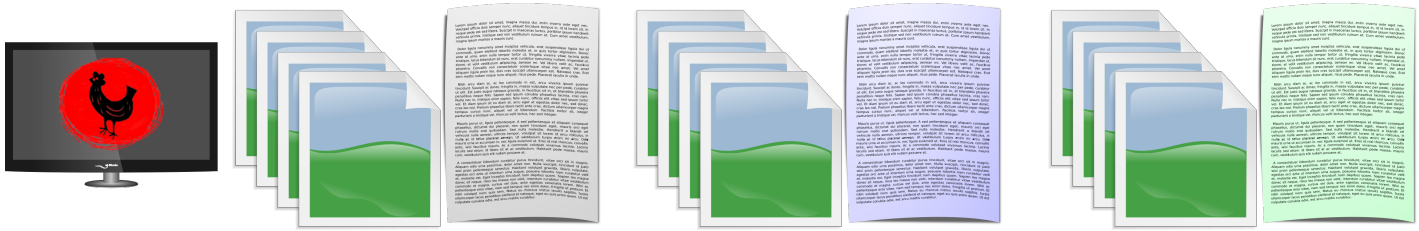
UMSI

In Proceedings of the Web Conference (WWW), 2019

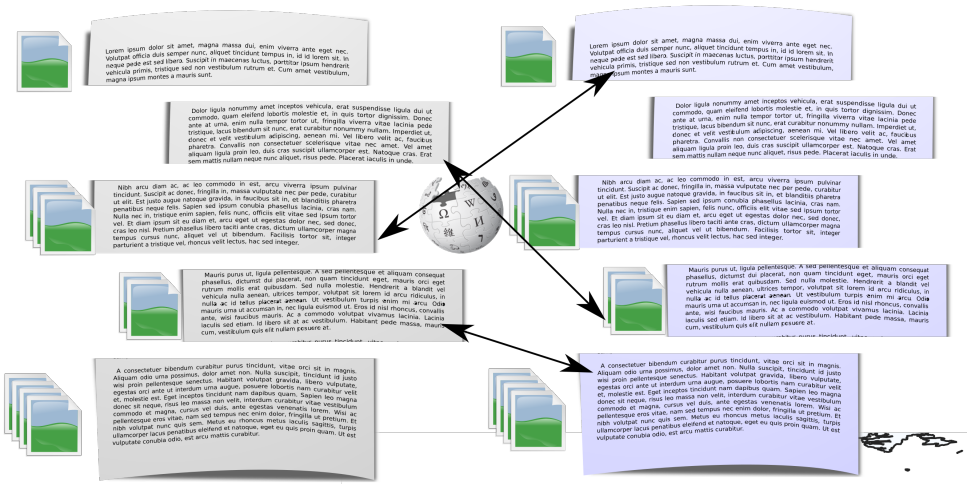
New resources for the community

- M3 model as off-the-shelf python library
  `pip install m3inference`
  https://github.com/euagendas/m3inference
- Inclusion probabilities for Europe and eventually other regions
  http://euagendas.org/inclusionprobs
- Live demo of M3 here: http://m3.euagendas.org/
- Everything is at https://github.com/euagendas/

Measuring the media agenda

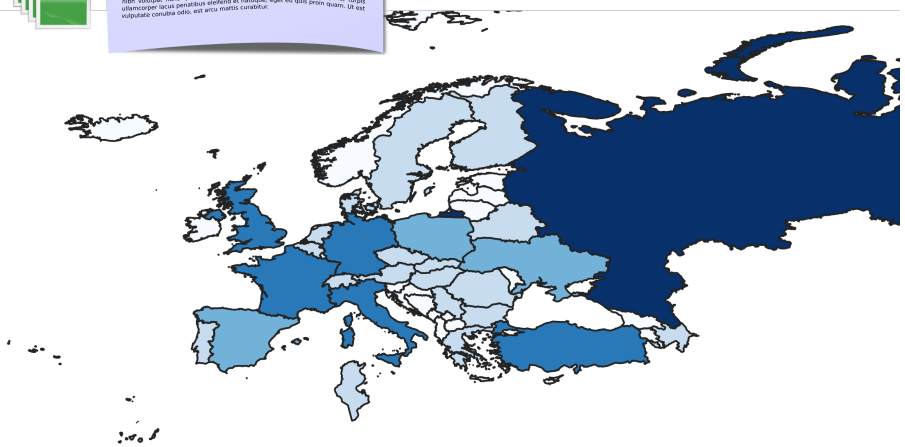


Weekly/nighly news across EU+Russia

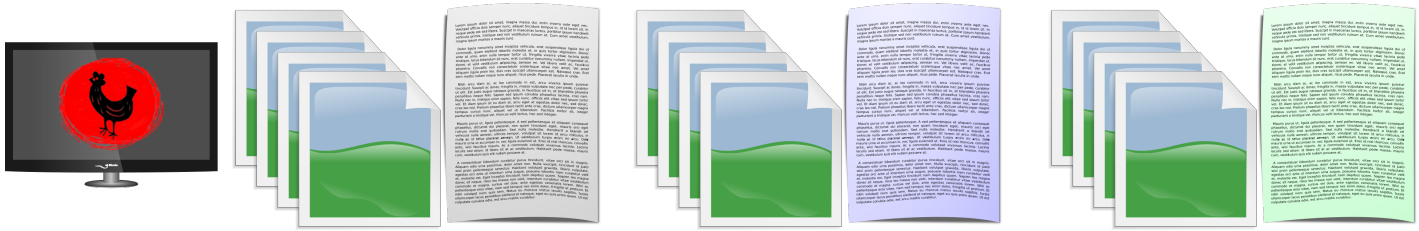


Segment and match news items

Compare similarity of news within and between countries



Measuring the media agenda

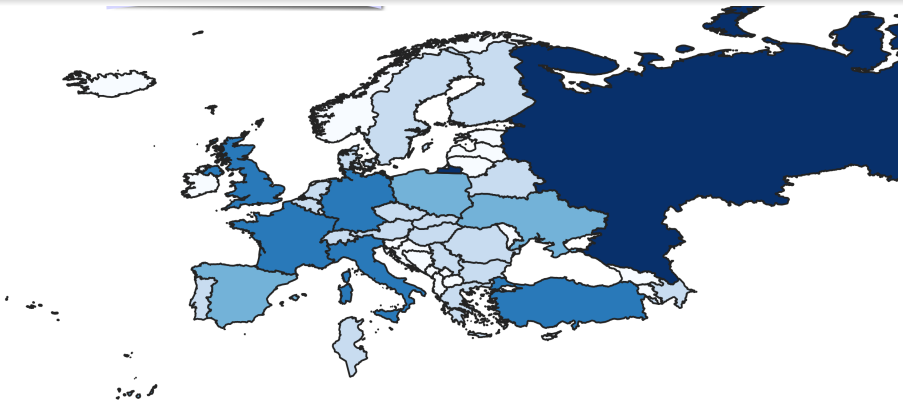


Weekly/nightly news across EU + Russia

Cross-national media agenda similarity

- Shared language
- Geographic proximity
- Past migration
- Economic similarity
- Temporal factors (e.g., GDP)

Compare similarity of news within and between countries



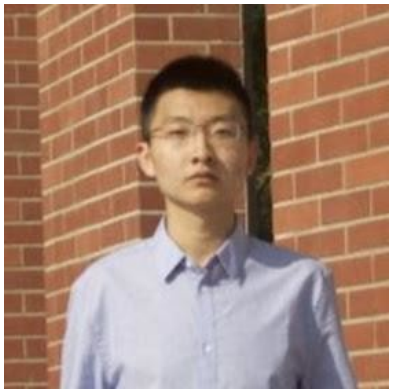
Thank you!

Scott Hale

David Jurgens

Fabian Flöck

Przemyslaw Grabowicz



Zijian Wang

David Adelani

Timo Hartmann

Chico Camargo

Research questions

RQ1 What differences exist in the media, political, and public agendas across nations and languages and how do they evolve over time in comparison to each other?

RQ2 To what extent do social media reflect a broader public agenda?

RQ3 How do the agendas of actors in a specific national or linguistic context influence how agendas are adopted in other contexts, and what explains these influences?