MSc Applied Information and Data Science

HSLU Lucerne University of Applied Sciences and Arts

Data Analytics for Data Scientists

Design of Experiments (DoE)

Lecture 11: Experiments in social media

2025

Prof. Dr. Jürg Schwarz

Program: 16:15 until 17:55

16:15	**Begin of the lesson**
	Lecture: Jürg Schwarz ◦ An example – Instagram and Flickr ◦ Experiments in social media? ◦ Bias ◦ Research with social media ◦ Sampling / Data collection
	Tutorial: Students / Jürg Schwarz / Assistants ◦ Working on the exercise ◦ Support by Jürg Schwarz / Assistants
17:55	**End of the lesson**

Detailed table of contents → Slide 21

An example

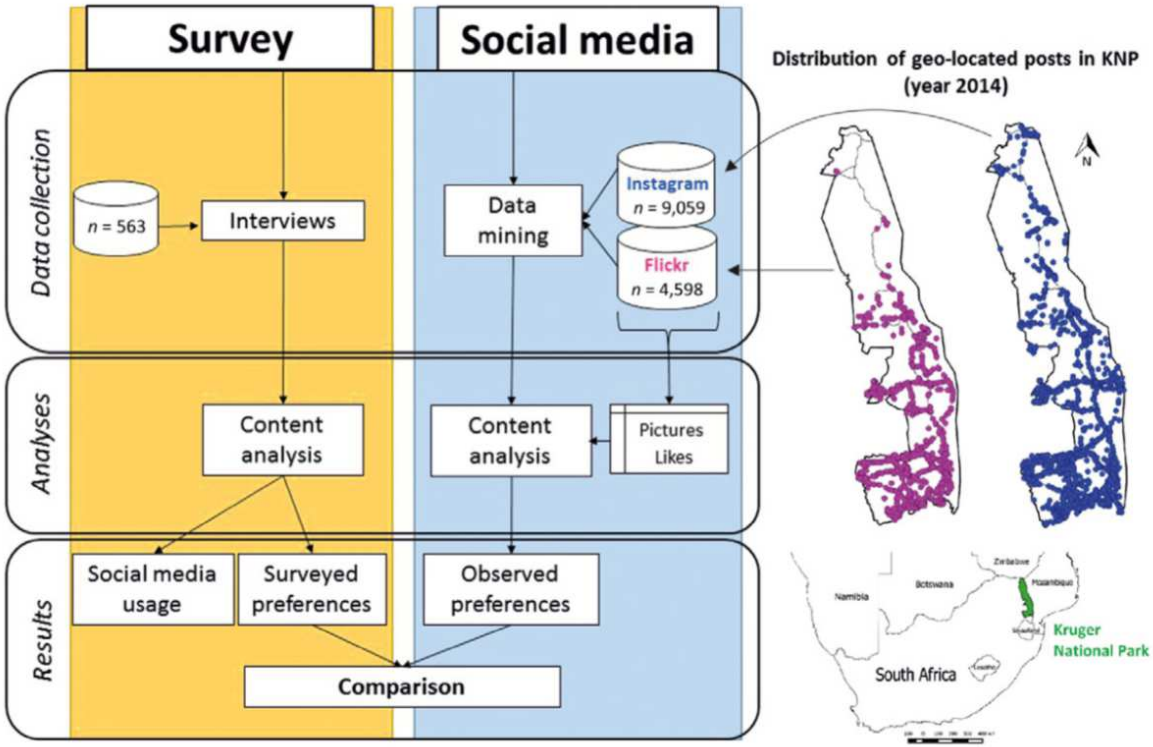
Social media as a data source?

Data from «classic» survey in comparison with social media (Instagram and Flickr)

Research question

What are the preferences of tourists in the Kruger National Park regarding biodiversity?

Study design



Survey

Random sample of **563 tourists** who were surveyed → Interviews

Social Media (Instagram and Flickr)

Publicly accessible, geotagged contributions (image / text) → API

Instagram: 9,059 images / 4,616 users

Flickr: 4,598 images / 108 users

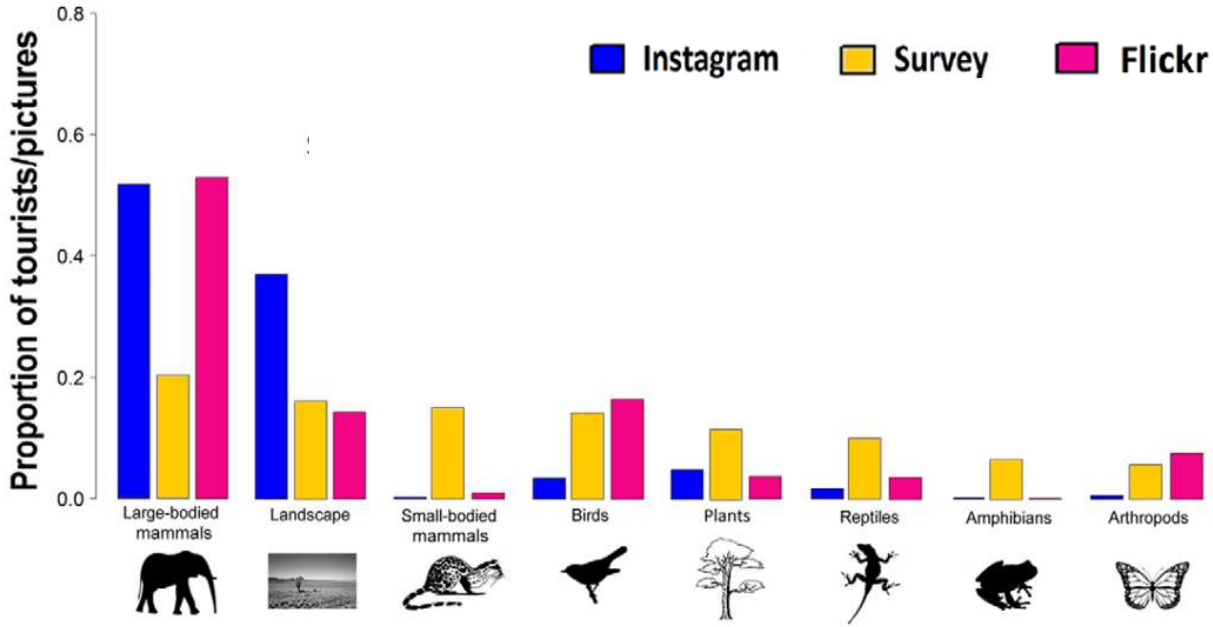
Total: **13,657 images**

Hausmann et al. (2018)

Operationalization

- Survey: Questionnaire that included these questions (answer: Yes / No)
  - Specific interest regarding biodiversity
    - Large mammals (average weight > 5 kg) / small mammals (average weight < 5 kg)
    - Landscape and seascape
    - …
- Social media: Downloading images via API (manual classification of images)

Results



Overall, there are no differences between the survey and the data from Instagram and Flickr.

Experiments in social media?

What are social media? ... it gets complicated!

How is social media different from traditional media like print, television and radio and from other new media like podcasts, websites, etc.?

The variety of standalone and integrated social media services currently available presents classification challenges, but there are some common characteristics:

Social media …

- are interactive internet-based applications.
- are websites or apps designed and maintained by the social media organization on which users create service-specific, custom profiles.
- support user-generated content such as images, text, videos, and status updates.
- enable users to connect with one another through various interaction mechanisms, such as follows and likes (X), sharing and viewing reels (Instagram), engaging with trends (TikTok), or forming friendship connections (Facebook).
- support various means for members to engage with one another in the form of collaboration, community building, participation, sharing, linking and other means.

Structure of communication in social media

The structure of communication is an important aspect
when conducting experiments in social media:

- One – e.g. X
  Specific information is sent from one source to «everyone»
- Two – e.g. SMS / WhatsApp / …
  Exchange of information between two individuals via a certain medium.
- Many – e.g. Facebook / WhatsApp / …
  Information is exchanged in a group via social media actively or passively.

Mode of interaction in social media, especially in mobile communication

In mobile communication, the mode of interaction is also changed during a session and also «multitasking» can occur:

Users can speak, text, email, video chat, and post items on social media channels or blogs.

The communication / interaction is not always carried out in the same mode:

For example, a telephone call can be answered with a text message or a telephone call is made during an ongoing mail exchange.

Quantitative approaches to social media

Survey

- Basics of DoE and quantitative-empirical research → Lectures 01 to 06
  Note in particular: Bias → Slide 8 ff

Field experiment / Social Media as a Research Laboratory

- Basics of DoE and quantitative-empirical research → Lectures 01 to 06
  Shift to A/B testing → Lecture 08
  Shift to and focus on dynamic-algorithmic methods → Lecture 08

Procedures in the field of social computing (not covered in this module)

- Social network analysis
  The principles of empirical social research (1890s)
  Émile Durkheim / Ferdinand Tönnies
- Sentiment analysis
  The principles of natural language processing (NLP) (1950s)
  Alan Turing → «Computing Machinery and Intelligence» ↔ Turing Test
- More …

Bias

Definition / Properties

Definition: Bias (population bias / selection bias / …) essentially means that the population being examined does not correspond to the defined population.

Properties

If the defined population refers to the entire population and their subgroups, social media is by definition subject to bias.

This is mainly due to the following two elements. Certain manifestations of these elements can be observed mainly (in some cases exclusively) in social media.

Element	Data from a general survey	Data from social media
Sampling frame	Phone book, email addresses, household addresses, etc.	Posts, news, likes, etc. Profile data, administrative data
Sampling procedure	Random sample → Probability sampling	Administrative data / self-selection → Nonprobability sampling

Schober et al. (2016)

Sources of bias in social media

- Activity bias I

  In a study data from users are collected who are ...

  - active (only) at the time of the study / data collection / storage.
  - are active only once or a few times.

- Activity bias II

  A few users are very active on social media,
  while most users use social media only passively.
  Therefore, the data obtained ...

  - (often) relates to a specific topic / person.
    Example Facebook data with 40,000 users: 7% is productive / 50% is completely passive
  - is (often) limited to a certain period.
    Example A storm on X where the cause itself is not included.

- Activity with bots (web crawlers / social bots / ...)

  Programs that behave like users or react to specific triggers.

- Commercial pages / fan pages / fake accounts / death online / memorial state / …

Öhman & Watson (2019) / Baeza-Yates (2018)

Standardization of bias in social media

Three types → Bias arises through ...

- Medium & platform
  Restrictions on access to data / (unknown) ways and methods on how data is stored
- Survey / sampling & representativeness
  Not taking into account the circumstances, e.g. in the case of a storm on X
- Ways of managing data and sources
  Reproducibility limited by restrictive access to data / by deletion of items

Types	Sub-Types	Categories
• Media & platform biases	• Functional biases & APIs • Population biases • Behavioral biases	• APIs restrictions (e.g. rate limits, filtering by content or geographical location • skewed to young and urban demographic groups, different demographics are • people behave differently on distinct platforms, humans alter their behavior
• Collection biases & representativeness	• Data querying & sampling • Temporal considerations • Context-specific biases • Size & Representativeness	• explicit and implicit bias inherent to a data collection approach • change in the usage of a platform, “Swiss cheese” decay of Twitter test • bias due to individual characteristics or privacy concerns, self-cleaning and se • bigger data are not always better data, issues with representativeness of
• Data sharing, reproducibility & other challenges	• Data sharing • Digital divide • Spam & Non-humans • Other issues	• proprietary nature of social media leads to problems with replicability and data • economic and social inequality with regard to access to, use of, or impact of ICT • non-humans, spammers, bots, fake accounts • online vs. offline behavior, legal obligation to remove deleted content

Olteanu (2016) → Table 2.1

Quantification of the population bias in relation to the population

Example from the U.S. – Pew Research Center: «Social Media Use in 2021»*

See newer examples in the Appendix → Slide 18ff

Gender, age, education, etc.

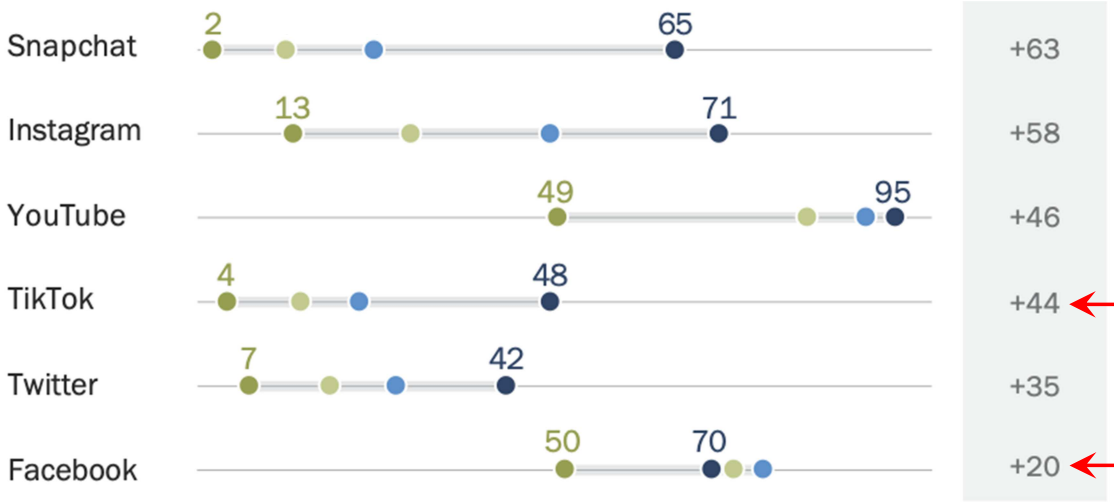
	YouTube	Facebook	Instagram	Pinterest	LinkedIn	Snapchat	Twitter	WhatsApp	TikTok	Reddit	Nextdoor
Total	81	69	40	31	28	25	23	23	21	18	13
Men	82	61	36	16	31	22	25	26	17	23	10
Women	80	77	44	46	26	28	22	21	24	12	16
White	79	67	35	34	29	23	22	16	18	17	15
Black	84	74	49	35	27	26	29	23	30	17	10
Hispanic	85	72	52	18	19	31	23	46	31	14	8
Ages 18-29	95	70	71	32	30	65	42	24	48	36	5
30-49	91	77	48	34	36	24	27	30	22	22	17
50-64	83	73	29	38	33	12	18	23	14	10	16
65+	49	50	13	18	11	2	7	10	4	3	8
<$30K	75	70	35	21	12	25	12	23	22	10	6
$30K-$49,999	83	76	45	33	21	27	29	20	29	17	11
$50K-$74,999	79	61	39	29	21	29	22	19	20	20	12
$75K+	90	70	47	40	50	28	34	29	20	26	20
HS or less	70	64	30	22	10	21	14	20	21	9	4
Some college	86	71	44	36	28	32	26	16	24	20	12
College+	89	73	49	37	51	23	33	33	19	26	24
Urban	84	70	45	30	30	28	27	28	24	18	17
Suburban	81	70	41	32	33	25	23	23	20	21	14
Rural	74	67	25	34	15	18	18	9	16	10	2

0% 20 40 60 80 100

Percent of US adults

*www.pewresearch.org (April 2025 → Last version is from April 2021)

Age gaps in Snapchat and Instagram use are relatively wide, less so for Facebook



Age Groups

18-29

30-49

50-64

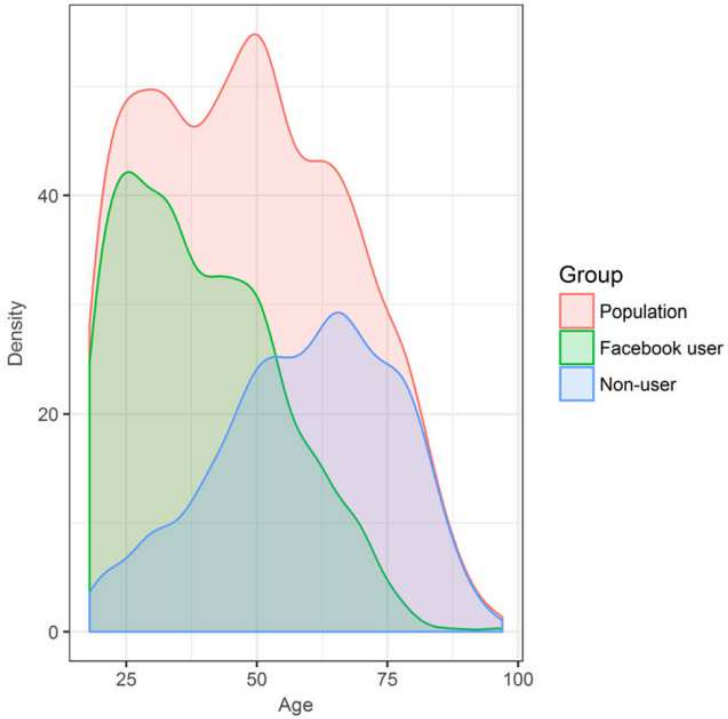
65+

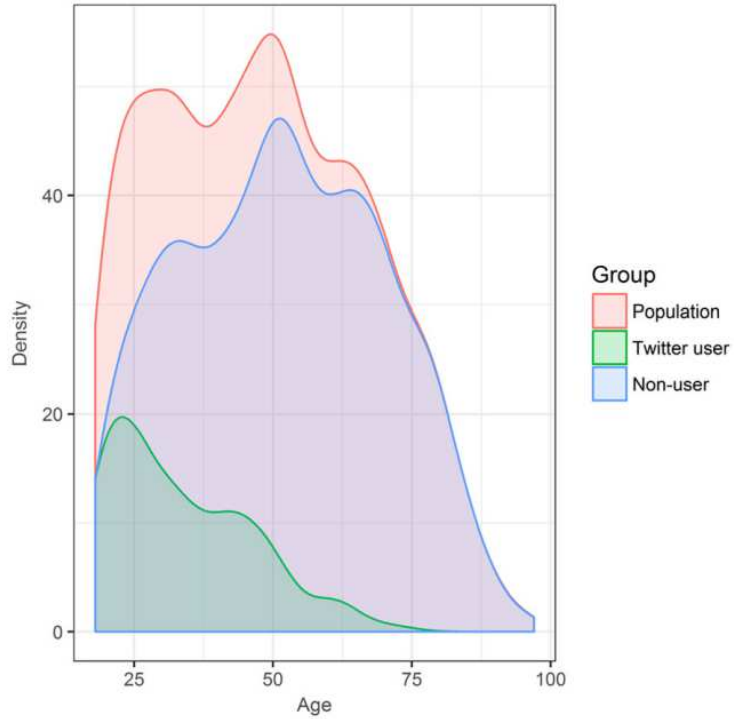
Difference between youngest and oldest

50% of age group 65+ use Facebook
70% of age group 18-29 use Facebook

www.pewresearch.org, 2021

Age distribution of Facebook and X (Twitter) users





Mellon & Prosser (2017)

Research with social media

Potential

Studies using data from social media have great potential for investigating research questions in social science / psychological research questions that are new or place special demands on the study design.

This is made possible by specific properties of data / information from social media, some of which are significantly different from those of classical surveys or are new:

- Rapid availability of data / information and continuous updating
- Simple and low-cost extraction processes compared to classical surveys

Example I for research with social media

Crowdsourcing helps locate earthquakes quickly and reliably.

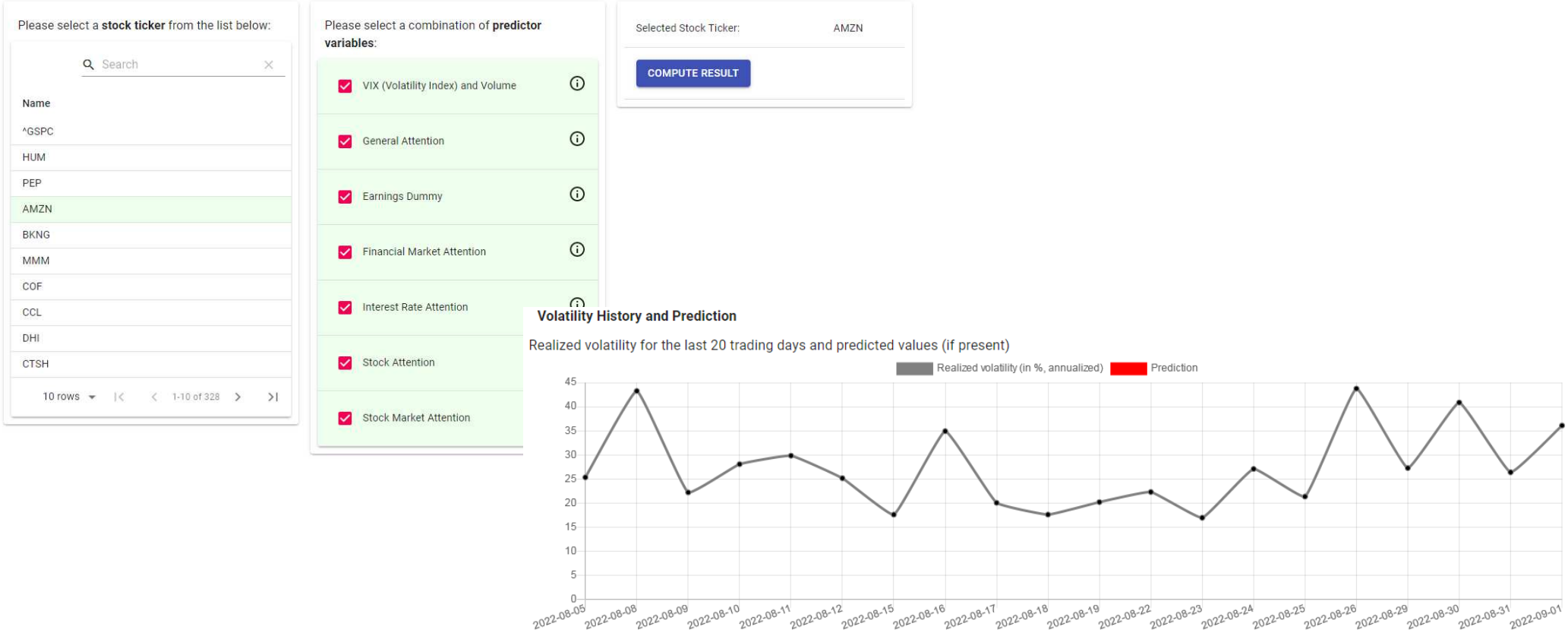
Earthquakes and their effects can be identified using tweets and similar information.
To do this, a system was created to collect reports from people.
The system was tested using real data from 2016 and 2017.
It found 50% of earthquake locations in just 103 seconds.
This was 76 seconds faster than GEOFON* and 271 seconds faster than the European-Mediterranean Seismological Centre.

*GEOFON program (geofon.gfz-potsdam.de) / Steed et al. (2019)

The impact of sentiment and attention measures on stock market volatility

We analyze the impact of sentiment and attention variables on the stock market volatility by using a novel and extensive dataset that combines social media, news articles, information consumption, and search engine data.

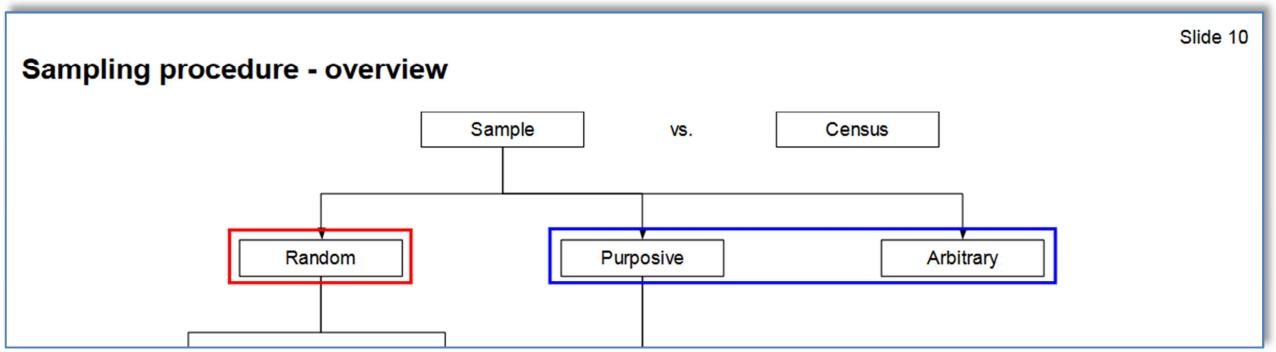
Application on the website www.sentivol.ch with example AMZN (Amazon stock price)*



*Retrieved: April 2025 / Site is limited to the years up to and including 2022 / Audrino, Sigrist, and Ballinari (2020)

Sampling / Data collection

General statement on sampling → Lecture 05: Sampling



Given: Bias (population bias / selection bias / …)

this results in

Sampling in social media involves *a priori* an arbitrary selection.

General observation concerning data collection → Lecture 10: Large data quantities

Given: Technical-administrative acquisition / collection of data from social media

this results in

Data from social media is most likely «found data».

Two methods of data collection

Using the functionalities of social media

All variants of arbitrary selection, such as

- Use individual snowball sampling
- Set up groups / accounts in social media → starting point for snowball sampling

Paid access to survey

Place ads in social media

- Insert a link to survey or to web page that contains a link to survey
- Use Facebook's advertising platform as a "digital census"

Use survey tools, such as

- SurveyMonkey : … share polls and surveys … with our Facebook Collector or post a survey questionnaire directly on your fan page with our Web Link Collector.
- Many more …

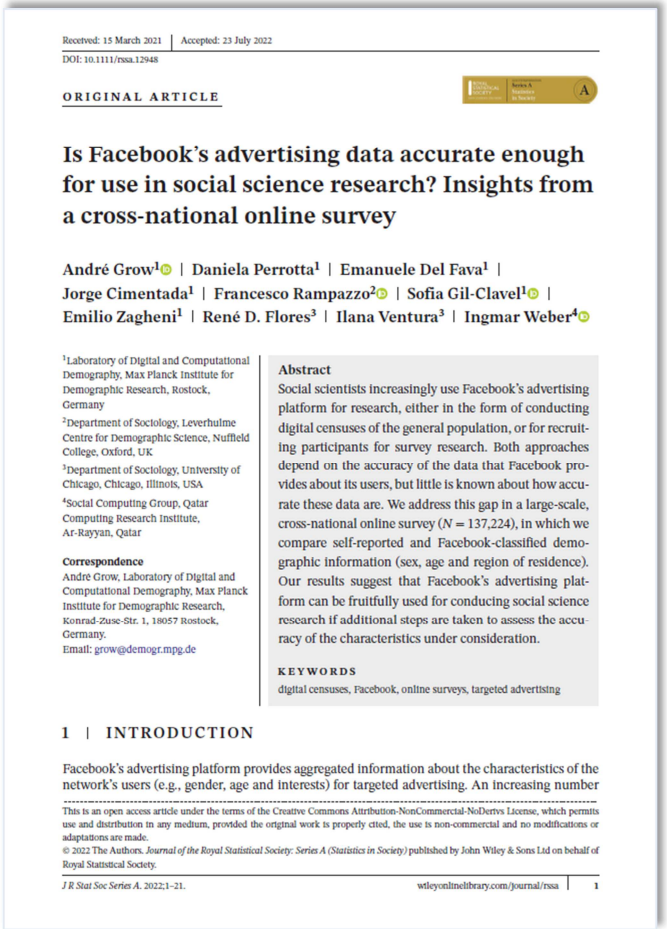
Include specialized platforms, such as

- Amazon Mechanical Turk: … crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs …. This could include anything … to more subjective tasks like survey participation, content moderation, …
- Many more …

An Example – Survey Data Collection with Facebook in the U.S.

Article from Grow et al. (2022)

Is Facebook's advertising data accurate enough for use in social science research?



Received: 15 March 2021 | Accepted: 23 July 2022
DOI: 10.1111/rssa.12948

ORIGINAL ARTICLE

Is Facebook's advertising data accurate enough for use in social science research? Insights from a cross-national online survey

André Grow[1] | Daniela Perrotta[1] | Emanuele Del Fava[1] | Jorge Cimentada[1] | Francesco Rampazzo[2] | Sofia Gil-Clavel[1] | Emilio Zagheni[1] | René D. Flores[3] | Ilana Ventura[3] | Ingmar Weber[4]

[1]Laboratory of Digital and Computational Demography, Max Planck Institute for Demographic Research, Rostock, Germany
[2]Department of Sociology, Leverhulme Centre for Demographic Science, Nuffield College, Oxford, UK
[3]Department of Sociology, University of Chicago, Chicago, Illinois, USA
[4]Social Computing Group, Qatar Computing Research Institute, Ar-Rayyan, Qatar

Correspondence
André Grow, Laboratory of Digital and Computational Demography, Max Planck Institute for Demographic Research, Konrad-Zuse-Str. 1, 18057 Rostock, Germany.
Email: grow@demogr.mpg.de

Abstract
Social scientists increasingly use Facebook's advertising platform for research, either in the form of conducting digital censuses of the general population, or for recruiting participants for survey research. Both approaches depend on the accuracy of the data that Facebook provides about its users, but little is known about how accurate these data are. We address this gap in a large-scale, cross-national online survey ($N = 137{,}224$), in which we compare self-reported and Facebook-classified demographic information (sex, age and region of residence). Our results suggest that Facebook's advertising platform can be fruitfully used for conducing social science research if additional steps are taken to assess the accuracy of the characteristics under consideration.

KEYWORDS
digital censuses, Facebook, online surveys, targeted advertising

1 | INTRODUCTION

Facebook's advertising platform provides aggregated information about the characteristics of the network's users (e.g., gender, age and interests) for targeted advertising. An increasing number

Abstract

Social scientists are increasingly using Facebook's advertising platform for research purposes, either to conduct digital censuses of the general population or to recruit participants for survey studies.

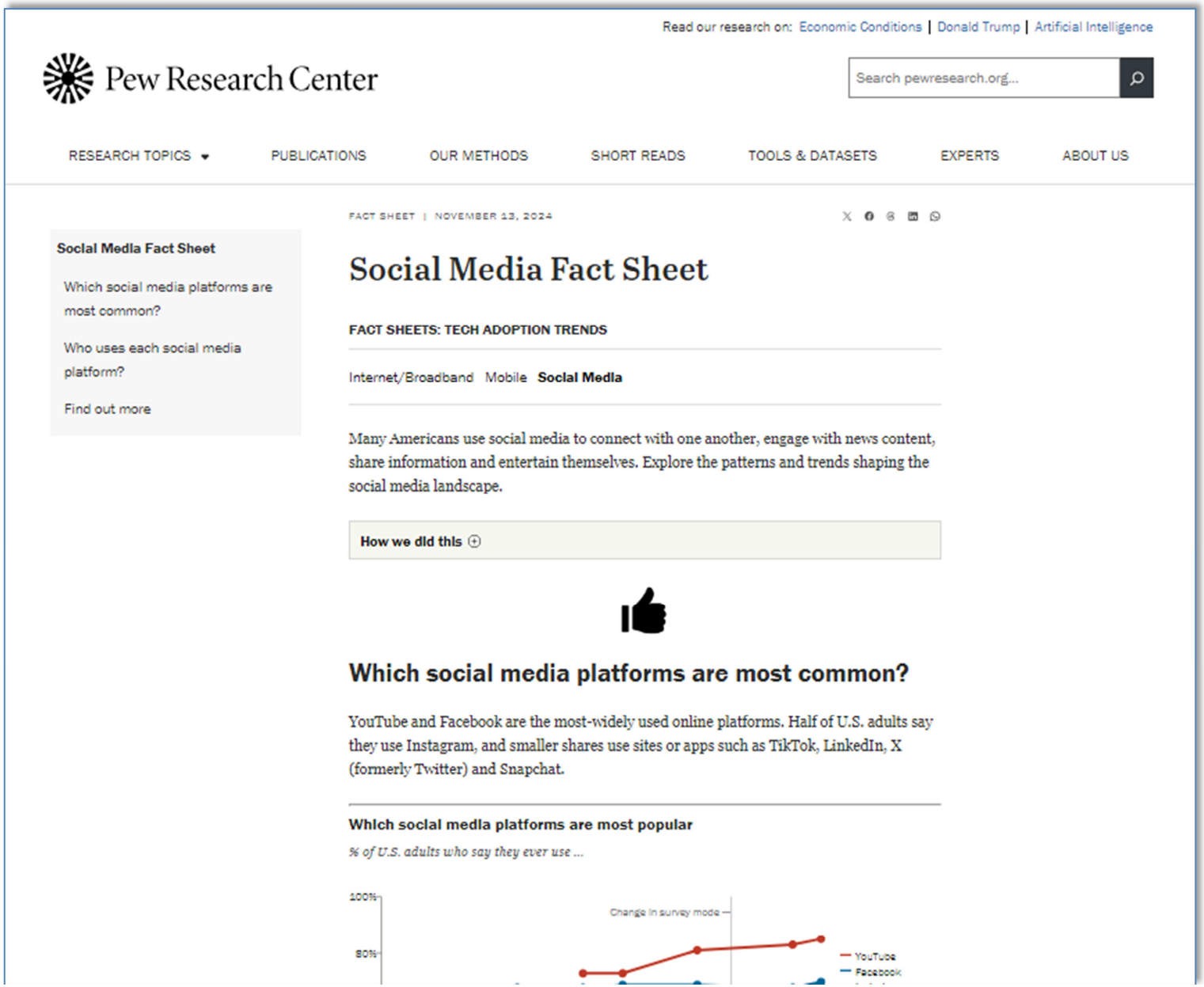
Both approaches rely on the accuracy of the demographic data that Facebook provides about its users; however, little is known about the actual quality of these data.

To address this gap, we conducted a large-scale, cross-national online survey (N = 137,224), comparing self-reported demographic information (sex, age, and region of residence) with classifications assigned by Facebook.

Our findings suggest that Facebook's advertising platform can serve as a valuable tool for social science research, provided that additional validation measures are implemented to assess the accuracy of the demographic characteristics in question.

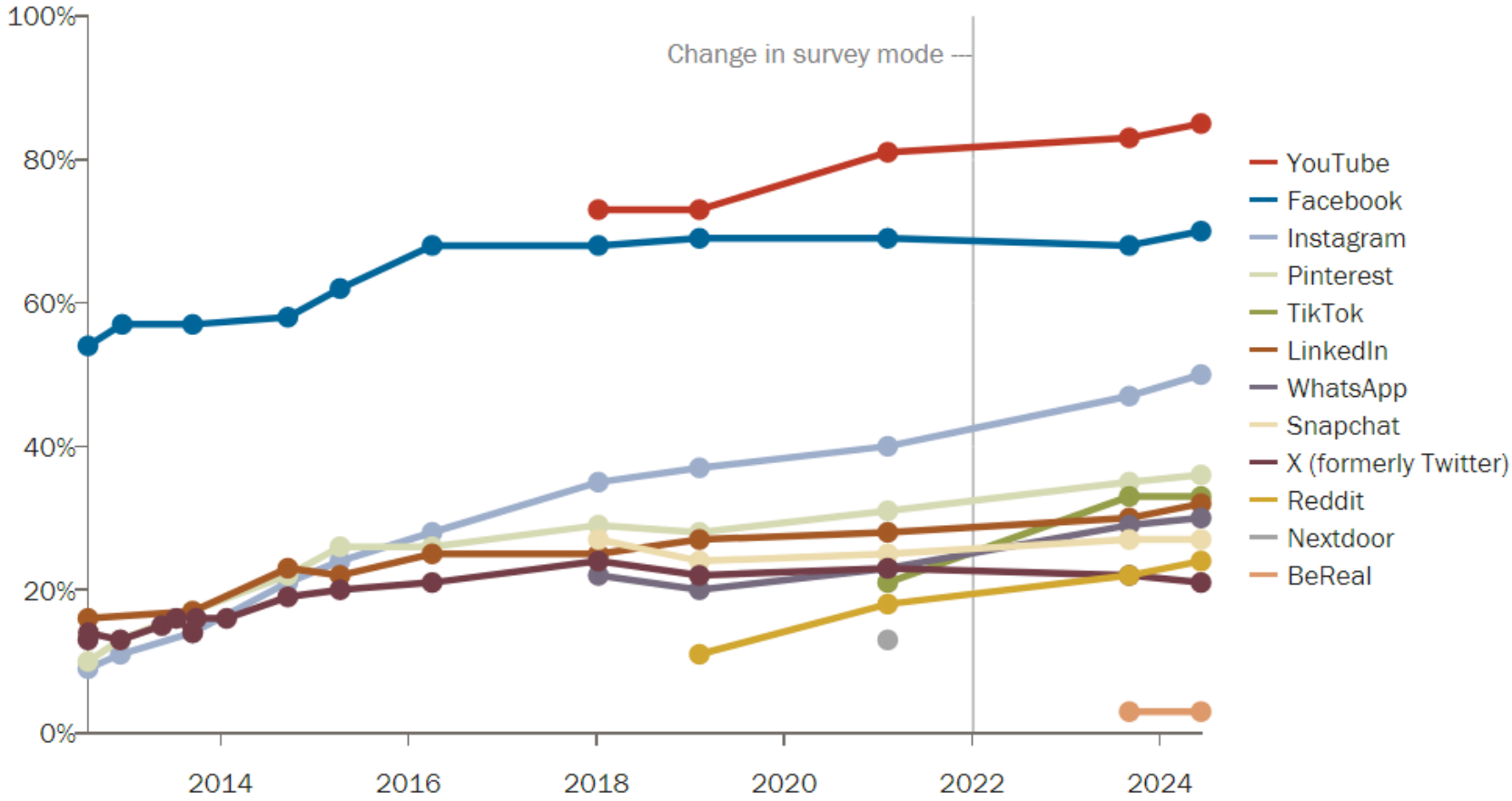
Appendix

General access via www.pewresearch.org/internet/fact-sheet/social-media



Which social media platforms are most popular*

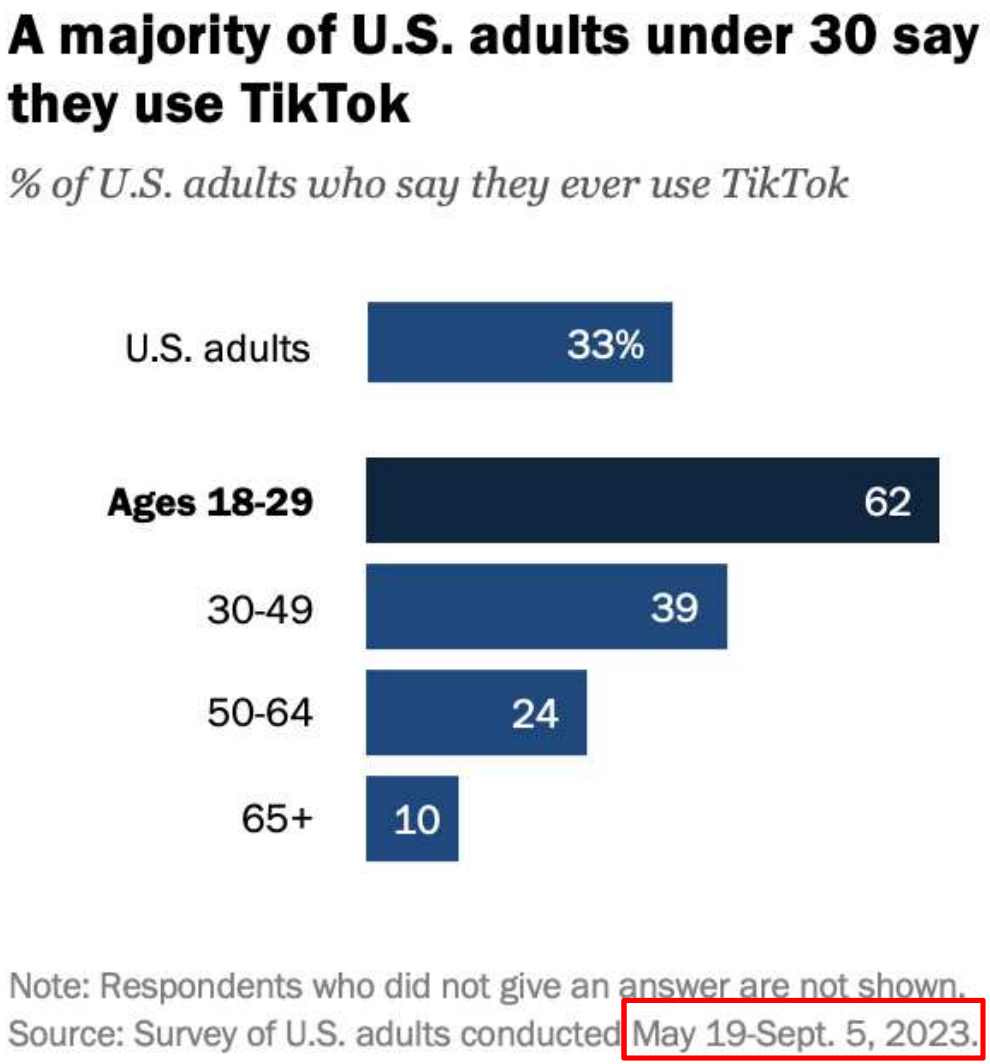
% of U.S. adults who say they ever use …



Source: Surveys of U.S. adults conducted 2012-2024.
Note: The vertical line indicates a change in mode. Polls from 2012-2021 were conducted via phone. In 2023, the poll was conducted via web and mail. In 2024, the poll was conducted via web, mail and phone. For more on the mode shift in 2023, read our Q&A. Refer to the topline for more information on how question wording varied over the years. Pre-2018 data is not available for YouTube, Snapchat or WhatsApp; pre-2019 data is not available for Reddit; pre-2021 data is not available for TikTok; pre-2023 data is not available for BeReal. Respondents who did not give an answer are not shown.

*www.pewresearch.org/internet/fact-sheet/social-media, April 2025

A majority of U.S. adults under 30 say they use TikTok*



*www.pewresearch.org, April 2025

A majority of U.S. adults under 30 say they use TikTok*

AGE | GENDER | RACE & ETHNICITY | HOUSEHOLD INCOME | EDUCATION | COMMUNITY

	Men	Women
Facebook	61	78
Instagram	44	55
LinkedIn	35	30
X (formerly Twitter)	25	17
Pinterest	19	51
Snapchat	23	31
YouTube	87	83
WhatsApp	28	32
Reddit	28	20
TikTok	26	39
BeReal	2	3

Note: Respondents who did not give an answer are not shown.

Source: Survey of U.S. adults conducted Feb. 1-June 10, 2024.

*www.pewresearch.org, April 2025

Table of contents