INTRODUCTION TO SOCIAL MEDIA AND BIG DATA FOR MIGRATION STUDIES

Google Trends data in migration studies

Ebru Şanlıtürk
sanlituerk@demogr.mpg.de

GOOGLE TRENDS DATA IN MIGRATION STUDIES

WHAT IS GOOGLE TRENDS?

Google Trends is a tool by Google, that shows the **relative** interest over time and/or by subregion for any selected query, time period and location.

(Trends Help, 2021)

(see: https://support.google.com/trends/answer/4365533?hl=en&ref_topic=6248052)

WHAT DO GOOGLE TRENDS DATA TELL US?

- Interest for a selected query over time

Search interest for a topic as a proportion of all searches on all topics on Google at the specified time and location

- Interest for a selected query by subregions

Search interest for a topic by subregions as a proportion of all searches on all topics on Google in that same place and time.

WHAT DO GOOGLE TRENDS DATA TELL US?

- Google Trends **does not** report the overall search volume for a selected query.

Google Ads – Keyword Planner is meant for insights into monthly and average search volumes, specifically for advertisers to assess the size of the audience (https://support.google.com/google-ads/answer/6325025)

- It gives us a measure of interest for a query normalized for the selected time and location.

(Trends Help, 2021)

NORMALIZATION

- Google Trends normalizes search data to make comparisons between terms easier. Search results are normalized to the selected time and location of a query as follows;

- “Each data point is divided by the total searches of the geography and time range it represents to compare relative popularity”
- This process is necessary to avoid the places with the most search volume to always rank the highest.
- “The resulting numbers are then scaled on a range of 0 to 100 based on a topic’s proportion to all searches on all topics”

(Trends Help, 2021)

NORMALIZATION

- Different regions that show the same search interest for a term don't always have the same total search volumes.

- The parameters we enter matter. 100 indicates the maximum search interest for a query, only for the time and location selected. Shortening and extending the selected time period may change the minimum and maximum interest points.

- Time adjustment for non-real time data

GOOGLE TRENDS IN THE LITERATURE

GOOGLE TRENDS DATA IN LITERATURE

- Epidemiology:
  - Online search data to *nowcast* outbreaks (**Flu Trends!**)

(Ginsberg, et al., 2009) (Pelat, Turbelin, Bar-Hen, Flahault, & Valleron, 2009) (Brownstein, Freifeld, & Madoff, 2009)

- Economics:
  - Online search data to forecast unemployment rate, economic activity, inflation rate

(Ettredge, Gerdes, & Karuga, 2005) (Askitas & Zimmermann, 2009) (Choi & Varian, 2009) (Guzman, 2011)

GOOGLE TRENDS DATA IN DEMOGRAPHY LITERATURE

- Demography
- Online search data to *forecast* abortions, fertility behaviour, suicides and causes of mortality

(Reis & Brownstein, 2010)

(Billari, D'Amuri & Marcucci, 2016)

(Wilde, Chen & Lohmann, 2020)

(McCarthy, 2010)

(Song, et al., 2014)

(Chang, Kwok, Cheng, Yip, & Chen, 2015)

(Solano, et al., 2016)

(Ricketts & Silva, 2017)

GOOGLE TRENDS DATA IN MIGRATION STUDIES

- Use in migration research

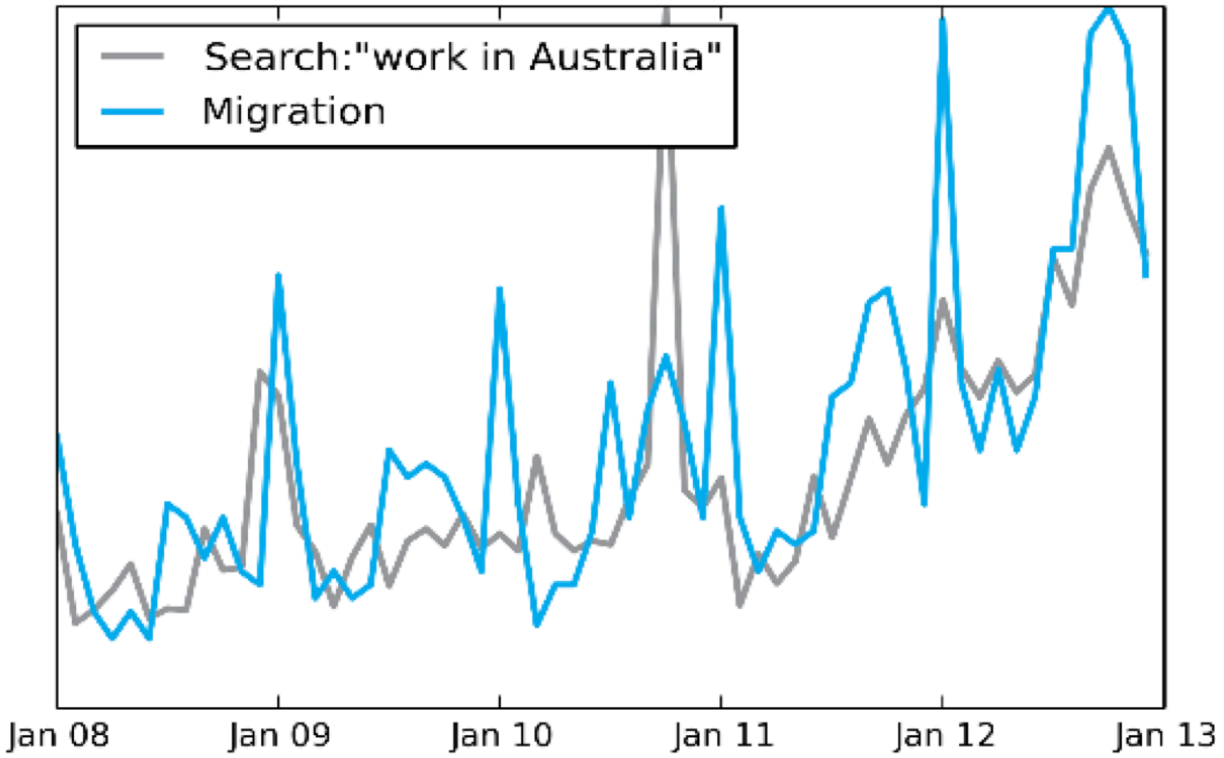
- Estimating migration flows
- Estimating migration stocks
- Now-casting and forecasting

GOOGLE TRENDS DATA IN MIGRATION STUDIES LITERATURE

- Migration from Latin America to Spain & Google search (Wladyka, 2013)
- UN Global Pulse 2014 – Estimating migration flows using online search data
- Internal migration & Bing search (Lin, Cranshaw & Counts, 2019)
- Syrian refugees & Google search (Connor, 2017)
- Predicting international migration with online search keywords (Böhme, Gröger & Stöhr, 2020)

UN Global Pulse (2014), p.13

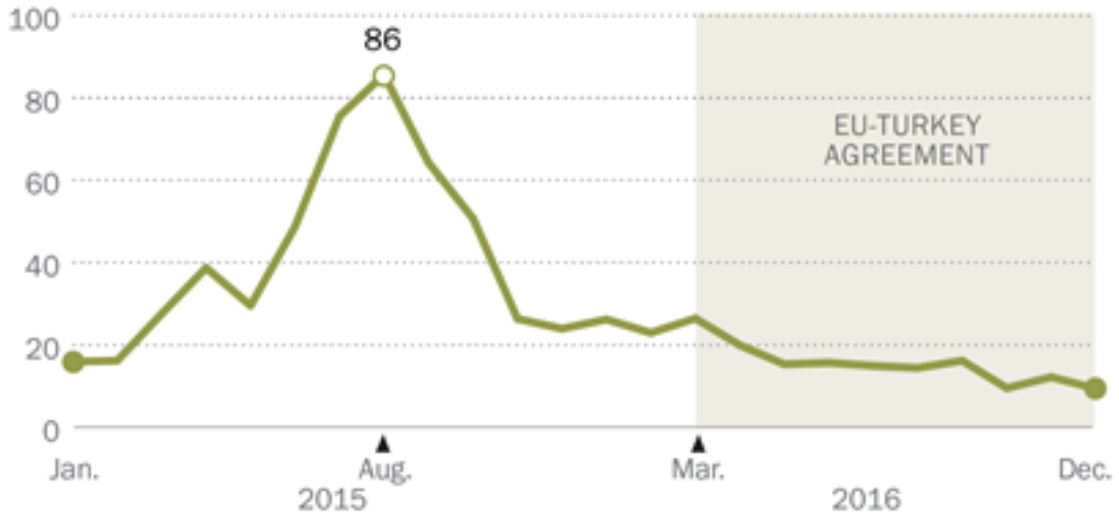
- **Online search as proxy for migration statistics:** The results of this study demonstrate the potential for online search volumes to be used as proxy for migration statistics. This implies that people interested in migrating conduct online searches to explore employment just prior to migrating, and thus search data could be used as proxy for intent to migrate.



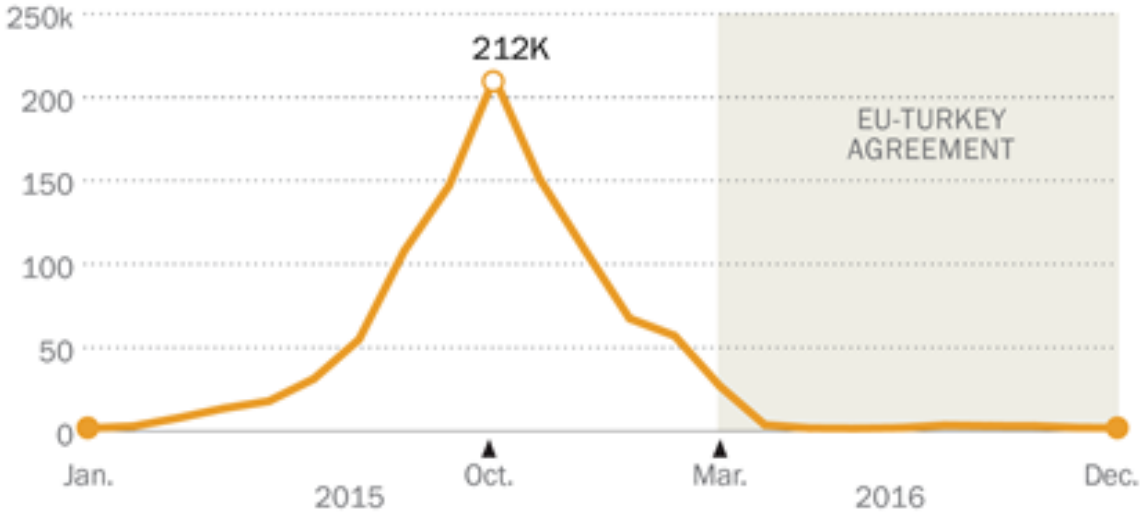
The graph above shows the trend in actual migration from Italy to Australia from January 2008 to December 2013 (blue line) and Google search activity from Italy for the query 'work in Australia' (grey line). The correlation value for migration from Italy to Australia with search query 'work in Australia' is r=0.74, p<0.001.

Surge of Arabic searches for 'Greece' in Turkey preceded surge in refugees arriving in Greece

Google Trends: Relative volume of Arabic-language Google searches for "Greece" by users in Turkey



Monthly arrivals (in thousands) of migrants into Greece



Note: Google trends data do not indicate the number of searches but instead are standardized data, displaying the relative change in searches over the time period on a 0 to 100 scale. Google trends are monthly averages based on weekly volume. Search data are for the term "Greece" in Arabic (اليونـــان). Arrivals into Greece are for all nationalities, not only Arabic speakers. See methodology for more details.

Sources: Pew Research Center analysis of Google Trends (accessed on March 3, 2016, at 1:17 p.m.) and United Nations High Commissioner for Refugees (UNHCR) data, accessed March 13, 2017.

"The Digital Footprint of Europe's Refugees"

PEW RESEARCH CENTER

Connor. (2017), Pew Research Center

Böhme, Gröger & Stöhr. (2020), Journal of Development Economics

- Google Trends Index explains significant amount of variance in unilateral and bilateral migration flow estimations.
  - Stronger with more widespread internet use.
  - Outperforms Gallup survey in the predicting migration intentions.
- Google Trends data can help to generate short-term predictions of current migration flows ahead of official data releases, which could help policy-makers **as well as aid organizations in case of humanitarian crises.**

Table 1
List of main keywords.

English	French	Spanish
applicant	candidat	solicitante
arrival	arrivee	llegada
asylum	asile	asilo
benefit	allocation sociale	beneficio
border control	controle frontiere	control frontera
business	entreprise	negocio
citizenship	citoyennete	ciudadania
compensation	compensation	compensacion
consulate	consulat	consulado
contract	contrat	contrato
customs	douane	aduana
deportation	expulsion	deportacion
diaspora	diaspora	diaspora
discriminate	discriminer	discriminar
earning	revenu	ganancia
economy	economie	economia
embassy	ambassade	embajada
emigrant	emigre	emigrante
emigrate	emigrer	emigrar
emigration	emigration	emigracion
employer	employer	empleador
employment	emploi	empleo
foreigner	etranger	extranjero
GDP	PIB	PIB
hiring	embauche	contratacion
illegal	illegal	ilegal
immigrant	immigre	inmigrante
immigrate	immigrer	inmigrar
immigration	immigration	inmigracion
income	revenu	ingreso
inflation	inflation	inflacion
internship	stage	pasantia
job	emploi	trabajo
labor	travail	mano de obra
layoff	licenciement	despido

UNDERSTANDING GOOGLE TRENDS DATA IN MIGRATION CONTEXT

GOOGLE TRENDS – UNDERSTANDING THE DATA

- Google Trends, while a big data source in itself, limits our access to aggregated and normalized data

- Google Trends gives us a proxy for the intended behavior, i.e. in the case of migration studies intention to move

- Google Trends allows us to form variable for intention to move measured at any given location and any given time
  - as known as Search Volume Index or Google Trends Index

OVERLOOK AT GOOGLE TRENDS DATA

- Data does not show the volume of Google searches but its popularity.
- Calculated and normalized by Google
- Data are anonymized, categorized, and aggregated.
- Sample data

OVERLOOK AT GOOGLE TRENDS DATA

- There are two types of Google Trends data that can be accessed:

- Real-time data covering the last seven days.
  - Time unit: hour
- Non-real time data (a separate sample from real-time data)
  - Between 2004 and up to 36 hours prior

GOOGLE TRENDS – DATA PROCESSING

- Google processes data prior to reporting Google Trends output.

- The data pre-processing includes;
  - filtering irregular activities (some may still remain),
  - sampling,
  - placing thresholds

GOOGLE TRENDS – DATA PROCESSING

- Google Trends data excludes;

- Search terms with low volume that cannot pass the threshold (appear as "0”)
- Repeated searches from the same person over a short period of time as irregular activity.
- Queries with apostrophes and other special characters.

(Trends Help, 2021)

REPRESENTATIVENESS

- Google Trends output is calculated based on a representative sample instead of the entire volume of Google searches. This is due to the too big volume of Google searches, exceeding billions of searches per day.
- We don’t know the exact sampling methodology used by Google.

- Even if you search for trends using the same parameters, you may get very slightly different results, due to the sample. These are statistically not significant – but can do a robustness check.

REPRESENTATIVENESS

- Beware of the representation bias while using digital trace data
- Google usage is mostly more widespread than use of a certain social media outlet, but is still bound by the same limitations
- In statistical analyses using an adjustment factor is encouraged
  - such as the Google search engine market share or internet penetration rate

NON-REAL TIME DATA – REPORTING

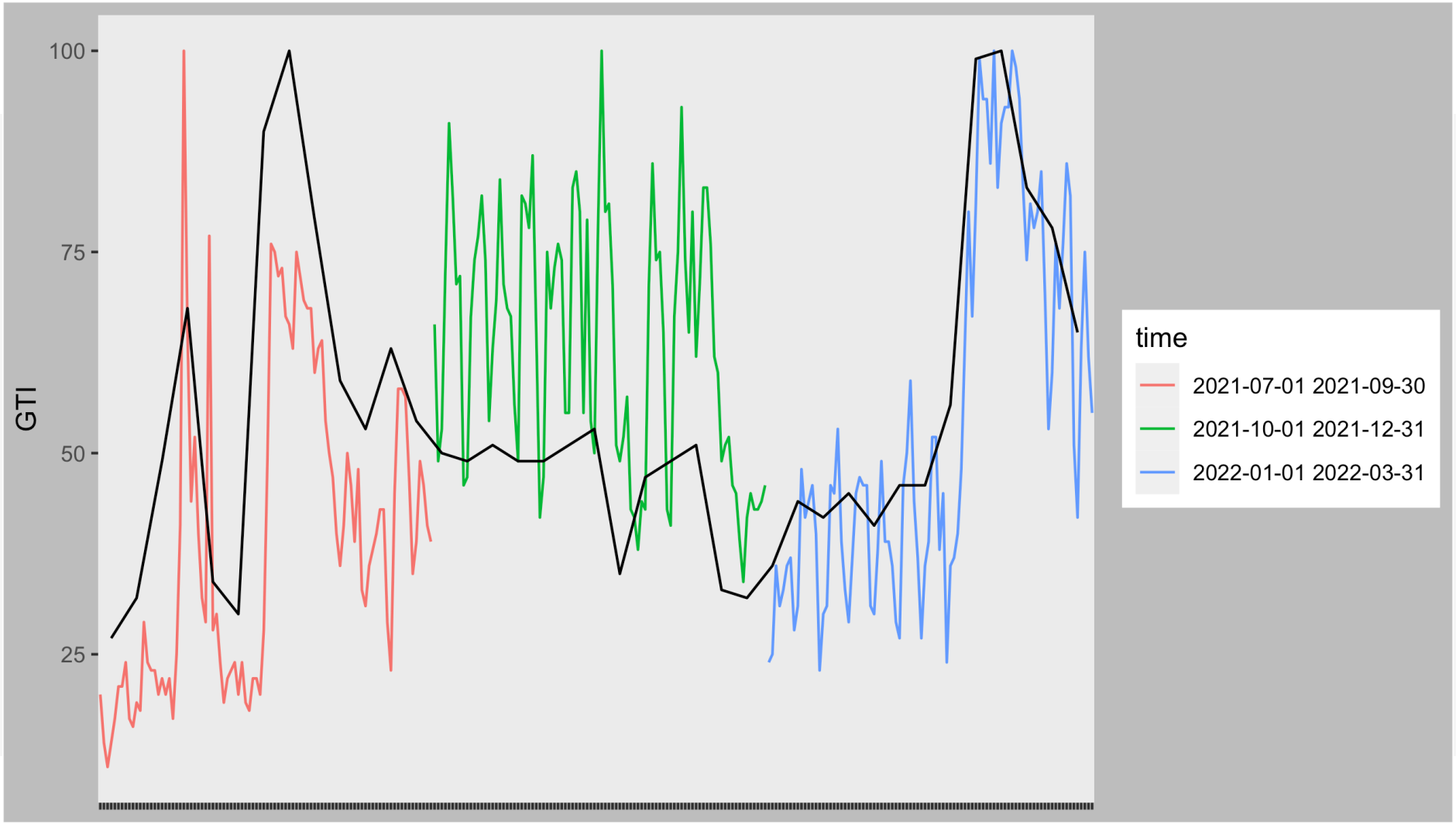
➢ Time unit of non-real time data reporting depends on the selected time period

- Up to 7 days hourly data
- **Up to 9 months** daily data
- Up to 5 years weekly data
- + 5 years monthly data

EXTENDING THE TIME PERIOD

- If we need daily data for longer than 9 months or weekly data for longer than 5 years, we need to download them separately.
- Normalization problem
- Google Trends normalizes the data for the given time period. Merging different time periods requires additional adjustments or normalization.

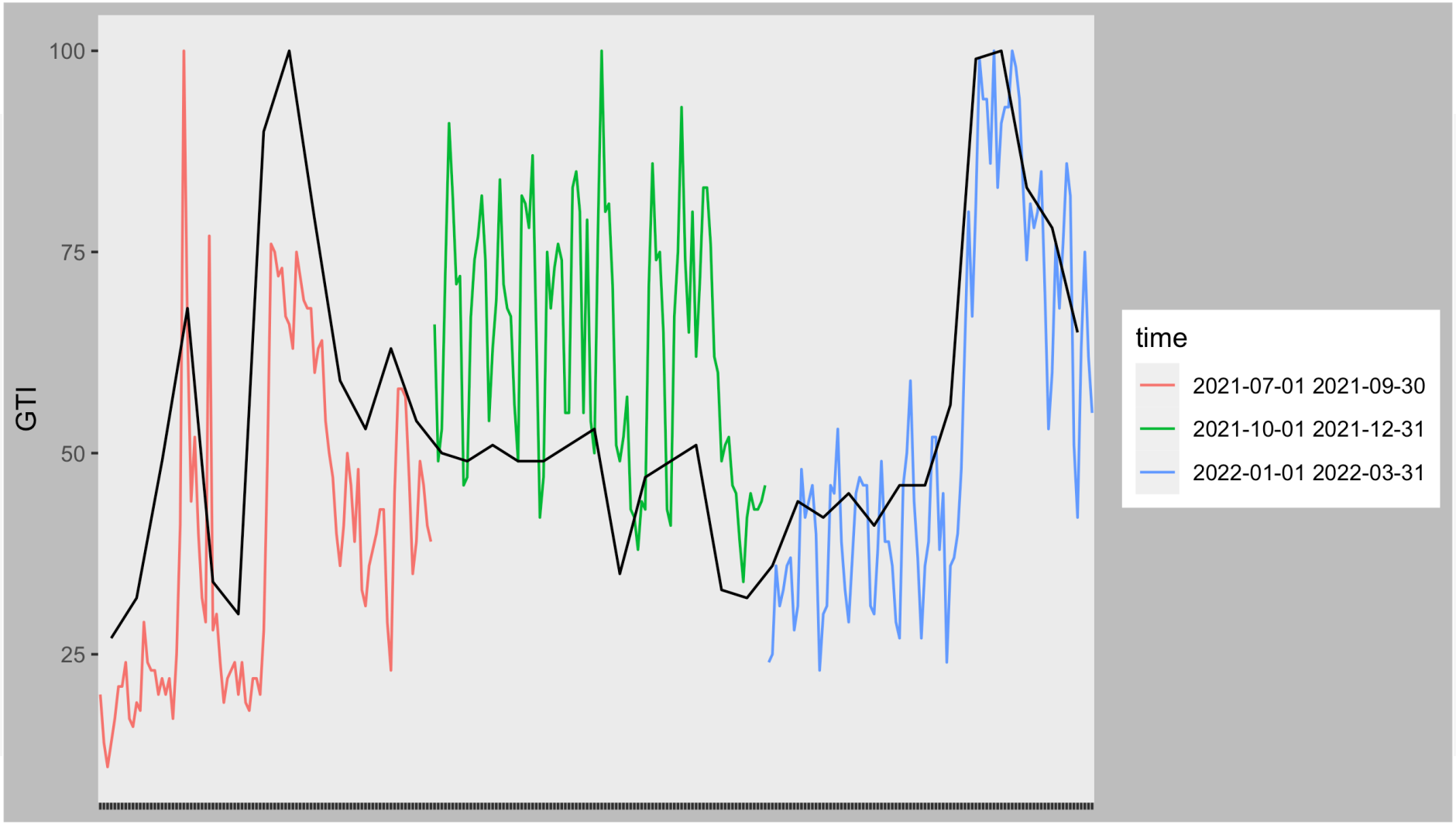
MERGED DAILY DATA



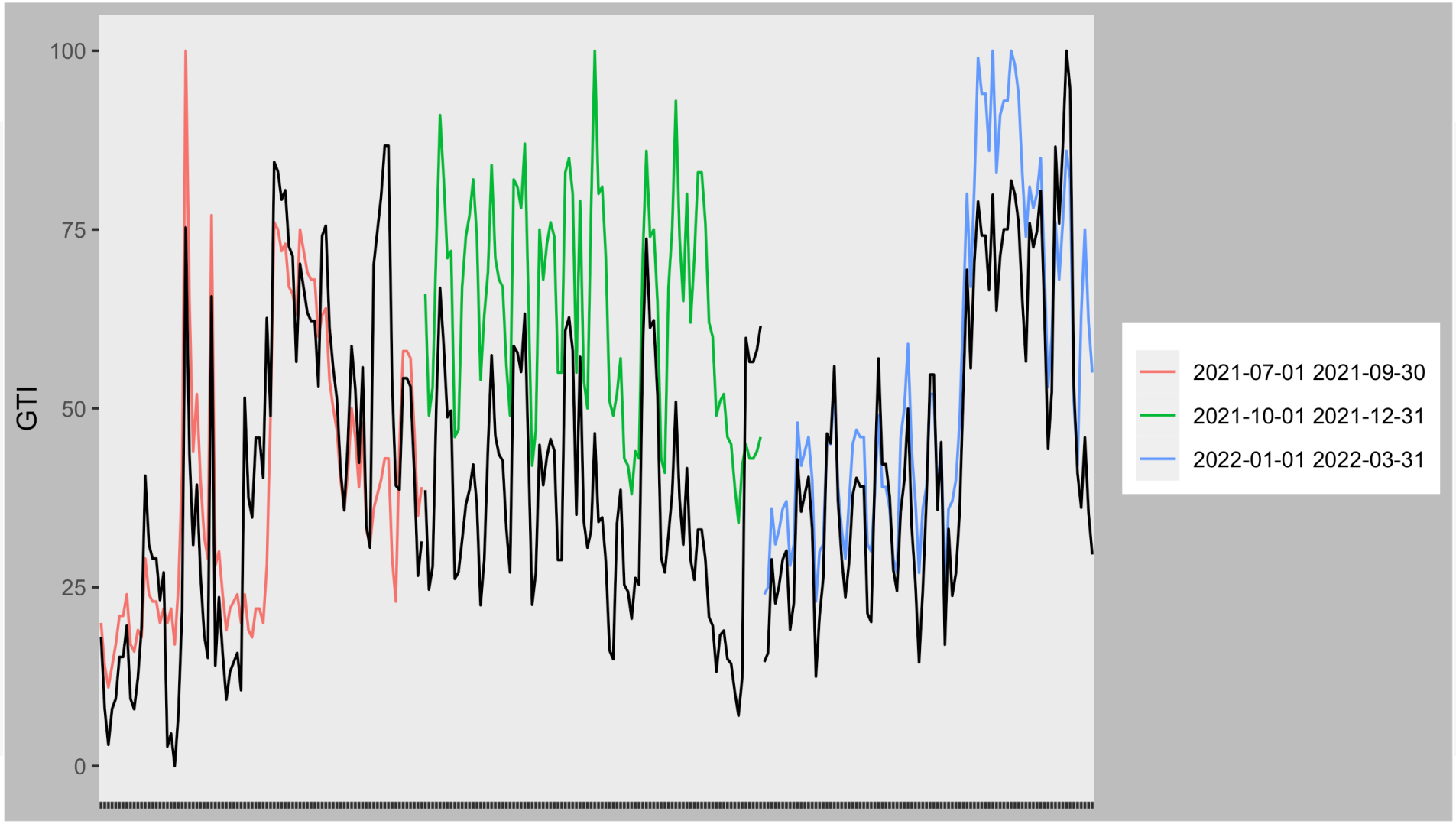
EXTENDING THE TIME PERIOD

- Create an adjustment factor
- Combine daily (or weekly) data sets
- Download weekly (or monthly) data set for the same time period
- Calculate the adjustment factor by the overlapping dates and apply the adjustment to the daily data of the same week (weekly data of the same month)
  - (Johansson, 2014; Risteski & Davcev, 2014)
- Rescale to 0-100 range

EXTENDING THE TIME PERIOD



EXTENDING THE TIME PERIOD



ALTERNATIVES?

YANDEX KEYWORD STATISTICS

- Useful for research on Russian-speaking communities
- Used in several countries apart from Russia, but Google is clearly the market leader
- https://wordstat.yandex.com

YANDEX KEYWORD STATISTICS

- Monthly data for 2 years, weekly data for 1 year
- Custom date selection is not possible
- Provides absolute numbers of searches as well as relative figures
- Provides a distinction between searches made on all and mobile devices

YANDEX KEYWORD STATISTICS

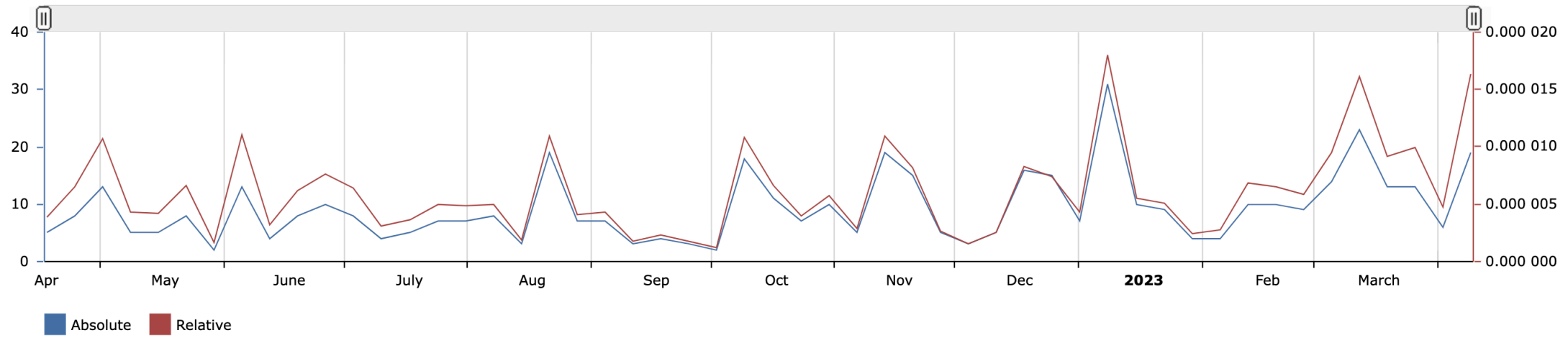
Direct Directory Metrica Advertising Network Market more Logout

Yandex
Keyword statistics

migration Submit
By keyword By region Query history United Kingdom

Impressions history for keyword "migration"

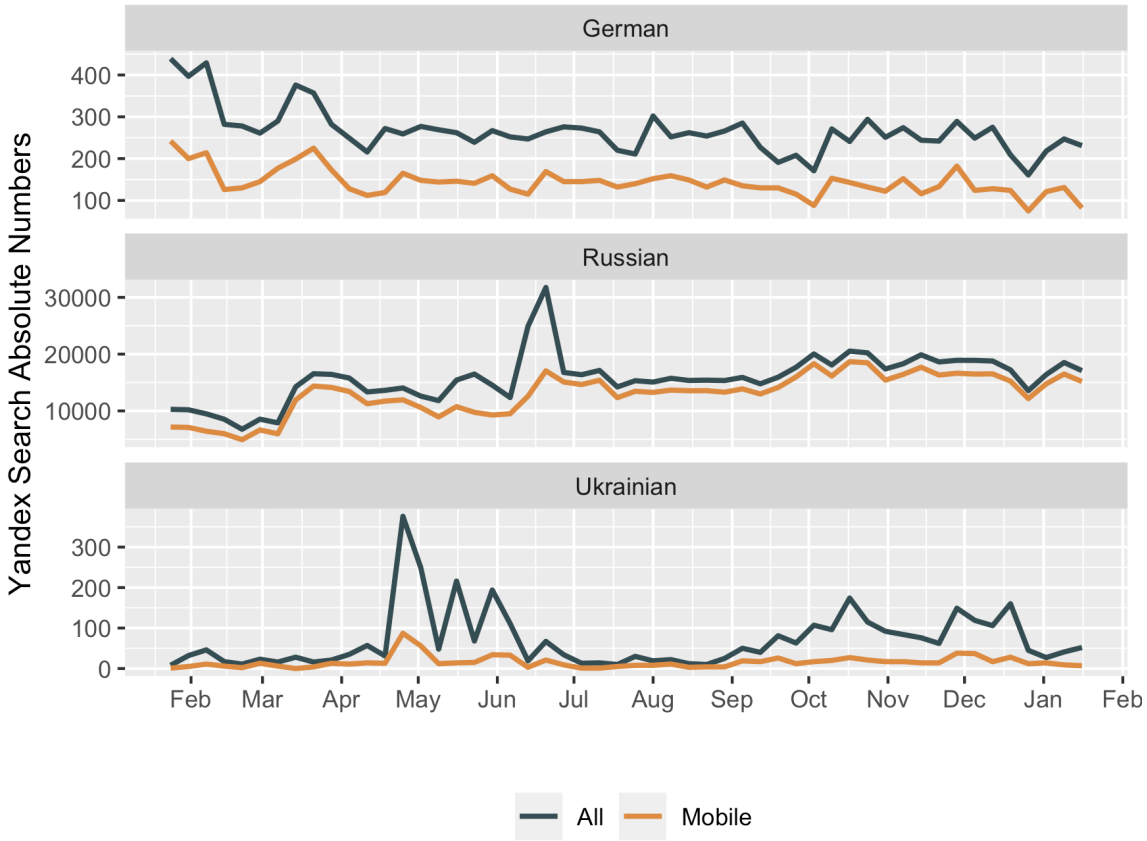
Group by: month week | All Desktop Mobile Phones only Tablets only



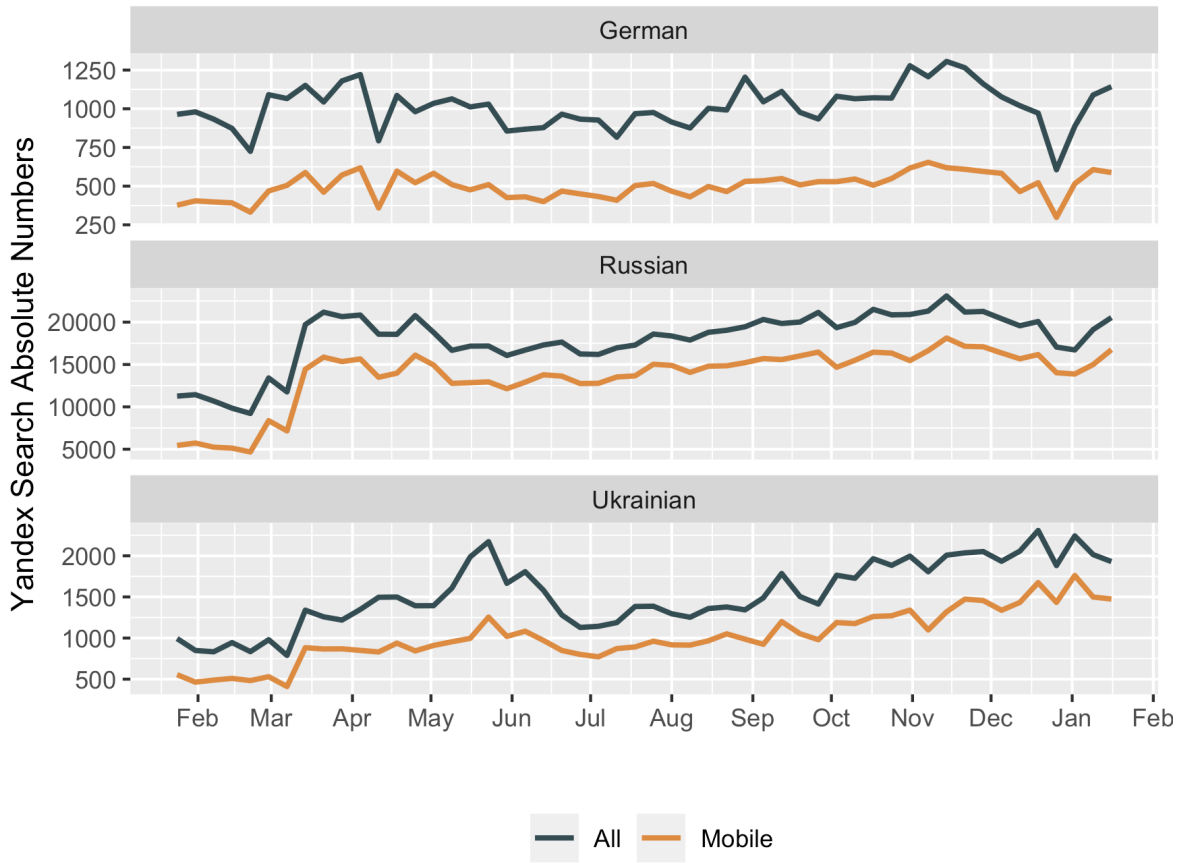
Period	Absolute	Relative	Period	Absolute	Relative
11.04.2022 - 17.04.2022	5	0.000 003 859 177	10.10.2022 - 16.10.2022	11	0.000 006 683 209
18.04.2022 - 24.04.2022	8	0.000 006 478 142	17.10.2022 - 23.10.2022	7	0.000 003 924 607
25.04.2022 - 01.05.2022	13	0.000 010 695 592	24.10.2022 - 30.10.2022	10	0.000 005 731 116
02.05.2022 - 08.05.2022	5	0.000 004 285 441	31.10.2022 - 06.11.2022	5	0.000 002 855 059

YANDEX KEYWORD STATISTICS

Health-related keywords



Job-related keywords



Sanliturk & Deimantas (2023)

BING SEARCH?

- See:

  - Lin, A. Y., Cranshaw, J., & Counts, S. (2019). Forecasting US Domestic Migration Using Internet Search Queries. Proceedings of the 2019 World Wide Web Conference (WWW'19), (pp. 13-17).

USING GTRENDSR TO RETRIEVE DATA

HOW TO MAKE A QUERY

- Determine the keyword(s)
- https://trends.google.com/ can be helpful in the brainstorming phase
- See which keywords produce a meaningful result

HOW TO MAKE A QUERY

- Being more specific with selected keywords helps narrow down the focus to the matter of interest

- When you determine the parameters, download the data using *gtrendsR* package by Massicotte & Eddelbuettel

  - For further information, see https://github.com/PMassicotte/gtrendsR

HOW TO MAKE A QUERY

Search term	Results
tennis shoes	Results can include searches containing both tennis and shoes in any order. Results can also include searches like "red tennis shoes," "funny shoes for tennis," or "tennis without shoes." No misspellings, spelling variations, synonyms, plural, or singular versions of your terms are included.
"tennis shoes"	Results include the exact phrase inside double quotation marks, possibly with words before or after, like "red tennis shoes."
tennis + squash	Results can include searches containing the words "tennis" OR "squash."
tennis -shoes	Results include searches containing the word "tennis," but exclude searches with the word "shoes."
center + centre + centere	Results include alternative spellings like “centre” or “centere,” and common misspellings like "centere." Trends considers each version of a word a different search, including misspellings.

Google. (2022), Trends Help: Search Tips for Trends, https://support.google.com/trends/answer/4359582?hl=en&ref_topic=4365530

HINTS

- Consider alternatve spellings for search queries
- Consider the use of accented characters
- Google may aggregate results for a query with and without accented characters for local language
  - Same filter / aggregation may not apply in other locations, make a few trials
- Consider regional dialects that may apply
- Be careful when using phrases as search queries
- Consider the justification of query selection

GTRENDSR

- gtrends(keyword = “”, geo = “”, time = “”, gprop = “”, hl = “”, low_search_volume = **TRUE** , compared_breakdown = FALSE)

- time – default is last 5 years
  - “now 7-d” (last seven days), “today 1-m” (past 30 days), “today 3-m” (past 90 days), “today 12-m” (past 12 months), “Y-m-d Y-m-d”
- compared_breakdown can only be used to compare multiple keywords in a single location.

GTRENDSR OUTPUT

- **Interest over time**
  - use “onlyInterest = TRUE”

- Interest by country (or region)
- Interest by dma (designated market area)
- Interest by city
- **Related topics**
- **Related queries**

GTRENDSR – COMMON ERRORS

- Error in get_widget(… : widget$status_code == 200 is not TRUE

- Make sure you use geo identifiers as given in the *countries* data
- Try downloading the developer version
  - devtools::install_github("PMassicotte/gtrendsR")
- It's possible that you have exceeded a limit with Google Trends, try dividing the sets included in your code (keywords, locations etc.)
- It's possible that you have exceeded a limit with Google Trends for the day

GTRENDSR – COMMON ERRORS

- gtrends(keyword = "asylum", time = "2022-01-01 2022-03-29",
  gprop = "web", hl = "en", low_search_volume = TRUE,
  onlyInterest = TRUE)

- Unless specified, geo is considered worldwide
- hl – language, important for related queries, related topics and location names

GTRENDSR

```
keys = c(”migration”, “residence permit”, “asylum”)
time = "2022-01-01 2023-03-28”

for (i in keys) {
trendsoutput = gtrends(keyword=i, gprop =“web”, geo=”US", time = time,
onlyInterest = TRUE, low_search_volume = FALSE)
  Sys.sleep(5)
  results [[i]] = trendsoutput$interest_over_time
}
```

REFERENCES – GOOGLE TRENDS & MIGRATION

- Böhme, M. H., Gröger, A., & Stöhr, T. (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, *142*, 102347.
- Connor, P. (2017). *The Digital Footprint of Europe's Refugees*. Pew Research Center.
- Lin, A. Y., Cranshaw, J., & Counts, S. (2019). Forecasting US Domestic Migration Using Internet Search Queries. *Proceedings of the 2019 World Wide Web Conference (WWW'19)*, (pp. 13-17).
- Johansson, E. (2014). *Creating daily search volume data from weekly and daily data*. Retrieved from: http://erikjohansson.blogspot.com/2014/12/creating-daily-search-volume-data-from.html
- 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)*.
- UN (2014). Estimating migration flows using online search data. *Global Pulse Project Series*, *4*, 1-2.
- Wladyka, D. (2013, October). *The Queries to Google Search as Predictors of Migration Flows from Latin America to Spai*n. University of Texas at Brownsville.

REFERENCES

- Askitas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *IZA Discussion Paper No. 4201*. Institute for the Study of Labor (IZA).
- Billari, F., D'Amuri, F., & Marcucci, J. (2016). Forecasting births using Google. *CARMA 2016: 1st International Conference on Advanced Research Methods in Analytics*. Valencia: Editorial Universitat Politècnica de València.
- Brownstein, J. S., Freifeld, C. C., & Madoff, L. C. (2009). Digital disease detection – Harnessing the Web for public health surveillance. *The New England Journal of Medicine*, 360(21), 2153–2157.
- Chang, S.-S., Kwok, S. S., Cheng, Q., Yip, P. S., & Chen, Y.-Y. (2015). The association of trends in charcoal-burning suicide with Google search and newspaper reporting in Taiwan: a time series analysis. *Social Psychiatry and Psychiatric Epidemiology*, 50(9), 1451-1461.
- Choi, H., & Varian, H. 8.-9. (2009, April 10). Predicting the present with Google Trends.
- Ettredge, M., Gerdes, J., & Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, *48*(11), 87-92.
- 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), 1012-1014.
- Guzman, G. (2011). Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of economic and social measurement*, 36(3), 119-167.

REFERENCES

- McCarthy, M. J. (2010). Internet monitoring of suicide risk in the population. *Journal of Affective Disorders*, 122(3), 277-279.
- Pelat, C., Turbelin, C., Bar-Hen, A., Flahault, A., & Valleron, A. J. (2009). More diseases tracked by using Google Trends. *Emerging infectious diseases*, 15(8), 1327-1328.
- Reis, B. Y., & Brownstein, J. S. (2010). Measuring the impact of health policies using Internet search patterns: the case of abortion. *BMC Public Health*, 10(1), 514.
- Ricketts, C. F., & Silva, C. G. (2017). An analysis of morbidity and mortality using Google Trends. *Journal of Human Behavior in the Social Environment*, 27(6), 559-570.
- Solano, P., Ustulin, M., Pizzorno, E., Vichi, M., Pompili, M., Serafini, G., & Amore, M. (2016). A Google-based approach for monitoring suicide risk. *Psychiatry research*, 246, 581-586.
- Song, T. M., M., Song, J., An, J. Y., Hayman, L. L., & Woo, J. M. (2014). Psychological and social factors affecting Internet searches on suicide in Korea: a big data analysis of Google search trends. *Yonsei Medical Journal*, 55(1), 254-263.
- Wilde, J., Chen, W., & Lohmann, S. (2020). COVID-19 and the future of US fertility: what can we learn from Google? (No. 13776). IZA Discussion Papers.

THANK YOU!

GOOGLE TRENDS VISUALIZATION