gesis Leibniz Institute for the Social Sciences

Potentials and Pitfalls of Social Media Data

Indira Sen & Katrin Weller

GESIS workshop - December 2022

Leibniz Association

gesis Leibniz Institute for the Social Sciences

0. Welcome and logistics

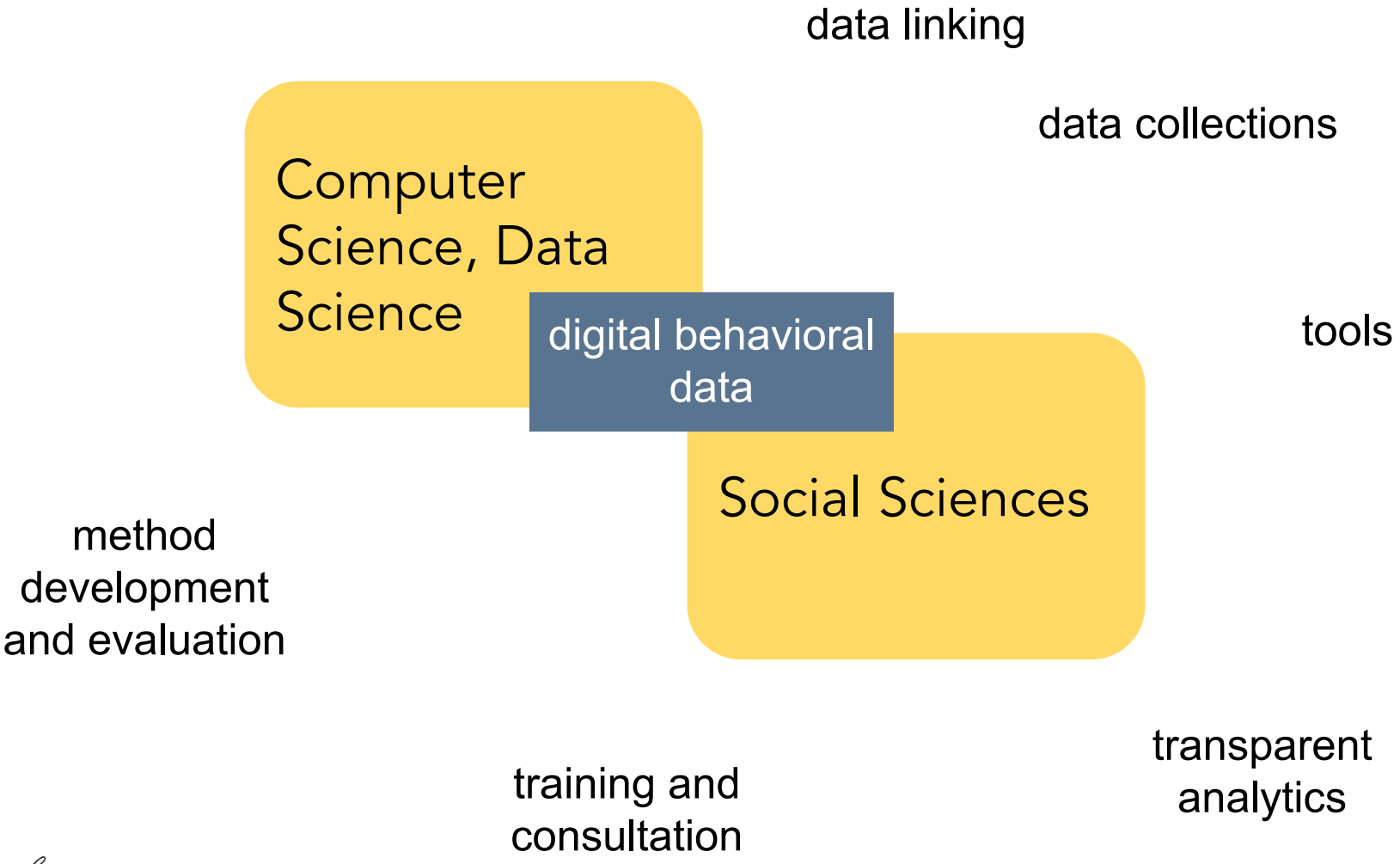
About us

Contact

- Indira: indira.sen@gesis.org, @indiiigosky

- Katrin: katrin.weller@gesis.org, @kwelle

Computational Social Science at GESIS



Acknowledgements

We thank our colleagues at GESIS, especially:

Claudia Wagner

Fabian Flöck

Mattia Samory

Arnim Bleier

Bernd Weiß

Leon Fröhling

Felix Soldner

… and the entire CSS department

About the workshop

Zoom Etiquette

- Microphones off when not speaking
- Camera preferably on
- Asking questions or making comments (both welcome!)
  - raise hand on zoom
  - type in zoom chat
  - => **We will try to get to you ASAP**

- We have planned for breaks in the schedule, but feel free to request for additional short breaks through chat

Resources

Materials will be shared via:
https://github.com/Indiiigo/social_media_data_research_2022

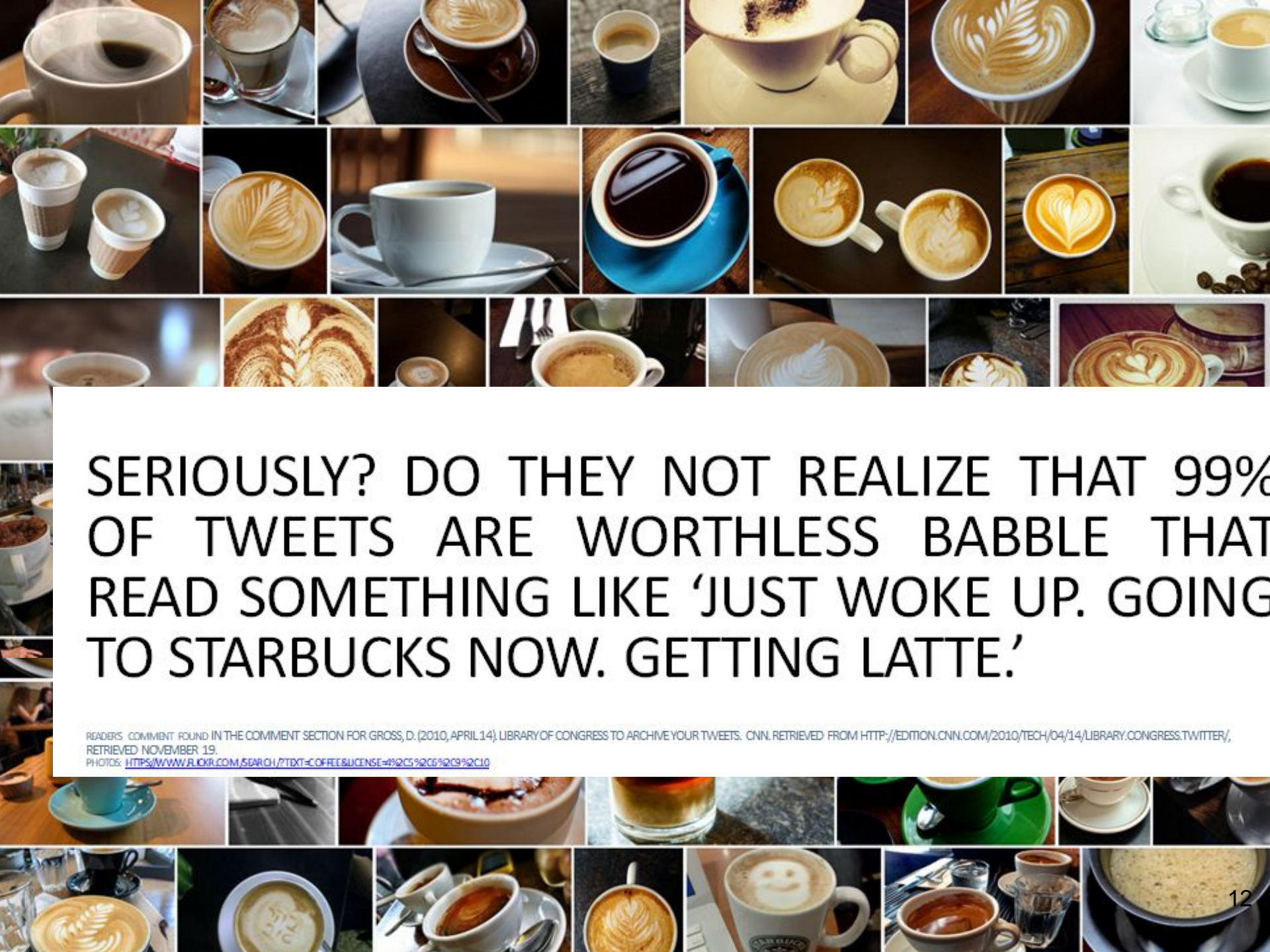
This includes:

- **Slides**
- List of **references** at the end of the last slide deck
- **Notebooks** with examples of code that can be executed
- **Discussion Board:**
  https://tinyurl.com/socialmediaworkshop22

Workshop goals

- provide an overview on current approaches in research **based on digital traces from social media**
- outline different steps in the research process when working with social media data, and provide **practical examples** for data collection, cleaning and analysis
- offer a **structured approach** to think about potential pitfalls and error sources in social media research, that can help to design, present, talk about research approaches

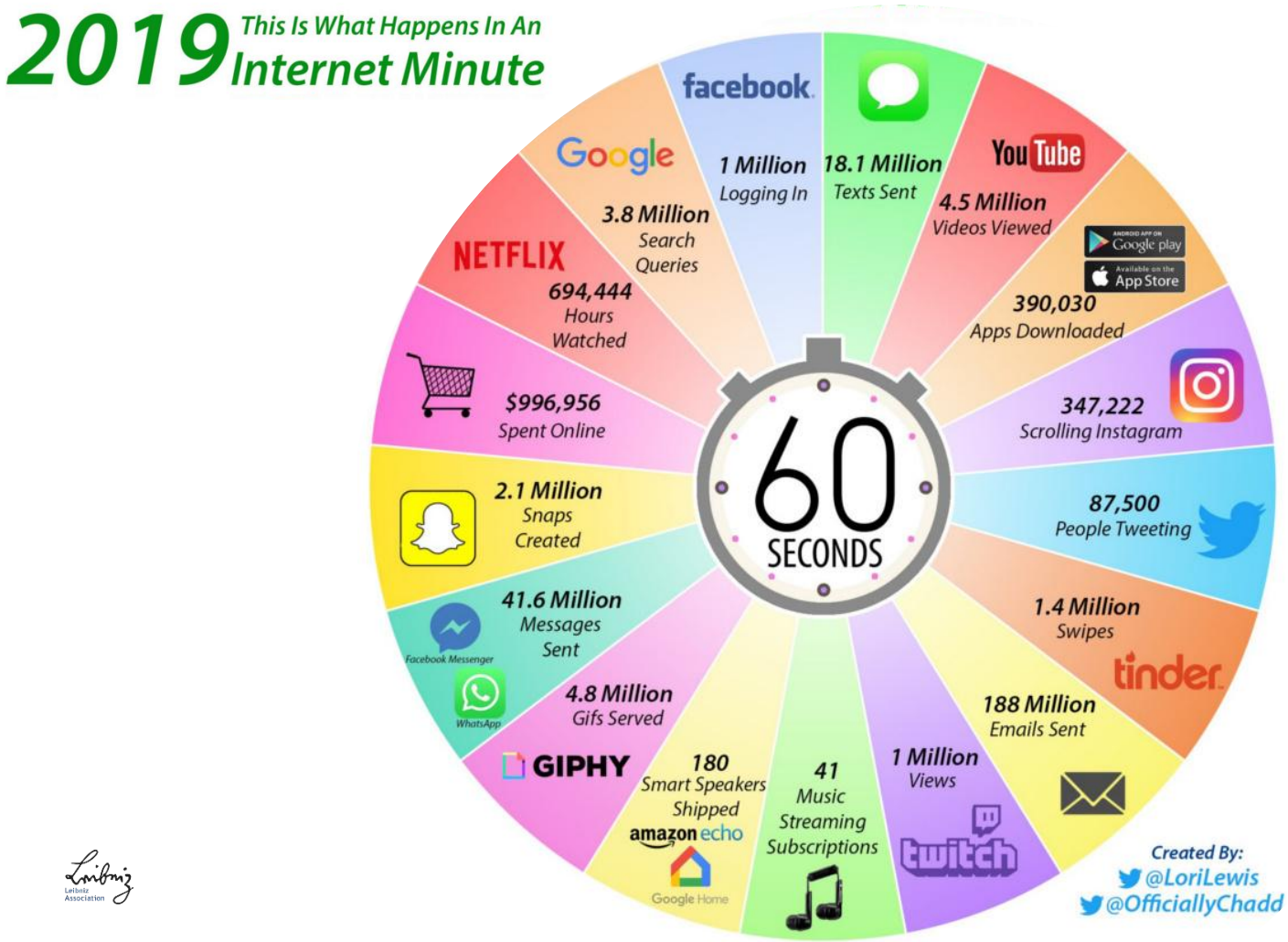
Motivation & Background



SERIOUSLY? DO THEY NOT REALIZE THAT 99% OF TWEETS ARE WORTHLESS BABBLE THAT READ SOMETHING LIKE 'JUST WOKE UP. GOING TO STARBUCKS NOW. GETTING LATTE.'

READER'S COMMENT FOUND IN THE COMMENT SECTION FOR GROSS, D. (2010, APRIL 14). LIBRARY OF CONGRESS TO ARCHIVE YOUR TWEETS. CNN. RETRIEVED FROM HTTP://EDITION.CNN.COM/2010/TECH/04/14/LIBRARY.CONGRESS.TWITTER/, RETRIEVED NOVEMBER 19.
PHOTOS: HTTPS://WWW.FLICKR.COM/SEARCH/?TEXT=COFFEE&LICENSE=4%2C5%2C6%2C9%2C10

Digital Traces: So much data! (...about people's behavior and attitudes)



gesis Leibniz Institute for the Social Sciences

And so much research!

Computer Science

Social Sciences

2010 From tweets to polls [O'Connor et al.]

2013 Text as data [Grimmer et al.]

2014 Predicting tie strength with social media [Gilbert et al.]

2013 Big Data in Survey Research: AAPOR Task Force Report [Japec et al.]

2019 Investigating commentator bias in football broadcasts [Merullo et al.]

2019 Combining surveys and digital traces [Stier et al., Pasek et al.]

Leibniz Association

No formal research field - no standard methods

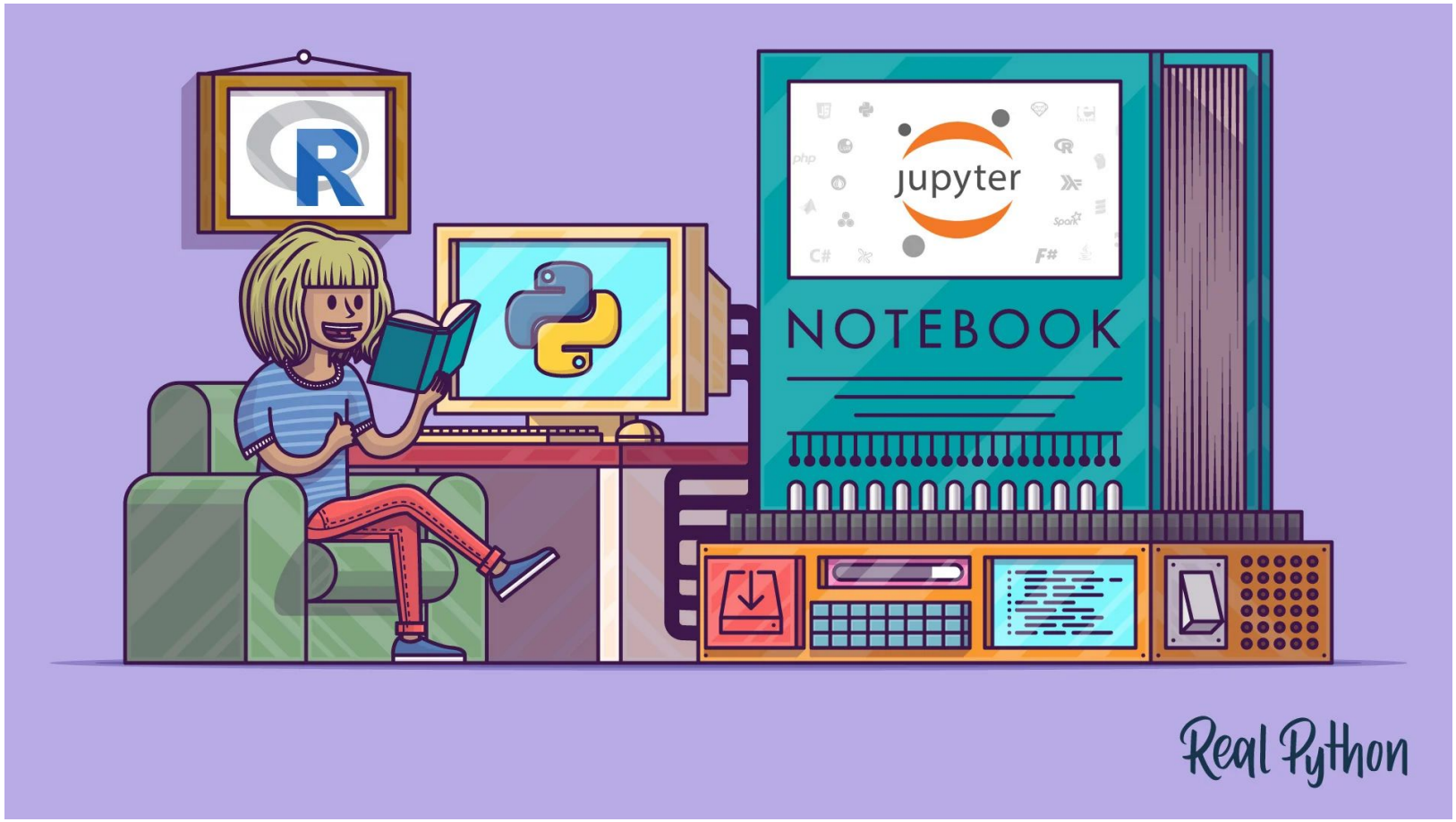
- multidisciplinary entry points
- diverse range of approaches (also mixed methods)
- many exploratory studies

how to get started?

Workshop goal

- Combine critical reflections and hand-on exercises
- get some first idea on how approaches work in practice
- provide reusable material

We will be using Notebooks for Coding



We will be using Notebooks for Coding

- Python + Google Colab: https://colab.research.google.com/

  → you need to login via a google account
- light-weight, easy to share, good for reproducing results
- but of course there are also other options

different exercises throughout the workshop

Agenda

- Session 1: Introduction to Research with Social Media Data (SMD)
- Session 2: SM Data Collection
- Session 3: SMD Preprocessing and Analysis
- Session 4: Potential Pitfalls of SMD
- Session 5: Identifying Pitfalls with help from surveys
- Session 6: Identifying Pitfalls in SMD
- Session 7: Mitigating Pitfalls
- Session 8: Documenting Pitfalls
- Session 9: Recap and Conclusions

Today's Schedule

Monday, 05.12.	
9:30-11:00	Introduction to Social Media Research
11:00-11:15	*Break*
11:15-12:00	Data Collection
12:00-12:30	*Break*
12:30-14:00	Hands on data collection

About you

Introductory Round

What brought you to this course?

We would like to learn more about

- Your interest in social media data (specific types of data or methods you might want to learn more about)
- Prior experience
- Your disciplinary background
- https://tinyurl.com/socialmediaworkshop22

Agenda

- Session 1: Introduction to Research with Social Media Data (SMD)
- Session 2: SM Data Collection
- Session 3: SMD Preprocessing and Analysis
- Session 4: Potential Pitfalls of SMD
- Session 5: Identifying Pitfalls with help from surveys
- Session 6: Identifying Pitfalls in SMD
- Session 7: Mitigating Pitfalls
- Session 8: Documenting Pitfalls
- Session 9: Recap and Conclusions

gesis Leibniz Institute for the Social Sciences

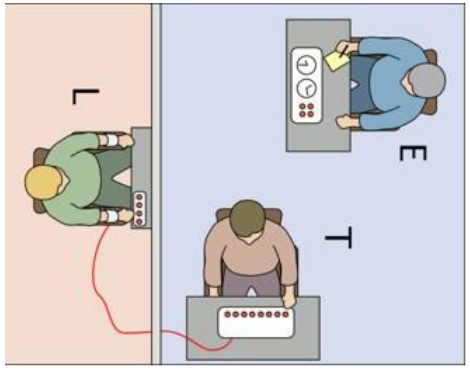
1. Introduction and Data Collection

Photo: Katrin Weller

Introduction: digital traces from social media

Sources of data for SocSci research

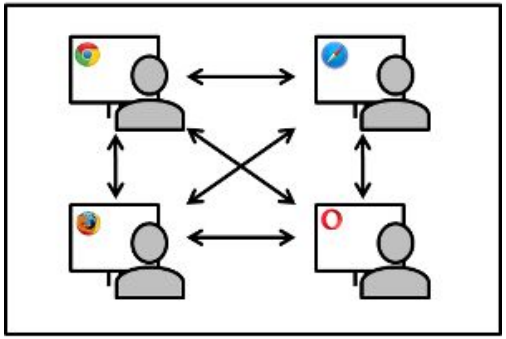
Surveys



Experiments

Observation

Online/digital surveys



Online/digital experiments

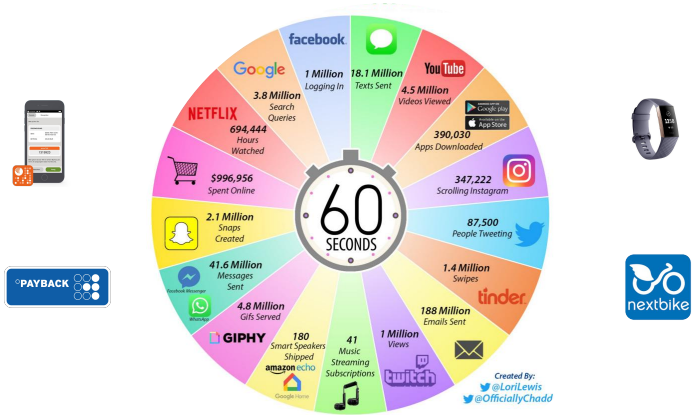
Online/digital observation

26

Two main sources of data

Digital traces of humans

"Data documenting the **interactions of users with digital devices or services**" [Howison et al., 2011]



Surveys (via Questionnaires)

"The collection of information from a **sample of individuals** through their **responses to questions**"
[Check & Schutt, 2012]

Two main sources of data

Digital traces of humans

"Data documenting the **interactions of users with digital devices or services**" [Howison et al., 2011]

Surveys (via Questionnaires)

"The collection of information from a **sample of individuals** through their **responses to questions**" [Check & Schutt, 2012]

Digital behavioral data

Social Data

Found data

Big data

Two main sources of data

Properties of...

Digital traces of humans	Surveys
- non-designed ("found") & not reactive to sp. stimulus	- designed in setup & specific stimulus
- self-selected & n=all per system	- usually probabilistically sampled f. target pop.
- very heterogeneous signals	- reaction space constrained
- large volume, cheap	- small scale, costly
- fine-grained (longitudinal, space)	- limited set of collections
- relational	- unlinked units

gesis Leibniz Institute for the Social Sciences

Computer Science

Social Sciences

2010 From tweets to polls [O'Connor et al.]

2013 Text as data [Grimmer et al.]

2014 Predicting tie strength with social media [Gilbert et al.]

2013 Big Data in Survey Research: AAPOR Task Force Report [Japec et al.]

2019 Investigating commentator bias in football broadcasts

2019 Combining surveys and digital traces [Stier et al., Pasek et al.]

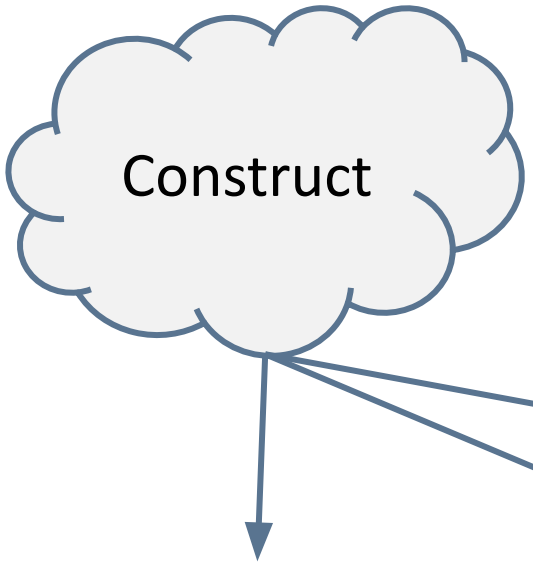
United by data: Social scientists and computer scientists are both interested in how digital traces can be used for inferring human behaviour & attitudes

Let's unite on Methods and Theory

- **Common vocabulary** that would improve communication and cross-development of measurement theories
- **Knowledge transfer re: overlapping approaches,** especially in addressing challenges and errors
- **Standards in documenting pipelines and reporting results**

Understanding Social Phenomena

Understanding Social Phenomena



- Elements of **information** that are sought by the researcher [Groves et al., 2011]
- Can range from simple (weight) to complicated (sexism)

Attitudes

Behaviours

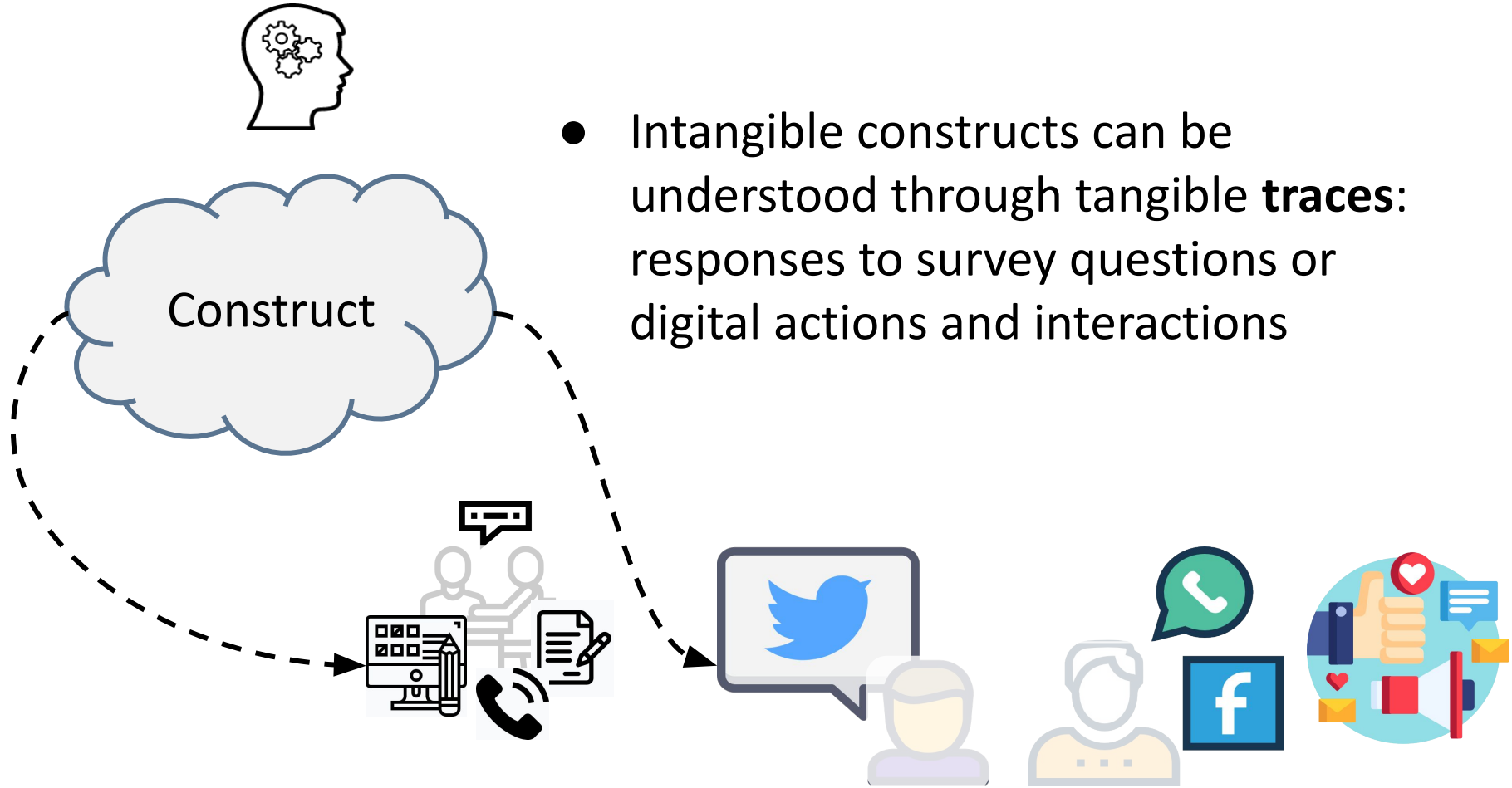
Characteristics

Understanding Social Phenomena

- The set of **people** to be studied [Groves et al., 2011]
- Can be easily defined (all the preschool students in a city) to more difficult (all refugees)
- Usually a national population, can also be any “system population” (-> platform study vs. “Social Sensing”)

Target Population

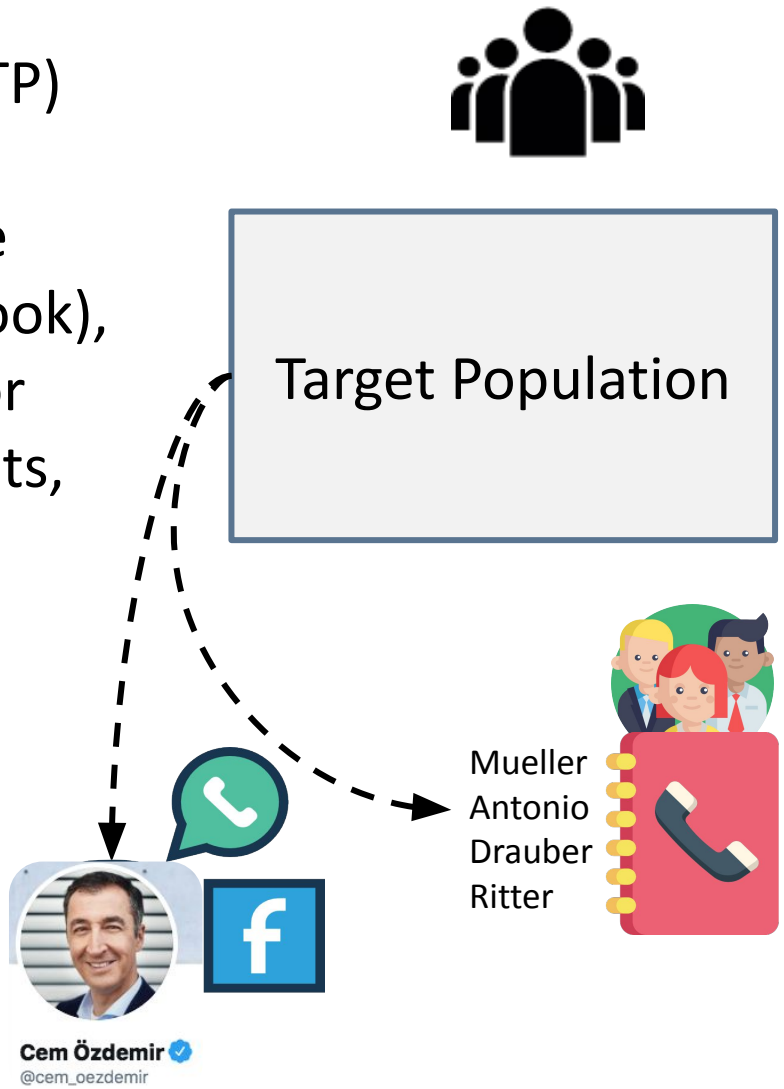
Understanding Social Phenomena



- Intangible constructs can be understood through tangible **traces**: responses to survey questions or digital actions and interactions

Understanding Social Phenomena

- **Users** of the target population (TP) are the *representation of TP elements* in a record. This can be offline records (census, phonebook), online platforms (profiles, IPs), or other digital forms (bank accounts, fitness trackers)



Studying social media as digital traces

- Researchers value social media as a new type of data
- Previously „ephemeral data“ become visible
- Immediate – quick reaction to events
- Structured
- „natural“ data

“What I find really interesting is that structure becomes manifest in internet communication. So it’s the first time in history actually that we can, that social structures between people become manifest within a technology. (...) They become visible, they become crawlable, they become analyzable.”

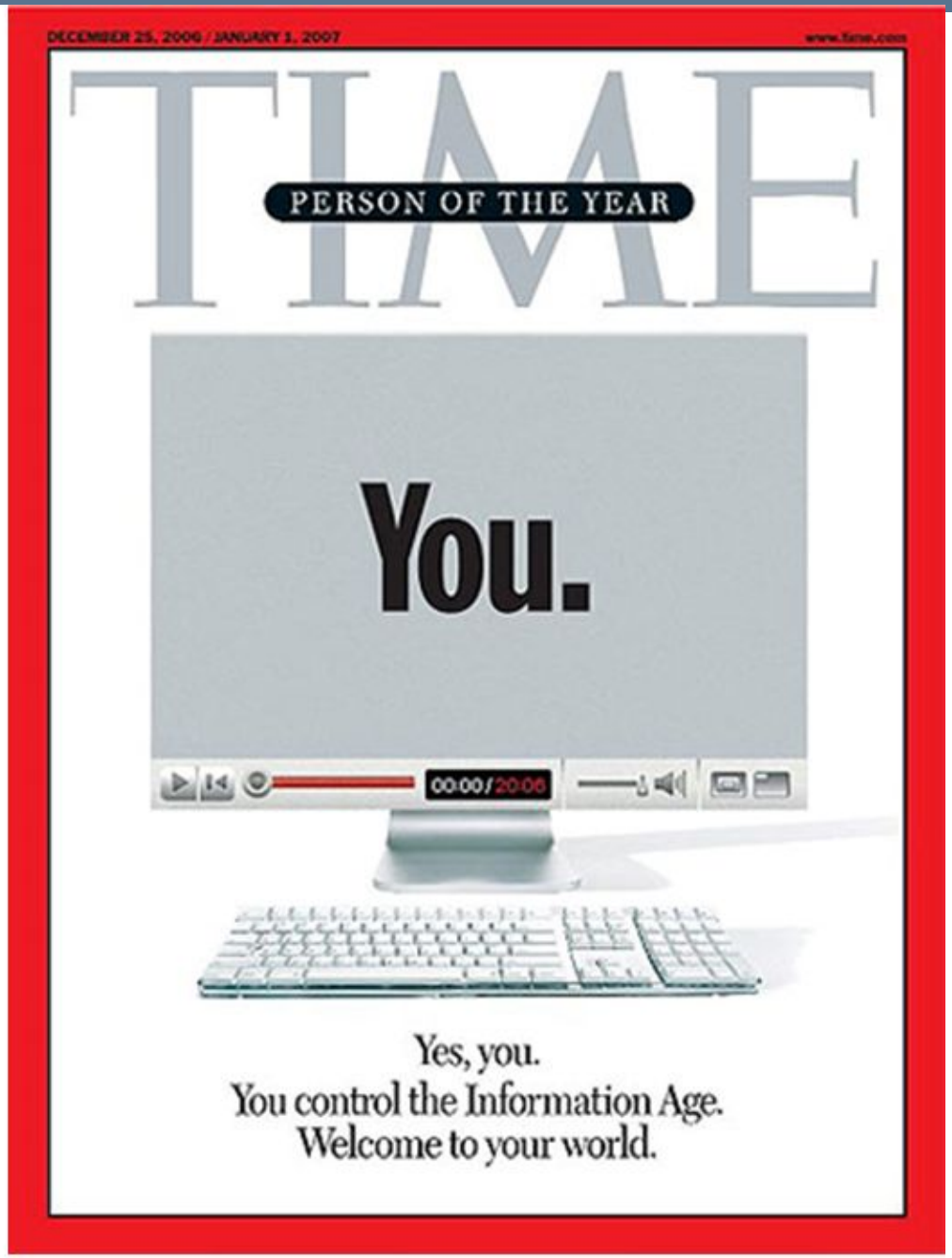
Kinder-Kurlanda, Katharina E., and Katrin Weller. 2014. "'I always feel it must be great to be a hacker!': The role of interdisciplinary work in social media research." In Proceedings of the 2014 ACM conference on Web Science, 91-98. New York: ACM.

Studying social media as digital traces

Data on social media platforms is not created for research purposes.

This leads to challenges in

- Accessibility
- Quality
- Interpretation
- Ethics


DECEMBER 25, 2006 / JANUARY 1, 2007
TIME
PERSON OF THE YEAR
You.
00:00 / 20:06
Yes, you.
You control the Information Age.
Welcome to your world.

2006

Social Media

Online platforms, that are typically build around user activities and connections/interactions between users.

Examples:

Social-Networking-Sites, Microblogs, Blogs (& podcasts), Media Sharing Sites (video sharing, photo sharing …), Social games, Social news, Q&A communities, Wikis, Product-based communities

Digital traces may also come from:

search engines, shopping portals, dating websites...

Categorizing social media platforms?

Web 2.0 (O'Reilly 2005)
Social Web
Social Networking Sites SNS
Social Media
Virtual Communities

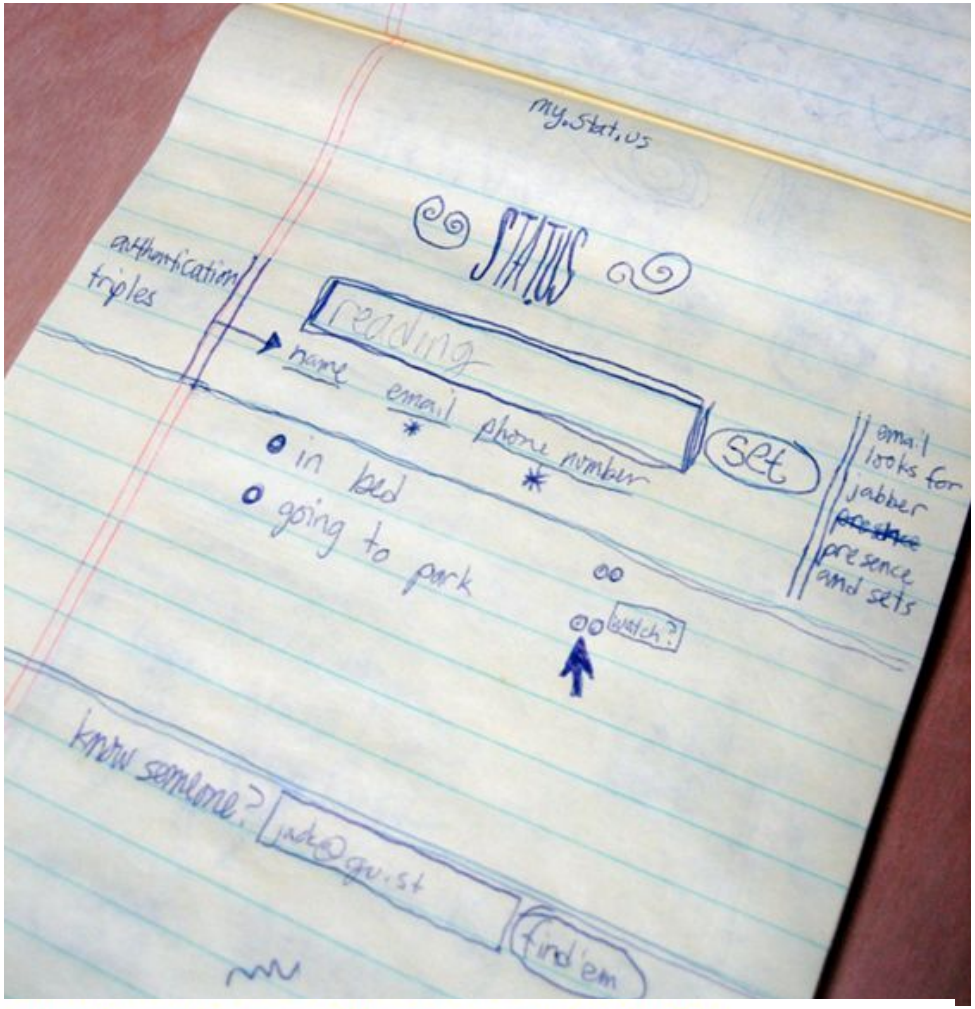
Prosumerism (Buzzetto-More 2013) [based on Toffler, 1980]
Produsage (Bruns 2008)

User-generated content
Online data / Internet data / Web data
Digital behavioural data

Categorizing social media platforms?

- What is the main type of „content“? (photos, texts, videos, users…)
- What are the types of interactions?
- What are the main activities and use cases? (stay informed, connect, bookmark, share, play, sell…)
- Openness of the platform?
- Professional or private use?
- International or local?

Platforms and affordances



https://www.flickr.com/photos/jackdorsey/182613360

Cross-platform connections

- Facebook post linking to a blog post
- Instagram photo shared through Twitter
- Tweet consisting of just a link to a Facebook post
- Newspaper article shared and commented through Facebook
- …

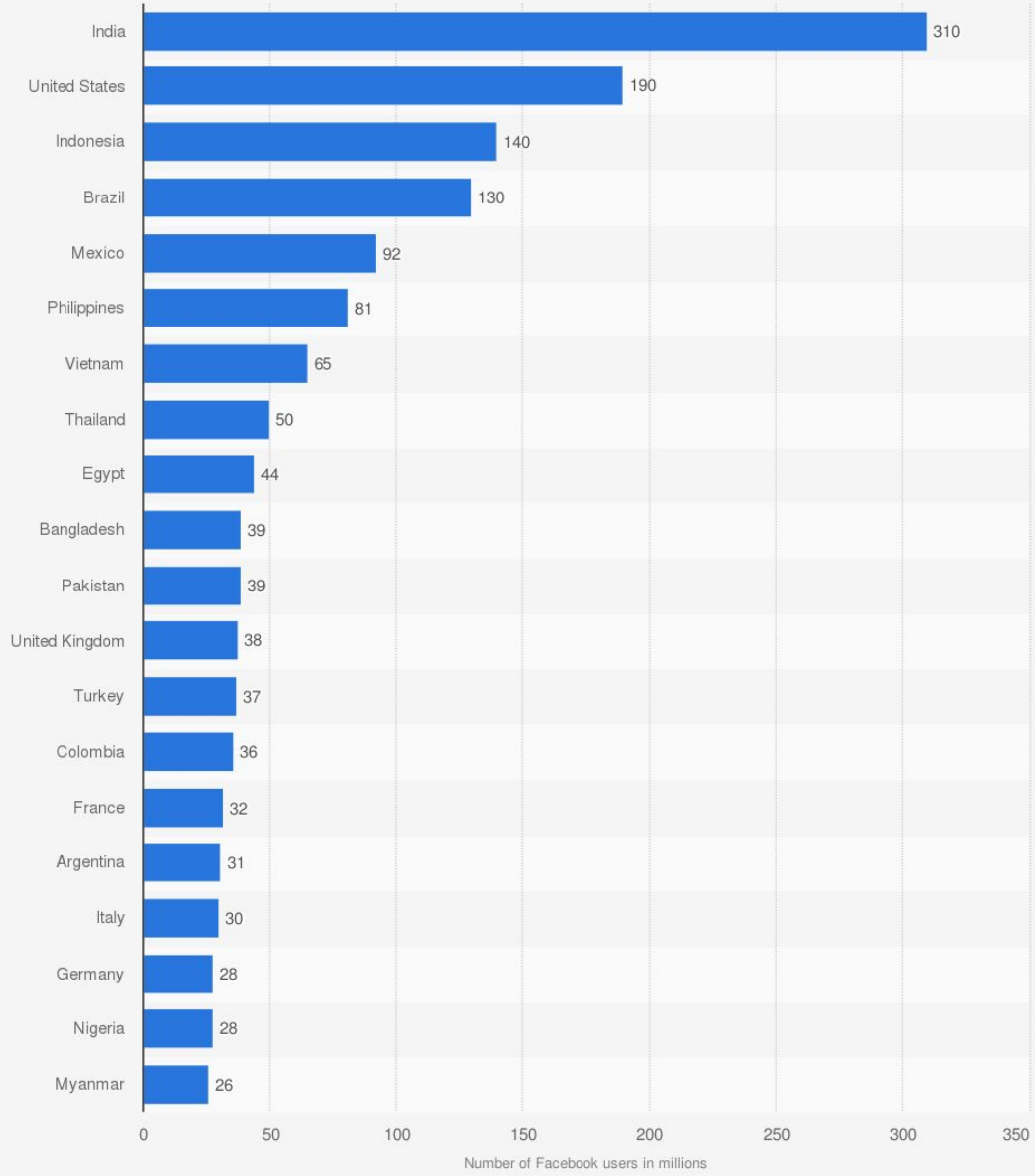
Platform affordances / functionalities

- Different interpretations of functionalities („Retweets are not endorsements“)
- Platform affordances can influence user behaviour - and the other way round
- Users creating new practices / standards

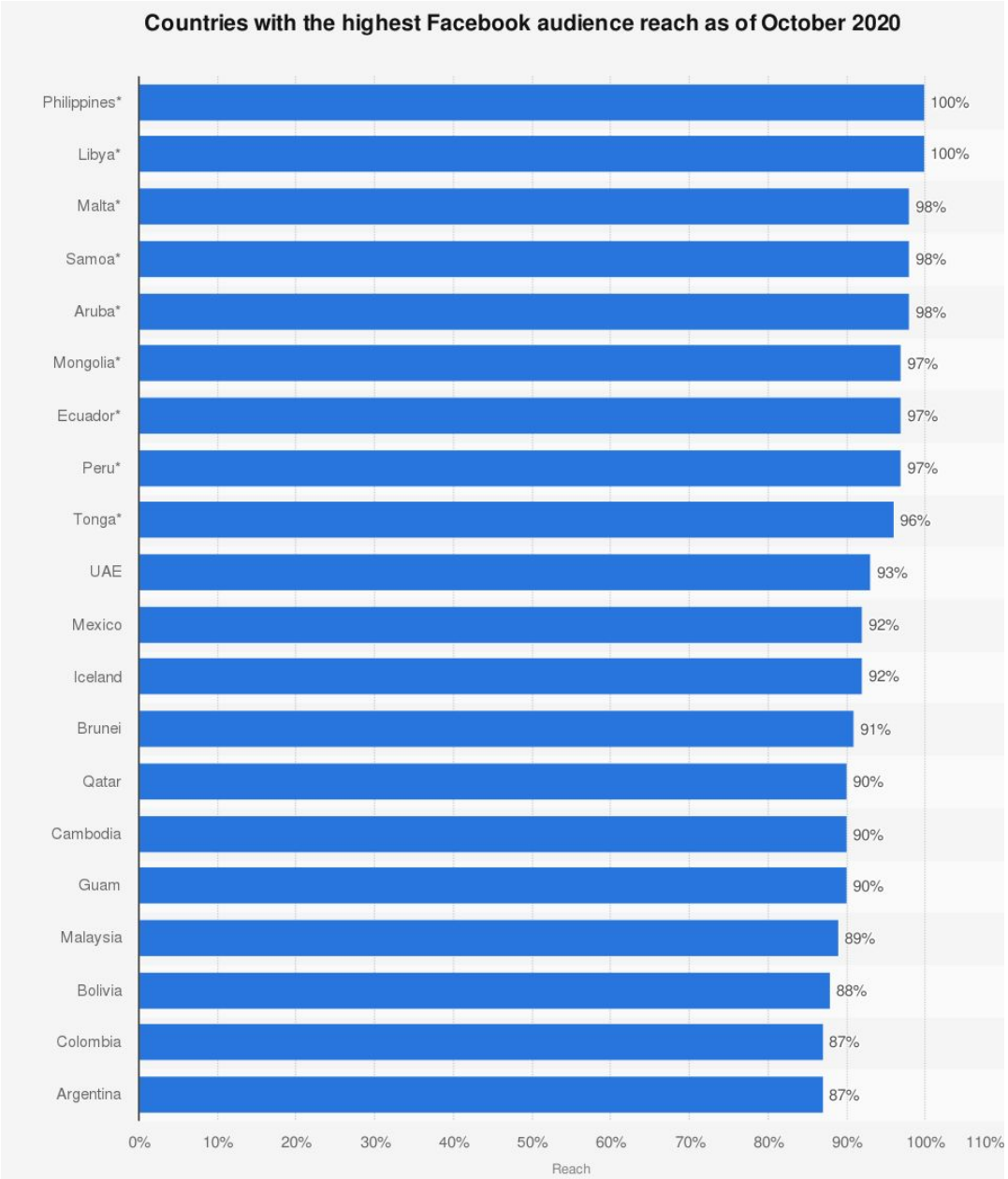
https://twitter.com/chrismessina/status/223115412?ref_src=twsrc%5Etfw

Social media users and usage

- Different user demographics per platform (often limited profile information)
- Different popular platforms per country
- Changes over time



https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/



https://www.statista.com/statistics/278435/percentage-of-selected-countries-internet-users-on-facebook/

User numbers provided by platforms

What is being measured?

- Daily active users
- Monthly active users
- „active“ users
- Everyone with an account

Other challenges when working with usage numbers

- Different contexts for usage (e.g. mobile)
- Different approaches to capture demographics
- “Active” and “passive” users
- User behaviour is constantly changing, people are exploring new practices.

non-human social media users

https://twitter.com/MarsCuriosity

32


A tale of two Libyas
Plus: Why the U.S. can't sit on the sidelines BY FAREED ZAKARIA
The GOP's misinformation campaign BY JOE KLEIN
Could your baby be depressed?
THE CULTURE
Word up: A dictionary of slang
TIME
YOUR DATA
FOR SALE
Everything about you is being tracked— get over it
BY JOEL STEIN
Owns a laptop
Age: 38-39
Likes: online news
Likes: Asian cuisi
Dislikes: cars
Likes: green living
Favorite celebrities:
Wi-fi warrior
Likes: business & financ
Sister is a la
Likes: parenting
Spent $180 on intim
Male
Mother: Rosalind Burd
Dislikes: autos & vehicles
Works at company with 5,000+ employees
Likes: movies
Smart-phone user
Sister: Lisa Stein Browning
Has had LASIK surgery
Likes: restaurants
Married
Likes: hiking
Likes: retail
Likes: newspapers
Magazine subscriber
Likes: cooking & recipes
Likes: online shopping
& actresses
Politically active
Religion: Jewish
hockey
Likes: fashion
Major life-insurance holder
Lives in New York City
Likes: music
What data-mining companies think they know about Joel Stein

2011

Research ethics

- Research ethics practices are also still evolving
- Main focus on privacy
- Lack of informed consent
- “participants” not expecting research activities on their data [Fiesler & Proferes, 2018]
- Different assumptions for different types of user groups (e.g. vulnerable groups)
- Ethically problematic research questions
- Potential starting point: AoIR ethics guidelines

your own challenges?

Next week, we will discuss examples of research designs with social media data. Especially social science questions that have been investigated with other types of data, but can now be explored with social media data.

Email us until Tuesday afternoon if you want your specific examples to be included. Vague ideas are totally okay!