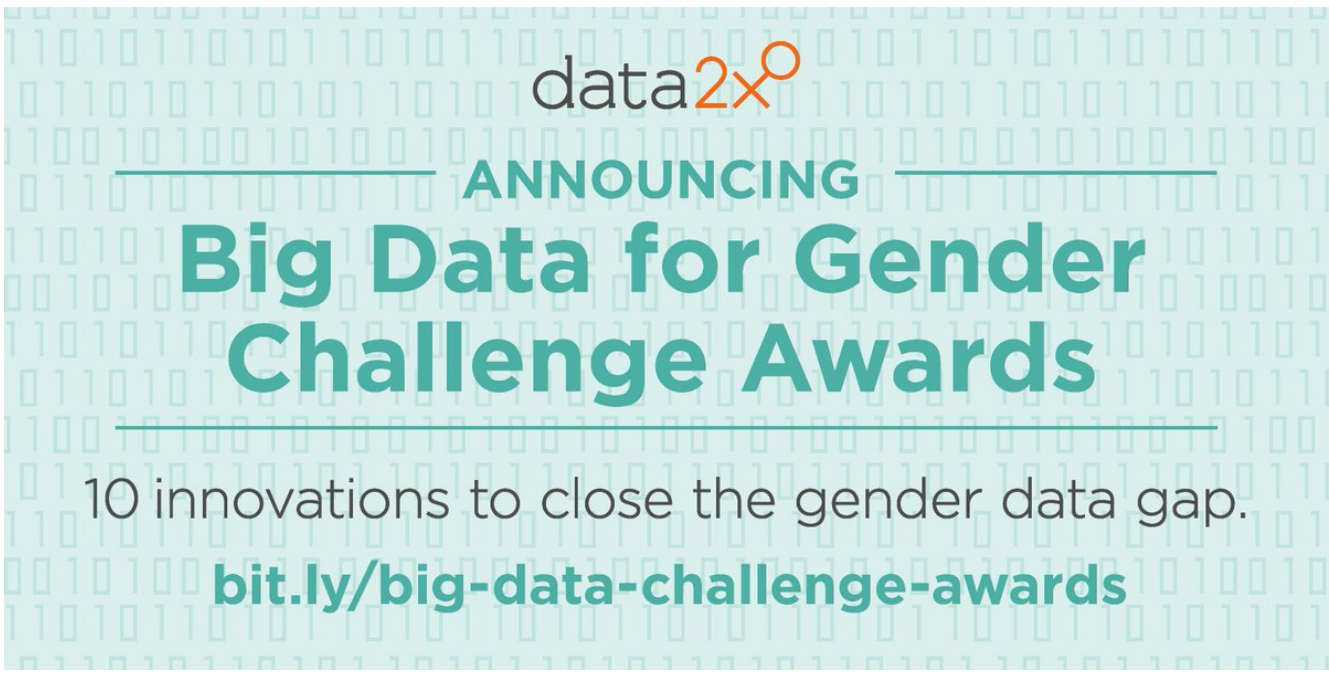
# Measuring global gender inequality indicators with large-scale online advertising data

Ridhi Kashyap

University of Oxford

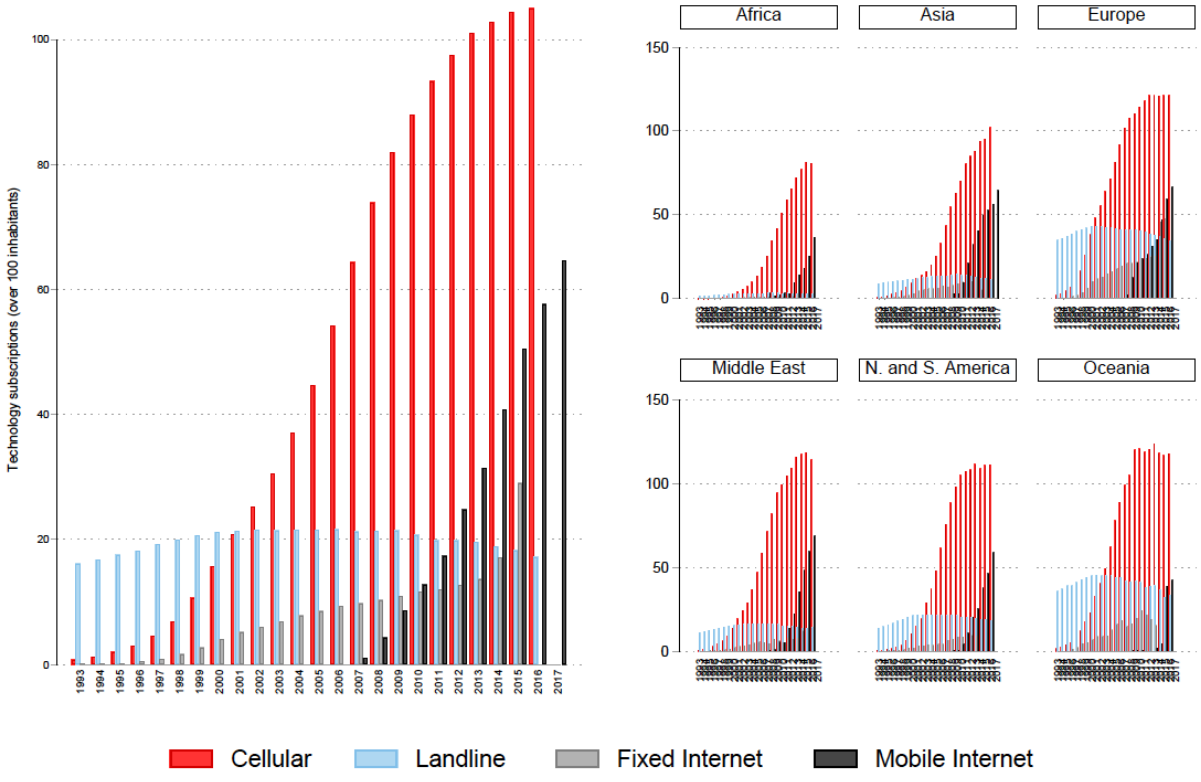
SICSS Bamberg
August 2, 2019

# Support




**The Digital Traces for the Gender Digital Divide** at the University of Oxford (Grant No. UNF-17-936).

# The Digital Revolution



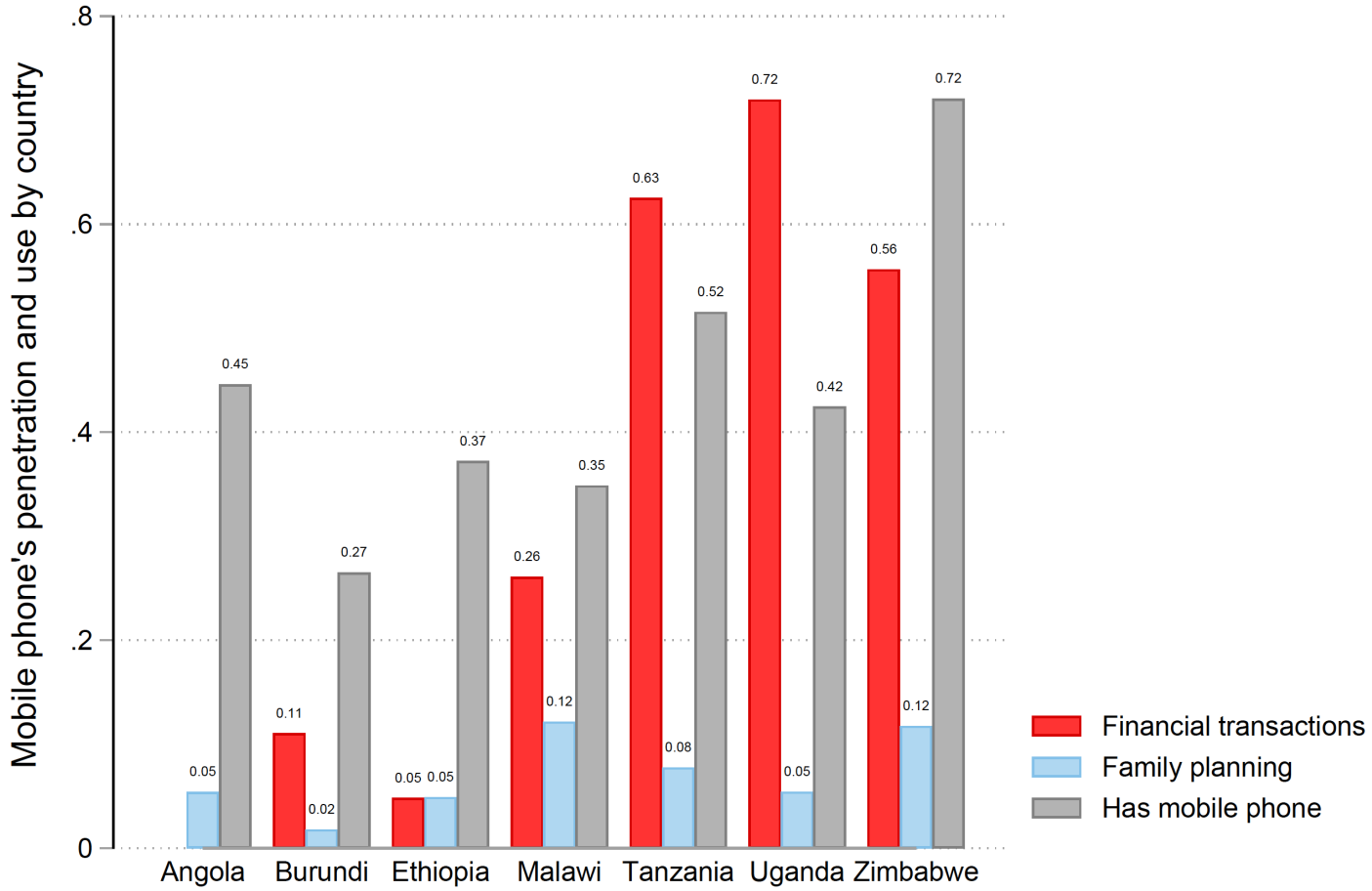
# The Digital Revolution

- The digital revolution is a:
  - **Social revolution:** technology that has the potential to empower marginalised populations and facilitate the attainment of sustainable development goals (SDGs).
  - **Data Revolution:** filling data gaps and measuring progress on the SDGs.

# Sustainable Development Goals

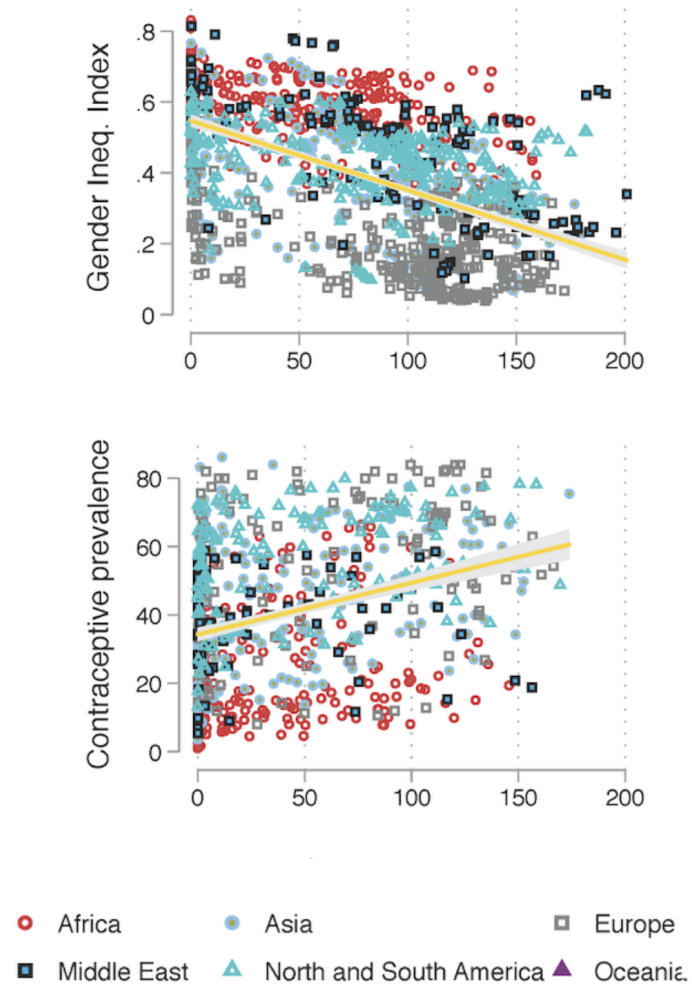
- Adopted by the UN in January 2016 as 17 goals to guide international development policy for the next 15 years.
- Successor to Millennium Development Goals that were adopted in 2000.

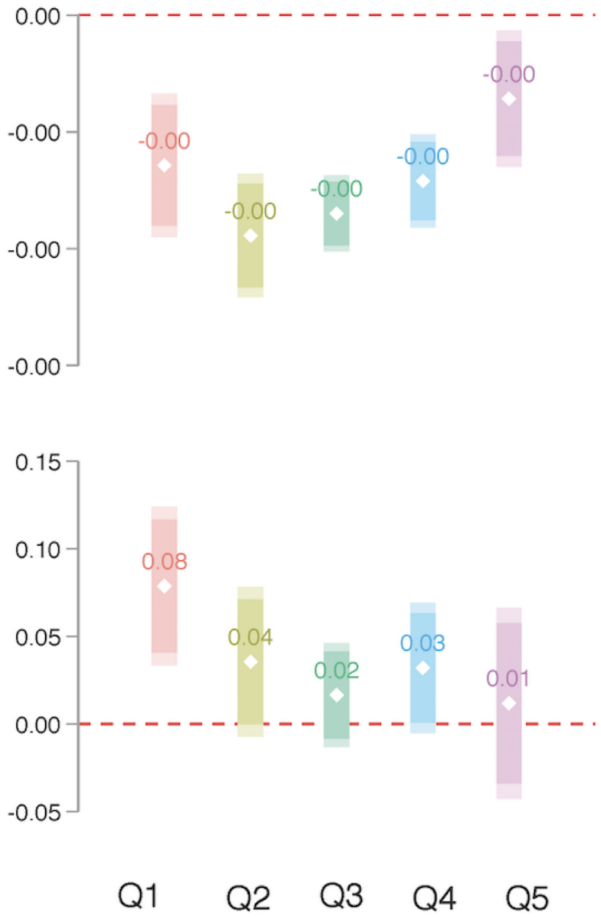
# The Digital Revolution and its Implications




[1] Rotondi, Kashyap, Pesando, Spinelli and Billari. "Leveraging mobile phones to attain sustainable development."

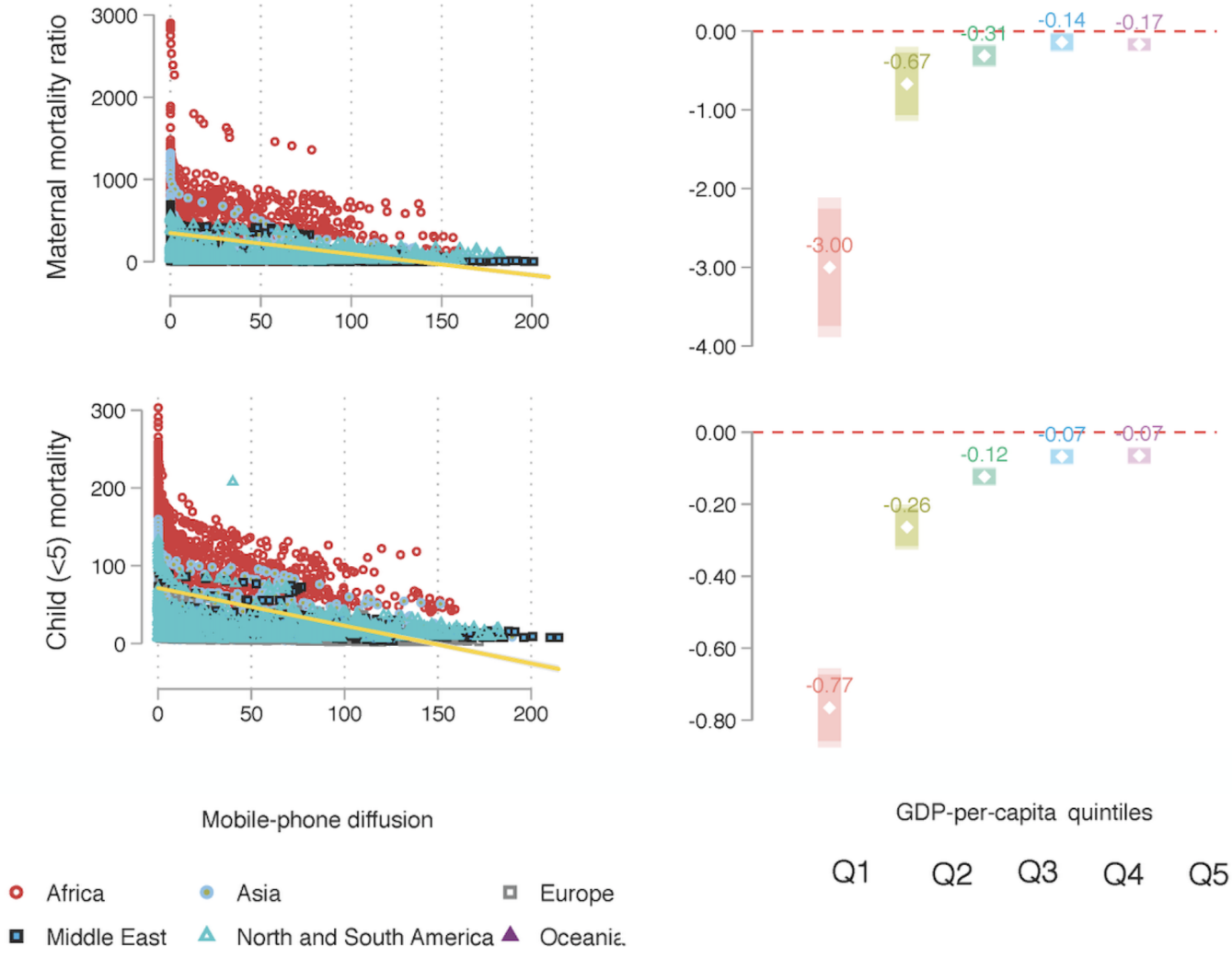
# The Digital Revolution and its Implications




[1] Rotondi, Kashyap, Pesando, Spinelli and Billari. "Leveraging mobile phones to attain sustainable development."

# The Digital Revolution and its Implications




[1] Rotondi, Kashyap, Pesando, Spinelli and Billari. "Leveraging mobile phones to attain sustainable development."

# The Digital Revolution and its Implications

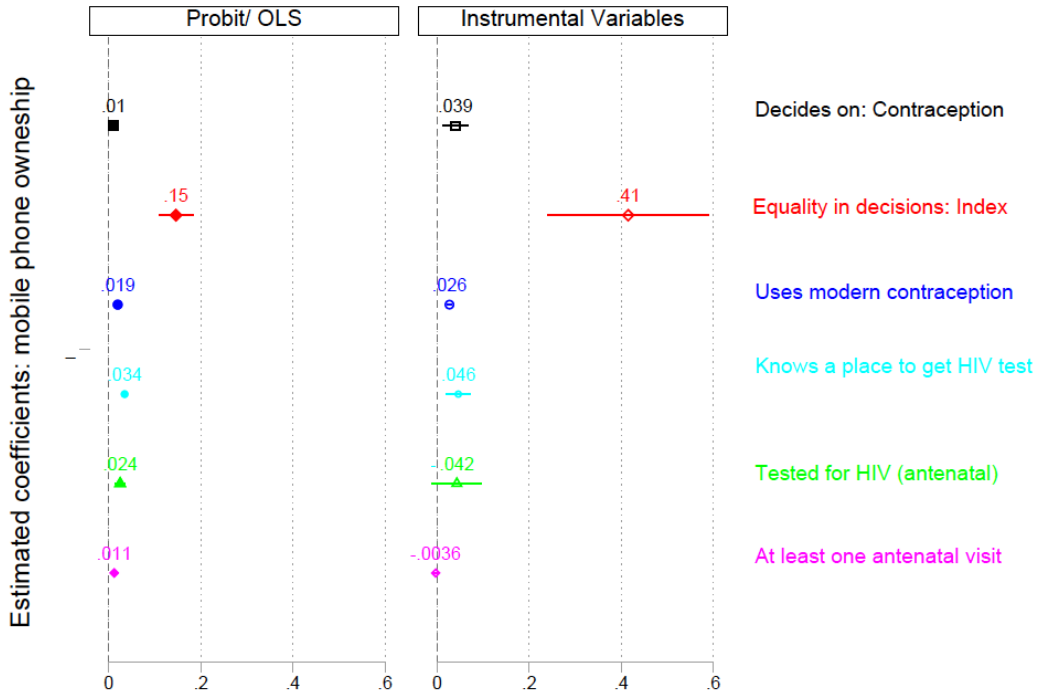



Figure: Individual-level effects of mobile-phone ownership on development outcomes in Sub-Saharan Africa

[1] Rotondi, Kashyap, Pesando, Spinelli and Billari. "Leveraging mobile phones to attain sustainable development."

# Digital Revolution as Data Revolution

## The "Big Data" Era

"An explosion in the **volume** of data, the **speed** with which data are produced, the **number of producers** of data, the dissemination of data, and the **range of things** on which there is data, **coming from new technologies** ..."

– *A World that Counts*, p. 6.

`http://www.undatarevolution.org`

# The Data Revolution

## The Vision

"The **integration of these new data with traditional data** to produce high-quality information that is more detailed, timely and relevant for many purposes and users, especially to **foster and monitor sustainable development**."

– *A World that Counts*, p. 6.

`http://www.undatarevolution.org`

# The Data Revolution

## Data Gaps

- **Coverage and Lags:** Too many countries still have poor data, data arrives too late and too many issues are still barely covered by existing data.
- **Disaggregation:** Large groups of people remain invisible. Gender-disaggregated data are significantly lacking.

# Gender Data Gaps

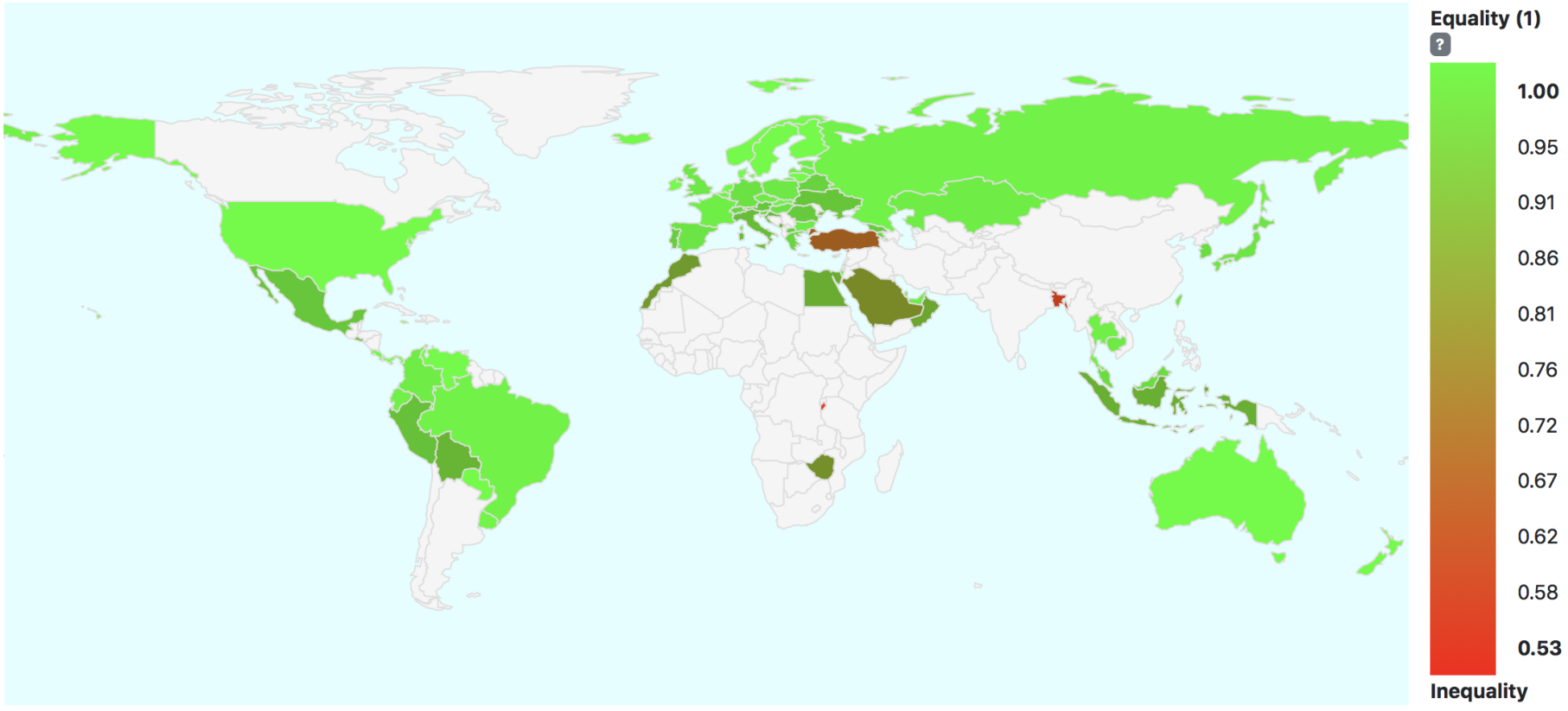



Figure: Gender gaps in internet use computed using data from International Telecommunications Union (ITU) available at `www.digitalgendergaps.org`

# Gender Data Gaps

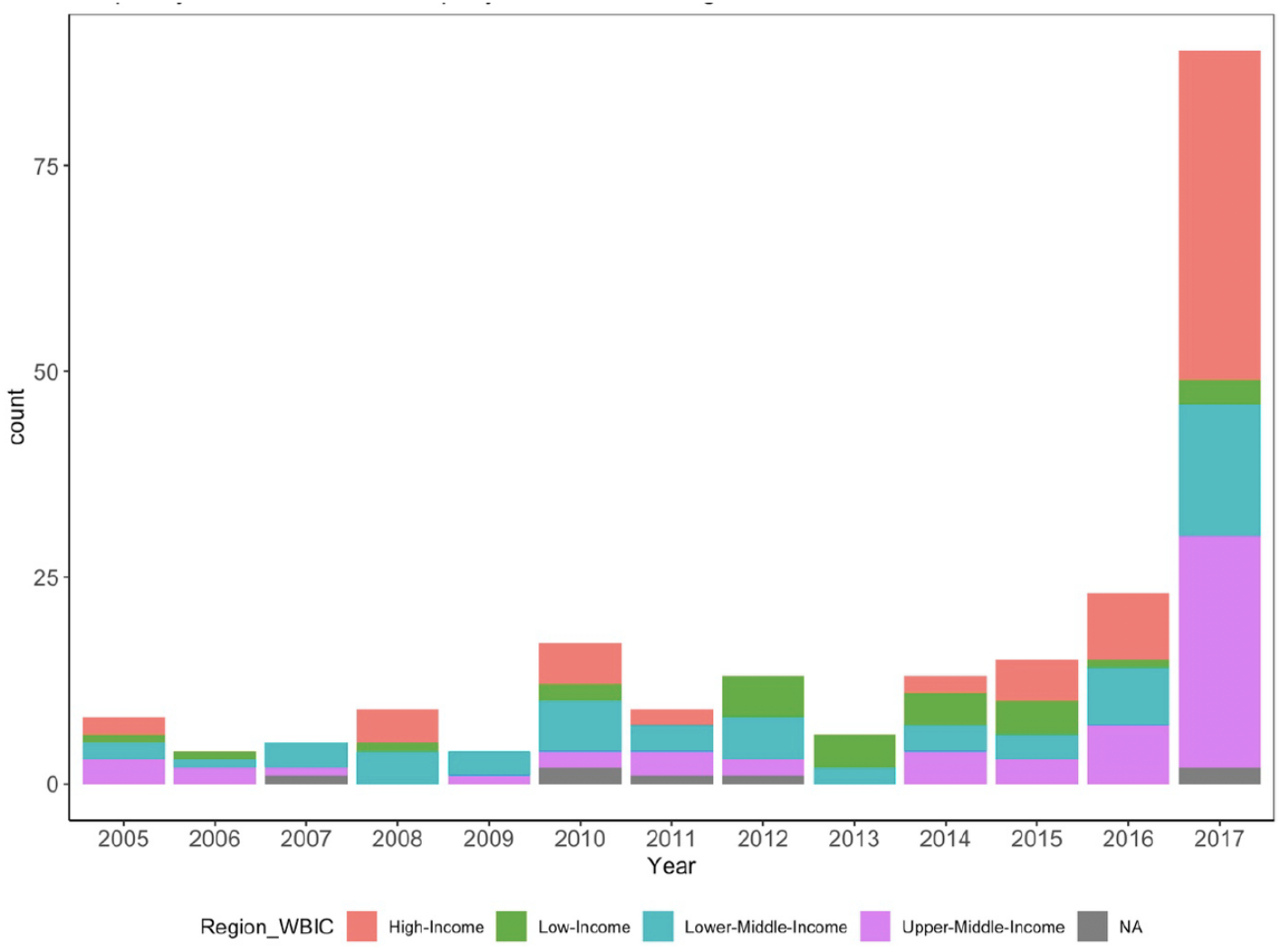



Figure: Number of countries by year of last available labour force survey and World Bank income region, ILOSTAT (2018)

# (Big) Data Innovation

- Could **digital trace data** from the web help fill these data gaps?
- Big data are created and collected by companies and governments for purposes other than research. Using this data for research, therefore, requires repurposing (Salganik 2017)
  - Custom-mades versus ready-mades
  - 3 **Vs**: Volume, Variety, and Velocity

[1] Salganik, Matthew J. Bit by bit: social research in the digital age. Princeton University Press, 2017.

# (Big) Data Innovation

- Could **digital trace data** from the web help fill these data gaps?
- Digital trace data offer promise for **nowcasting**: generating real-time predictions of social outcome indicators in the present (di Bella et al., 2016, Choi and Varian 2012).
- Custom-mades versus ready-mades (Salganik 2017)

---

[1] Hyunyoung Choic, Hal Varian: Predicting the Present with Google Trends. Economic Record, vol. 88, Iss. 1, pp. 2-9, 2012.

[2] di Bella, Enrico, Lucia Leporatti, and Filomena Maggino. "Big data and social indicators: Actual trends and new perspectives." Social Indicators Research 135, no. 3 (2018): 869-878.

[3] Salganik, Matthew J. Bit by bit: social research in the digital age. Princeton University Press, 2017.

# Ad Audience Estimates

- Ad audience estimates provided by online platforms act as **digital census** of the user base.
- How many users of 'x' characteristics (age, gender, location, etc) are on these platforms?
- **Pros:** Real-time or higher frequency measurement, topics and issues on which conventional data sources are lacking, anonymous, aggregated, no cost.
- **Cons:** Non-representative, black-box algorithms.

[1] Weber, Kashyap and Zagheni (2018), Using Advertising Audience Estimates to Improve Global Developments Statistics, *ITU Journal: ICT Discoveries*

# Today's Examples

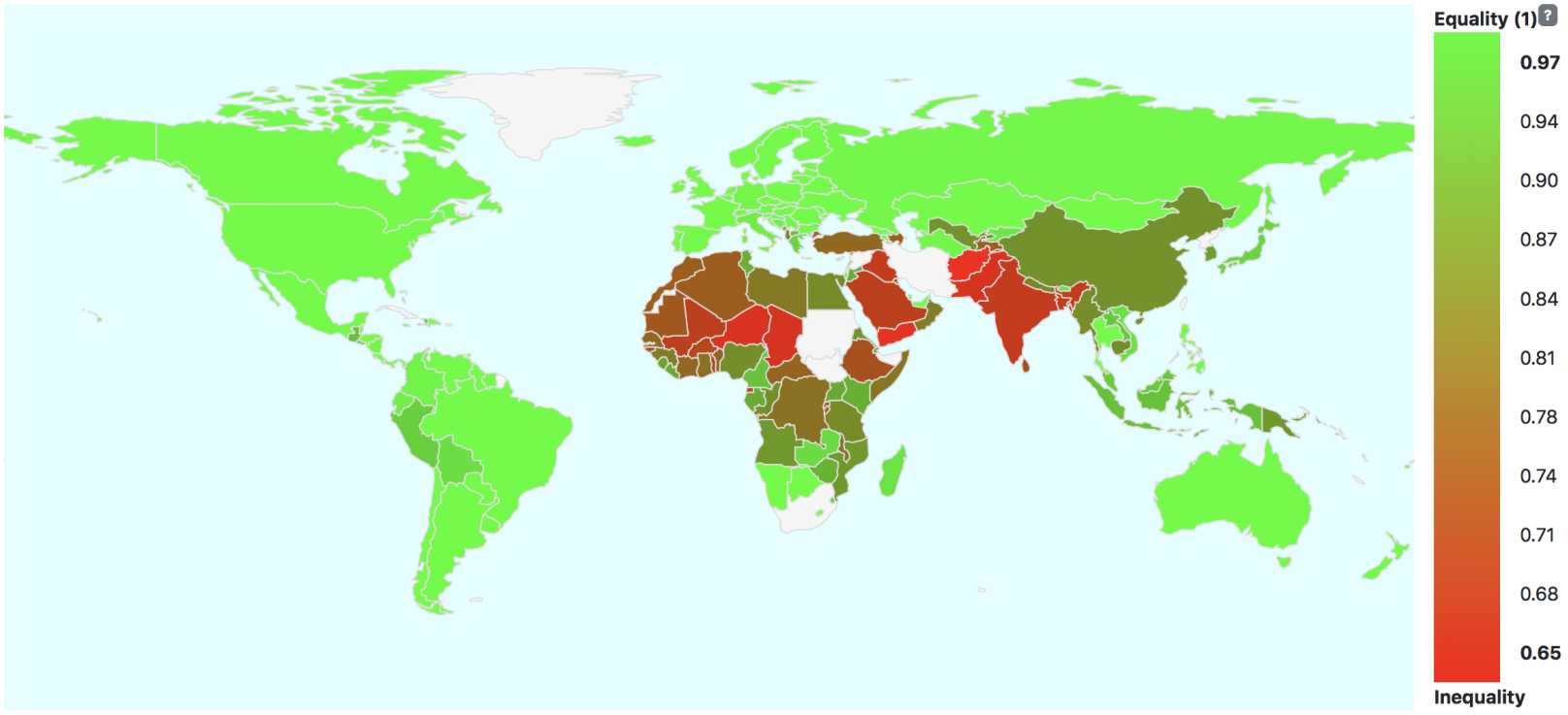



Figure: Gender gaps in internet use computed using data from Facebook (online model) available at `www.digitalgendergaps.org`

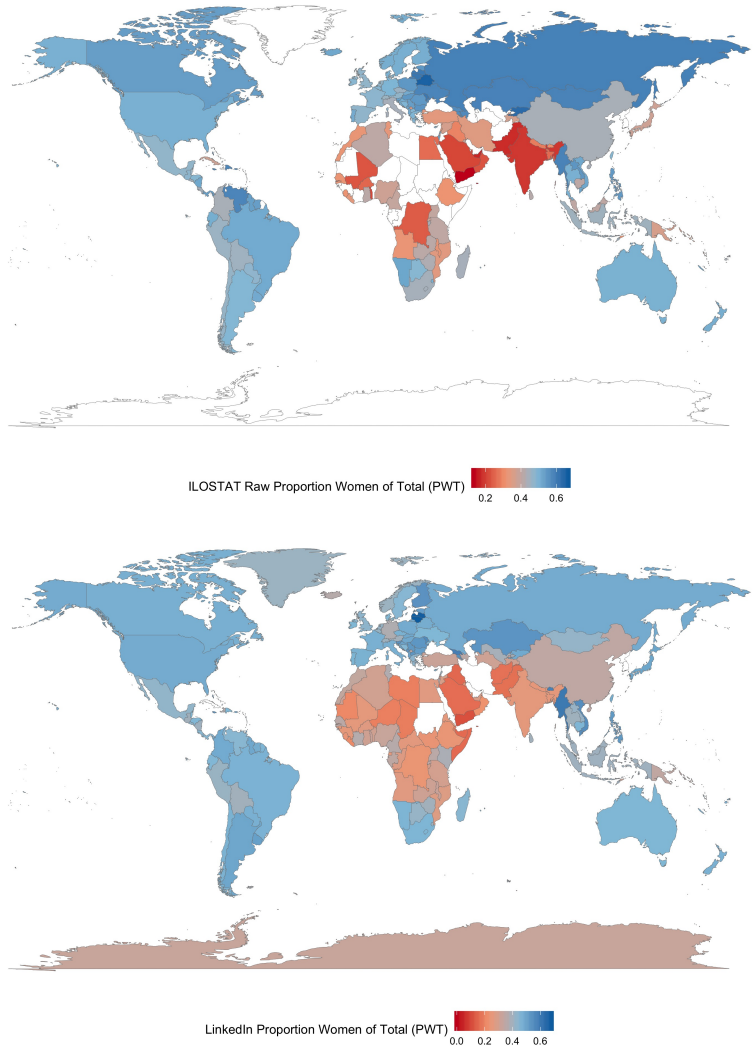
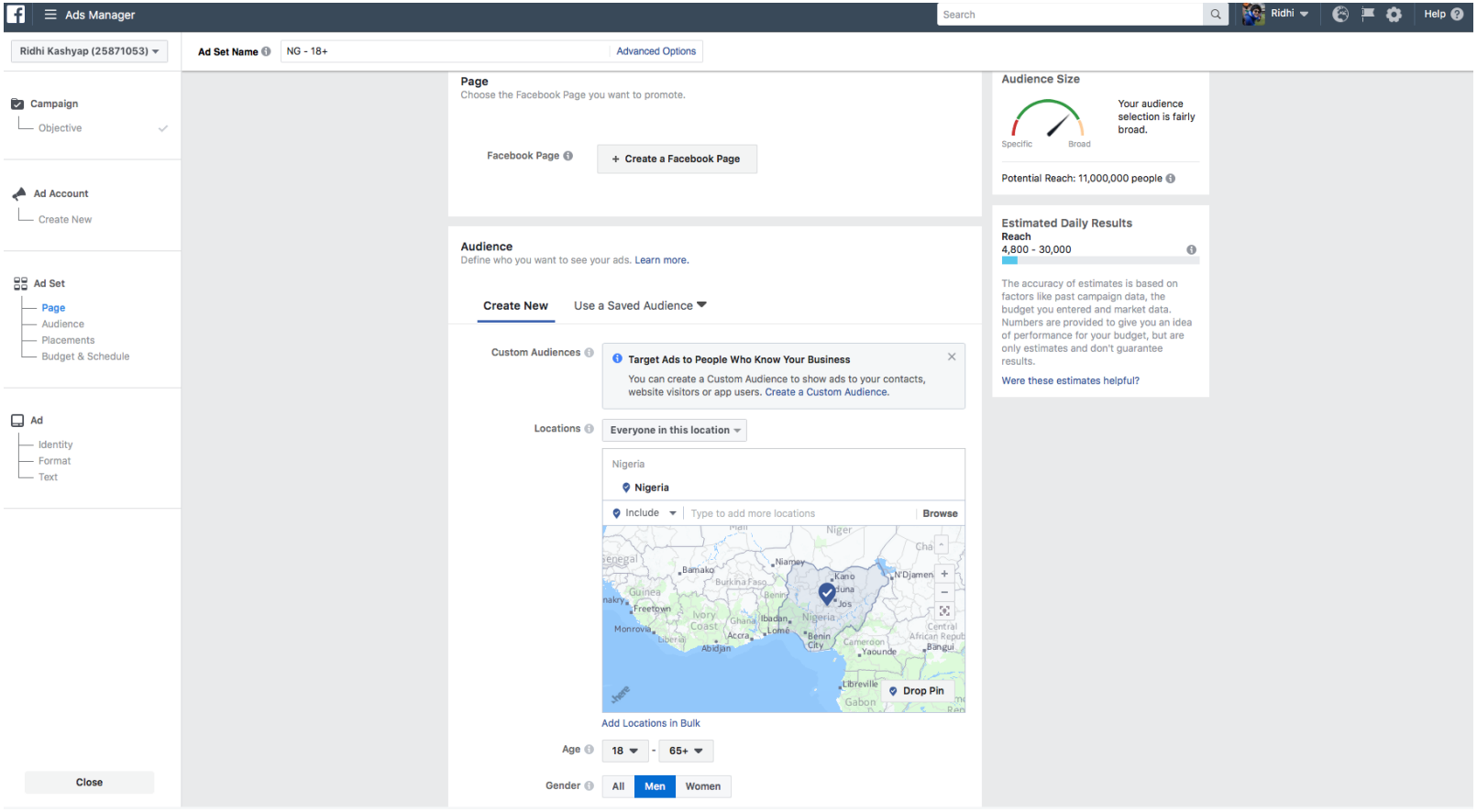
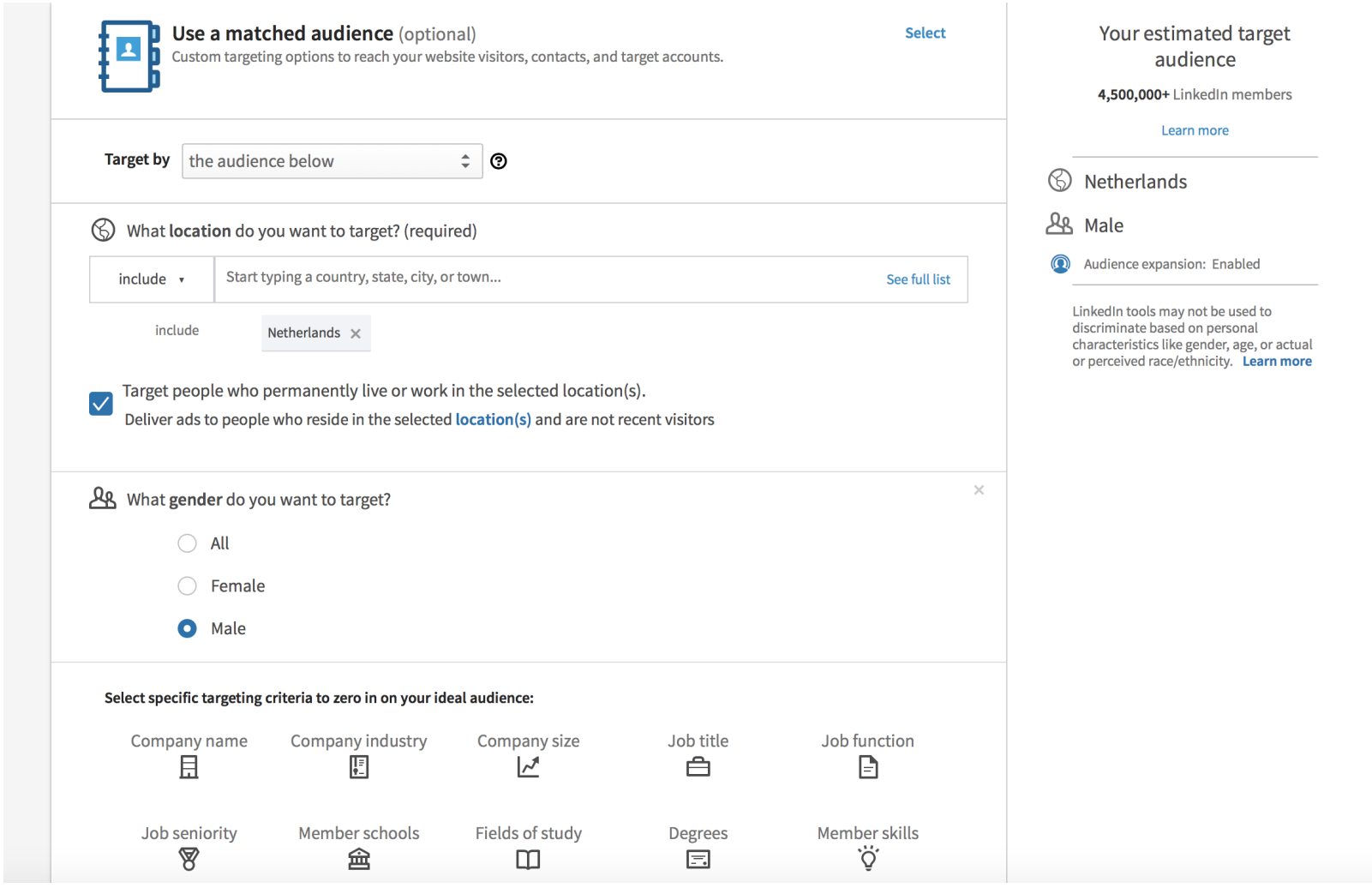



Figure: (a) professional gender gaps computed using ILOSTAT (Skill levels 3 and 4), and (b) professional gender gaps computed using LinkedIn.

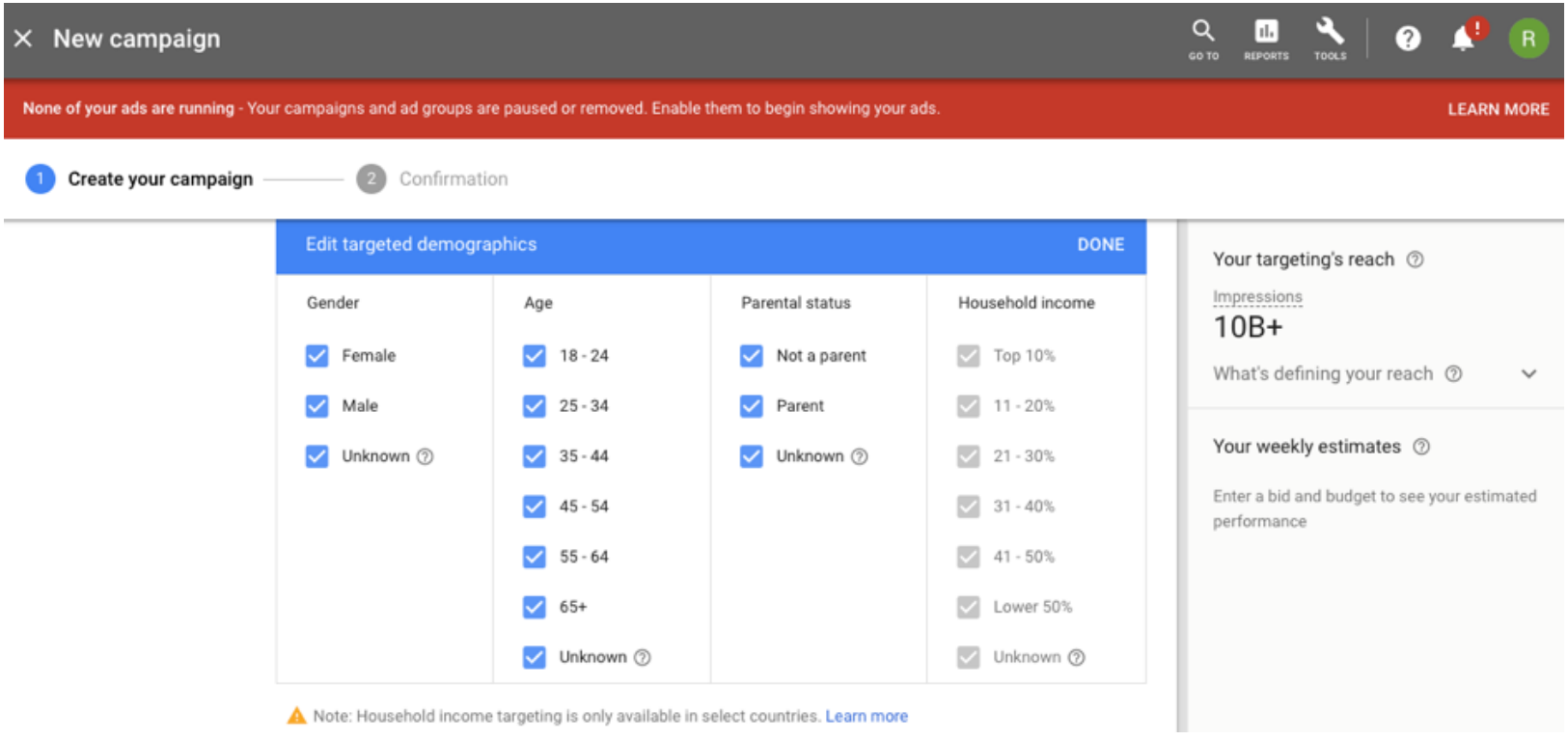
# Facebook Ad Audience Estimates



# LinkedIn Ad Audience Estimates



# Google AdWords Impressions



# Measuring Gender Gaps

- **Data sources**
  - Facebook: 2+ billion users
  - Google: 2+ billion users
  - LinkedIn: 560 million
- **Gender gaps**
  - Digital gender gaps: internet and mobile use, digital skills
  - Professional occupations, post-secondary education gaps

# Digital Gender Gaps

- Development indicator with data gap.
  - UN SDG Goal 5 pledges to "enhance the use of ... information and communication technology to promote the empowerment of women"
  - ITU/UNESCO: the lack of gender-disaggregated data on internet and mobile phone access is "one of the key barriers" for measuring progress in development goals that call for gender equality in access to the internet
- Digital inequality is an important dimension of inequality.
- De-biasing big data if women are underrepresented online.

# Data

- Our dataset comprises:

1. Online indicators on Facebook gender gaps for $\sim$ 200 countries
2. Offline indicators related to a country's overall level of development and gender gaps (e.g. education)
3. Data on gender gaps in internet use from surveys available from ITU

# Indicators

## Facebook Gender Gap Index

$$\text{FB GGI} = \frac{\text{Female to male gender ratio of users on Facebook}}{\text{Female to Male gender ratio of the population}}$$

## ITU Internet Gender Gap Index

- Using ITU data, we compute for $\sim$ 80 countries:

$$\text{Internet GGI} = \frac{\text{\% of female population using internet}}{\text{\% of male population using internet}} \quad (1)$$

- All gender gap indices capped at 1.

# Indicators

$$\text{FB GGI} \approx \text{Internet GGI}$$

- Online, Facebook indicators show strongest correlations with ITU Internet GGI
- Correlation of FB GGI 18+ with Internet GGI is 0.83, FB GGI 25-29, 0.81
- Stronger correlations than any other development indicator, including internet penetration (corr. $= 0.6$)

# Our Approach

- We fit OLS regression models that predict ITU internet gender gap measures.
- Three models:
  1. Online model: parsimonious single FB variable model
  2. Online-offline model: FB and other development variables
  3. Offline model: only development variables
- Variable selection using stepwise forward regression, also with five-fold cross-validation.
- Different measures of predictive fit: adjusted R-squared, mean absolute error, and SMAPE (Leave-One-Out cross validation)

# Internet Gender Gaps

| | Online Model | Online-Offline Model | Offline Model |
|---|---|---|---|
| Adjusted R-squared | 0.691 | 0.791 | 0.615 |
| Mean Abs. Error | 0.0325 | 0.0288 | 0.037 |
| SMAPE | 3.92% | 3.90% | 4.97% |
| F-statistics | 169 | 67.38 | 29.79 |
| df | 74 | 66 | 68 |
| N | 76 | 71 | 73 |
| # predicted countries | 152 | 127 | 132 |

Table: Summary of measures of predictive fit for three OLS regression models predicting internet gender gaps.

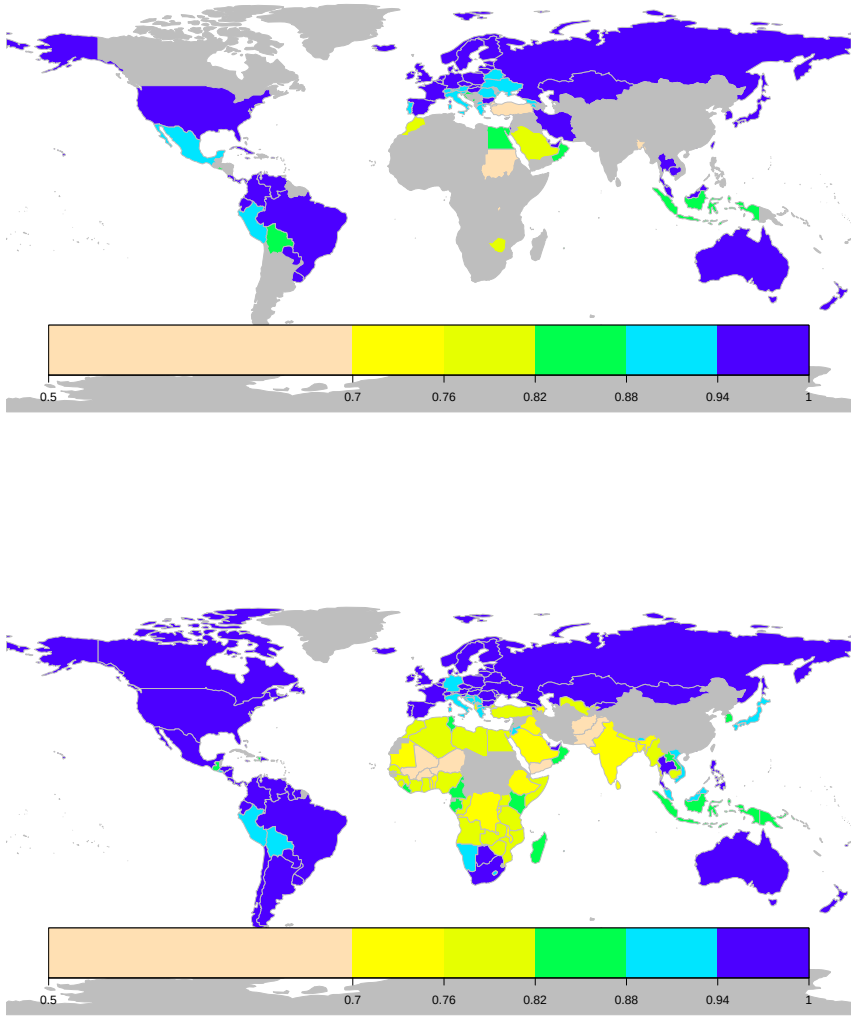



Figure: The internet gender gap index computed using (a) ITU survey data, and (b) predicted using Facebook data.

# Automated Nowcasting Platform

Daily nowcasts available at:
`www.digitalgendergaps.org`

Fatehkia, Masoomali, Ridhi Kashyap, and Ingmar Weber. "Using Facebook ad data to track the global digital gender gap." *World Development* 107 (2018): 189-209.

# Google AdWords and Digital Gender Gaps

- Can we leverage Google's **advertisement impression estimates (AdWords)** to predict internet gender gaps?
  - Facebook reaches 60% of internet users worldwide. Google claims that "the Google Display Network reaches 90% of Internet users worldwide".
- What kinds of digital skills are Facebook and Google use proxies for?

# FB GGI (18+) v ADW GGI (18+)

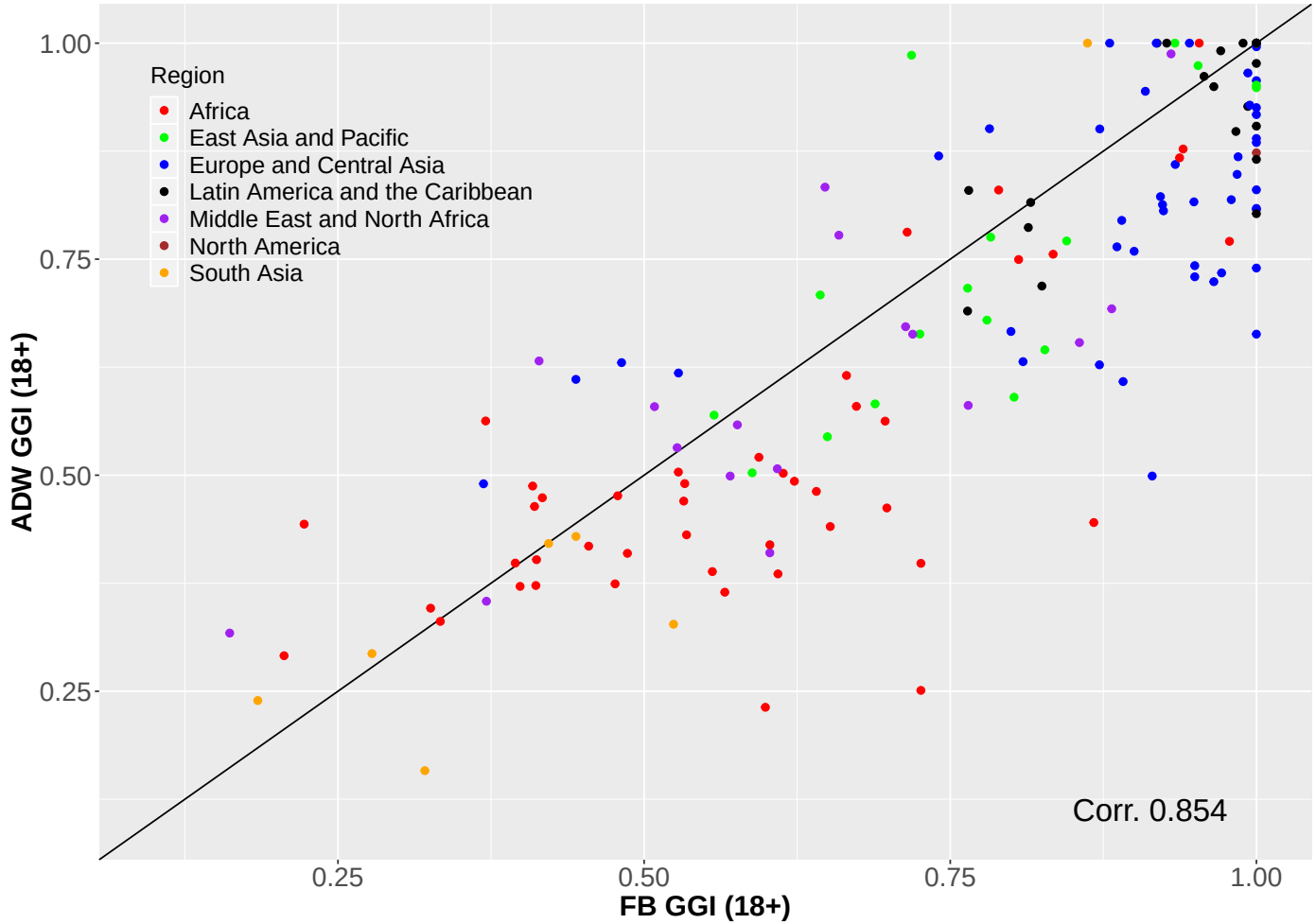



Figure: The FB GGI (ages 18+) against the ADW GGI (ages 18+). Each point indicates a country and points are colour coded by world region. The line is the x=y diagonal.

[1] Kashyap, Tamime, Fatehkia, Weber. "Monitoring digital gender inequality with Facebook and Google advertising audience estimates."

# Digital Skills Gender Gaps

- Facebook indicators are better able to predict low-level digital skill gender gaps.

| | Number of Countries in Dataset | FB GG age 18+ | Number of Countries in Dataset | ADW GG age 18+ |
|---|---|---|---|---|
| DS GGI-Copying or moving file or folder | 46 | 0.723 | 44 | 0.614 |
| DS GGI-Using copy and paste tools | 32 | 0.787 | 31 | 0.722 |
| DS GGI-Sending e_mails with attached files | 16 | 0.820 | 15 | 0.774 |
| DS GGI-Using basic arithmetic formula in spreadsheet | 41 | 0.526 | 39 | 0.462 |
| DS GGI-Connecting installing new devices | 19 | 0.205 | 17 | 0.421 |
| DS GGI-Finding, downloading and installing software | 39 | 0.438 | 37 | 0.406 |

**Table:** Correlations between FB and ADW GGI and different digital skills gender gaps (DS GGI).

[1] Kashyap, Tamime, Fatehkia, Weber. "Monitoring digital gender inequality with Facebook and Google advertising audience estimates."

# Professional Gender Gaps

- Using ad audience estimates from LinkedIn we can generate:

$$\text{LinkedIn PWT} = \frac{\text{Number of women on LinkedIn with characteristic}}{\text{Number of women \& men on LinkedIn with characteristic}}$$

- From ILOSTAT, we can compute:

$$\text{ILOSTAT PWT} = \frac{\text{Number of women in levels 3 or 4 skilled occupations}}{\text{Number of women \& men in level 3 or 4 skilled occupations}}$$

[1] Kashyap, Verkroost, Tamime, Fatehkia, Weber. "Measuring global gender inequality indicators with large-scale online advertising data."

# Education Gender Gaps

$$\text{ADW Educ GGI} = \frac{\text{Female to male gender ratio of impressions (filtered*)}}{\text{Female to male gender ratio of the population}}$$

- Filtered by those actively searching and planning for post-secondary education.
- Post-secondary education GGI, educational attainment GGI, literacy GGI

[1] Kashyap, Verkroost, Tamime, Fatehkia, Weber. "Measuring global gender inequality indicators with large-scale online advertising data."

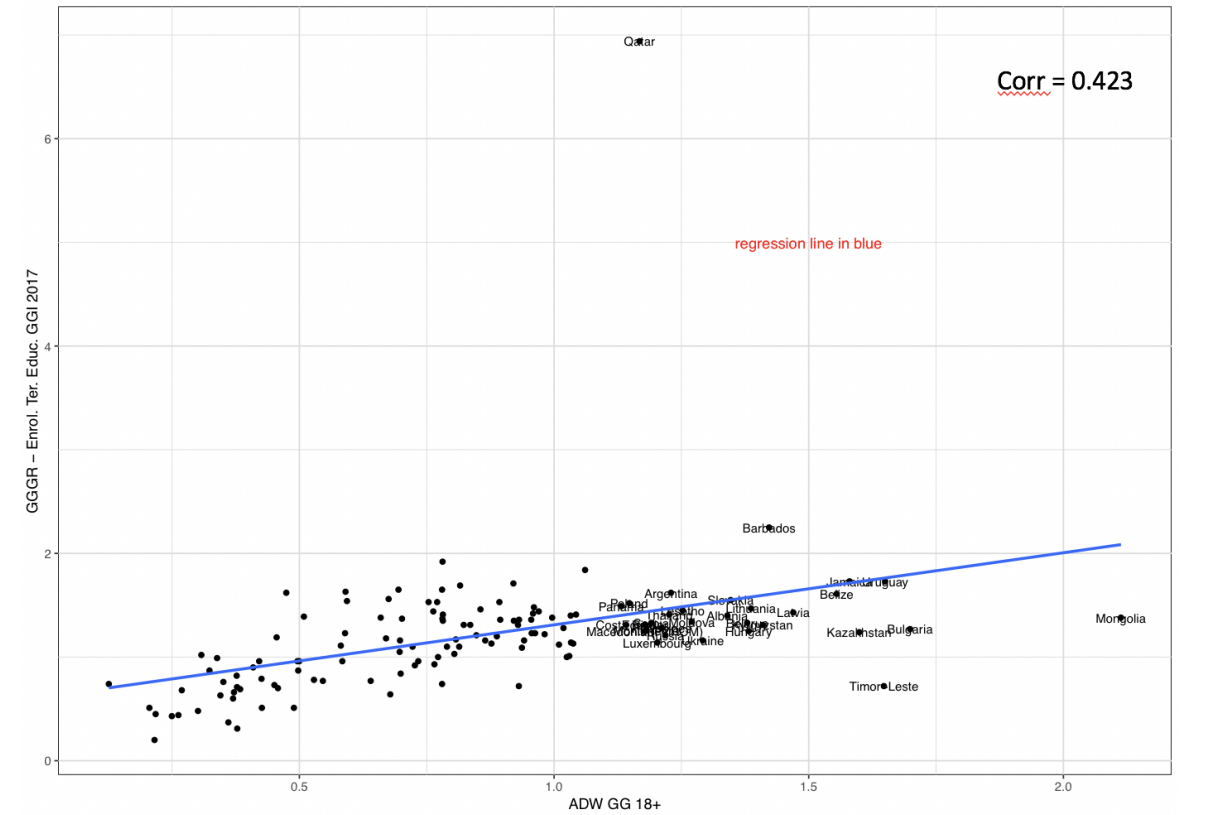
# Professional Gender Gaps

| LinkedIn Group PWT | Pearson's Corr. | Intercept | P-value | Coefficient | P-value | $R^2$ | Adj. $R^2$ | SMAPE | N (obs) | N (pred) |
|---|---|---|---|---|---|---|---|---|---|---|
| Overall (18+) | 0.798 | 0.103 | <0.000 | 0.834 | <0.000 | 0.636 | 0.634 | 11.456 | 168 | 264 |
| **Age Groups** | | | | | | | | | | |
| 18-24 | 0.679 | 0.177 | <0.000 | 0.595 | <0.000 | 0.461 | 0.458 | 14.136 | 158 | 203 |
| 18-24 & 35-54 | 0.677 | 0.207 | <0.000 | 0.587 | <0.000 | 0.458 | 0.455 | 13.996 | 161 | 215 |
| 18-24 & 35-54 & 55+ | 0.676 | 0.211 | <0.000 | 0.596 | <0.000 | 0.457 | 0.454 | 14.026 | 161 | 215 |
| 18-24 & 55+ | 0.673 | 0.191 | <0.000 | 0.612 | <0.000 | 0.454 | 0.450 | 14.818 | 158 | 208 |
| 25-34 | 0.669 | 0.194 | <0.000 | 0.559 | <0.000 | 0.448 | 0.444 | 13.375 | 161 | 215 |
| **Seniority Groups** | | | | | | | | | | |
| Senior | 0.799 | 0.151 | <0.000 | 0.700 | <0.000 | 0.638 | 0.636 | 11.604 | 165 | 222 |
| Unpaid & Senior | 0.797 | 0.152 | <0.000 | 0.700 | <0.000 | 0.635 | 0.632 | 11.668 | 165 | 222 |
| Senior & Partner | 0.796 | 0.152 | <0.000 | 0.704 | <0.000 | 0.634 | 0.632 | 11.688 | 165 | 222 |
| Training & Senior | 0.796 | 0.150 | <0.000 | 0.700 | <0.000 | 0.634 | 0.632 | 11.707 | 165 | 222 |
| Unpaid & Senior & Partner | 0.794 | 0.151 | <0.000 | 0.701 | <0.000 | 0.631 | 0.628 | 11.754 | 165 | 222 |
| **Industry Groups** | | | | | | | | | | |
| PST & HHS & AER & WRT & FIA & CON & REA & OSA | 0.807 | 0.193 | <0.000 | 0.586 | <0.000 | 0.651 | 0.649 | 11.737 | 150 | 188 |
| PST & HHS & AER & WRT & CON & REA & OSA | 0.806 | 0.195 | <0.000 | 0.575 | <0.000 | 0.650 | 0.648 | 11.505 | 150 | 187 |
| PST & HHS & AER & WRT & CON & OSA | 0.806 | 0.194 | <0.000 | 0.574 | <0.000 | 0.650 | 0.648 | 11.517 | 150 | 187 |
| PST & HHS & AER & WRT & FIA & CON & OSA | 0.806 | 0.192 | <0.000 | 0.584 | <0.000 | 0.650 | 0.647 | 11.760 | 150 | 188 |
| PST & HHS & AER & WRT & CON & OSA & ASS | 0.805 | 0.196 | <0.000 | 0.572 | <0.000 | 0.649 | 0.646 | 11.376 | 150 | 187 |

Figure: Summary of correlations and single-variable regression models predicting ILOSTAT GGI using LinkedIn GGIs

[1] Kashyap, Verkroost, Tamime, Fatehkia, Weber. "Measuring global gender inequality indicators with large-scale online advertising data."

# Educational Gender Gaps




[1] Kashyap, Verkroost, Tamime, Fatehkia, Weber. "Measuring global gender inequality indicators with large-scale online advertising data."

# Education Gender Gaps with Google AdWords

| | Educational att. GG | Post-sec Edu GGI (Wittg) | Post-sec Edu GGI (GGGR) |
|---|---|---|---|
| Adjusted R-squared | 0.347 | 0.404 | 0.173 |
| Mean Abs. Error | 0.042 | 0.205 | 0.268 |
| SMAPE | 4.47% | 23.8% | 22.5% |
| F-statistics | 74.708 | 114.178 | 28.608 |
| df | 138 | 166 | 131 |
| N | 140 | 168 | 133 |

Table: Summary of measures of predictive fit for three single-variable online OLS regression models with AdW GGI variables predicting education gender gaps

[1] Kashyap, Verkroost, Tamime, Fatehkia, Weber. "Measuring global gender inequality indicators with large-scale online advertising data."

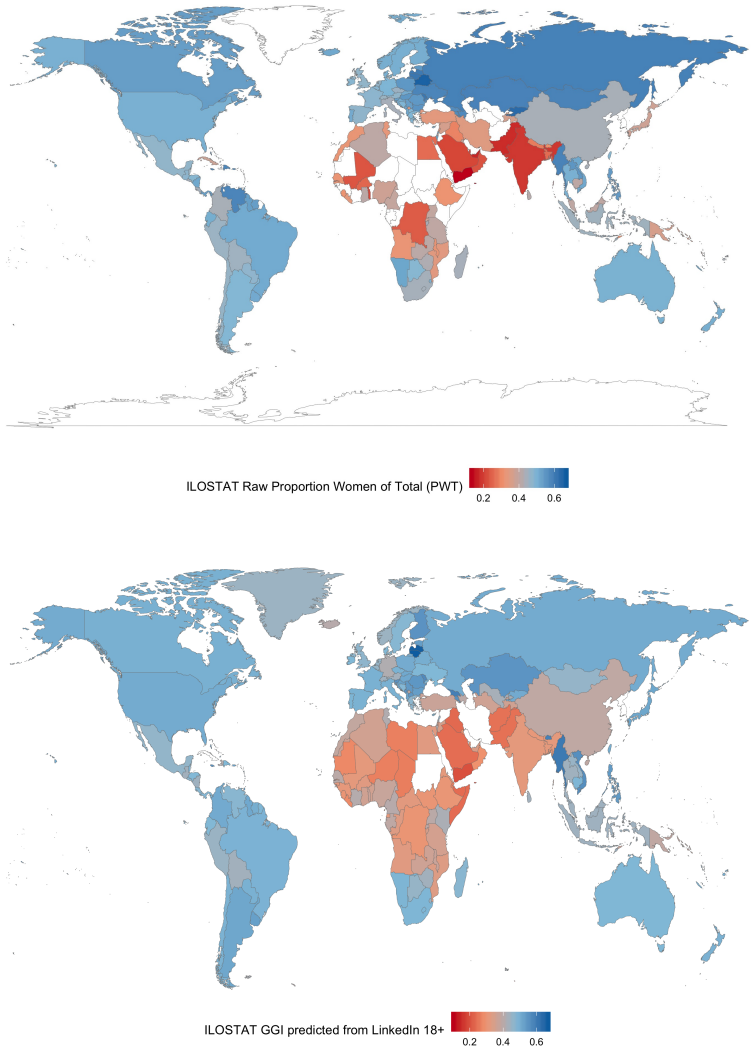
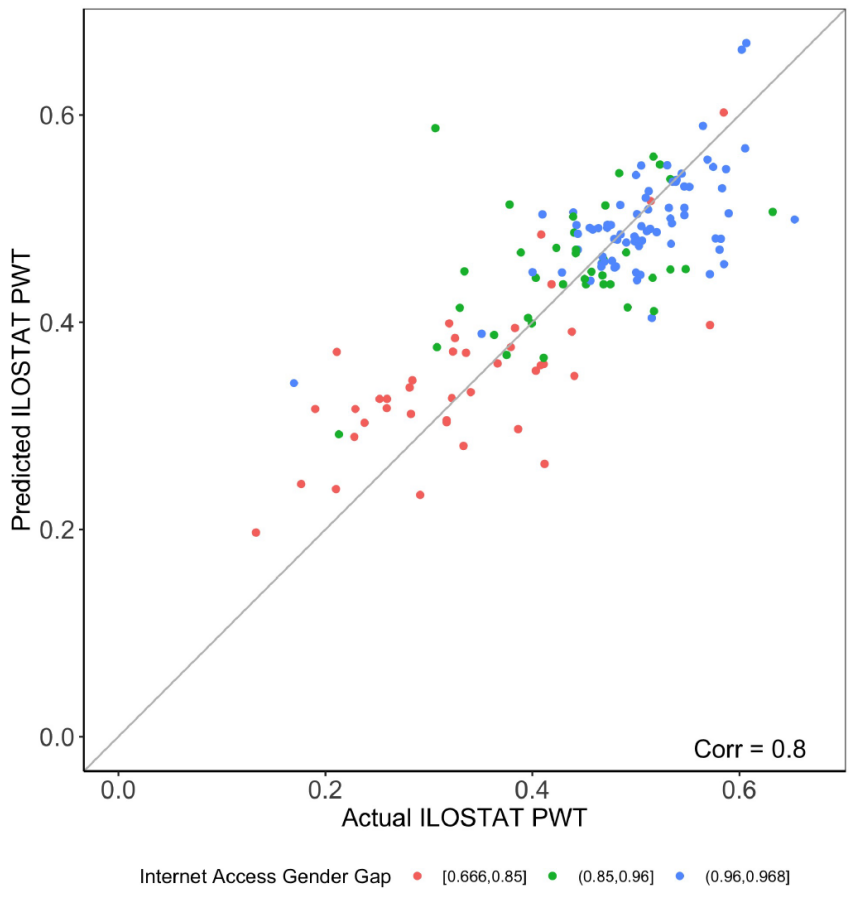



Figure: (a) professional gender gaps computed using ILOSTAT (Skill levels 3 and 4), and (b) professional gender gaps predicted using LinkedIn 18+ GGI.

# Biases

- Digital censuses tend to overpredict gender equality in settings with larger gender gaps in internet access.



# Conclusions

- Digital revolution has implications for the attainment and measurement of SDG 5.
- Using digital censuses, we can generate gender gap indicators.
  - Expanding geographical coverage including less developed countries
  - Finer temporal (and spatial) resolution
- Good complement, but not substitute
- Interesting biases in online data sources.
- Combining across multiple data sources with different strengths, weaknesses and temporalities.

# Thank you!

ridhi.kashyap@nuffield.ox.ac.uk