

gesis

Leibniz-Institut
für Sozialwissenschaften



Chancen und Herausforderungen in der Forschung mit digitalen Verhaltensdaten

Johannes Breuer



Vorstellung

- Senior Researcher bei GESIS – Leibniz-Institut für Sozialwissenschaften
 - ▶ Datenarchiv für Sozialwissenschaften
 - Team Data Linking & Data Security
- davor: Mitarbeit in verschiedenen Forschungsprojekten zur Nutzung und Wirkung digitaler Medien (Uni Köln, Uni Hohenheim, WWU Münster, IWM Tübingen)
- Promotion in Psychologie (Uni Köln), Studium der Medienwissenschaften (Uni Köln)
- More: <https://www.johannesbreuer.com/>

Was ist GESIS?

- Institut der Leibniz-Gemeinschaft
- Standorte: Köln & Mannheim
- größte Infrastruktureinrichtung für Sozialwissenschaften in Deutschland
- großes Angebot an Angeboten und Services für Forschende (richtet sich auch an Studierende)
- mehr zu Angeboten & Struktur von GESIS unter www.gesis.org
- Social Media
 - ▶ Twitter: https://twitter.com/gesis_org
 - ▶ Facebook: <https://www.facebook.com/gesis.org/>

- Recherchieren** Angebot
- Studien planen**
 - GESIS Survey Guidelines
 - Projektberatung
 - Stichproben
 - Items und Skalen
 - Kognitive Pretests
 - Sozialstrukturelle Merkmale
 - Übersetzung
 - Data Management Plans
- Daten erheben**
 - GESIS Survey Guidelines
 - Projektberatung
 - Interviewertraining
 - Ziehung Telefonstichproben
 - GESIS Panel
 - Survey Operations
 - Erhebung Digitaler Verhaltensdaten**

- Daten analysieren**
 - ALLBUS
 - Amtliche Mikrodaten
 - Internationale Umfragen
 - PIAAC
 - Wahlstudien
 - Weitere Sekundärdaten
 - EUROLAB
 - Datenharmonisierung
 - Analyse Digitaler Verhaltensdaten**
- Archivieren und registrieren**
 - Datenarchivierung
 - Datenservices
 - Datenregistrierungsagentur da|ra
 - CESSDA Training
 - Daten-Repositorium datorium
 - Volltextserver SSOAR

- Publikationen**
 - GESIS-Schriftenreihe
 - GESIS Papers
 - Zeitschrift HSR
 - Zeitschrift mda
 - ISI
 - Online Journal Survey Methods
 - Archiv
- Veranstaltungen**
 - GESIS Training
 - GESIS Vortragsreihen
 - GESIS Tagungen
 - Computational Social Science Events
 - Veranstaltungsarchiv

Gliederung

1. Was sind digitale Verhaltensdaten und was kann man mit ihnen machen?
2. Wie kommt man an digitale Verhaltensdaten ran?
3. Welche besonderen Herausforderungen gibt es bei digitalen Verhaltensdaten und wie kann man mit diesen umgehen?
4. Where to learn more?

Was sind digitale Verhaltensdaten und was kann man mit ihnen machen?

Definition

“digital trace data can be roughly defined as ‘**records of activity** (trace data) undertaken through an **online information system** (thus, digital)’ ([Howison, Wiggins, & Crowston, 2011](#)) and can be collected from a multitude of technical systems, such as **websites, social media platforms, smartphone apps, or sensors**” ([Stier et al., 2019](#))

Begrifflichkeiten

- Digitale Verhaltensdaten (digital behavioral data)
 - ▶ Definition von Verhalten?
 - ▶ Bsp.: kommentieren vs. Liking/Following -> Einstellungen
- Digitale Spurdaten (digital trace data)
 - ▶ Intentionalität?
 - ▶ [Menchen-Trevino \(2013\)](#): participation traces (z.B. Kommentare/Posts) vs. transactional data (z.B. Logins)
- Big Data
 - ▶ Kriterien?
 - Größe in GB/TB?
 - Nicht mehr lokal auf Rechner analysierbar?
 - Um Faktor x größere Zahl an Variablen im Vergleich zu Fällen

Arten von digitalen Verhaltensdaten

- **Social-Media-Daten:** Twitter, Facebook, Reddit, Instagram, YouTube etc.
- **Tracking-Daten:** Webtracking, App-Nutzung...
- **Sensordaten:** z.B. GPS-Daten von Smartphones
- [Menchen-Trevino \(2013\)](#):
 - ▶ **Horizontal** trace data: Breite (z.B. alle Tweets zu einem bestimmten Hashtag)
 - ▶ **Vertical** trace data: Tiefe (umfassende Daten zur Nutzung für begrenzte Gruppe von Nutzer_innen)

Datenformate

- Text (z.B. Posts, Tweets, Kommentare)
- Netzwerkdaten (z.B. Follower auf Twitter)
- Log Files (z.B. besuchte Webseiten)
- Tabular Data (z.B. exportierte Sensordaten -> Zeitreihen)
- Bilder (z.B. Fotos)
- Videos

Analyseeinheiten

- Nutzer_innen
 - ▶ individuell
 - ▶ gruppiert (z.B. nach Regionen) -> Aggregatdaten
- Inhalte
 - ▶ Videos, Fotos, Songs...
 - ▶ Hashtags
 - ▶ Webseiten

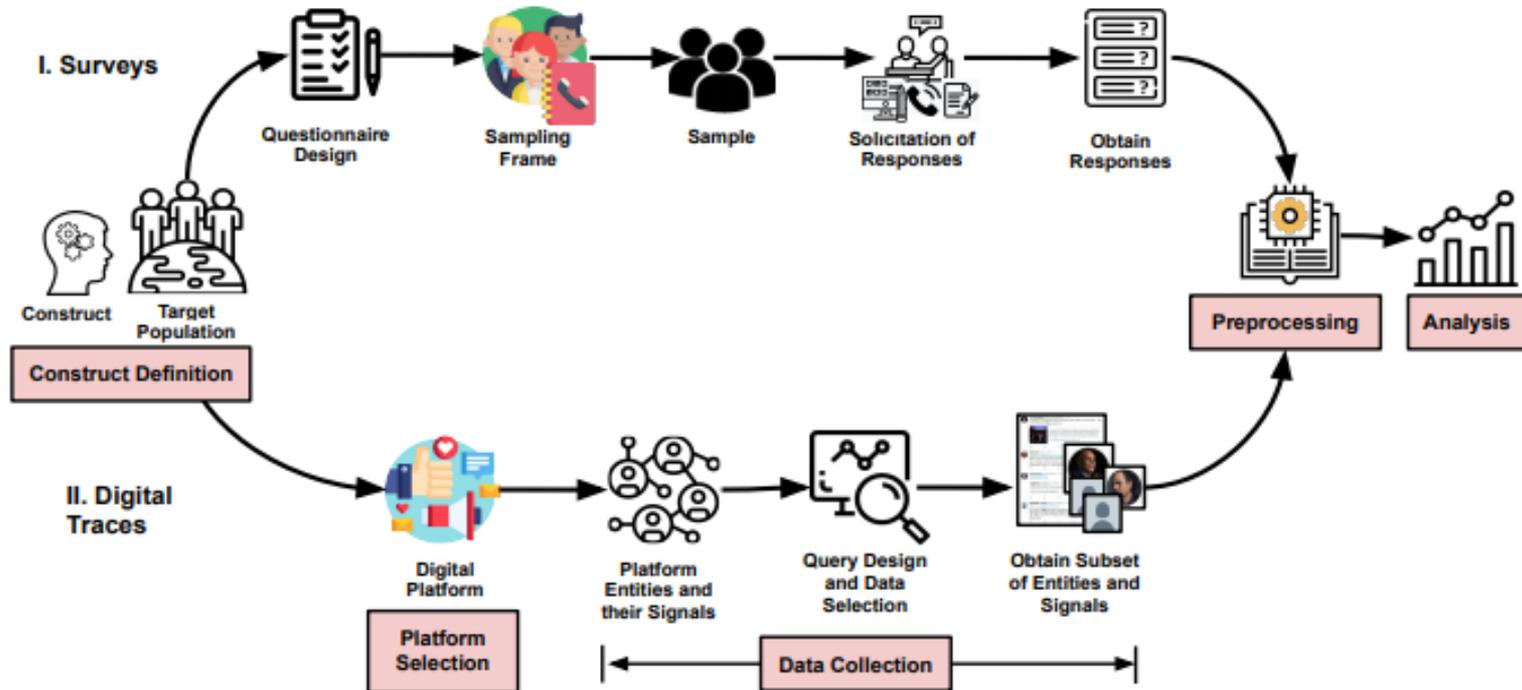
Setting the scene...

- die Nutzung digitaler Medien/Geräte durchdringt immer mehr Lebensbereiche
- diese Nutzung generiert eine große Menge an Daten
- viele dieser Daten sind auch für die (sozial-)wissenschaftliche Forschung interessant

Warum digitale Verhaltensdaten?

- Verfügbarkeit
 - ▶ Aber: Zugang häufig begrenzt/eingeschränkt
- Verwertbarkeit
 - ▶ Entwicklung von Rechenleistung (computing power)
 - ▶ (Weiter-)Entwicklung von Analysemethoden
- Limitationen von Selbstauskünften (self-report)
 - ▶ soziale Erwünschtheit
 - ▶ Schwierigkeiten beim Erinnern
 - Bsp. Mediennutzung

Befragungen vs. DVD



aus: [Sen et al. \(2019\): A Total Error Framework for Digital Traces of Humans](#)

Wie kommt man an digitale Verhaltensdaten ran?

Datenzugänge

- **Prinzipiell 3 Wege** (Breuer, J., Bishop, L., & Kinder-Kurlanda, K. (2019). The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. *New Media & Society*, Accepted for publication):
 1. Daten selbst erheben/sammeln
 2. Direkte Kooperationen mit Anbietern/Betreibern von Plattformen/Services
 3. Daten kaufen: Data Reseller oder Marktforschungsunternehmen
- Wahl des Zugangswegs abhängig von verfügbaren Ressourcen (Fähigkeiten, Geld, Zeit)

Daten selbst sammeln

- Erneut grundlegend 3 Optionen
 1. Application Programming Interfaces (APIs) von Plattformen/Services
 2. Web Scraping
 3. „Datenspende“ der Nutzer_innen
- Jede Variante hat eigene Vor- und Nachteile

APIs

Vorteile	Nachteile
<ul style="list-style-type: none">• werden von vielen Plattformen/Services Angeboten (z.B. Twitter, Spotify, FitBit)• zumeist gut dokumentiert• viele Tools verfügbar, mit denen auch ohne Programmierkenntnisse auf APIs zugegriffen werden kann• liefern strukturierte Daten (häufig in Form von JSON-Files)	<ul style="list-style-type: none">• nicht für wissenschaftlichen Datenzugang entwickelt• in aller Regel Begrenzungen für die Anzahl der Abfragen (und diese können sich ändern)• Zugang kann jederzeit von Anbietern begrenzt oder sogar komplett beendet werden (Bsp.: Facebook API nach Cambridge-Analytica-Skandal)

APIs

- [Freelon \(2018\): Computational Research in the Post-API Age](#)
 - ▶ Gestiegene Bedeutung von Web Scraping als Zugangsweg aufgrund von Einschränkungen von APIs
- Siehe auch Artikel von [Bruns \(2019\)](#) und Replik von [Puschmann \(2019\)](#) zu den Konsequenzen der Schließung/Einschränkung von APIs für die Wissenschaft

Web Scraping

Vorteile	Nachteile
<ul style="list-style-type: none">• flexibel• nicht (unmittelbar) von Entscheidungen von Anbietern (von Plattformen/Services) abhängig	<ul style="list-style-type: none">• technisch anspruchsvoller (in aller Regel zumindest grundlegende Programmierkenntnisse nötig)• Änderungen in Webseitenstrukturen können Probleme verursachen• unstrukturierte Daten, die aufbereitet werden müssen• fortlaufende Diskussion über rechtliche Situation (Golla & Schönfeld, 2019; RatSWD, 2019)

Datenspende

- Nutzer_innen können ihre Daten von vielen Plattformen/Services (z.B. Facebook, Twitter, Google) oder ihre Browser-History exportieren
 - ▶ Alternative: Browser-Plugins (siehe z.B. [Haim & Nienierza, 2019](#) für Facebook-Daten)
- diese Daten können sie dann Forschenden zur Verfügung stellen
- [Halavais \(2019\)](#): Möglichkeit, um Einschränkungen durch Terms of Service (z.B. für APIs) zu vermeiden
- Beispiele:
 - ▶ [OpenSCHUFA](#) von Algorithm Watch
 - ▶ [Open Humans](#)

Datenspende

Vorteile	Nachteile
<ul style="list-style-type: none">• Informierte Einwilligung (Informed Consent)• Transparenz für Studienteilnehmer_innen• Keine Einschränkung durch Rate Limits oder Terms of Service von APIs	<ul style="list-style-type: none">• Nicht einfach umzusetzen<ul style="list-style-type: none">• Teilnehmer_innen brauchen Anleitung• Lösungen für sicheres Teilen benötigt• Lösung für Anonymisierung der Daten nötig (Bsp.: Freunde der Teilnehmer_innen, die in Posts/Kommentaren markiert sind)

Welche besonderen
Herausforderungen gibt
es bei digitalen
Verhaltensdaten und
wie kann man mit
diesen umgehen?

Beispiele spezieller Herausforderungen



Abbildung: <https://www.gesis.org/angebot>

Studienplanung

- Auswahl des Datenzugangswegs
 1. Forschungsinteresse
 2. Verfügbare Ressourcen
 - Fähigkeiten/Wissen
 - Zeit
 - Geld
- Nutzung vorhandener Datensätze?
 - ▶ [Documenting the Now](#)
 - ▶ [TweetSets](#)
 - ▶ [TweetsKB](#)
 - ▶ Daten aus Repositorien: [GESIS Data Search](#)

Datenerhebung

- im Falle von Rate Limits: Aufteilung der Sammlung über Zeit und/oder Accounts
- im Fall von Individualdaten von/über Nutzer_innen: nach Möglichkeit informierte Einwilligung der Personen einholen
 - ▶ einfacher bei direkter Rekrutierung von Teilnehmer_innen über Plattformen oder andere Wege/Kanäle
 - ▶ schwierig, wenn Daten nur über API gesammelt werden
- Kombination mit anderen Daten: Befragungen, Interviews...

Datenauswertung

- Analysemethode(n) passend zu Fragestellung und Datentypen wählen
- Beispiele:
 - ▶ Inhaltsanalyse
 - ▶ Netzwerkanalyse
 - ▶ Text Mining & Natural Language Processing
 - ▶ Machine Learning
 - Supervised
 - Unsupervised

Daten archivieren & teilen

- Nachnutzung & Transparenz: Open Science
- Datenschutz: Persönliche Daten
 - ▶ Anonymisierung, z.B. durch:
 - Reduktion
 - Aggregation
- Terms of Service sollten beachtet werden
- “As open as possible, as restrictive as necessary”
- Empfehlungen und Beispiele zur Archivierung von Social-Media-Daten: [Bishop & Gray \(2017\)](#); [Kinder-Kurlanda et al. \(2017\)](#); [Mannheimer & Hull \(2017\)](#); [Thomson \(2016\)](#); [Williams et al. \(2017\)](#)

Where to learn more?

Open Humans

- <https://www.openhumans.org/>
- Projekte zur Exploration eigener Daten (z.B. Twitter, Fitness-Tracker)
- Community von Forscher_innen und Bürgerwissenschaftler_innen
- Möglichkeit, eigene Daten mit anderen zu teilen

Digital Methods Initiative

- <https://wiki.digitalmethods.net/Dmi/DmiAbout>
- Uni Amsterdam (UvA)
- Fokus eher auf beschreibenden, qualitative Methoden
- Digital Methods Winter School
- Entwicklung zahlreicher Tools für verschiedene Plattformen, Inhalte und Zwecke:
<https://wiki.digitalmethods.net/Dmi/ToolDatabase>

Social Media Lab

- <https://socialmedialab.ca/>
- Ryerson University (Toronto)
- Entwicklung von Tools zur Sammlung und Analyse von Social-Media-Daten:
<https://socialmedialab.ca/apps/>
- *Social Media & Society* Conference:
<https://socialmediaandsociety.org/>

GESIS-Angebote

- Angebote zur Erhebung und Auswertung von DVD (Fokus auf quantitativen Methoden)
- [Social Media Monitor](#)
 - ▶ Social-Media-Aktivitäten von Politiker_innen und Organisationen
 - ▶ Fokus auf Bundestagswahl 2017
- [GESIS Training](#)
 - ▶ [Workshops](#), z.B.:
 - [Automatic Sampling and Analysis of YouTube Data](#) (Köln, 10. & 11.02.2020)
 - [Research Factory for Text Mining in the Social Sciences](#) (23.-25.03.2020, Köln)
 - [Linking Twitter & Survey Data](#) (23. & 24.06.2020, Köln)
 - ▶ [Spring Seminar 2020: Digital Behavioral Data](#) (3 einwöchige Kurse; alle in Köln)
 - [Fundamentals of Data Analysis with Python](#) (02.- 06.03.2020)
 - [A Practical Introduction to Machine Learning in Python](#) (09. – 13.03.2020)
 - [Social Network Analysis with Digital Behavioral Data](#) (16. – 20.03.2020)

Vielen Dank für Ihre Aufmerksamkeit!

gesis

Leibniz-Institut
für Sozialwissenschaften

Leibniz
Leibniz
Gemeinschaft