



Doing Research with **Social Media Data** in Psychology

André Bittermann

Leibniz Institute for Psychology (ZPID)
Trier, Germany



Images created with DALL-E / Midjourney



Dr. André Bittermann, Dipl.-Psych.

Acting Head of
Big Data Research Unit at [ZPID](#)

Computational Psychology, Text as Data, Bibliometrics

Current social media projects:

- [Academic #TwitterMigration to Mastodon](#)
- [Twitter Data for Unobtrusive Measurement of Academics' Well-Being](#)



Leibniz-Institut für
Psychologie



Public Open Science Institute for Psychology

Located in **Trier**
(oldest city in Germany,
close to Luxembourg)

Formerly known as Leibniz-
Zentrum für
Psychologische
Information und
Dokumentation

Our Products

PubPsych

PSYINDEX

Open Test Archive

PsychTopics

PsychAuthors

KLARpsy

PreReg

PsychArchives

FDZ am ZPID

PsychOpen

Today's Topics

1. Introduction and Overview
2. Data Collection and Analysis
3. Responsible Social Media Research

1. Introduction and Overview

How Trump Consultants Exploited the Facebook Data of Millions

By Matthew Rosenberg, Nicholas Confessore and Carole Cadwalladr

March 17, 2018



<https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html>

https://commons.wikimedia.org/wiki/Category:Donald_Trump_by_year#/media/File:Donald_Trump_official_portrait.jpg

DAS MAGAZIN LETZTE AUSGABEN

Das Magazin N°48 – 3. Dezember 2016

LOGIN ePAPER

PNAS
Vol. 110 | No. 15

Abstract
Results
Conclusions
Data Availability
Acknowledgments
Supporting Information
References

Ich habe nur gezeigt, dass es die Bombe gibt

Der Psychologe Michal Kosinski hat eine Methode entwickelt, um Menschen anhand ihres Verhaltens auf Facebook minutiös zu analysieren. Und verhalf so Donald Trump mit zum Sieg.

f t e

PORTRÄT: LAUREN BAMFORD

RESEARCH ARTICLE | SOCIAL SCIENCES | 

Private traits and attributes are predictable from digital records of human behavior

[Michal Kosinski](#) , [David Stillwell](#), and [Thore Graepel](#) [Authors Info & Affiliations](#)

Edited by Kenneth Wachter, University of California, Berkeley, CA, and approved February 12, 2013 (received for review October 29, 2012)

March 11, 2013 | 110 (15) 5802-5805 | <https://doi.org/10.1073/pnas.1218772110>

 704,429 | 1,407

<https://web.archive.org/web/20180125202753/https://www.dasmagazin.ch/2016/12/03/ich-habe-nur-gezeigt-dass-es-die-bombe-gibt/>

OBSERVATORY

Facebook Knows You Better Than Anyone Else

By Douglas Quenqua

Jan. 19, 2015

<https://www.nytimes.com/2015/01/20/science/facebook-knows-you-better-than-anyone-else.html>

“It needed access to just 10 likes to beat a work colleague, 70 to beat a roommate, 150 to beat a parent or sibling, and 300 to beat a spouse.”

PNAS

Vol. 112 | No. 4

Significance

Abstract

Data Availability

Acknowledgments

Supporting Information

References

RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES | 8




Computer-based personality judgments are more accurate than those made by humans

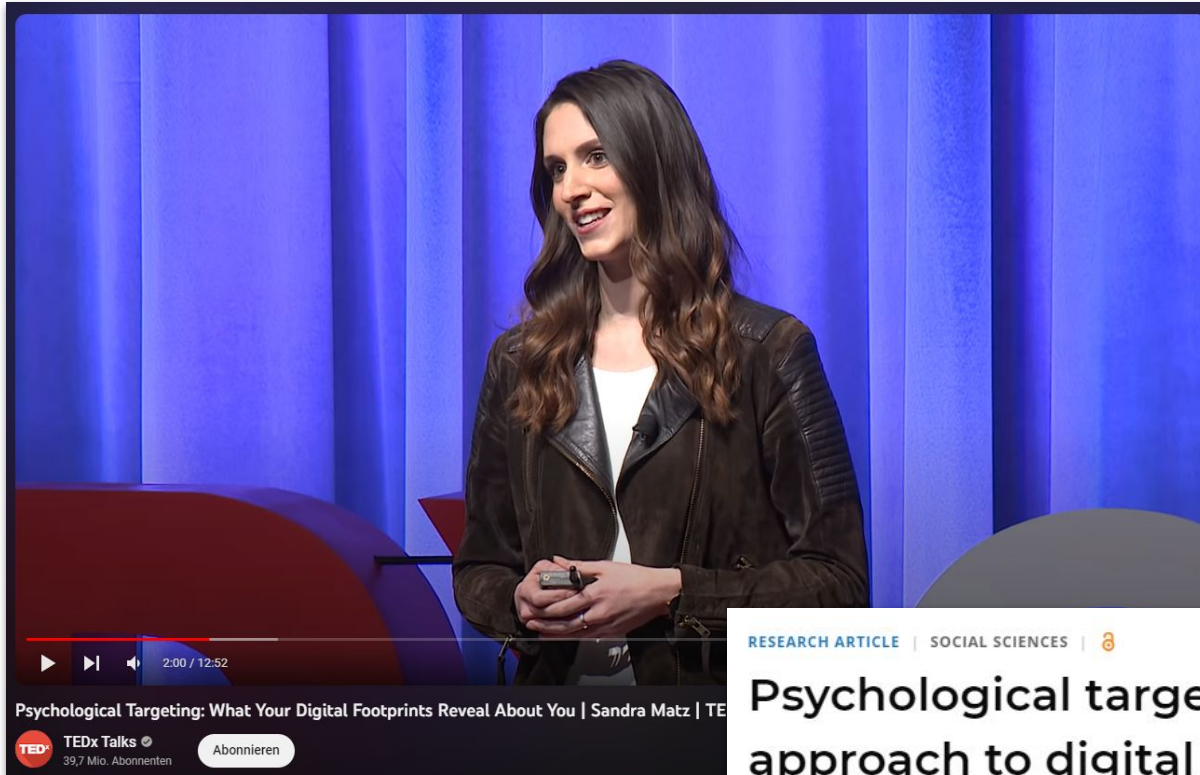
Wu Youyou , Michal Kosinski, and David Stillwell [Authors Info & Affiliations](#)

Edited by David Funder, University of California, Riverside, CA, and accepted by the Editorial Board December 2, 2014 (received for review September 28, 2014)

January 12, 2015 | 112 (4) 1036-1040 | <https://doi.org/10.1073/pnas.1418680112>

 397,456 | 515







https://www.youtube.com/watch?v=Mkl_TrPmKgA

“Recent research, however, shows that **people’s psychological characteristics** can be accurately predicted from their **digital footprints**, such as their Facebook Likes or Tweets.”

RESEARCH ARTICLE | SOCIAL SCIENCES | 8



Psychological targeting as an effective approach to digital mass persuasion

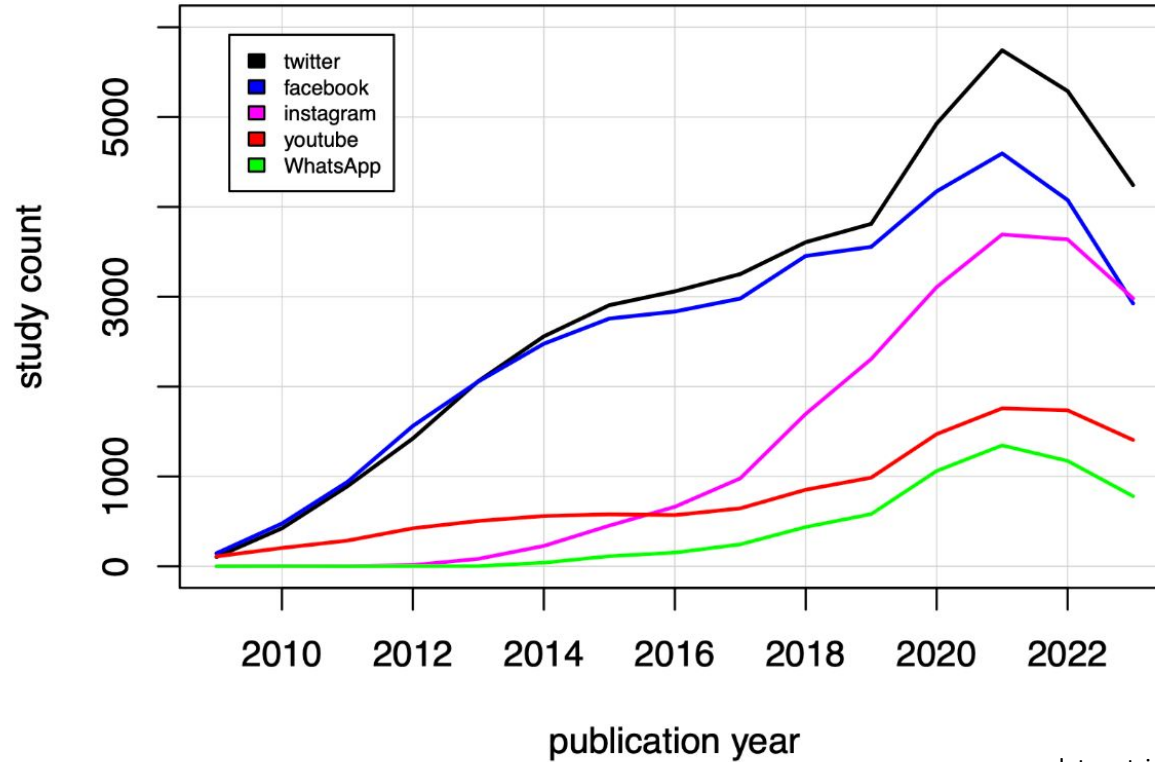
S. C. Matz , M. Kosinski , G. Nave, and D. J. Stillwell [Authors Info & Affiliations](#)

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved October 17, 2017 (received for review June 17, 2017)

November 13, 2017 | 114 (48) 12714-12719 | <https://doi.org/10.1073/pnas.1710966114>

Platform	Study Example
Facebook	Marengo, D., & Montag, C. (2020). Digital phenotyping of big five personality traits via Facebook data mining: A meta-analysis. https://doi.org/10.24989/dp.v1i1.1823
Instagram	Thömmes, K., & Hübner, R. (2022). Why people press “like”: A new measure for aesthetic appeal derived from Instagram data. https://doi.org/10.1037/aca0000331
Twitter	Cohrdes, C. et al. (2021). Indications of depressive symptoms during the COVID-19 pandemic in Germany: Comparison of national survey and Twitter data. https://doi.org/10.2196/27140
LinkedIn	Utz, S. & Breuer, J. (2019). The relationship between Networking, LinkedIn Use, and retrieving informational benefits. https://doi.org/10.1089/cyber.2018.0294
Snapchat	Bayer, J. et al. (2015). Sharing the small moments: ephemeral social interaction on Snapchat. https://doi.org/10.1080/1369118x.2015.1084349
TikTok	Haensch, A. et al. (2023). Seeing ChatGPT through Students’ eyes: An analysis of TikTok data. https://doi.org/10.48550/arxiv.2303.05349
Reddit	Zomick et al. (2019). Linguistic analysis of schizophrenia in Reddit posts. http://dx.doi.org/10.18653/v1/W19-3009
YouTube	Lam, N. H. T. et al. (2017). Exploring the role of YouTube in disseminating psychoeducation. https://doi.org/10.1007/s40596-017-0835-9
WhatsApp	Rosenfeld, A. et al. (2018). A study of WhatsApp usage patterns and prediction models without message content. https://doi.org/10.4054/DemRes.2018.39.22
Telegram	Scheffler, T. et al. (2021). The Telegram chronicles of online harm. https://doi.org/10.5334/johd.31
Tumblr	Cavazos-Rehg, P. A. et al. (2017). An analysis of Depression, Self-Harm, and Suicidal Ideation content on Tumblr. https://doi.org/10.1027/0227-5910/a000409
Mastodon	La Cava, L. et al. (2022). Information consumption and boundary spanning in decentralized online social networks: the case of Mastodon users. https://doi.org/10.1016/j.osnem.2022.100220
Bluesky	Jeong, U. et al. (2023). User Migration across Multiple Social Media Platforms. https://doi.org/10.48550/arXiv.2309.12613
Threads	Jeong, U. et al. (2023). User Migration across Multiple Social Media Platforms. https://doi.org/10.48550/arXiv.2309.12613

Social media research over the years



data retrieved from LENS.org (November 2023)

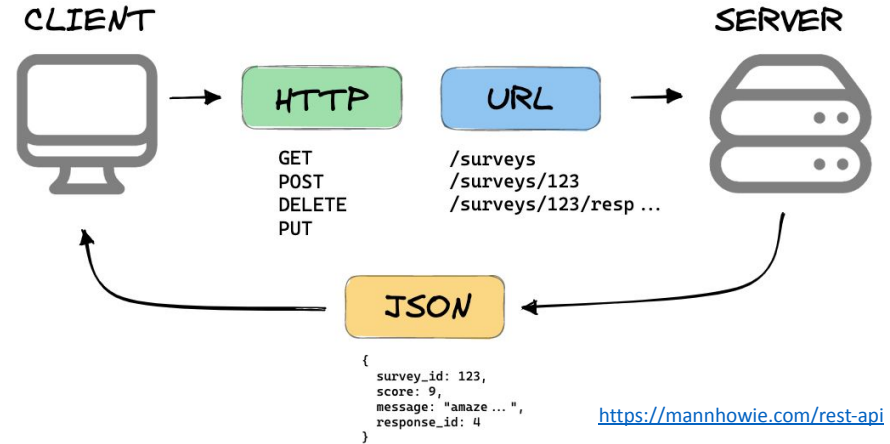
2. Collection and Analysis of Social Media Data

Data is best collected via **API**
(Application Programming Interface)

Some platforms provide free API access (e.g. Mastodon). For some (e.g. Twitter/X) you need to apply (and pay*).

WHAT IS A REST* API?

*REpresentational State Transfer



*Twitter/X is basically dead for research: 10K tweets cost \$100/month, 1M tweets \$5000/month (until April 2023: 10M/month for free)

For example, in R you can send POST request to an API endpoint with the `httr` package.

What data can be retrieved?
→ Check API documentation

```
get_client <- function(instance = "mastodon.social") {  
  url <- prepare_url(instance)  
  
  auth1 <- httr::POST(httr::modify_url(url = url, path =  
    "/api/v1/apps"), body = list(  
    client_name = "rtoot package",  
    redirect_uris = "urn:ietf:wg:oauth:2.0:oob",  
    scopes = "read write follow"  
  ))  
  
  client <- httr::content(auth1)  
  client <- client[c("client_id", "client_secret")]  
  client$instance <- instance  
  class(client) <- "rtoot_client"  
  client  
}
```

<https://github.com/gesistsa/rtoot/blob/main/R/auth.R>

Or you can use convenience functions in software packages, for example in `rtoot*` for Mastodon posts.

*posts on Mastodon used to be called “toots”

```
rtoot::get_timeline_public(instance = "fediscience.org")
```

	id	created_at	content
1	111454417409644774	2023-11-22 13:23:58	<p>Lo relevante del informe de la crisis climática de Oxfam ...
2	111454417329357701	2023-11-22 13:24:02	<p>Regardez ça, essayez de ne pas vous étouffer. <a h...
3	111454417311690981	2023-11-22 13:23:59	<p>Ouais j'ai mis un distributeur de croquettes pour que m...
4	111454417211096305	2023-11-22 13:23:56	<p>A medical nutrition formula for the treatment of relapsi...
5	111454417206670587	2023-11-22 13:22:51	<p>8-bit lorte saxofon musik KLIP "Receptformyelser samt ti...
6	111454416924710086	2023-11-22 13:23:57	<p>How to Shape a Beret</p>
7	111454416730905994	2023-11-22 13:23:53	<p>RIVM varianten update. Het groene vlak (Pirola oft...
8	111454416580525744	2023-11-22 13:23:47	<p>EURLex tip to check if a judgement has been cited in ot...





Twitter (meta)data (using rtweet package and V1.1 API)

```
> names(timelines_all)
```

[1] "user_id"	"status_id"	"created_at"	"screen_name"	"text"
[6] "source"	"display_text_width"	"reply_to_status_id"	"reply_to_user_id"	"reply_to_screen_name"
[11] "is_quote"	"is_retweet"	"favorite_count"	"retweet_count"	"quote_count"
[16] "reply_count"	"hashtags"	"symbols"	"urls_url"	"urls_t.co"
[21] "urls_expanded_url"	"media_url"	"media_t.co"	"media_expanded_url"	"media_type"
[26] "ext_media_url"	"ext_media_t.co"	"ext_media_expanded_url"	"ext_media_type"	"mentions_user_id"
[31] "mentions_screen_name"	"lang"	"quoted_status_id"	"quoted_text"	"quoted_created_at"
[36] "quoted_source"	"quoted_favorite_count"	"quoted_retweet_count"	"quoted_user_id"	"quoted_screen_name"
[41] "quoted_name"	"quoted_followers_count"	"quoted_friends_count"	"quoted_statuses_count"	"quoted_location"
[46] "quoted_description"	"quoted_verified"	"retweet_status_id"	"retweet_text"	"retweet_created_at"
[51] "retweet_source"	"retweet_favorite_count"	"retweet_retweet_count"	"retweet_user_id"	"retweet_screen_name"
[56] "retweet_name"	"retweet_followers_count"	"retweet_friends_count"	"retweet_statuses_count"	"retweet_location"
[61] "retweet_description"	"retweet_verified"	"place_url"	"place_name"	"place_full_name"
[66] "place_type"	"country"	"country_code"	"geo_coords"	"coords_coords"
[71] "bbox_coords"	"status_url"	"name"	"location"	"description"
[76] "url"	"protected"	"followers_count"	"friends_count"	"listed_count"
[81] "statuses_count"	"favourites_count"	"account_created_at"	"verified"	"profile_url"
[86] "profile_expanded_url"	"account_lang"	"profile_banner_url"	"profile_background_url"	"profile_image_url"



A look at selected **metadata** in Twitter data

```
> head(sort(table(timelines_all$source), decreasing = TRUE), 10)
```

Twitter for iPhone	Twitter for Android	Twitter Web App	Twitter Web Client	Twitter for iPad	TweetDeck
5372663	2327232	2086068	1916866	458516	158755
Facebook	Twitter for Websites	Buffer	Instagram		
94144	93897	80794	75040		

```
> head(sort(table(timelines_all$location), decreasing = TRUE), 10)
```

	London, England	London	New York, NY	Chicago, IL	Melbourne, Victoria
2133624	213172	212371	115099	94988	88694
Los Angeles, CA	UK	United Kingdom	Manchester, England		
80192	78489	78033	76727		

Data collection challenges (see [Murphy, 2017](#))

1. Data preprocessing

Slang, Emojis, multilingual data, non-Latin characters

2. Data veracity and representativity

Twitter had an API that provided a “sample” of all current tweets. It was not clear, how the data was sampled. Research has shown that this sample is not representative of all tweets ([Morstatter et al., 2013](#); [Tromble et al., 2017](#))

3. Changes in API and data structure

Data may be provided in different formats over time, making it difficult to merge data sets.

4. Bots and fake accounts

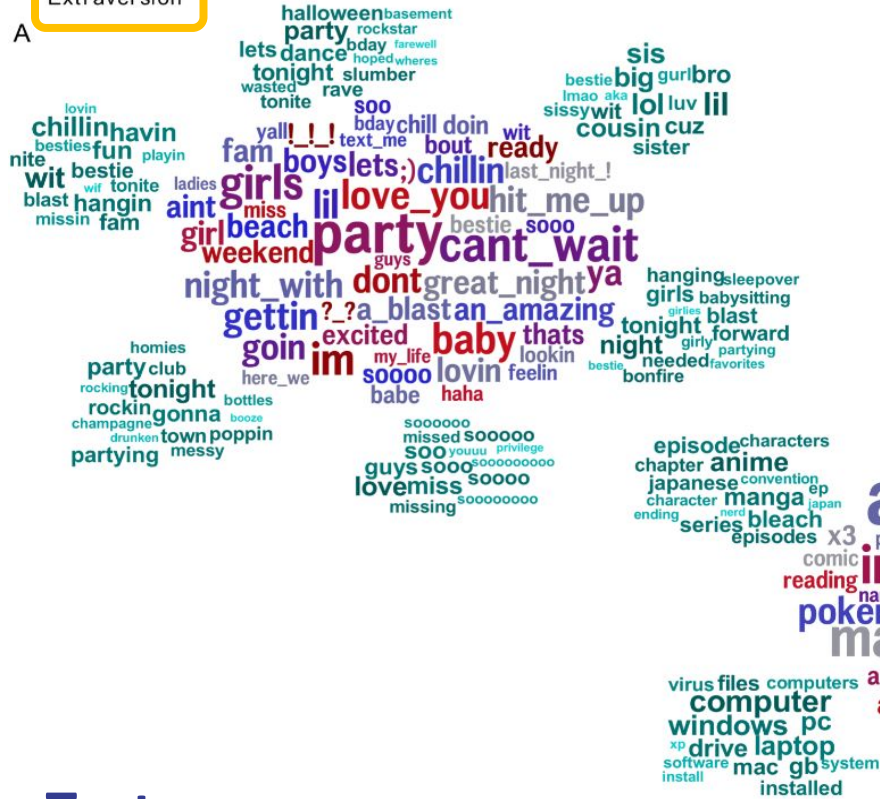
5. Matching with external data (e.g., surveys)

Analyzing Social Media Data

Type	Approach	Examples
Text	Natural Language Processing	<ul style="list-style-type: none">● Topic Modeling● Sentiment Analysis● Named Entity Recognition
Network Data	Social Network Analysis	<ul style="list-style-type: none">● Centrality Measures● Community Detection● Layout Algorithms for Visualization
Images, Videos	Computer Vision	<ul style="list-style-type: none">● Object Detection● Facial Recognition● Motion Analysis

A Extraversion

A



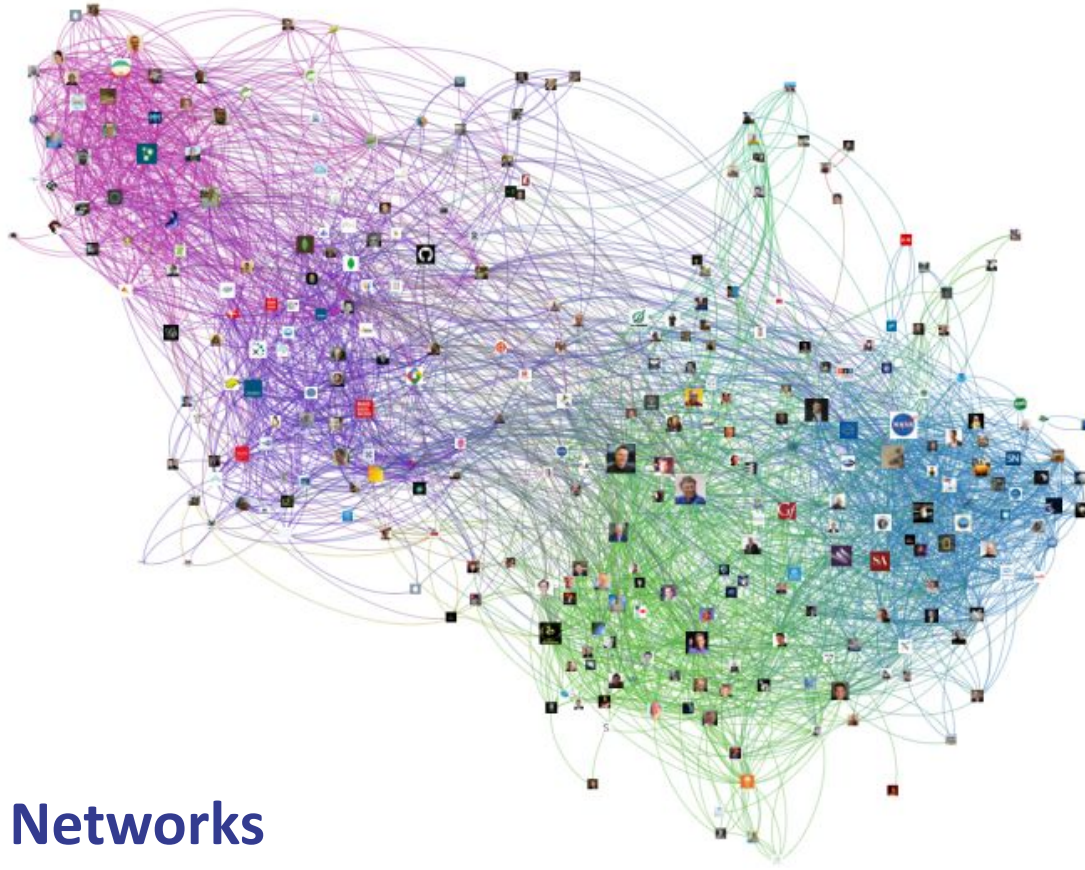
Text

“We analyzed 700 million words, phrases, and topic instances collected from the Facebook messages of 75,000 volunteers, who also took standard personality tests [...]”

Schwartz et al. (2013)

Introversion





“I use my Twitter account almost exclusively for professional networking. **My friend network had four major communities in it**”

<https://allthingsgraphed.com/2014/11/02/twitter-friends-network/>

Networks

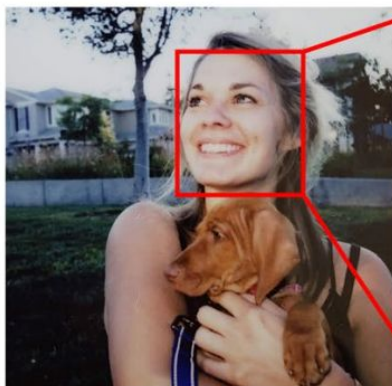


Co-Author network of “Der Spiegel” magazine

https://dkriesel.com/blog/2016/1229_video_und_folien_meines_33c3-vortrag_spiegelmining

Inference of properties
(here: departments) of
“gray” authors on the
basis of the properties of
their co-authors.

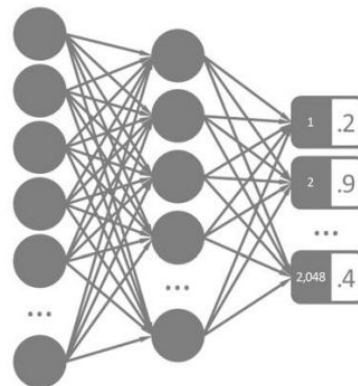
Networks



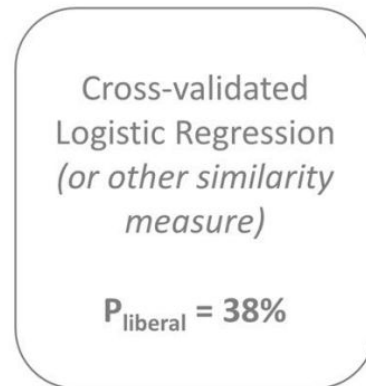
Detect face (Face++)



Crop and resize
(224 x 224 pixels)



Extract 2,048 face
descriptors (VGGFace2)



Compare with liberal
and conservative faces

“A **facial recognition algorithm** was applied to naturalistic images of 1,085,795 individuals [...].
Political orientation was correctly classified in 72% of liberal–conservative face pairs [...]”

[Kosinski \(2021\)](#)

Images

Two current studies at ZPID



RESEARCHERS

HAPPY

~~WIFE~~



HAPPY

~~LIFE~~

RESEARCH (OUTCOMES)

Academic #TwitterMigration to Mastodon: The Role of Influencers and the Open Science Movement

H1: Researchers who are under higher **social influence from #TwitterMigration influencers** are more likely to migrate to Mastodon than researchers who are under lower social influence.

Social Impact Theory (Latané, 1981):

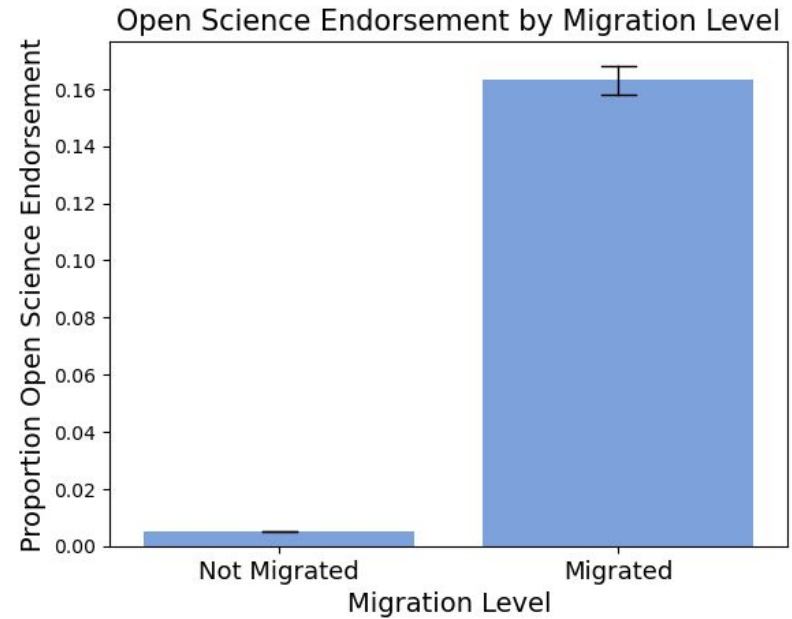
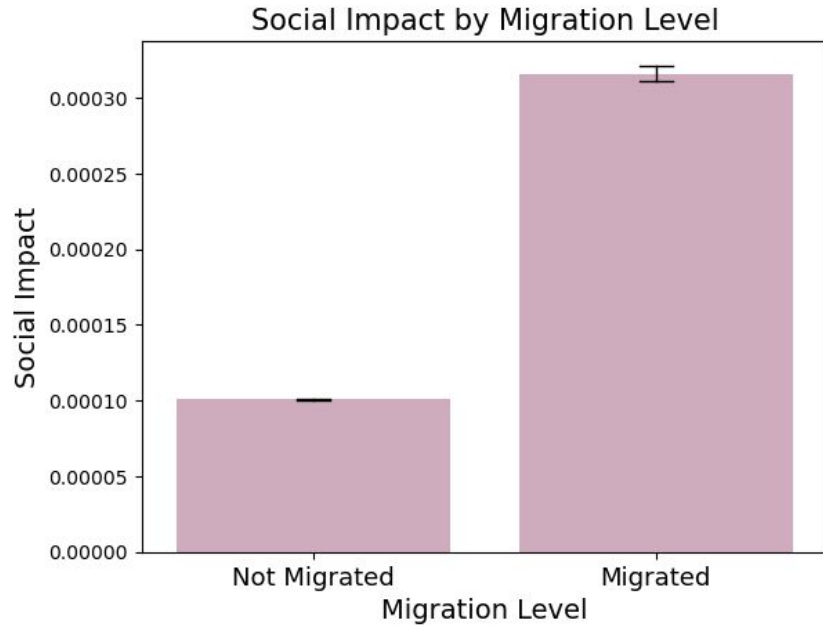
- Strength
- Immediacy
- Number of sources

H2: Researchers who endorse the **open science movement** are more likely to migrate to Mastodon than researchers who do not.



<https://doi.org/10.23668/psycharchives.13062>

Preliminary Results



Can 280 Characters Speak for Researchers? Leveraging Twitter Data for Unobtrusive Measurement of Academics' Occupational Well-Being



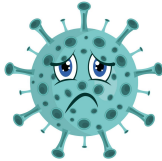
<https://osf.io/zf6kh>



Well-being derived from literature



Well-being inferred from Twitter dataset

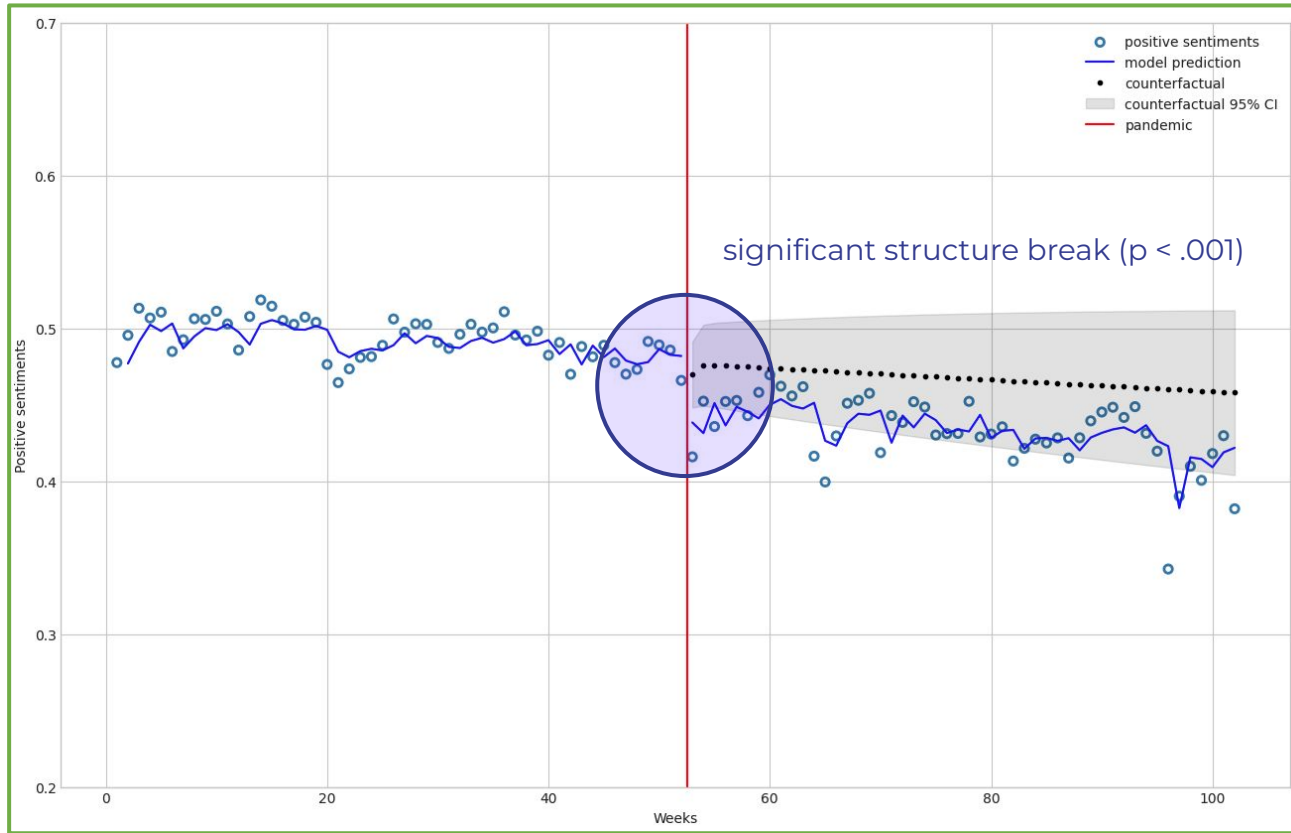


H1: Well-being scores derived from tweets are lower for the period during the COVID-19 pandemic than before the pandemic

(Alfawaz et al., 2021; KNAW, 2022; LNVH, 2021; Ramos, 2021.; Subramanya et al., 2020)

Results: H1

Positive Sentiments



3. Responsible Social Media Research



Zimmer (2010):

“In 2008, a group of researchers publicly released profile data collected from the **Facebook accounts of an entire cohort of college students** from a US university.

While good-faith attempts were made to hide the identity of the institution and protect the privacy of the data subjects, **the source of the data was quickly identified**, placing the privacy of the students at risk.”

“[This] illuminates two central ethical concerns with the “Tastes, Ties, and Time” project: the failure to properly mitigate what amounts to **violations of the subjects’ privacy**, and, thus, the **failure to adhere to ethical research standards**.”



Social Media Ethics

e.g., Gold (2020), Williams et al. (2017)

- May publication of data expose users to harm?
- Are users vulnerable and contents sensitive?
- Was the post/account deleted at time of writing?

Handling privacy issues while doing open science

- Provide “rehydratable” datasets
- A good example:
github.com/JonasRieger/corona100d
- **Note that social media is dynamic!**
(e.g., deletion of posts)



Excerpt from the ethics application for our #TwitterMigration study:

“The study is in line with

- the recommendations of the Council for Social and Economic Data for Big Data in the social, behavioral and economic sciences ([RatSWD, 2019](#))
- the handout of the German Education Research Data Network on the use of social media data in educational research ([Bayer et al., 2021](#))
- the revised “[Social Media Guidelines](#)” of the associations of market and social research in Germany
- the [Declaration of Helsinki](#),
- the [ethics guidelines of the APA](#),
- the [developer guidelines of Twitter](#),
- the Rhineland-Palatinate State Data Protection Act ([LDSG](#)),
- and the EU General Data Protection Regulation ([GDPR](#)).”

Thank you for your attention!



abi@leibniz-psychology.org



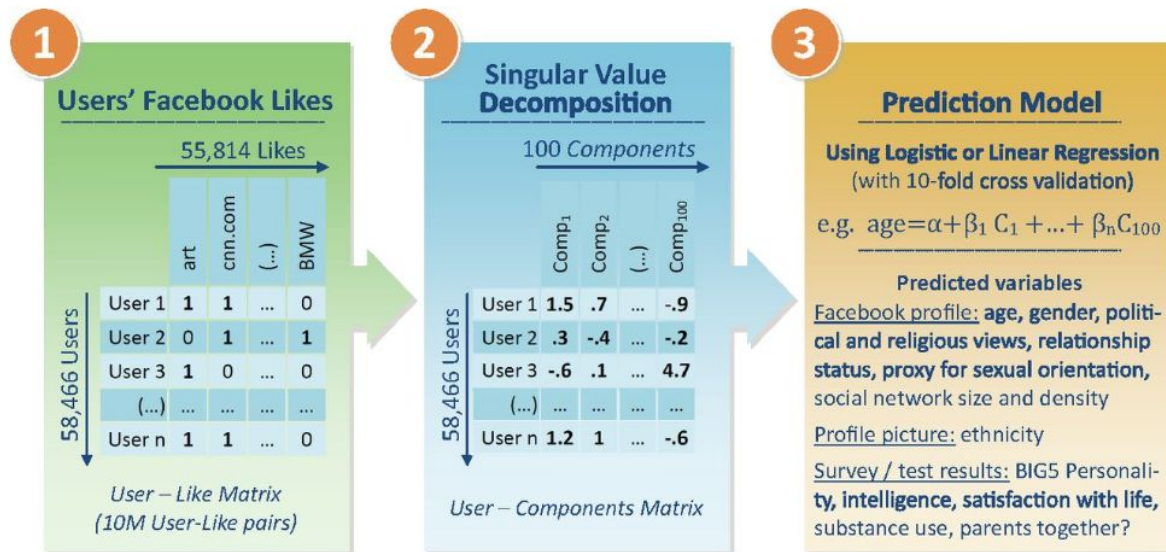
github.com/abitter



@abitter@fediscience.org

[@abitter.bsky.social](https://bsky.app/@abitter.bsky.social)

Appendix



The study is based on a sample of 58,466 volunteers from the United States, obtained through the myPersonality Facebook application (www.mypersonality.org/wiki), which included their Facebook profile information, a list of their Likes ($n = 170$ Likes per person on average), psychometric test scores, and survey information. Users and their Likes were represented as a sparse user-Like matrix, the entries of which were set to 1 if there existed an association between a user and a Like and 0 otherwise. The dimensionality of the user-Like matrix was reduced using singular-value decomposition (SVD) (24). Numeric variables such as age or intelligence were predicted using a linear regression model, whereas dichotomous variables such as gender or sexual orientation were predicted using logistic regression. In both cases, we applied 10-fold cross-validation and used the $k = 100$ top SVD components. For sexual orientation, parents' relationship status, and drug consumption only $k = 30$ top SVD components were used because of the smaller number of users for which this information was available.

Private traits and attributes are predictable from digital records of human behavior

Kosinski et al. (2013)

[https://en.wikipedia.org/wiki/File:Elon_Musk_\(3018710552\).jpg](https://en.wikipedia.org/wiki/File:Elon_Musk_(3018710552).jpg)



The Verge

TECH / TWITTER - X / ELON MUSK

More than two million users have flocked to Mastodon since Elon Musk took over Twitter

The Twitter alternative has skyrocketed in popularity, leaping from 300,000 monthly active users to 2.5 million between October and November.

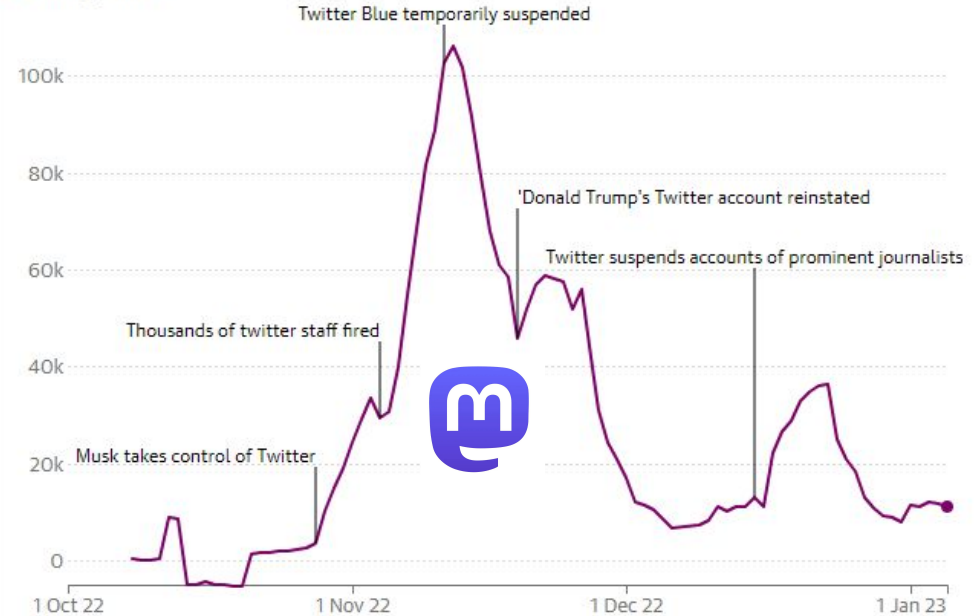
By [Jay Peters](#), a news editor who writes about technology, video games, and virtual worlds. He's submitted several accepted emoji proposals to the Unicode Consortium.

Dec 20, 2022, 7:31 AM GMT+1 | [47 Comments](#) / [47 New](#)

Change in total users registered on Mastodon servers

Showing the seven day rolling average in the increase or decrease in total Mastodon users. Only showing days for which there is complete data.

● Change in users



Guardian graphic | Source: Joinmastodon.org, Wayback Machine

Twitter as a research tool

Focus: Large-scale sampling, inferring psychological characteristics from digital footprints, geolocation

Examples:

Detecting depression

e.g., Cavazos-Rehg et al. (2016); Seabrook, Kern et al. (2018)

Predicting personality traits

e.g., Qiu et al. (2012); Quercia, Kosinski et al. (2012)

Religious differences in happiness

e.g., Chen & Huang (2019); Ritter et al. (2013)

Twitter as a research object

Focus: Characteristics of Twitter usage and online behavior, dynamics of interpersonal communication, social networks

Examples:

Social media use and mental health

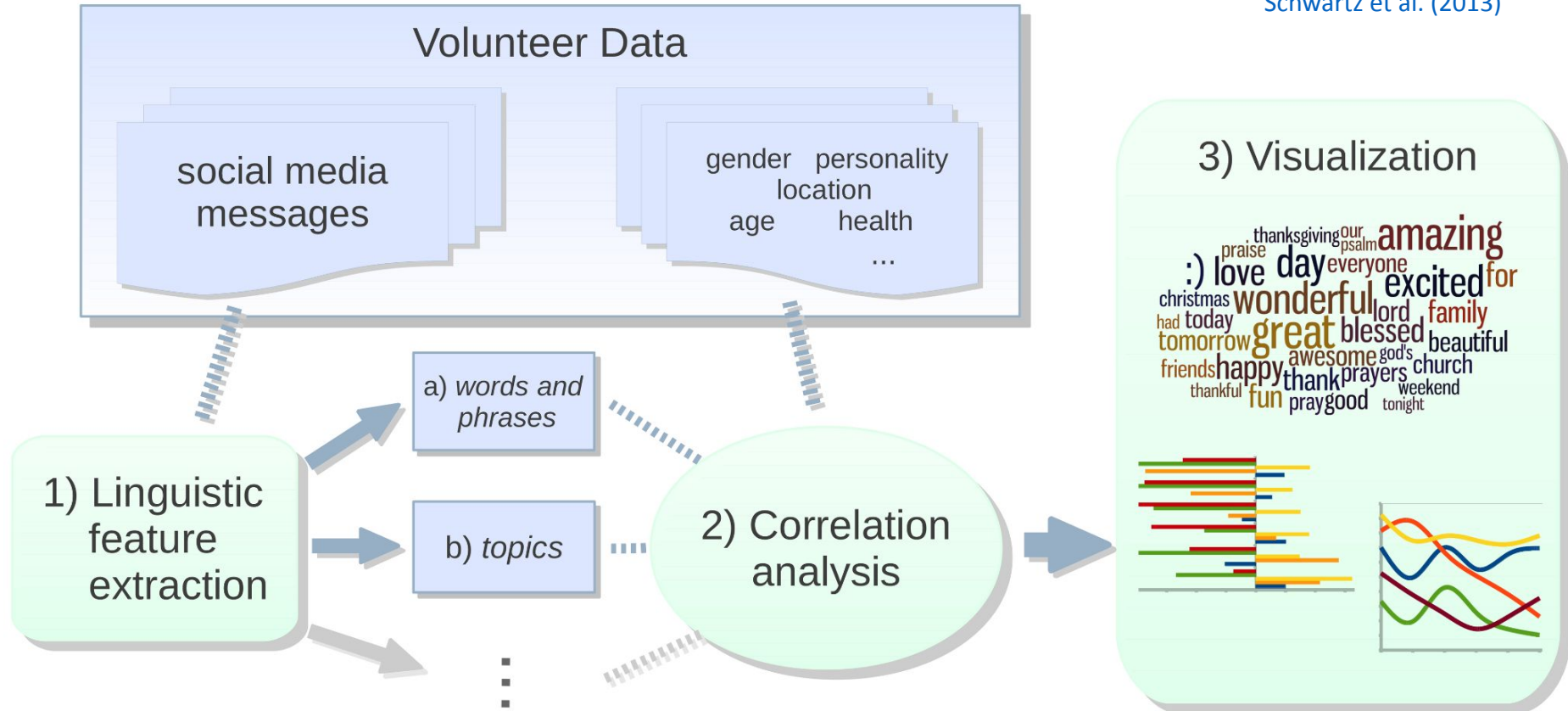
e.g., Appel et al. (2019); Ivie et al. (2020)

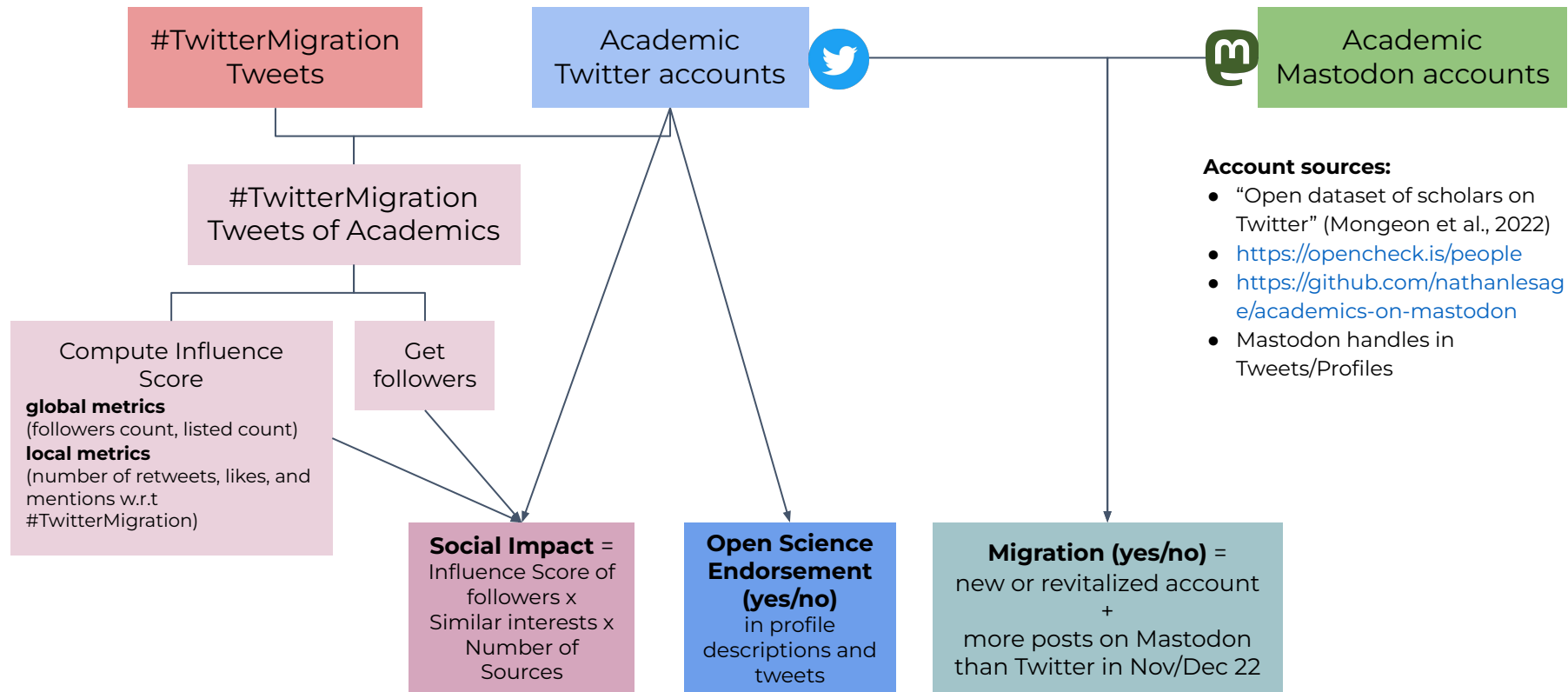
Cyberbullying in social networks

e.g., Alim (2015); Simão et al. (2021)

Online trolling behavior

e.g., Akhtar & Morrison (2019); Synnott et al. (2017)



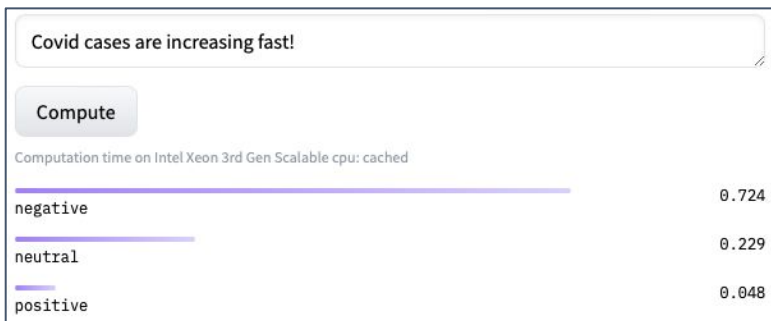


$$\text{Logit}(P(\text{Migration})) = \beta_0 + \beta_1 * (\text{Social Impact}) + \beta_2 * (\text{Open Science Endorsement}) + \beta_3 * (\text{Follower Count}) + \beta_4 * (\text{Following Count}) + \beta_5 * (\text{Account Age}) + \beta_6 * (\text{Tweet Count}) + \epsilon$$



Emotional Well-being

- affective tone of the tweets
- **data-driven Sentiment Analysis**
(Time LM-22 model; Luoreiro et al., 2022)



```
In [5]: print(scores)  
[0.724, 0.229, 0.048]
```

Methods

Cognitive Well-being



- How do researchers refer to well-being related topics?
- Combination of
 - Sentiment Analysis
 - **transformer-based Topic Modeling**
(BERTopic; Grootendorst, 2022)



Reliability of measuring personality traits using digital trace data

Overall, the research suggests a moderate to high level of accuracy.

Accuracy of Predictions

- 86.2% accuracy on Facebook and 88.5% on Twitter datasets for measuring personality traits ([Christian et al., 2021](#))
- 78% accuracy using the Myers-Briggs Type Indicator (MBTI) ([Jaysundara et al., 2022](#))

Predictive Power of Digital Footprints:

- Ranging from 0.29 for agreeableness to 0.40 for extraversion ([Azucar et al., 2018](#))
- Another study confirms that digital traces can predict psychosocial characteristics accurately ([Settanni et al., 2018](#)).