

# Title: Introduction to Social Media and Big Data for Migration Studies

## Organizers:

- [Jisu Kim](#), MPIDR, Germany, [kim@demogr.mpg.de](mailto:kim@demogr.mpg.de)\*
- [Ebru Şanlıtürk](#), MPIDR, Germany, [sanlituerk@demogr.mpg.de](mailto:sanlituerk@demogr.mpg.de)

The organizers are research scientists in the [Laboratory of Digital and Computational Demography](#) at the Max Planck Institute for Demographic Research (MPIDR) in Germany and they are all active in the research area of [Migration and Mobility](#).

[Jisu](#) holds a PhD in Data Science from Scuola Normale Superiore in Italy. She has been working on exploring and establishing novel methods to improve relevant statistics of international migration using social media data. Her research focuses on the intersection of migration sciences, economics of migration, complex social networks, statistical models and data-driven algorithms.

[Ebru](#) holds a PhD in Public Policy and Administration from Bocconi University in Italy. During her PhD, she specialized on demography and migration, completing her dissertation on the digital and computational approaches to migration studies. Her research interests are use of digital data in migration studies and demography, big data in migration studies and computational methods for the analysis of migration and mobility patterns.

## Objectives and goals

Main sources of data in the study of demography have usually been collected from traditional data, i.e., census, survey, and register data. Although they provide useful information, they come with some limitations. First of all, the cost associated with collecting traditional data is high. It is also time consuming and provides low quality information on some of the important demographic information such as seasonal migrants and short-term stays<sup>1</sup>, and studies show evidence that migrants are prone to higher non-response rates in surveys<sup>2</sup>. Asylum seekers are also mostly absent from national surveys, unless they are specifically targeted for surveys aiming to help refugee policies, yet they are known to use social media channels extensively<sup>3,4</sup>. Online communication technologies help migrants and transform migrant

---

<sup>1</sup> Sumption, Madeleine. "How useful are survey data for analyzing immigration policy?." *Data & Policy* 2 (2020).

<sup>2</sup> Bernhardt, Robert, and Phani V. Wunnava. "The CPS Citizenship Question and Survey Refusals: Causal and Semi-Causal Evidence Featuring a Two-Stage Regression Discontinuity Design." (2020).

<sup>3</sup> Dekker, Rianne, et al. "Smart refugees: How Syrian asylum migrants use social media information in migration decision-making." *Social Media+ Society* 4.1 (2018): 2056305118764439.

<sup>4</sup> Merisalo, Maria, and Jussi S. Jauhiainen. "Asylum-Related Migrants' Social-Media Use, Mobility Decisions, and Resilience." *Journal of Immigrant & Refugee Studies* 19.2 (2021): 184-198.

networks<sup>5</sup>, making it necessary for social scientists to study digital migration networks with a closer focus.

Today, we are provided with innovative and novel data from social media and search platforms such as Facebook, Twitter and Google, where the data comes in various formats such as text, audio, image, and video. The amount of data that is produced by individual users is increasing significantly, often referred to as big data. In recent years, many researchers have taken advantage of such data to study various types of social phenomena such as migration<sup>6,7,8</sup>, public health<sup>9</sup>, gender gaps<sup>10</sup>, political polarisation<sup>11</sup>, and much more. Big data, obtained from social media, provides researchers information on hard-to-reach populations that is relatively easy and less costly to collect. Such big data can be collected both in real time and retrospectively, offering the opportunity for the analysis of more recent (or real time) events with a comparative perspective to the past. Social media data also includes detailed information and allows for monitoring the mobility with geo-tagging features. However, these innovative data also come with different types of limitations, biases and challenges, requiring researchers to be equipped with different types of skills and approaches to deal with these types of data.

In this workshop, we aim to introduce two types of big data, i.e., Google Trends and Twitter data. This workshop will be the great opportunity for the researchers in the study of migration but also demography in general to get familiar with two popular sources of data. We plan to begin the session by introducing the data format, related literature, empirical findings, advantages and critical challenges of such data. We then plan to have more interactive sessions on how to retrieve these data and write and search inquiries.

While the innovative data sources such as social media provide us with great volumes of data for research, representativeness of such data is still questionable and dependent on the context. Therefore, we adopt the general strategy of combining innovative data sources with traditional data, commonly used in the literature. Thus, we will mention how to combine

---

<sup>5</sup> Dekker, Rianne, and Godfried Engbersen. "How social media transform migrant networks and facilitate migration." *Global Networks* 14.4 (2014): 401-418.

<sup>6</sup> Kim, Jisu, et al. "Digital footprints of international migration on twitter." *International Symposium on Intelligent Data Analysis*. Springer, Cham, 2020.

<sup>7</sup> Sîrbu, Alina, et al. "Human migration: the big data perspective." *International Journal of Data Science and Analytics* 11.4 (2021): 341-360.

<sup>8</sup> Böhme, Marcus H., André Gröger, and Tobias Stöhr. "Searching for a better life: Predicting international migration with online search keywords." *Journal of Development Economics* 142 (2020): 102347.

<sup>9</sup> Köksal, Selin, Luca Maria Pesando, Valentina Rotondi, and Ebru Şanlıtürk. "Harnessing the Potential of Online Searches for Understanding the Impact of COVID-19 on Intimate Partner Violence in Italy." (2021).

<sup>10</sup> Fatehkia, Masoomali, Ridhi Kashyap, and Ingmar Weber. "Using Facebook ad data to track the global digital gender gap." *World Development* 107 (2018): 189-209.

<sup>11</sup> Garimella, Venkata Rama Kiran, and Ingmar Weber. "A long-term analysis of polarization on Twitter." *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. No. 1. 2017.

Google Trends and Twitter data with traditional data sources<sup>12,13,14</sup>, both to test their reliability and understand the adjustments necessary to make meaningful analyses.

Some of the literature that will be studied in this workshop are:

- Böhme, Marcus H., André Gröger, and Tobias Stöhr. "Searching for a better life: Predicting international migration with online search keywords." *Journal of Development Economics* 142 (2020): 102347.
- Sanliturk, Ebru, and Francesco Billari. Search for a New Home: Refugee Stock and Google Search. (2021)
- Sîrbu, Alina, et al. "Human migration: the big data perspective." *International Journal of Data Science and Analytics* 11.4 (2021): 341-360.
- Kim, Jisu, et al. "Digital footprints of international migration on twitter." *International Symposium on Intelligent Data Analysis*. Springer, Cham, 2020.
- Zagheni, Emilio, et al. "Inferring international and internal migration patterns from twitter data." *Proceedings of the 23rd International Conference on World Wide Web*. 2014.
- Hsiao, Yuan, et al. Modeling the bias of digital data: an approach to combining digital and survey data to estimate and predict migration trends. No. WP-2020-019. Max Planck Institute for Demographic Research, Rostock, Germany, 2020.

## Topics to be covered (tentative programme)

### Google Trends

- What is Google Trends data? How can it be used in social sciences and migration studies?
- How to use R to obtain a Google Trends data set suitable for a given research?
- Advantages and shortcomings of the data and how to address them.
- Comparison with traditional migration data (stock and flow data) and research examples.

### Twitter

- What is Twitter data? What is the information that can be leveraged from Twitter?
- What information is available to be used in social sciences and migration studies?
- How to use R or/and Python to obtain data from Twitter API?
- Advantages and shortcomings of the data and how to address them.
- Comparison with traditional migration data (stock and flow data) and research examples.

---

<sup>12</sup> Hsiao, Yuan, et al. *Modeling the bias of digital data: an approach to combining digital and survey data to estimate and predict migration trends*. No. WP-2020-019. Max Planck Institute for Demographic Research, Rostock, Germany, 2020.

<sup>13</sup> Zagheni, Emilio, et al. "Inferring international and internal migration patterns from twitter data." *Proceedings of the 23rd International Conference on World Wide Web*. 2014.

<sup>14</sup> Risteski, D., & Davcev, D. (2014, October). Can we use daily Internet search query data to improve predicting power of EGARCH models for financial time series volatility. In *Proceedings of the International Conference on Computer Science and Information Systems (ICSIS'2014)*, October 17–18, 2014, Dubai (United Arab Emirates).

## Expected outcomes

Google Trends website by Google offers basic comparisons and visualisations for online search interest on selected queries. In this workshop we will present how to download and process Google Trends data in a way that is suitable for research and statistical analysis, using migration studies as the main field of focus.

Twitter is one of the most popular microblogging platforms worldwide. A tweet on Twitter includes a text message, together with the user information, time, location etc. As for the Twitter data, we will be presenting the newest version of the Twitter API<sup>15</sup> that has been released this year.

This workshop will be presenting methods using both R and Python. To be more specific, the first session on Google Trends will be taught using R and the package *gtrendsR*<sup>16</sup>. The second session on Twitter, will be taught in Python but the R counterpart will also be introduced and shown to allow participants to choose freely depending on their preferences.

By the end of this workshop, we expect the participants to learn to:

- Obtain basic knowledge of both Google Trends and Twitter data
- Know how to access Google Trends data using statistical software
- Know how to process Google Trends data to make it suitable for research
- Know how to access the data and learn Twitter API.
- Know how to apply for the Academic research on Twitter for better access to the data
- Know how to write search queries
- Extract interesting new research ideas using these data

## Target audience

This workshop aims to bring together social scientists who would like to get acquainted with social media sources for research and fundamentals of data collection. In that sense, this workshop is targeting the social scientists who are not familiar with accessing Twitter and Google trends data in a programmatic way and would like to explore them. This workshop will target primarily researchers in their early career stage, especially the young PhD students, master's students, but it is not limited to young researchers.

We, therefore, expect the participants to obtain basic knowledge of the two platforms, i.e., Google Trends and Twitter, and on how to access such data.

---

<sup>15</sup> Application Programming Interface (API)

<sup>16</sup> Massicotte, Philippe, Dirk Eddelbuettel, and Maintainer Philippe Massicotte. "Package 'gtrendsR'." (2016).

**Requirements:**

The participants are required to bring their laptops to the workshop. Prior to the workshop, it is necessary to download and install both R and Python. To participate in the hands-on session, it is required that the participants should have a minimum knowledge of either R or Python. They are also required to read and follow the instructions and materials that will be sent out before the workshop.

**Format:**

- Preferred day: Wednesday, April 6, 2022
- Duration: ½ day (preferably in the afternoon)
- Audience: we estimate around 15~20 participants (in-person) and 20~30 participants (online)
- Willing to convert to an online? Yes