

# Introduction to Social Big Data for research

**Jisu Kim, Ph.D  
ASA, August 2023**



MAX PLANCK INSTITUTE  
FOR DEMOGRAPHIC  
RESEARCH

# About the instructor

**Jisu Kim**

Ph.D in Data Science

Research scientist at Digital and Computational Demography department, Max Planck Institute for Demographic Research

Twitter: @kr\_jisu

# Materials

Workshop materials can be found here:

<https://github.com/jisukimmmm/ASA2023>



## Reading

Sîrbu, A, et al. "Human migration: the big data perspective." International Journal of Data Science and Analytics 11.4 (2021): 341-360.

Link: <https://link.springer.com/article/10.1007/s41060-020-00213-5>

# Outline

- ▶ Limitations of traditional data and social media data as an alternative data
- ▶ Twitter data format
- ▶ Twitter and related researches
- ▶ Web-scraping
- ▶ Ethics and privacy

# Study of migration

Immigrants touch upon multidimensional aspects of both the host country and the home country.

- ▶ Economics:
    - ▶ Jobs
    - ▶ Unemployment
  - ▶ Society:
    - ▶ Integration
    - ▶ Friends
    - ▶ Well-being
    - ▶ Population density
    - ▶ Fertility
  - ▶ Culture:
    - ▶ Food
    - ▶ Language
    - ▶ Inter-marriage
    - ▶ more...
  - ▶ Politics
- and many more...

Immigration

Trump put up walls to immigrants, with stinging rhetoric and barriers made of steel and regulation

The screenshot shows a news article from a German political website. The main headline reads "U.S. and Haiti work to address migration challenges". Below the headline, there is a sub-headline: "Home | News & Events | U.S. and Haiti work to address migration challenges". At the bottom of the page, there is a navigation bar with links: "GERMAN GENERAL ELECTION | ELECTIONS | POLITICS | REFUGEES IN GERMANY | ASYLUM | MIGRATION POLICY". On the right side of the page, there is a sidebar with a small image of a person and some text. The overall layout is typical of a news website.

U.S. and Haiti work to address migration challenges

Home | News & Events | U.S. and Haiti work to address migration challenges

GERMAN GENERAL ELECTION | ELECTIONS | POLITICS | REFUGEES IN GERMANY | ASYLUM | MIGRATION POLICY

German election: How do political parties view migrants' issues?



# Migration research

## Who is an **Immigrant**?:

- ▶ “A person who moves to a country other than that of his or her usual residence for a period of at least a year.<sup>1</sup>”
- ▶ “Whose movement across borders-whether legal or illegal- is essentially permanent<sup>2</sup>”
- ▶ “is defined on the ground of the place of birth (foreign-born) or of the citizenship (foreigners)<sup>3</sup>”

---

<sup>1</sup>United Nation

<sup>2</sup>World Bank

<sup>3</sup>OECD

# Limitation of traditional data sources

Census, survey, register data

- ▶ Costly
- ▶ Outdated
- ▶ Time consuming
- ▶ Inconsistent
- ▶ Unavailable
- ▶ Lack of data on emigration
- ▶ Incomplete answers/misunderstanding questions etc.
- ▶ Immigrants are often underrepresented traditional data sources.
- ▶ limited in hard-to-reach contexts and societies.

# Social media



# Big data

is “information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation”<sup>4</sup> Laney, 2001. It can be described in 3 Vs which are:

- ▶ Volume: As the name suggests, the size of the data is Big, hence the volume of the data.
- ▶ Velocity: Big data such as Twitter allow us to stream data at real-time. The rate at which we obtain data is faster than the traditional data sources.
- ▶ Variety: Traditional data are mostly structured data. Big data, on the other hand, come in various forms. It can be videos, photos, texts, and audios. It requires a thorough data processing before extracting information/knowledge from it.

---

<sup>4</sup><https://www.gartner.com/en/information-technology/glossary/big-data>

## Alternative data sources: Big data

Twitter, Facebook, Yahoo, Web...

- ▶ Free
- ▶ Granular data
- ▶ Large scale data
- ▶ Continuously generated
- ▶ Information/opinion shared by users from an uncontrolled environment
- ▶ various forms of data: video, image, text, audio etc.

# Required information

- Geo-location
- Information on time
- Personal identifier

So how do we get this Big data?

CHRIS STOKEWALKER

BUSINESS 28.10.2022 02:28 PM

# Elon Musk's Twitter Will Be Chaos

The entrepreneur's laundry list of ideas includes scrapping content moderation, charging subscription fees, and even branching out beyond social media.



PHOTOGRAPH: CARINA JOHANSEN/GETTY IMAGES

5

<sup>5</sup>source:

<https://www.wired.co.uk/article/elon-musk-twitter-deal-chaos>



14 / 61

# What is he going to do???



**Elon Musk** ✅ @elonmusk · 17h

Please note that Twitter will do lots of dumb things in coming months.

...

We will keep what works & change what doesn't.

34.9K

44.3K

435.4K



**Elon Musk** ✅ @elonmusk · Nov 7

Any name change at all will cause temporary loss of verified checkmark

...

12K

16.9K

142.1K



**Elon Musk** ✅ @elonmusk · Nov 5

Twitter will soon add ability to attach long-form text to tweets, ending absurdity of notepad screenshots

...

35.4K

67.3K

610.9K



**Elon Musk** ✅ @elonmusk · Nov 5

Followed by creator monetization for all forms of content

...

9,032

17.3K

220K



**Elon Musk** ✅ @elonmusk · Nov 5

Trash me all day, but it'll cost \$8

...

112.1K

126K

1.3M



## Price

"The basic tier costs \$100 per month but allows researchers to collect only 10,000 tweets per month-a mere 0.3% of what could previously be collected for free in one day. The Enterprise tier, which ranges from \$42,000 to \$210,000 per month, is not affordable for researchers"<sup>6</sup>

---

<sup>6</sup><https://independentechresearch.org/letter-twitters-new-api-plans-will-devastate-public-interest-research/>

# Don't cry...

- ▶ archive.com
- ▶ [https://www.trackmyhashtag.com/blog/  
free-twitter-datasets/](https://www.trackmyhashtag.com/blog/free-twitter-datasets/)
- ▶ <https://data.world/datasets/twitter>
- ▶ ...

## Twitter data format

- ▶ Tweet object
- ▶ User object
- ▶ Entity object

## Tweet object

**“id”**: "1050118621198921728"

**“text”**: "Apply now for our week-long open online course on Digital and Computational Demography hosted by @MPIDRnews scientists ..."

**“Context annotations”**:

**‘name’**: ‘Interests and Hobbies Category’, **‘description’**: ‘A grouping of interests and hobbies entities, like Novelty Food or Destinations’, **‘entity’**: ‘id’: '852291840472629248’, ‘name’: ‘Online education’, ‘description’: ‘Online education’

**“created\_at”**: "202x-09-xxTxx:xx:xx.000z"

**“conversation\_id”**: "1435336531519197188"

**“lang”**: “en”

**“geo”**: e.g., “coordinates”: [-73.999xx, 40.7416xxx]

**“place\_id”**: "01a9a39529b27f36"

## Context annotation

3 - TV Shows 4 - TV Episodes 6 - Sports Events 10 - Person 11 - Sport 28 - NFL Football Game 35 - Politicians 38 - Political Race 40 - Sports Series 47 - Brand 48 - Product 49 - Product Version 54 - Musician 55 - 56 - Actor 58 - Entertainment Personality 60 - Athlete 67 - Interests and Hobbies 68 - Hockey Game 71 - Video Game 85 - Book Genre 86 - Movie 87 - Movie Genre 88 - Political Body 89 - Music Album 92 - Sports Personality 93 - Coach ...and more<sup>7</sup>

---

<sup>7</sup><https://developer.twitter.com/en/docs/twitter-api/annotations/overview>



# Entity object

```
"hashtags": ["tag": "BuildWhatsNext"],  
"mentions": "tag": "@TwitterDev...",  
"url": "https://t.co/z5RhIVxJFK",
```



Twitter API @TwitterAPI · Feb 9

We have added the sort\_order parameter to the search endpoints in the Twitter API v2 which gives developers the option of returning Tweets based on recency or relevancy. Check out the details here [👉](#)



twittercommunity.com

Introducing the sort\_order parameter for search en...

Today, we're sharing a small, but important enhancement to the search functionality in the ...

16

20

73



```
"expanded_url": "https://twittercommunity.com/t/  
updates-to-retweets-lookup-and...",  
"description": "Thanks for your feedback on the v2 Retweets and  
Likes endpoints. We've heard you. Starting today, you can retrieve  
the complete list of accounts that have Liked or Retweeted a  
Tweet..."
```

## Collecting tweet object

- ▶ Streaming keywords/hashtags: e.g. #migrants, #refugees
- ▶ Streaming for specific geo-locations: place, country, point radius, bounding box
- ▶ Collecting tweets for a specific user: by user ID or user name
- ▶ Searching for historic tweets for a keyword

Query:

- ▶ from: twitter user
- ▶ –is : excluding retweets
- ▶ place\_country: country
- ▶ has: e.g. media, geo, images
- ▶ conversation\_id: returns all tweets in the conversation thread with conversation id. xxx
- ▶ lang: en, kr, ch, etc.

# User object

```
"id": "2244994945",
"name": "Twitter Dev",
"username": "TwitterDev",
"location": "127.0.0.1",
"verified": true,
"protected": false,
```

The image shows a Twitter profile card for the account "Twitter Dev" (@TwitterDev). The profile picture is a blue circle with a white Twitter bird icon. The bio reads: "The voice of the #TwitterDev team and your official source for updates, news, and events, related to the #TwitterAPI.". The location is listed as "127.0.0.1". The account is verified, indicated by a blue checkmark. The follower count is 522.3K, and the user has 2,024 following. The profile was joined in December 2013. There are standard Twitter interaction buttons like three dots, a bell, and a following button.

Twitter Dev

@TwitterDev

The voice of the #TwitterDev team and your official source for updates, news, and events, related to the #TwitterAPI.

127.0.0.1 developer.twitter.com/en/community Born March 21

Joined December 2013

2,024 Following 522.3K Followers

```
"description": "The voice of the #TwitterDev team and your official source for updates, news, and events, related to the #TwitterAPI.",
"url": "https://t.co/3ZX3TNiZCY",
"profile_image_url": "https://pbs.twimg.com/profile_images/1267175364003901441/tBZNFAgA_normal.jpg",
"created_at": "2013-12-14T04:35:55.000Z"
```

# Collecting user object

- ▶ Searching by user: user ID, user name,  
<https://followerwonk.com/bio>
- ▶ Social network: followers and friends

Showing 1 - 50 of 50,000 results (order by relevance)

screen name	real name	tweets	following	followers	account age	Social Authority
@BarackObama   Barack Obama	Dad, husband, President, citizen.	16,292	588,419	130,116,236	14.86 years	91
@justinbieber   Justin Bieber	JUSTICE. The album out now	31,377	286,547	114,068,969	12.59 years	93
@katyperry   KATY PERRY	Love. Light.	11,461	236	108,792,375	12.69 years	88
@rihanna   Rihanna		10,587	968	103,314,798	12.08 years	92
@Cristiano   Cristiano Ronaldo	This Privacy Policy addresses the collection and use of personal information - <a href="http://www.cristianoronaldo.com/terms">http://www.cristianoronaldo.com/terms</a>	3,726	58	95,241,254	11.38 years	92

Bio word cloud of users Oprah follows

To help make sense of the "biography" field of each Twitter user, we've assembled this word cloud which shows you the most frequently occurring words.

author – oprah – life – love – now – producer – actress – own – host – founder –  
#ownambassador – out – actor – twitter – show – official – bestselling – father – ceo – husband – live

Two word bio cloud

bestselling author

Location word cloud of users Oprah follows

Similar to the above word cloud, here we show you the relative frequency of words used in the "location" field of users Oprah follows.

los angeles – ca – new york – ny – usa – chicago – washington

# Twitter data and migration research

Information from Twitter data that can be used in migration research:

- ▶ Geo-tagged tweets
- ▶ Self-declared location information from profile
- ▶ Social networks: followers and friends
- ▶ Name
- ▶ Language

Anything else?

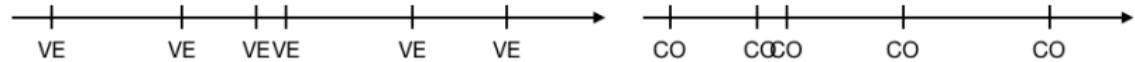
# Potential issues

- ▶ Tourists
- ▶ Travel blogs
- ▶ Bots
- ▶ ...

## Definitions of migrants on Twitter in the literature

- ▶ “A Twitter user has the nationality that others believe you have.” (Huang et al., 2014)
- ▶ “Any individual leaving Venezuela during the time window of observation.” (Mazzoli et al., 2020)
- ▶ “Anyone who tweeted exclusively from Venezuela in the time period between Feb. 1 and April 30 2017.” (Hausmann et al., 2018)
- ▶ “Migrants are users that are identified as people who moved to a different country for at least one of the 4-month periods.” (Zagheni et al., 2014)
- ▶ “A migrant is a person that has the residence different from the nationality.” (Kim et al., 2020)

# Identifying migrants (Mazzoli et al., 2020)



**Figure:** Mock example of geo-tagged tweet timeline

# Inferring nationalities of Twitter users (Huang et al., 2014)



- ▶ Twitter page: xx
- ▶ Homepage: XX
- ▶ Location: Germany
- ▶ Time zone: CET
- ▶ Interface language: English
- ▶ Tweets languages: (1) English: 97% (2) Italian: 2% (3) Korean: 1%
- ▶ Follower locations: (1) US: 151 (2) KR: 72 (3) DE: 3
- ▶ Following locations: (1) US: 80 (2) KR: 40 (3) IT: 30
- ▶ Tweets from (1) IT: 30 (2) KR: 29

# Identifying migrants on Twitter (Kim et al., 2020)

"A migrant is a person that has the residence different from the nationality."

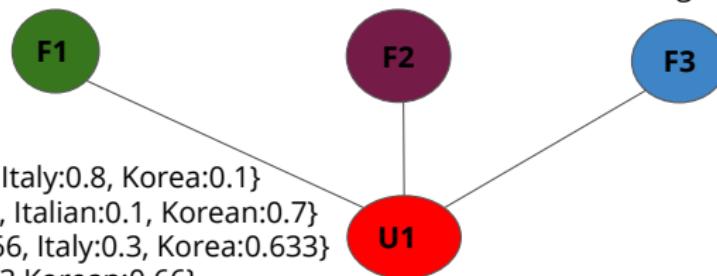
- ▶ Country of residence: "the country with the longest length of stay"
- ▶ Country of nationality: "the ensemble of features that make a person feel like they belong to a certain country"

# Identifying migrants on Twitter (Kim et al., 2020)

$\text{loc}^{F1} = \{\text{France}: 0.2, \text{Italy}: 0.8\}$   
 $\text{lang}^{F1} = \{\text{Italian}: 1\}$

$\text{loc}^{F2} = \{\text{Italy}: 0.1, \text{Korea}: 0.9\}$   
 $\text{lang}^{F2} = \{\text{Korean}: 1\}$

$\text{loc}^{F3} = \{\text{Korea}: 1\}$   
 $\text{lang}^{F3} = \{\text{Korean}: 1\}$



$\text{loc}^{U1} = \{\text{France}: 0.1, \text{Italy}: 0.8, \text{Korea}: 0.1\}$   
 $\text{lang}^{U1} = \{\text{French}: 0.2, \text{Italian}: 0.1, \text{Korean}: 0.7\}$   
 $\text{floc}^{U1} = \{\text{France}: 0.066, \text{Italy}: 0.3, \text{Korea}: 0.633\}$   
 $\text{flang}^{U1} = \{\text{Italian}: 0.33, \text{Korean}: 0.66\}$

# Validation

Are we correctly identifying migrants?  
Are there good ground truth data to compare?

## Validation-Gold standard data (Huang et al., 2014)

	Pre.	Rec.	F1
QA	86.67%	95.37%	90.81%
ARA	82.96%	71.16%	76.56%
WES	70.86%	70.62%	70.64%
SA	93.35%	90.48%	91.89%
IN	82.19%	71.13%	76.00%
OTH	78.76%	40.72%	53.54%
UN	30.78%	15.13%	20.16%

**Table:** The average Precision, Recall, and F1 scores for each nationality group Huang et al., 2014

# Validation-Official statistics (Kim et al., 2020)

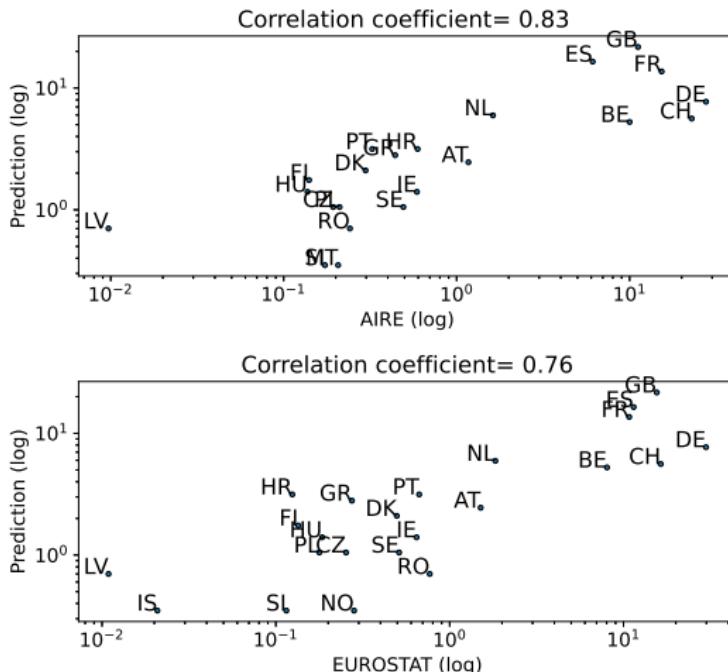


Figure: Correlation between predicted data and official statistics (Kim et al., 2020)

# Inferring international and internal migration patterns from Twitter data (Zagheni et al., 2014)

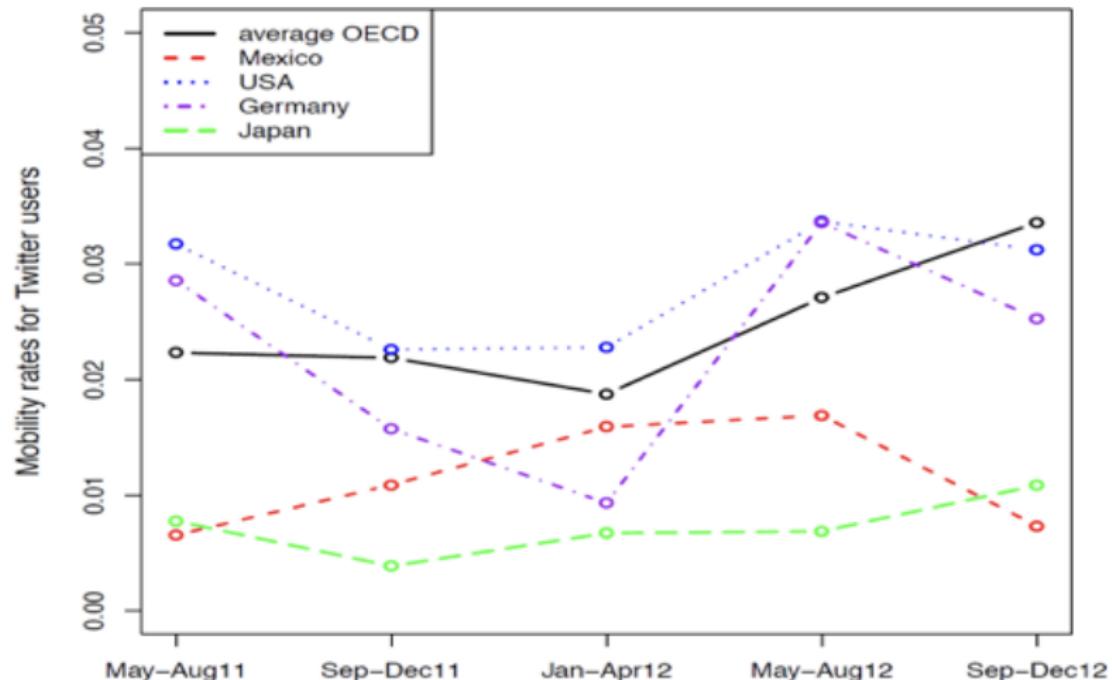


Figure: Mobility rates for Twitter users Zagheni et al., 2014

## Difference-in-Differences

- ▶ Out-migration rates clearly an overestimate
- ▶ Non-representative user set
- ▶ Selection bias is changing over time
- ▶ Focus on between-country differences

---

$$\hat{\delta}_c^t = (m_c^t - m_{oecd}^t) - (m_c^{t-\Delta} - m_{oecd}^{t-\Delta})$$

→ Diff-in-diff estimator to evaluate relative changes in trends  
(Zagheni et al., 2014)

Cont.

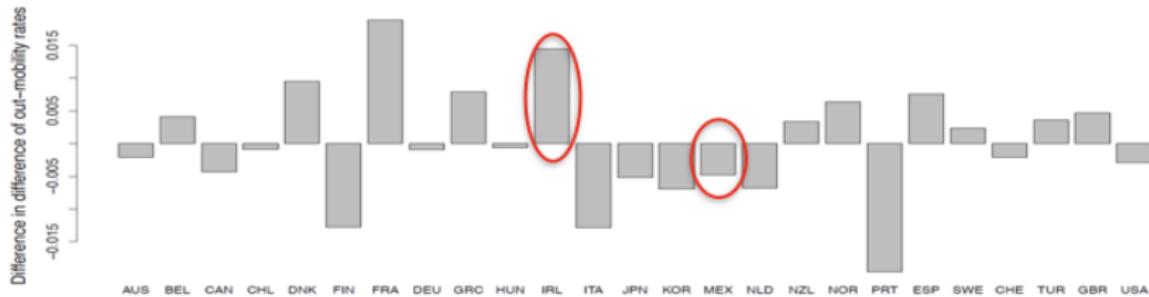


Figure: (Soft) Validation:Ireland out-migration rate grew by 2.2% 2011 -i 2012, more than most countries (Irish Central Statistics Office) Mexico also sees a reduction in out-migration (Pew Research Center)

## Other migration related researches

- ▶ Cultural integration (Kim et al., 2021)
- ▶ Spatial integration (Mazzoli et al., 2020; Lamanna et al., 2018)
- ▶ Sentiment analysis (Öztürk et al., 2018; Arcila-Calderón et al., 2021)
- ▶ Social integration (Kim et al., 2023; Kim et al., 2022)

# Cultural integration (Kim et al., 2021)

Q. *How much do migrants absorb the culture of their destination society? Do they lose connection with their home country?*

Traditional studies focus on elements such as:

- ▶ Language proficiency
- ▶ Marital status
- ▶ Role of media ...

	Low OA	High OA
Low DA	Marginalisation	Separation
High DA	Assimilation	Integration

Table: Theories of integration and their relation to OA and DA Kim et al., 2021

## Origin and Destination attachment indexes (Kim et al., 2021)

$$H(h) = \frac{-\sum_c P_h(c) \log P_h(c)}{\log(|P_h(c)|)} \quad (1)$$

$$OA(u) = \frac{\# C_n(u) \text{ hashtags}}{\# \text{ total hashtags}} = \frac{HT(u, C_n(u))}{HT(u)} \quad (2)$$

$$DA(u) = \frac{\# C_r(u) \text{ hashtags}}{\# \text{ total hashtags}} = \frac{HT(u, C_r(u))}{HT(u)} \quad (3)$$

```
'Salvini':  
{'CH': 1,  
'CL': 1,  
'CZ': 2,  
'DE': 2,  
'ES': 3,  
'FR': 2,  
'GB': 1,  
'IT': 544,  
'LU': 1,  
'NL': 2,  
'PL': 1,  
'TH': 2,  
'TR': 1,  
'US': 6}  
/569
```

Entropy score=0.11  
∴ Italian specific

## OA & DA indexes

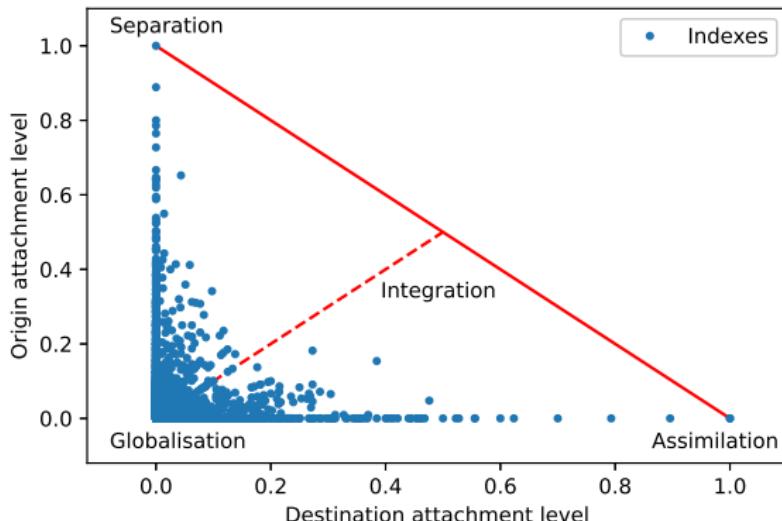


Figure: Relationship between OA & DA

# Italian emigrants in overseas

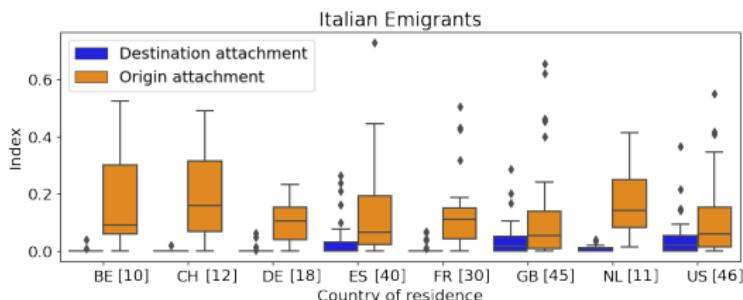
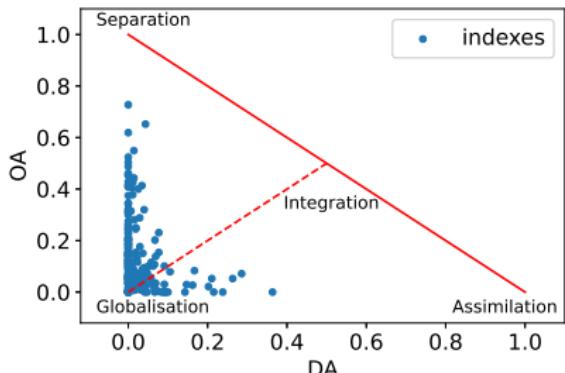
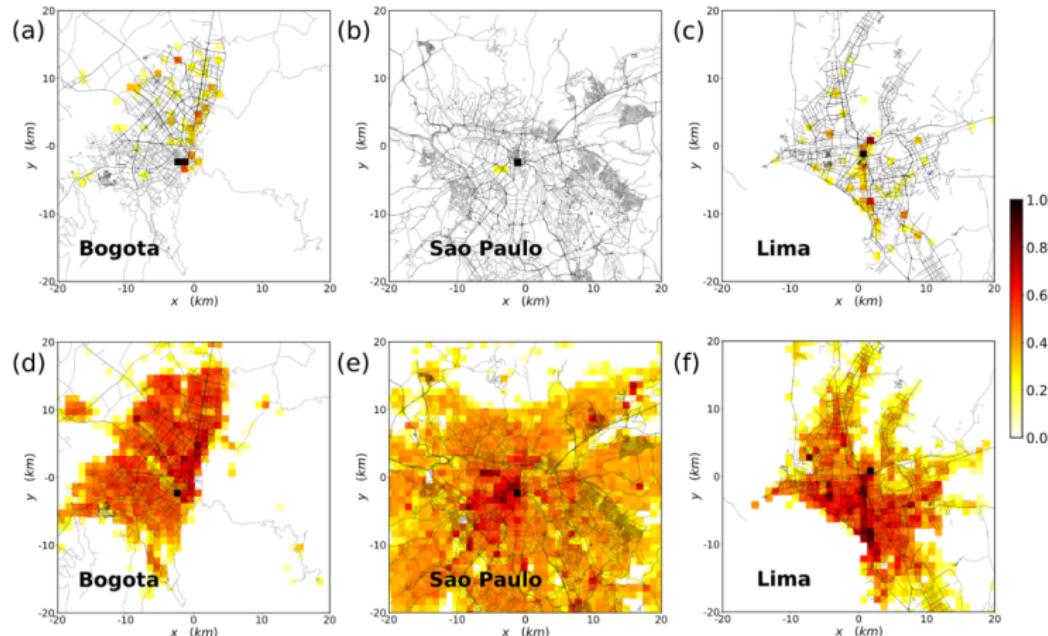


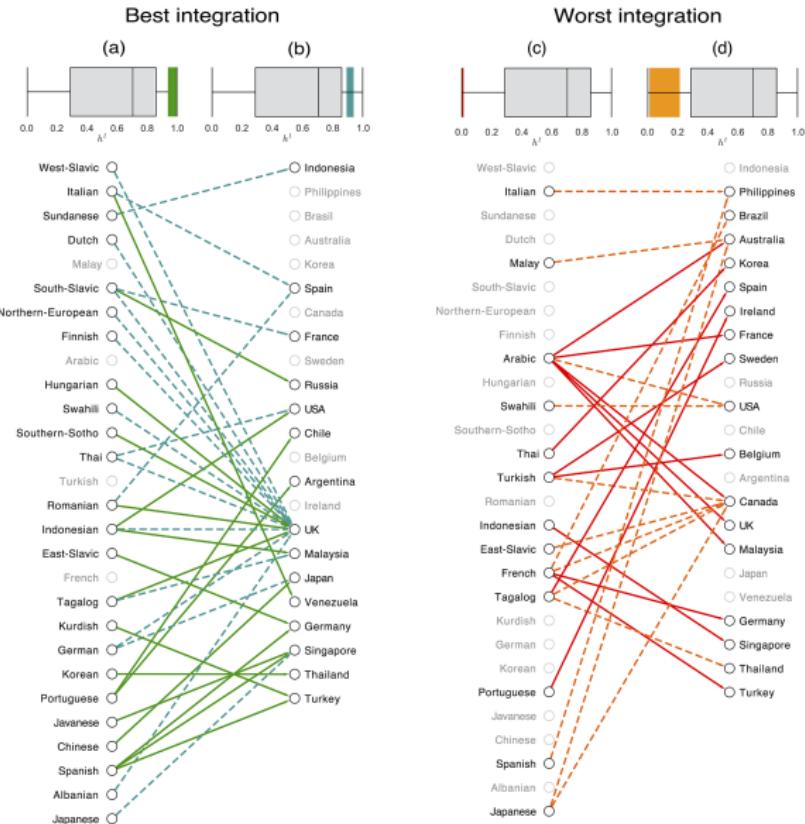
Figure: Case study of Italian emigrants in overseas

# Spatial integration (Mazzoli et al., 2020)



Spatial integration in Colombia, Brazil and Peru

# Language integration network (Lamanna et al., 2018)

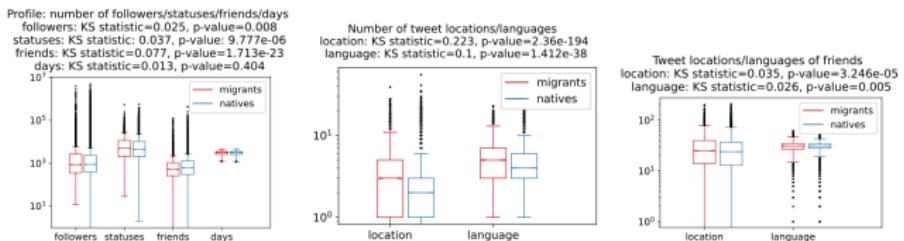


# Sentiment analysis (Arcila-Calderón et al., 2021)

Hate speech towards migrants/refugees? What are the main topics behind the messages in Twitter in Spain with hate speech against refugees?

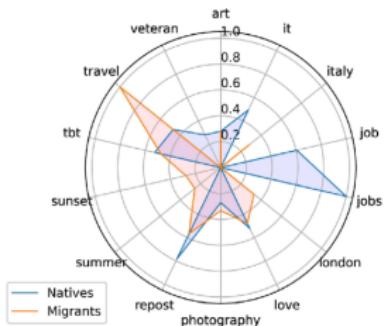
Topic	Most Common Words and Their Frequency within the Topic (the Underlined Words are the Most Determinant for the Labelling of the Topic)
Pull effect and consequences	"immigrants" (0.008) + "spain" (0.007) + "boat" (0.005) + " <u>effect</u> " (0.005) + " <u>pull</u> " (0.004) + "goes" (0.004) + "valencia" (0.004) + " <u>consequences</u> " (0.004) + "europe" (0.004) + " <u>concentration</u> " (0.004)
Pull effect and not welcoming "illegals"	"immigrants" (0.013) + "people" (0.008) + "spain" (0.008) + " <u>pull</u> " (0.007) + " <u>effect</u> " (0.007) + "illegal" (0.007) + " <u>protectyourborders</u> " (0.007) + " <u>spaniards</u> " (0.006) + "harbor" (0.006) + " <u>aquariusnotwelcome</u> " (0.006)
Not welcoming and terrorism	"go" (0.008) + "spain" (0.008) + "people" (0.006) + " <u>aquariusnotwelcome</u> " (0.005) + "country" (0.005) + " <u>boko</u> " (0.005) + " <u>haram</u> " (0.005) + "immigrants" (0.005) + "boat" (0.004) + "have" (0.004)
Smugglers and NGOs	"immigrants" (0.023) + "spain" (0.010) + "spaniards" + " <u>come</u> " (0.008) + " <u>mafias</u> " (0.008) + "people" (0.007) + "go" (0.006) + "illegal" (0.006) + " <u>ngos</u> " (0.006) + "boat" (0.006)
Money and jobs	"refugees" (0.011) + "immigrants" (0.009) + " <u>pay</u> " (0.007) + "spain" (0.007) + "countries" (0.005) + "boat" (0.005) + "spaniards" (0.005) + "people" (0.005) + " <u>solution</u> " (0.005) + "work" (0.005)
Entrance to Europe	"spain" (0.018) + "immigrants" (0.009) + "europe" (0.006) + "boat" (0.006) + "valencia" (0.005) + "spaniards" (0.005) + "government" (0.005) + "immigration" (0.005) + " <u>north</u> " (0.004) + " <u>millions</u> " (0.005)

# Characteristics of migrants on Twitter (Kim et al., 2022)



**Fig. 2** Left: Distributions of profile features: number of followers, tweets published (statuses) and friends and number of days since the account was created until 2018, respectively. Centre: Distribution of

tweet locations and languages. Right: Distribution of tweet locations and languages of friends



**Figure:** Distributions of profile features and top hashtags used by migrants and natives

# Social integration (Kim et al., 2023)

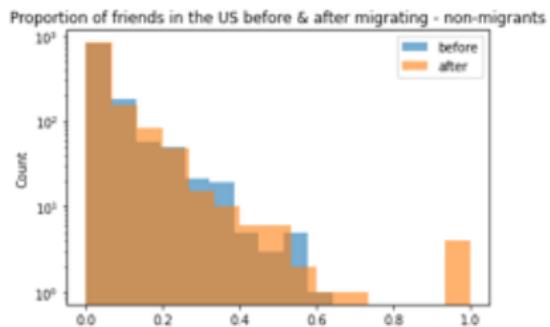
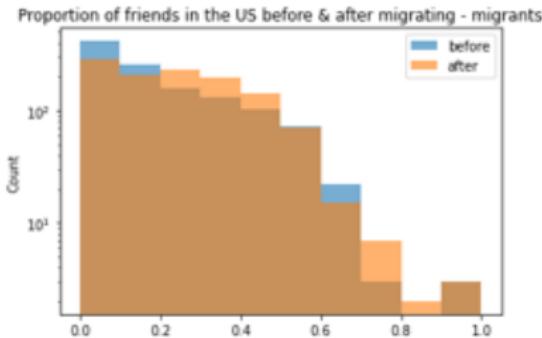


Figure: Distribution of fraction of friends residing in the United States before and after migration for migrants (left) and for propensity-score matched non-migrants (right).

# Web-scraping

## IMDb Charts

### IMDb Top 250 Movies

IMDb Top 250 as rated by regular IMDb voters

250 Titles

Sort by Ranking ↑



#### 1. The Shawshank Redemption

1994 2h 22m R

★ 9.3 ⚡ Rate



#### 2. The Godfather

1972 2h 55m R

★ 9.2 ⚡ Rate



#### 3. The Dark Knight

2008 2h 32m PG-

★ 9.0 ⚡ Rate



#### 4. The Godfather I

1974 3h 22m R

★ 9.0 ⚡ Rate



The Real Yellow Pages®

barbers

Auto Services Beauty Home Services Insurance

Barbers in Los Angeles, CA

All BBB Rated A+ A

View all businesses that are OPEN 24 Hours



#### #1 Barber Booking App - Find The Best Barbers

https://www.bookingbarber.com • The Easiest Way To Search For Top Ranked Barbers Near You. Book Your Next Haircut Today.

► Visit Website



#### ROWDTLA - Free 2 Hour Parking

https://www.rowdtla.com • Shop for Mens & Womens Apparel, Home Goods, Shoes, Kids Clothing & More All in One Place.

► Visit Website



#### Best Barbershop in Los Angeles - Best Men's Haircut Los Angeles

https://www.californiahairgroup.com/barbershop/pasadena • California Crossing upscale full service men's hair salon with car detailing while you wait. Come see...

► Visit Website



#### Barbers

Barbers, Beauty Salons

★★★★★ [1]

Wednesday, July 06, 2022

TODAY'S PAPER

## The New York Times

See

4



Associated Press / Getty Images

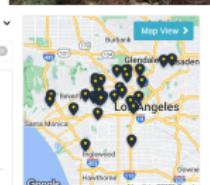


#### What 'The Bear' Gets Right About Chicago

The show celebrates a kind of audience... innocence and independence... that's often forgotten by the media. Maybe that's why it's so fitting to be important. It's a reminder.



Epstein



#### Places Near Los Angeles, CA with Barbers

Huntington Park (9 miles)

Glendale (10 miles)

West Hollywood (12 miles)

Arcadia (13 miles)

San Gabriel (14 miles)

Montebello (14 miles)

Pasadena (14 miles)

Inglewood (14 miles)

Calver City (14 miles)

Beverly Hills (15 miles)

#### More Types of Beauty Services in Los Angeles

Body Piercing

Hair Stylists

Massage Services

Nail Salons



## Hands-on example

<https://colab.research.google.com/>

# Ethics and privacy

BUT

**Bias in the data:**

“Twitter users are younger, more educated and more likely to be Democrats than general public”<sup>8</sup>

What about in other countries?

**Sample size:** Started BIG, ended small...

---

<sup>8</sup><https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/>

# Pros and cons

## Pros:

- ▶ Access
- ▶ Open box (Twitter)
- ▶ Longitudinal: access to historical data
- ▶ Aggregated data (FB)

## Cons:

- ▶ Bias
- ▶ Sample size
- ▶ Black box (FB)

## Issue:

- ▶ **Ethics and Privacy**

What can we do to mitigate the issues related to Ethics and Privacy?

# Ethics and Privacy

Twitter API is an open-access

Eavesdropping someone's conversation in a bus → can I expose this information?

Does this mean that we can do anything with the data?

→ Quick answer= NO!

# GDPR

General Data Protection Regulation was enacted in 2016: An EU law that aims to protect and regulate data and privacy.

7 principles<sup>9</sup>:

- ▶ **Lawfulness, fairness and transparency**
- ▶ Purpose limitation
- ▶ Data minimisation
- ▶ Accuracy
- ▶ **Storage limitation**
- ▶ **Integrity and confidentiality (security)**
- ▶ Accountability

For more info: <https://gdpr.eu>

---

<sup>9</sup><https://www.onetrust.com/blog/gdpr-principles/>

# Consent

Article 4(11) defines consent: “Consent of the data subject means any freely given, specific, informed and unambiguous indication of the data subject’s wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her.”<sup>10</sup>

Consent also means that “the data subject knows your identify, what data processing activities you intend to conduct, the purpose of it and that they can withdraw their consent at any time.”<sup>11</sup>

---

<sup>10</sup><https://gdpr.eu/gdpr-consent-requirements/>

<sup>11</sup>Idem.

# Confidentiality

- ▶ Pseudonymization: jisu → DKjxid73
- ▶ Anonymization: jisu → \*\*\*\*\*
- ▶ Data transformation: By modifying the format, value, or structure of data

## Data storage limitation

How, where and for how long can one store sensitive and personal data?

- Before storing the data, think of pseudonymizing, anonymising or transforming the data
- Secure data storage: need to store data where it is safe and well protected where it can prevent possible attacks or loss of data
- The length of the time that you are planning to store the data is required by the GDPR.

## When sharing Tweets

Sharing your Twitter is permitted as long as they are for non-commercial purposes. BUT you should not share entire tweets directly!

So how? → By dehydrating your data!

“Dehydrated” data set - each tweet is reduced to its unique ID number, and a list of these IDs is saved as a text document.<sup>12</sup>

Check out also this webpage: <https://covid.dh.miami.edu/2020/06/11/hydrating-tweetsets/>

---

<sup>12</sup> <https://scholarslab.github.io/learn-twarc/06-twarc-command-basics#dehydrated-and-rehydrated-data-sets> ▶

## Take home message

“Social big data can be proposed to fill some of the gaps and complement traditional data types but cannot replace them”.

-  Arcila-Calderón, Carlos et al. (2021). "Refugees Welcome? Online Hate Speech and Sentiments in Twitter in Spain during the Reception of the Boat Aquarius". In: *Sustainability* 13.5, p. 2728.
-  Hausmann, Ricardo et al. (2018). *Measuring venezuelan emigration with twitter*. Tech. rep. Kiel Working Paper.
-  Huang, Wenyi et al. (2014). "Inferring nationalities of twitter users and studying inter-national linking". In: *Proceedings of the 25th ACM conference on Hypertext and social media*, pp. 237–242.
-  Kim, Jisu et al. (2020). "Digital footprints of international migration on twitter". In: *International Symposium on Intelligent Data Analysis*. Springer, pp. 274–286.
-  Kim, Jisu et al. (2021). "Home and destination attachment: study of cultural integration on Twitter". In: *arXiv preprint arXiv:2102.11398*.
-  Kim, Jisu et al. (2022). "Where do migrants and natives belong in a community: a Twitter case study and privacy risk analysis". In: *Social Network Analysis and Mining* 13.1, p. 15.

-  Kim, Jisu et al. (July 2023). "Online social integration of migrants: Evidence from Twitter". In: *Migration Studies*, mnad017. ISSN: 2049-5846. DOI: 10.1093/migration/mnad017. eprint: <https://academic.oup.com/migration/advance-article-pdf/doi/10.1093/migration/mnad017/50827695/mnad017.pdf>. URL: <https://doi.org/10.1093/migration/mnad017>.
-  Lamanna, Fabio et al. (2018). "Immigrant community integration in world cities". In: *PLoS one* 13.3, e0191612.
-  Laney, Doug (2001). *3D data management: Controlling data volume, velocity and variety*. URL: <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-%20Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>.
-  Mazzoli, Mattia et al. (2020). "Migrant mobility flows characterized with digital data". In: *PLoS one* 15.3, e0230264.
-  Öztürk, Nazan et al. (2018). "Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis". In: *Telematics and Informatics* 35.1, pp. 136–147.



Zagheni, Emilio et al. (2014). "Inferring international and internal migration patterns from twitter data". In: *Proceedings of the 23rd International Conference on World Wide Web*, pp. 439–444.

## Supplement materials

F1-score measures the accuracy;

$$F1score = 2 * \frac{precision * recall}{precision + recall}$$

- ▶ Precision: How many selected items are relevant?

$$\frac{Truepositive}{Truepositive + Falsepositive}$$

- ▶ Recall: How many relevant items are selected?

$$\frac{Truepositive}{Truepositive + Falsenegative}$$

The score ranges from its worst score of 0 to its best score 1.