Going beyond self-reports: A computational look at processing, enriching, and analyzing digital traces

Joshua Claassen
German Center for Higher Education Research and Science Studies (DZHW)
Leibniz University Hannover

CS3 Meeting
July 14, 2026

I am grateful to my supervisor Jan Karem Höhne
and my co-authors Laura Boeschoten, Karin van Es,
Dennis Nguyen, Melanie Revilla, and Bella Struminskaya
for their great support in conducting this research.

Introduction I

Internet is used by **5.56 Billion** individuals worldwide

This number has increased by **2.5%** during the last year

Internet users spent on average **6-7 hours** per day online

Numbers according to the "Digital 2025: Global Overview Report": https://wearesocial.com/es/blog/2025/02/digital-2025-la-guia-esencial-del-estado-global-de-lo-digital/

Introduction II

ESS, Round 11 (2023)

On a typical day, about how much time do you spend using the internet on a computer, tablet, smartphone or other device, whether for work or personal use?

Please give your answers in hours and minutes.

TYPE IN DURATION: [ ][ ] [ ][ ]

hours minutes

Eurobarometer, 98.2 (2023)

Could you tell to what extent you watch television via the internet?

- Everyday/Almost everyday
- Two or three times a week
- About once a week
- Two or three times a month
- Less often
- Never

ESS CRONOS-3, Wave 1 (2024)

I know how to check the truthfulness of the information or content I find on the internet.

- Not at all true of me
- Not very true of me
- Somewhat true of me
- Mostly true of me
- Very true of me
- I don't understand what this means

Introduction III

nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-021-01117-5

Check for updates

A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use

Douglas A. Parry[1]✉, Brittany I. Davidson[2,3], Craig J. R. Sewall[4], Jacob T. Fisher[5], Hannah Mieczkowski[6] and Daniel S. Quintana[7,8,9]

There is widespread public and academic interest in understanding the uses and effects of digital media. Scholars primarily use self-report measures of the quantity or duration of media use as proxies for more objective measures, but the validity of these self-reports remains unclear. Advancements in data collection techniques have produced a collection of studies indexing both self-reported and log-based measures. To assess the alignment between these measures, we conducted a pre-registered meta-analysis of this research. Based on 106 effect sizes, we found that self-reported media use correlates only moderately with logged measurements, that self-reports were rarely an accurate reflection of logged media use and that measures of problematic media use show an even weaker association with usage logs. These findings raise concerns about the validity of findings relying solely on self-reported measures of media use.

Digital Trace Data

Web/API Scraping	Web Tracking	Data Donation
Measurement of digital behavior by extracting infor-mation from programming interfaces/websites	Measurement of digital behavior through tracking applications installed on participants' devices	Measurement of digital behavior through participants sharing personal data from digital platforms
Aggregate-level data	**Participant-level data**	**Participant-level data**
Publicly available content (e.g., website content)	Website visits, search terms, and app usage	Digital platform behavior (e.g., viewing histories)
Cross-device data	Device-specific data	Cross-device data
Case Studies 1 & 2	Case Study 1	Case Study 2

Research Problem

- Existing research has mostly focused on the collection of digital traces
  - *Standardized workflows and best practices for collecting traces are already available* (Bosch & Revilla, 2022; Boeschoten et al., 2022; Carrière et al., 2025)
  - *Ample research on participation willingness and bias in web tracking and data donation studies* (Keusch et al., 2022; Revilla et al., 2021; Strycharz et al., 2024)
  - ***But:** Lack of research when it comes to further processing and analyzing traces*

- To interpret traces, which usually lack context, they must be …
  - *… augmented with auxiliary information* (Lindberg, 2023)
  - *… classified with respect to the digital behavior measured* (Müller & Bach, 2023)

Research Question & Goal

- **Research Question (RQ):** How can digital traces be transformed into meaningful measurements of digital behavior?

- **Goal:** Developing best practices, methodological recommendations, and standardized workflows for processing and analyzing traces

Case Study 1: *Claassen, J., & Revilla, M. (in progress). Innovating the measurement of parental digital behavior: A case study using web tracking data.*

The WEB DATA OPP project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 849165)

Melanie Revilla is Principal Investigator (PI)

We also thank Netquest (www.netquest.com) for the data collection in their Spanish panel

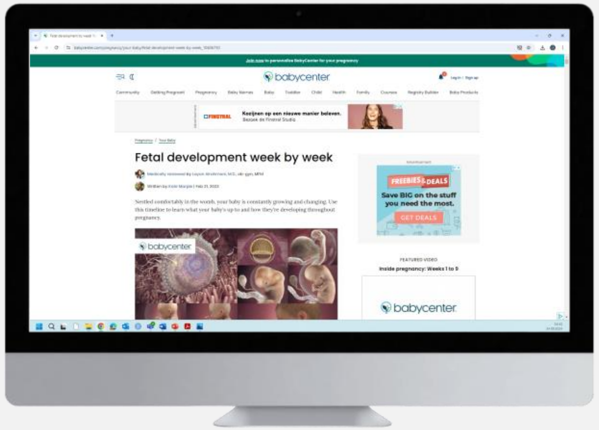
Parental Digital Behavior (PDB)

- Transition to parenthood introduces manifold challenges (Loudon et al., 2016)
  - *Coping with the physical and emotional stress of pregnancy and childbirth*
  - *Monitoring and ensuring child health, nutrition, and development*
- About 90% of parents use the internet to navigate these challenges (Jaks et al., 2019)
- PDB is related to real-world parenting, such as ...
  - *... providing educational support* (Livingstone et al., 2018)
  - *... deciding on child vaccination* (Charron et al., 2020)
- Previous studies are limited by primarily using survey self-reports

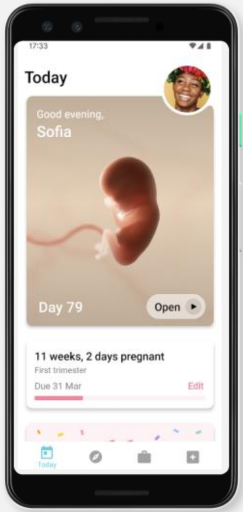
Research Question

- **RQ**: How can digital traces of website visits, app usage, and online searches be transformed into measurements of PDB?

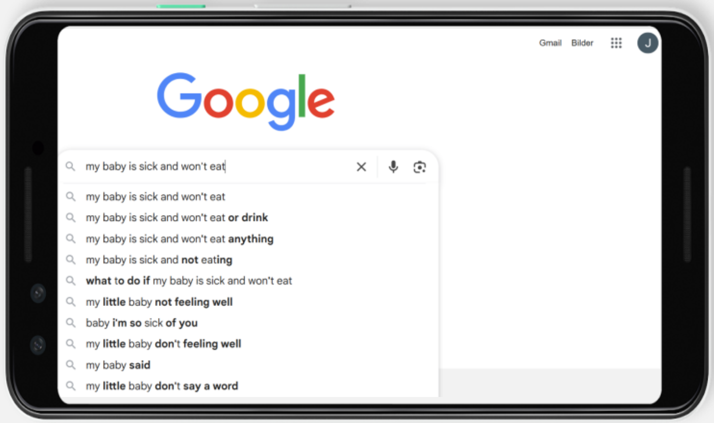
Method: Web Tracking



Website visits



App usage



Online searches

Method: Collection of Traces

- Data (N = 800) was collected in the Netquest Web Tracking Panel in Spain
    - *We selected participants in different stages of (expectant) parenthood*
    - *Study underwent comprehensive external ethics review and was approved by the ERC*

- Netquest retrospectively extracted all traces potentially related to (expectant) parenthood for a period of up to 24 weeks in 2023/24
    - ***<u>Data minimization</u>**: Extraction was based on approx. 700 websites, 1,200 apps, 700 search terms, as well as 27 URL keywords that we previously specified*

- In total, this resulted in approx. 184,000 traces
    - ***<u>Traces are enriched</u>** by scraping the HTML titles (website visits) as well as app descriptions and store categories (app usage)*

Method: Transformation of Traces

Multiple possible analytical decisions:

(1) What **type of trace** is considered?
- *Website visits, app usage, and/or online searches*

(2) What **classification method** is employed?
- *In our case: Llama 3.2 3B or Mistral 7b (i.e., two open-weight LLMs)*

(3) What **dimension of PDB** is focused on?
- *Overall, Health, General/Social services/Activities, Products, or Entertainment/Education*

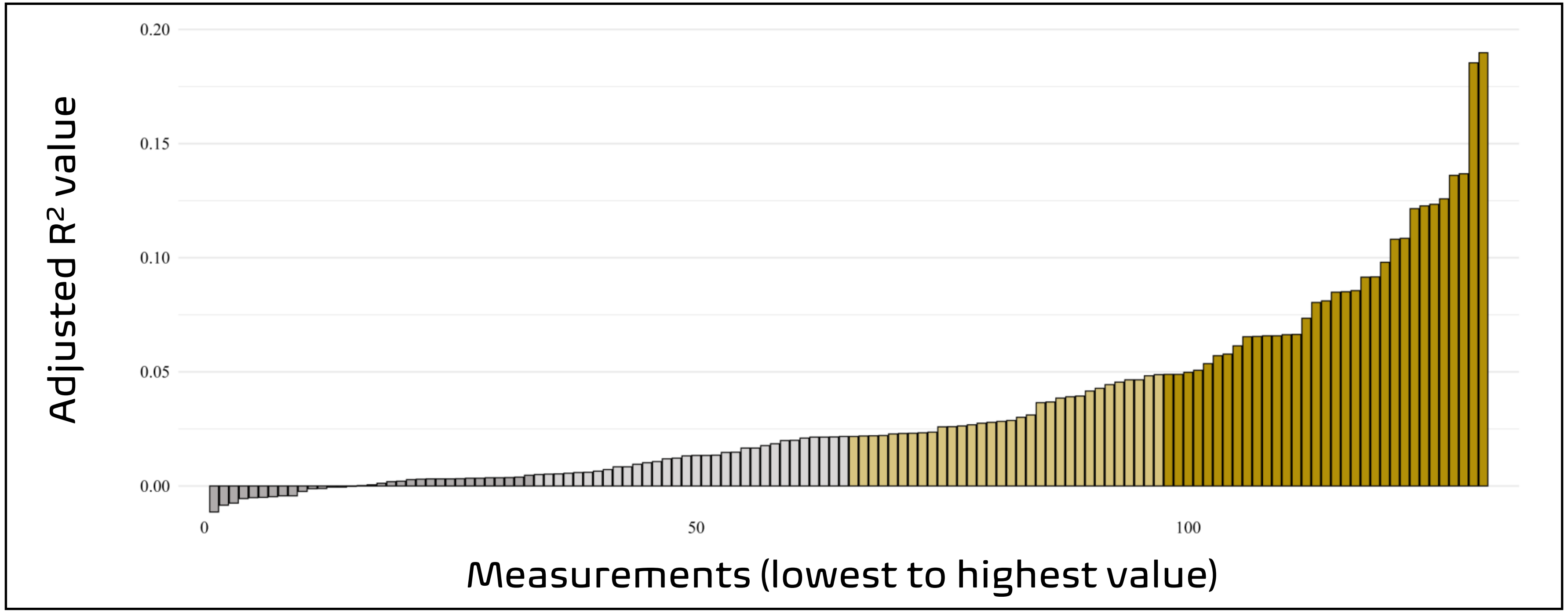
(4) What **aggregation rule** is applied?
- *Absolute frequency, absolute duration, relative frequency, or relative duration*

Method: Evaluation of PDB Measurements

- Analytical decisions result in 160 possible combinations (or measurements)
  - *Of these, **130 measurements** can be empirically realized*
- We estimate 130 linear regression models, with the measurements as dependent variables and three blocks of independent variables (IVs)
  - *Block 1: IVs on self-reported (expectant) parenthood*
  - *Block 2: IVs on sociodemographic characteristics*
  - *Block 3: IVs controlling for potential errors in the web tracking data*
- Each measurement is evaluated in terms of its "explainability"
  - *Explainability is defined as the adjusted R^2 values of the regression models*
- Impact of analytical decisions is assessed by estimating a "meta" regression
  - *Linear regression model with measurements as observations ($N = 130$), adjusted R^2 values as dependent variable, and analytical decisions as independent (dummy) variables*

Results: Explainability of PDB Measurements

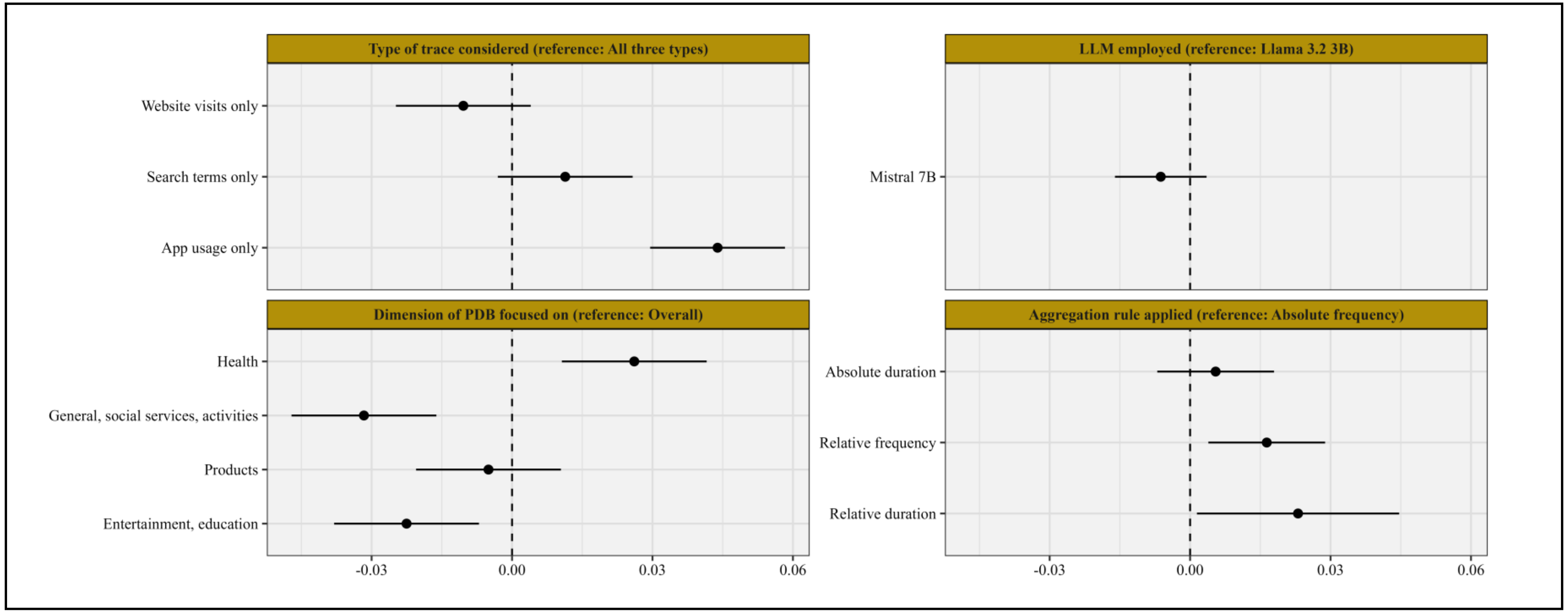
Fig 1. Adjusted R^2 values of all 130 realized PDB measurements



Note. Measurements are ordered from lowest to highest adjusted R^2 value. Bars are colored according to their measurement's quartile, ranging from dark silver (1st quartile) to dark gold (4th quartile).

Results: Impact of Analytical Decisions

Fig 2. Linear regression with adjusted R^2 as dependent variable



Note. Black dots = unstandardized coefficients. Black horizontal lines = 95% confidence intervals. N = 130. The four analytical decisions were included as dummy variables (reference category in parentheses, respectively).

Conclusion Case Study 1

- Overall, explainability varies considerably
  - *Across all 130 measurements, the adjusted R^2 values range between -0.01 and 0.19*
  - *Only ten measurements reach values above 0.10*
- Some analytical decisions are systematically associated with explainability
  - *Using **app usage only**, focusing on **health**, and applying **relative aggregation rules** leads to higher explainability*
  - *In contrast, the **LLM employed** appears interchangeable (in the context of our analyses)*

Case Study 2: *Claassen, J., Boeschoten, L., Nguyen, D., Struminskaya, B., & van Es, K. (in progress). Turning traces into measurements: How to utilize digital traces donated from Netflix to study video streaming behavior.*

Video Streaming Behavior (VSB)

- Viewing TV series and movies constitutes an important leisure activity
  - *Average daily television time in the US is about three hours* (Chan & Liang, 2023; Robinson & Martin, 2010)
  - *Series and movies can impact viewers' attitudes and well-being* (Scharrer & Warren; 2022; Till et al., 2019)

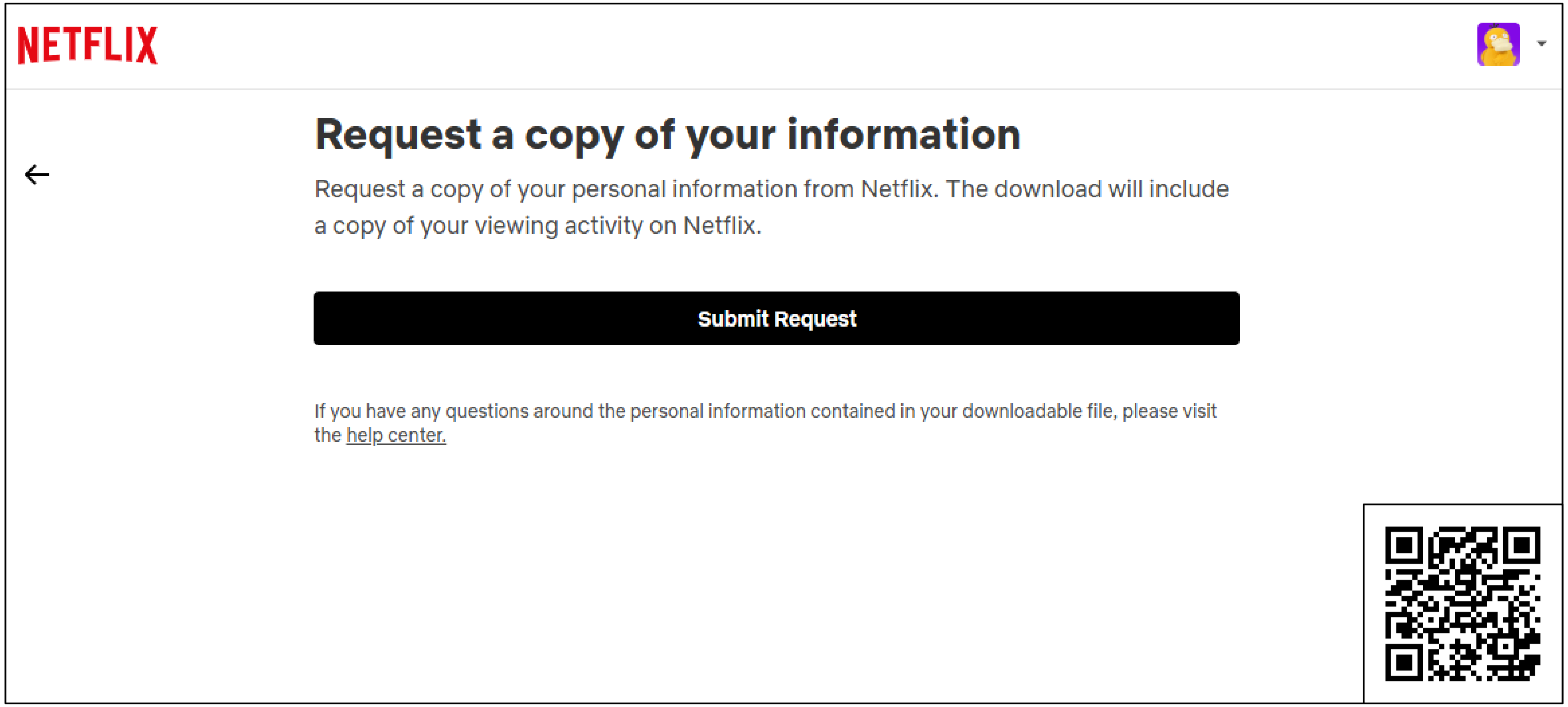
- Traditional viewing channels are supplemented by video streaming services
  - *Streaming services, such as Netflix, foster new viewing habits* (Tana et al., 2019; Kaur & Ashfaq, 2023)
  - *At the same time, they are "black boxes", so that empirical knowledge on VSB is still limited* (Van Es, 2024)

- Different VSB attributes are highly relevant to media/communication research
  - *Binge-watching* (Flayelle et al., 2020; Ilyas et al., 2023)
  - ***Genre patterns*** (Palomba, 2021; Scharrer & Blackburn, 2018)
  - *Diversity exposure: representation of societal groups and production countries* (Avery et al., 2021; Lotz et al., 2022)

Research Question

- **<u>RQ</u>**: How can digital traces from Netflix users be transformed into individual-level genre measurements?

Method: Digital Data Donation



Method: Trace Collection and Workflow

- Data donation (N = 126) in Dutch Ipsos I&O Panel in 2023 (Van Es et al., 2025)
  - *Participants requested and downloaded their Data Download Packages (DDPs) from Netflix*
- PORT locally extracted relevant traces from the DDPs (Boeschoten et al., 2023)
  - *This includes viewing histories (i.e., **titles of series episodes/movies viewed** with **timestamps**)*
- In total, this resulted in 617,951 traces
  - *These contain 66,033 unique traces (i.e., different series episodes/movies)*
  - *Traces must be enriched as they do **not contain any information** on **genres** and **runtimes***
- Development of a three-step workflow to link, enrich, and transform traces
  - *Pre-processing step ("extraction of title information") is not considered in this presentation*
  - *Workflow is connected to The Movie Database (TMDB) API and automated in Python*
  - *It is evaluated by calculating **quality indictors** for each workflow step*

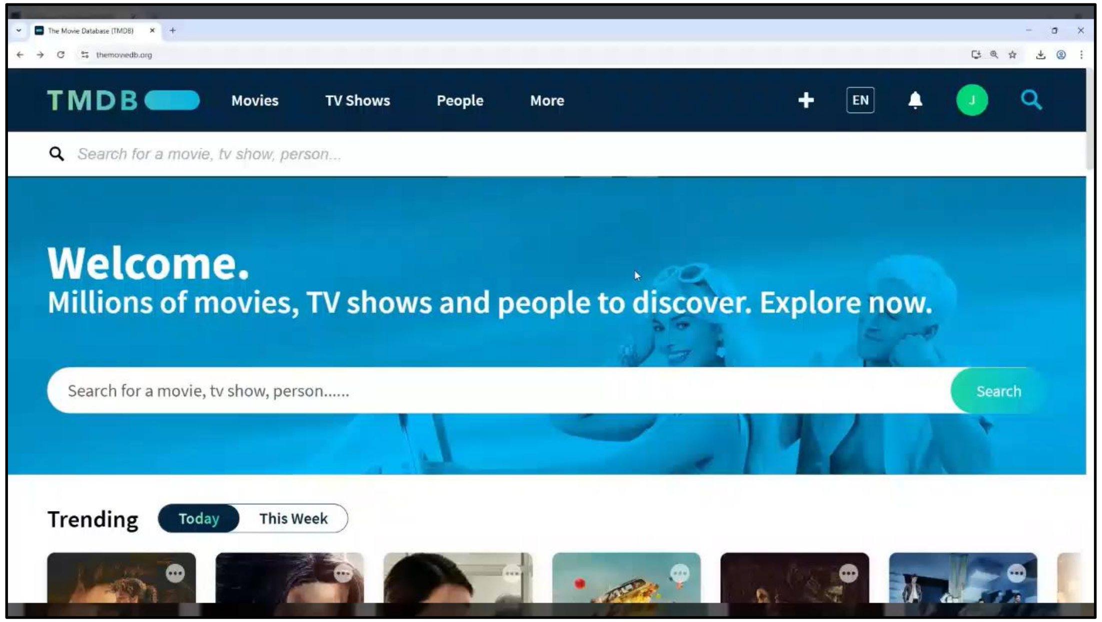
Workflow: (1) Linkage with TMDB Identifier I

Series: "The Office", Season: 2, Title: "Motivation", Episode: 4

Which one is it?!

Workflow: (1) Linkage with TMDB Identifier II

Series: "The Office", Season: 2, Title: "Motivation", Episode: 4



TMDB ID: 2996

Workflow: (2) Enrichment with Auxiliary Information

TMDB-ID: 2996, Series: "The Office", Season: 2, Title: "Motivation", Episode: 4

Genre	Runtime	Cast	Country
Comedy	27 minutes	Ricky Gervais (male), Martin Freeman (male), …	United Kingdom

Workflow: (3) Transformation into Genre Measurements

	Duration	Frequency
Absolute measurement	Michael views **2 hours** of "comedy" content per week.	Michael views **4 episodes/movies** of "comedy" content per week.
Relative measurement	Michael spends **50 percent** of his Netflix **time** on "comedy" content.	Of all **episodes/movies** viewed by Michael, **65 percent** contain "comedy" content.

Threshold for counting a "view" was viewing at least **0%** of an episode's/movie's runtime (alternatively: **70%** or **95%**).

Results: (1) Linkage with TMDB Identifier

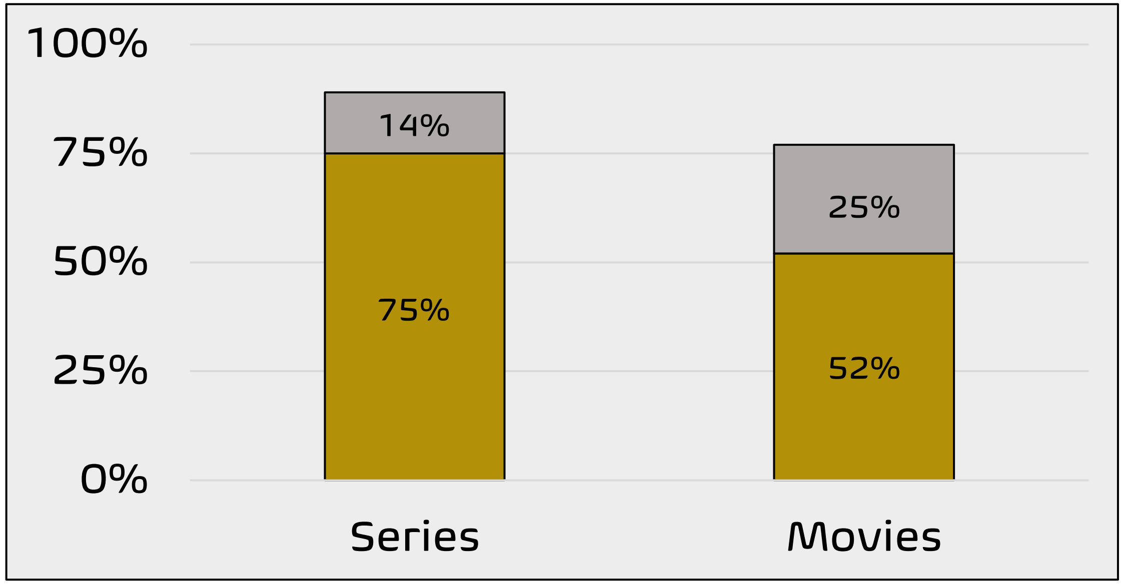
Quality indicator

Share of exact and probability-based matches, considering only unique traces

Calculation

$$\frac{exact\ matches + prob\ matches}{exact\ matches + prob\ matches + non\ matches}$$

Fig 3. Share of exact (gold) and probability-based (silver) matches (in %)



Results: (2) Enrichment with Auxiliary Information

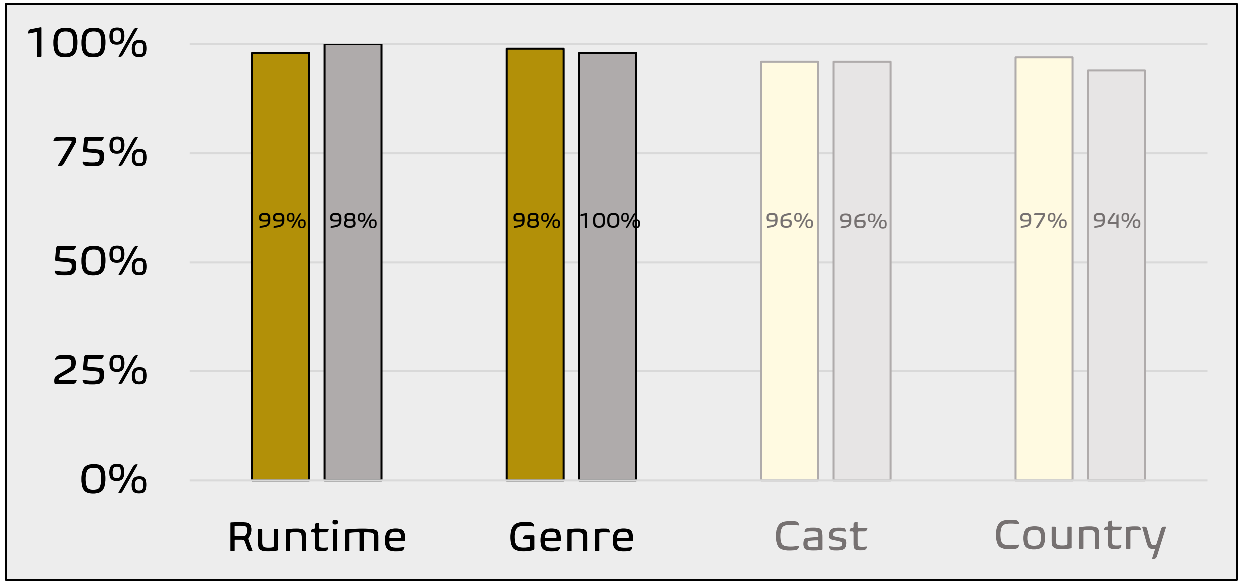
Quality indicator

Share of enriched traces, considering only matched traces

Calculation

$$\frac{enriched\ traces}{enriched\ traces + non\ enriched\ traces}$$

Fig 4. Share of enriched series (gold) and movie traces (silver) (in %)



Results: (3) Transformation into Genre Measurements

Quality indicator

Robustness to variations
in analytical decisions:

<u>0%</u> vs. <u>70%</u> vs. <u>95%</u> threshold
<u>Duration</u> vs. <u>frequency</u>
<u>Absolute</u> vs. <u>relative</u> measurement

Calculation

$$r_{xy} = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

Fig 5. Pairwise correlation coefficients between measurements of "comedy" genre

	abs. dur. (0%)	abs. dur. (70%)	abs. dur. (95%)	abs. freq. (0%)	abs. freq. (70%)	abs. freq. (95%)	rel. dur. (0%)	rel. dur. (70%)	rel. dur. (95%)	rel. freq. (0%)	rel. freq. (70%)	rel. freq. (95%)
abs. dur. (0%)	1.00	1.00	0.98	0.89	0.89	0.89	0.60	0.60	0.59	0.60	0.59	0.58
abs. dur. (70%)	1.00	1.00	0.99	0.88	0.89	0.89	0.61	0.60	0.60	0.60	0.59	0.58
abs. dur. (95%)	0.98	0.99	1.00	0.83	0.85	0.87	0.60	0.59	0.60	0.58	0.57	0.57
abs. freq. (0%)	0.89	0.88	0.83	1.00	0.98	0.96	0.65	0.64	0.62	0.71	0.70	0.68
abs. freq. (70%)	0.89	0.89	0.85	0.98	1.00	0.99	0.65	0.65	0.64	0.71	0.70	0.69
abs. freq. (95%)	0.89	0.89	0.87	0.96	0.99	1.00	0.66	0.66	0.66	0.71	0.70	0.70
rel. dur. (0%)	0.60	0.61	0.60	0.65	0.65	0.66	1.00	0.99	0.97	0.95	0.94	0.92
rel. dur. (70%)	0.60	0.60	0.59	0.64	0.65	0.66	0.99	1.00	0.99	0.94	0.95	0.94
rel. dur. (95%)	0.59	0.60	0.60	0.62	0.64	0.66	0.97	0.99	1.00	0.92	0.93	0.95
rel. freq. (0%)	0.60	0.60	0.58	0.71	0.71	0.71	0.95	0.94	0.92	1.00	0.98	0.96
rel. freq. (70%)	0.59	0.59	0.57	0.70	0.70	0.70	0.94	0.95	0.93	0.98	1.00	0.98
rel. freq. (95%)	0.58	0.58	0.57	0.68	0.69	0.70	0.92	0.94	0.95	0.96	0.98	1.00

Conclusion Case Study 2

- Linking traces to the TMDB identifier represents the biggest error source
  - *Between 10% and 25% of traces could not be linked at all (-> missing data error)*
  - *Probability-based linkages (15% to 25%) may result in erroneous genre attributions*
- Analytical decisions differ in their impact on research outcomes
  - *Using durations, frequencies, and different viewing thresholds results in similar measurements*
  - ***But**: Absolute measurements represent genre exposure (-> media effects research)*
  - ***In contrast**: Relative measurements represent genre preferences (-> research on media choices)*

Take-aways from the Two Case Studies

- Traces obtained via web tracking and data donation have similar characteristics
  - *They only show a small part of the full picture because they lack context*
  - ***Internal context:*** *Content characteristics of the trace (e.g., genre) -> Web/API scraping*
  - ***External context:*** *Real-world situation in which trace was produced -> Survey self-reports*

- Multitude of analytical decisions directly shape research outcomes
  - *Researchers analyzing digital trace data should* ***document*** *all decisions made*
  - *Whenever possible:* ***Robustness checks*** *with alternative combinations of analytical decisions*

Many thanks for your attention!

@jclaass.bsky.social
claassen@dzhw.eu

Literature I

- Avery, L., R., Stanton, A. G., Ward, L. M., Trinh, S. L., Jerald, M. C., & Cole, E. R. (2021). Remixing the script? Associations between black-oriented media consumption and black women's heteropatriarchal romantic relationship beliefs. Journal of Black Psychology, 47(7), 593-625. https://doi.org/10.1177/00957984211021236
- Boeschoten, L., Ausloos, J., Möller, J. E., Araujo, T., & Oberski, D. L. (2022). A framework for privacy preserving digital trace data collection through data donation. Computational Communication Research, 4(2), 388-423. https://doi.org/10.5117/CCR2022.2.002.BOES
- Boeschoten, L., de Schipper, N. C., Mendrik, A. M., Van der Veen, E., Struminskaya, B., Janssen, H., & Araujo, T. (2023). Port: A software tool for digital data donation. Journal of Open Source Software, 8(90), Article 5596. https://doi.org/10.21105/joss.05596
- Bosch, O. J., & Revilla, M. (2022). When survey science met web tracking: Presenting an error framework for metered data. Journal of the Royal Statistical Society (Series A), 185, 408-436. https://doi.org/10.1111/rssa.12956
- Carrière, T. C., Boeschoten, L., Struminskaya, B., Janssen, H. L., de Schipper, N. C., & Araujo, T. (2025). Best practices for studies using digital data donation. Quality & Quantity, 59, 389-412. https://doi.org/10.1007/s11135-024-01983-x
- Chan, L. S., & Liang, Z. (2023). Television continues to cultivate attitudes towards homosexuality, but only among politically conservative individuals and women: Evidence from the U.S. General Social Survey. International Journal of Communication, 17, 6812-6832. https://ijoc.org/index.php/ijoc/article/view/21439
- Charron, J., Gautier, A., & Jestin, C. (2020). Influence of information sources on vaccine hesitancy and practices. Médecine et Maladies Infectieuses, 50, 727-733. https://doi.org/10.1016/j.medmal.2020.01.010
- Flayelle, M., Maurage, P., Di Lorenzo, K. R., Vögele, C., Gainsbury, S. M., & Billieux, J. (2020). Binge-watching: What do we know so far? A first systematic review of the evidence. Current Addiction Reports, 7, 44-60. https://doi.org/10.1007/s40429-020-00299-8

Literature II

- Ilyas, U., Maqsood, R., Ahsan, H., Rauf, A., & Afzal, S. (2023). Binge-watching as behavioral addiction: A systematic review. Applied Psychology Review, 2(2), 39-65. https://doi.org/10.32350/apr.22.03
- Jaks, R., Baumann, I., Juvalta, S., & Dratva, J. (2019). Parental digital health information seeking behavior in Switzerland: A cross-sectional study. BMC Public Health, 19, Article 225. https://doi.org/10.1186/s12889-019-6524-8
- Kaur, M. H., & Ashfaq, D. R. (2023). The impact of Netflix on viewer behaviour and media consumption: An exploration of the effects of streaming services on audience engagement and entertainment preferences. Journal of Media, Culture and Communication, 3(4), 9-23. https://doi.org/10.55529/jmcc.34.9.23
- Keusch, F., Bähr, S., Haas, G. C., Kreuter, F., Trappmann, M., & Eckman, S. (2022). Non-participation in smartphone data collection using research apps. Journal of the Royal Statistical Society (Series A), 185(S2), 225-245. https://doi.org/10.1111/rssa.12827
- Lindberg, A. (2023). Analysis chaining: Conceptual and empirical framing of digital traces. In R. M. Davison (ed.), Handbook of Qualitative Research Methods for Information Systems (pp. 360-375). Edward Elgar Publishing.
- Livingstone, S., Blum-Ross, A., Pavlick, J., & Ólafsson, K. (2018). In the digital home, how do parents support their children and who supports them? Parenting for a Digital Future: Survey Report 1. Department of Media and Communications, The London School of Economics and Political Science, London, UK. http://eprints.lse.ac.uk/id/eprint/87952
- Lotz, A. D., Eklund, O., & Soroka, S. (2022). Netflix, library analysis, and globalization: Rethinking mass media flows. Journal of Communication, 72(4), 511-521. https://doi.org/10.1093/joc/jqac020
- Loudon, K., Buchanan, S., & Ruthven, I. (2016). The everyday life information seeking behaviors of first-time mothers. Journal of Documentation, 72(1), 24-46. https://doi.org/10.1108/JD-06-2014-0080
- Müller, P., & Bach, R. L. (2023). Populist alternative news use and its role for elections: Web-tracking and survey evidence from two campaign periods. News Media & Society, 25(10), 2663-2683. https://doi.org/10.1177/14614448211032969

Literature III

- Palomba, A. (2021). How consumers' personalities, lifestyles and demographics predict SVOD genre and SVOD platform consumption. Journal of Contemporary Marketing Science, 4(3), 362-384. https://doi.org/10.1108/JCMARS-06-2021-0020
- Parry, D. A., Davidson, B. I., Sewall, C. J. R., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. Nature Human Behaviour, 5, 1535-1547. https://doi.org/10.1038/s41562-021-01117-5
- Revilla, M., Couper, M. P., Paura, E., & Ochoa, C. (2021). Willingness to participate in a metered online panel. Field Methods, 33(2), 202-216. https://doi.org/10.1177/1525822X20983986
- Robinson, J. P., & Martin, S. (2010). IT use and declining social capital? More cold water from the General Social Survey (GSS) and the American Time-Use Survey (ATUS). Social Science Computer Review, 28(1), 45-63. https://doi.org/10.1177/0894439309335230
- Scharrer, E., & Blackburn, G. (2018). Cultivating conceptions of masculinity: Television and perceptions of masculine gender role norms. Mass Communication and Society, 21(2), 149-177. https://doi.org/10.1080/15205436.2017.1406118
- Scharrer, E., & Warren, S. (2022). Adolescents' modern media use and beliefs about masculine gender roles and norms. Journalism & Mass Communication Quarterly, 99(1), 289-315. https://doi.org/10.1177/10776990211035453
- Strycharz, J., Meppelink, C., Zarouali, B., Araujo, T., & Voorveld, H. (2024). The blind spot in data donations: Who is (not) willing to donate digital data in social scientific research. Computational Communication Research, 6(2), 1-27. https://doi.org/10.5117/CCR2024.2.3.STRY
- Tana, J., Eirola, E., & Nylund, M. (2019). When is prime-time in streaming media platforms and video-on-demands services? New media consumption patterns and real-time economy. European Journal of Communication, 35(2), 108-125. https://doi.org/10.1177/0267323119894482

Literature IV

- Till, B., Vesely, C., Mairhofer, D., Braun, M., & Niederkrotenthaler, T. (2019). Reports of adolescent psychiatric outpatients on the impact of the TV series “13 Reasons Why”: A qualitative study. Journal of Adolescent Health, 64, 414-416. https://doi.org/10.1016/j.jadohealth.2018.11.021
- Van Es, K. (2024). Exploring Netflix myths: Towards more media industry studies and empirical research in studying video-on-demand. Critical Studies in Television. https://doi.org/10.1177/17496020241297618
- Van Es, K., Nguyen, D., Boeschoten, L., & de Schipper, N. (2025). Exploring VoD services through data donations: A view from Netflix. Media Industries, 12(1). https://doi.org/10.3998/mij.7629