Wikipedia as Big Data for Political Research

Dr. Theresa Gessler

University of Zurich | http://theresagessler.eu | @th_ges

Introduction

Structure

- Wikipedia
- Wikipedia Data and Political Research
- Measurement

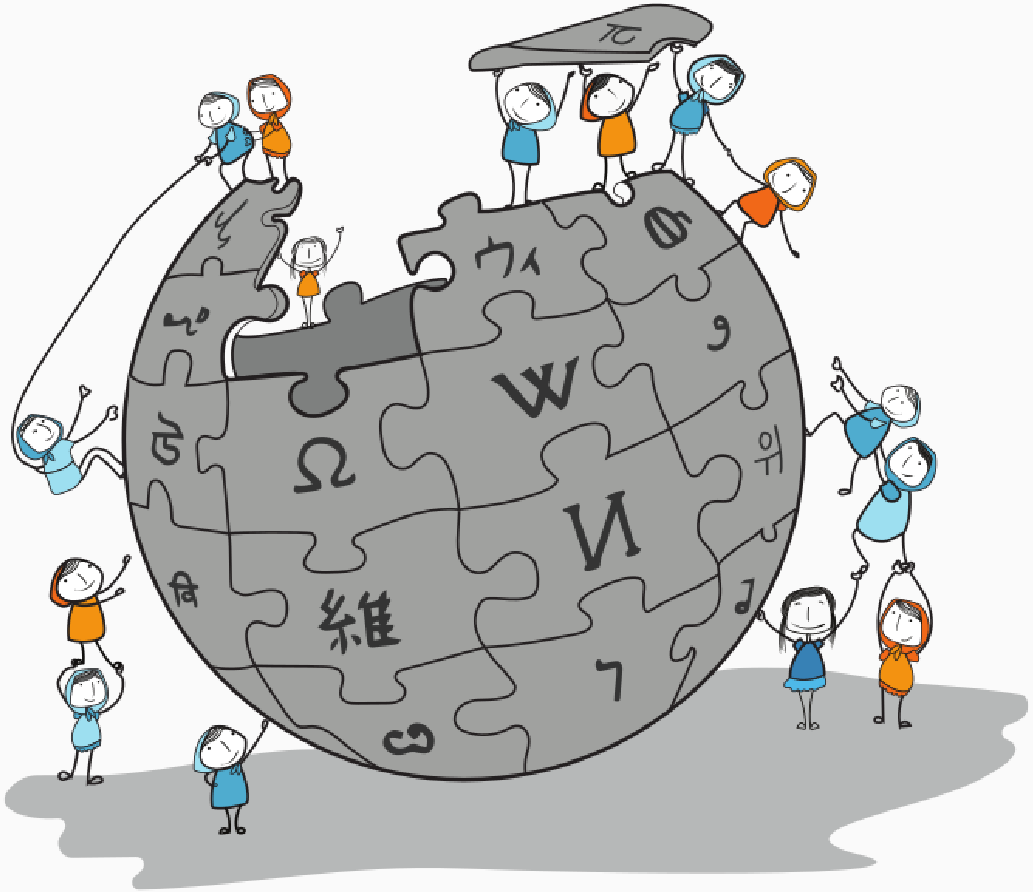
TL;DR

- Wikipedia offers new possibilities of measurement...
  - representation and bias
  - interest and attention
  - framing and stereotypes

- two types of inference, similar to survey research (Groves Fowler, et al., 2009)
  - measurement
  - representation

- Wikipedia reveals differences in portrayal of and interest in female and male politicians

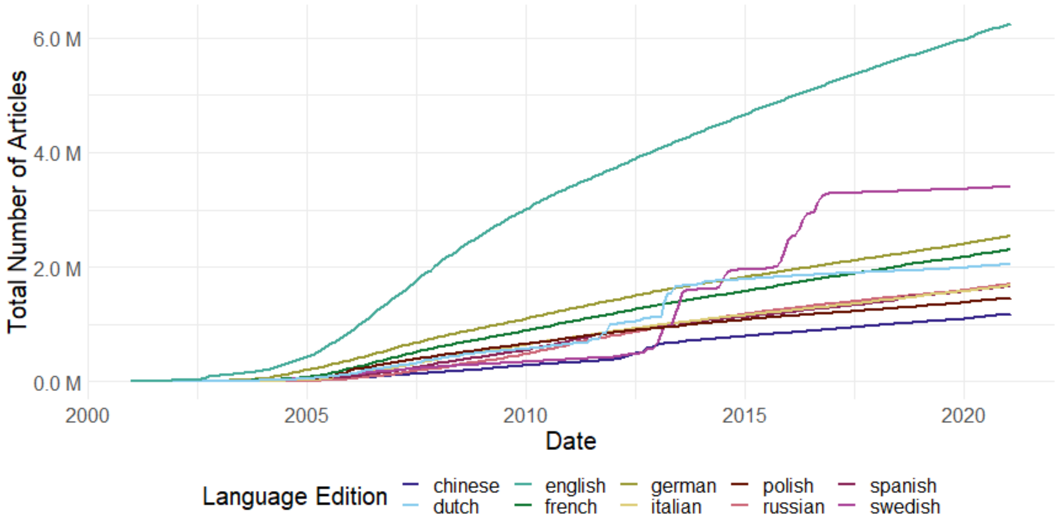
Wikipedia



- large collaborative encyclopedia founded in 2001
    - comprehensive topical coverage
    - equivalent to 30000 volumes printed encyclopedia
- **volunteer project shaped by its editors**
    - innovation
    - bias
- multiple **language versions**, among the most visited websites in many countries
    - used by large segments of population
- readership surveys (Singer Lemmerich, et al., 2017; Lemmerich Saez-Trumper, et al., 2019) show **extrinsic motivations** like media reports or conversations, as well as intrinsic motivations like learning

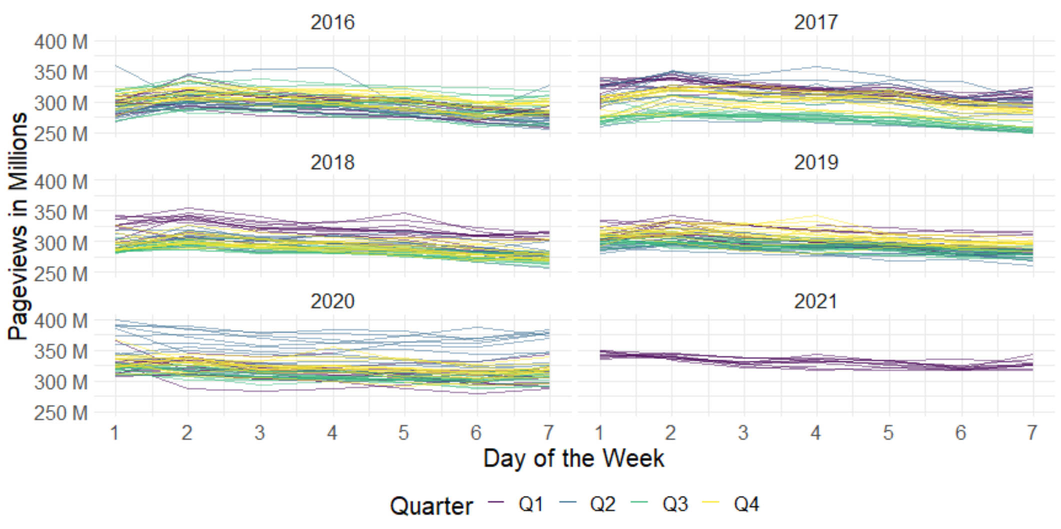
Wikipedia

Articles



- over 300 language editions
- large editions count between 1 and 6 million articles

Meta Data



- ~ 10 billion monthly pageviews for English Wikipedia
- cyclical viewing patterns

Wikipedia Data & Political Research

Wikipedia as a data source

Level	Data	Granularity
Articles	Article text Hyperlinks, Backlinks Page Categories	Full articles (current and historical)
Edits	version-to-version text change Wikipedia contributor Profile of registered contributors; IPs of non-registered contributors Tags associated with edit	individual edit
Page views	Page view data for individual articles	Daily Counts (since 2015)
Navigation	Aggregate dyadic clickstream counts for dyads with views>10	Monthly Aggregate Datasets

Wikipedia as a data source

- **digital trace data** (Howison Wiggins, et al., 2011)
  - found data
  - event-based data
  - longitudinal data
- **digital trace data in political science**
  - social media → widespread measurements of political behavior and information among small group
  - browsing behavior → promising but costly measurement
- **Wikipedia research**
  - often focused on technical aspects (Schroeder and Taylor, 2015)
  - sociological research with focus on editor communities (Okoli Mehdi, et al., 2012), rather than readers

→ Wikipedia data differs significantly from data typically used in the social sciences

Political Science applications

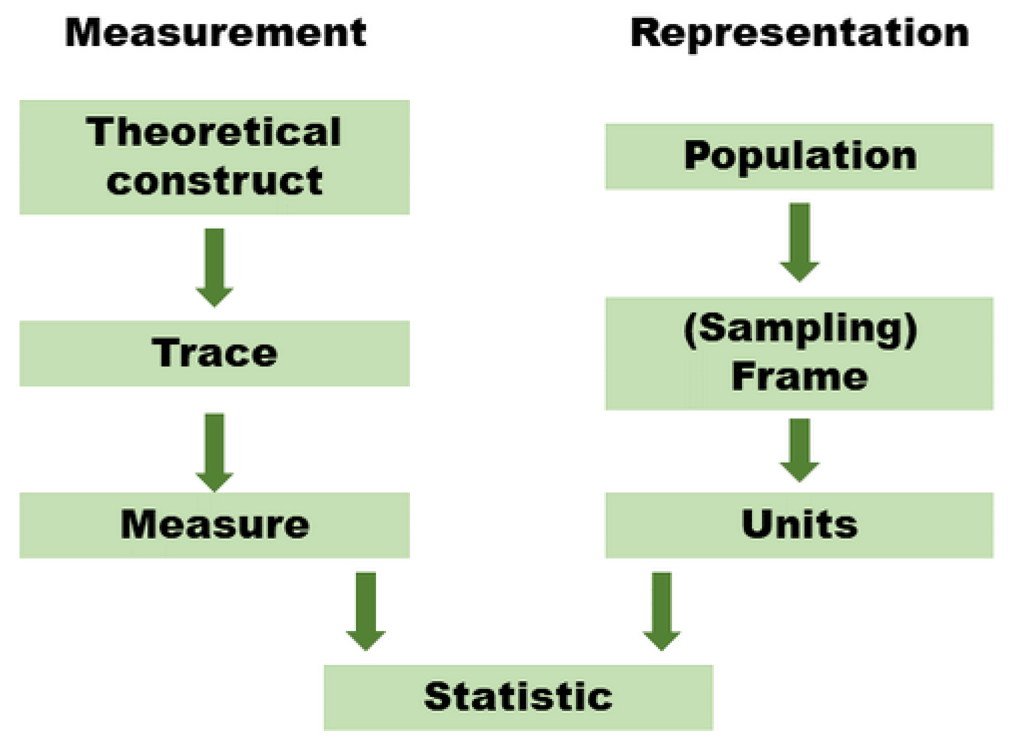
- **articles**
    - factual studies (Brown, 2011; Herrmann and Döring, 2021), also with Wikidata (Göbel and Munzert, 2021)
    - studies on bias in articles (Pradel, 2020; Langrock and Gonzalez-Bailon, 2020)
- **edits**:
    - sociology: community-centered and edit war studies (Neff Laniado, et al., 2013; Yasseri Sumi, et al., 2012; Shi Teplitskiy, et al., 2019)
    - strategic editing (Göbel and Munzert, 2018)
- **page views** as indicator for
    - interest (Atkinson and DeWitt, 2019; Margolin Goodman, et al., 2016)
    - vote choice and support (Yasseri and Bright, 2016; Salem and Stephany, 2021; Smith and Gustafson, 2017)
    - information-seeking and exposure (Pan and Roberts, 2020; Hobbs and Roberts, 2018)
- **navigation data**
    - mostly in computer science studies on navigation (Dimitrov Lemmerich, et al., 2018)
    - knowledge bubbles (Menghini Anagnostopoulos, et al., 2019)

Wikipedia Data and Measurement

Wikipedia Data and Measurement

Two types of inferences following 'Total Survey Error' Framework (Groves Fowler, et al., 2009; Sen Flöck, et al., 2021; Amaya Biemer, et al., 2020)

- **measurement**: generating statistics that reflect data
- **representation**: generalizing to a population of interest



Wikipedia Data and Measurement

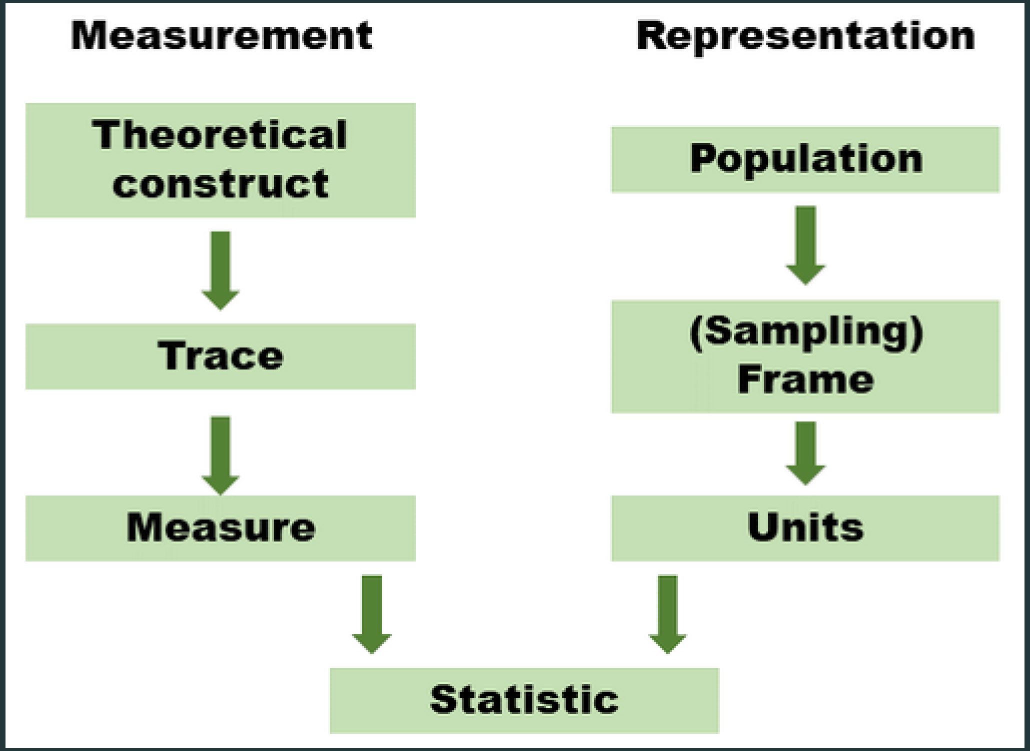
Example: Gender bias in perceptions of politicians

- gender bias in media reporting about politics
    - quantitative and qualitative biases
    - online spaces as opportunities and challenges
- systematic and structured data on legislators (see also Göbel and Munzert, 2021)

→ **How does the online representation of female and male politicians differ?**

→ **How do Wikipedia readers use the articles on female and male politicians?**

Measurement



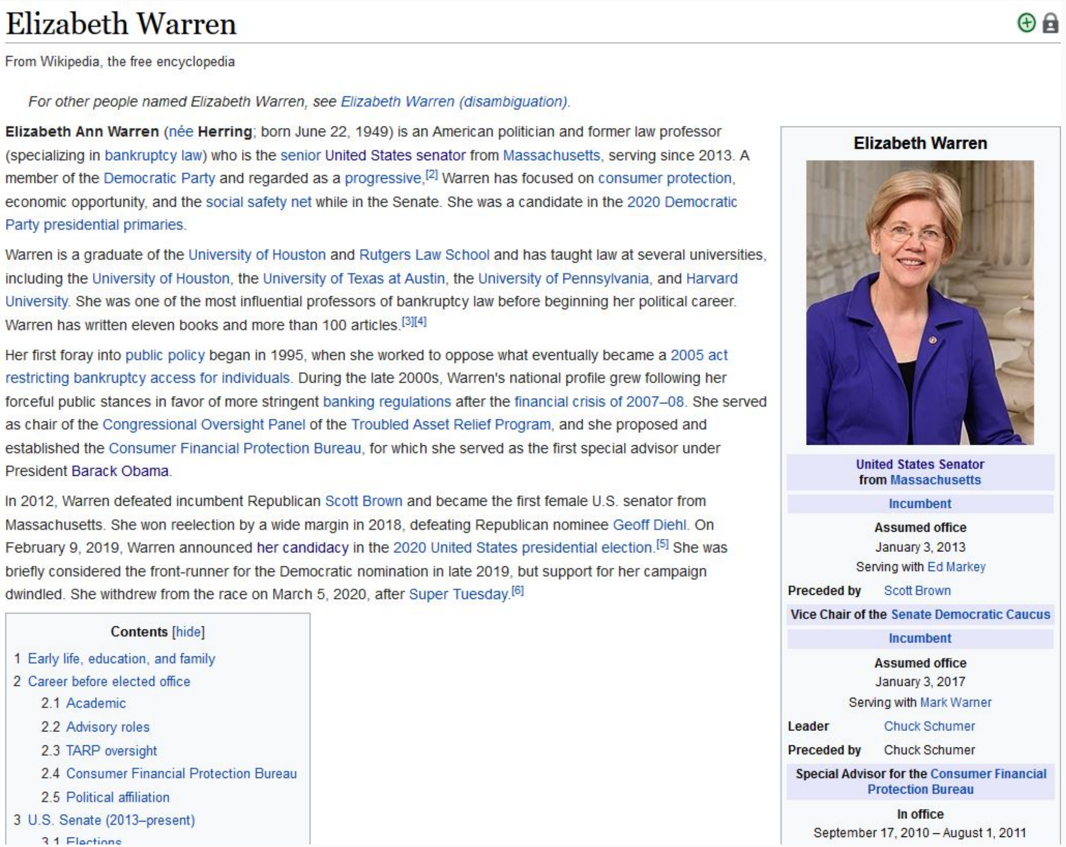
Measurement

Theoretical constructs & their measurement

- **representation and bias**
  - how are topics and people represented?
  - which biases exist in these representations?
- **interest and attention**
  - what are users interested in?
- **framing and stereotypes**
  - what shapes usage patterns?

Measurement

Theoretical constructs & their measurement



Elizabeth Warren

From Wikipedia, the free encyclopedia

For other people named Elizabeth Warren, see Elizabeth Warren (disambiguation).

Elizabeth Ann Warren (née **Herring**; born June 22, 1949) is an American politician and former law professor (specializing in bankruptcy law) who is the senior United States senator from Massachusetts, serving since 2013. A member of the Democratic Party and regarded as a progressive,[2] Warren has focused on consumer protection, economic opportunity, and the social safety net while in the Senate. She was a candidate in the 2020 Democratic Party presidential primaries.

Warren is a graduate of the University of Houston and Rutgers Law School and has taught law at several universities, including the University of Houston, the University of Texas at Austin, the University of Pennsylvania, and Harvard University. She was one of the most influential professors of bankruptcy law before beginning her political career. Warren has written eleven books and more than 100 articles.[3][4]

Her first foray into public policy began in 1995, when she worked to oppose what eventually became a 2005 act restricting bankruptcy access for individuals. During the late 2000s, Warren's national profile grew following her forceful public stances in favor of more stringent banking regulations after the financial crisis of 2007–08. She served as chair of the Congressional Oversight Panel of the Troubled Asset Relief Program, and she proposed and established the Consumer Financial Protection Bureau, for which she served as the first special advisor under President Barack Obama.

In 2012, Warren defeated incumbent Republican Scott Brown and became the first female U.S. senator from Massachusetts. She won reelection by a wide margin in 2018, defeating Republican nominee Geoff Diehl. On February 9, 2019, Warren announced her candidacy in the 2020 United States presidential election.[5] She was briefly considered the front-runner for the Democratic nomination in late 2019, but support for her campaign dwindled. She withdrew from the race on March 5, 2020, after Super Tuesday.[6]

Contents [hide]

Elizabeth Warren

United States Senator from Massachusetts

Incumbent

Assumed office
January 3, 2013
Serving with Ed Markey

Preceded by Scott Brown

Vice Chair of the Senate Democratic Caucus

Incumbent

Assumed office
January 3, 2017
Serving with Mark Warner

Leader Chuck Schumer

Preceded by Chuck Schumer

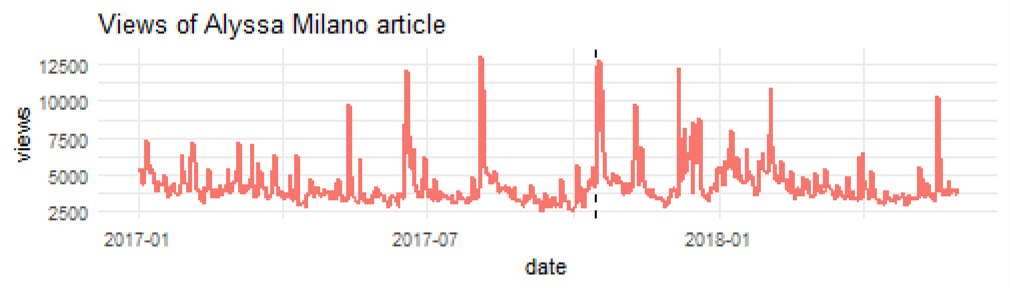
Special Advisor for the Consumer Financial Protection Bureau

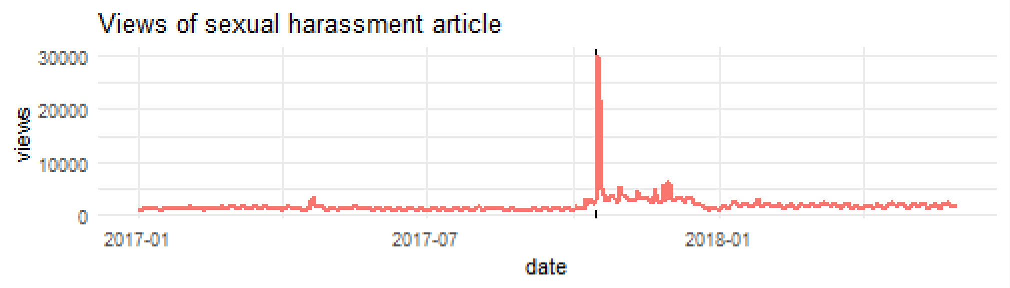
In office
September 17, 2010 – August 1, 2011

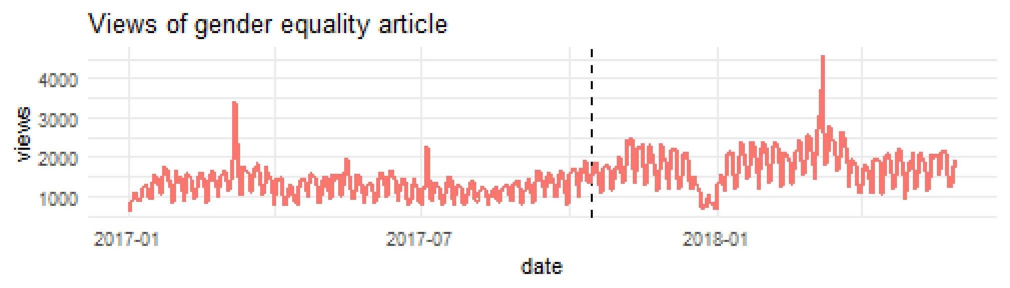
- conceptualizing **representation and bias**
  - knowledge about person / issue
  - bias in representation
- **traces**: article content (qualitative, quantitative), article embedding in link networks
- **measurement**: text-as-data and network methods, considering unique properties of Wikipedia article format
  - biographical focus
  - selective mentions and omissions

Measurement

Theoretical constructs & their measurement


Views of Alyssa Milano article


Views of sexual harassment article


Views of gender equality article

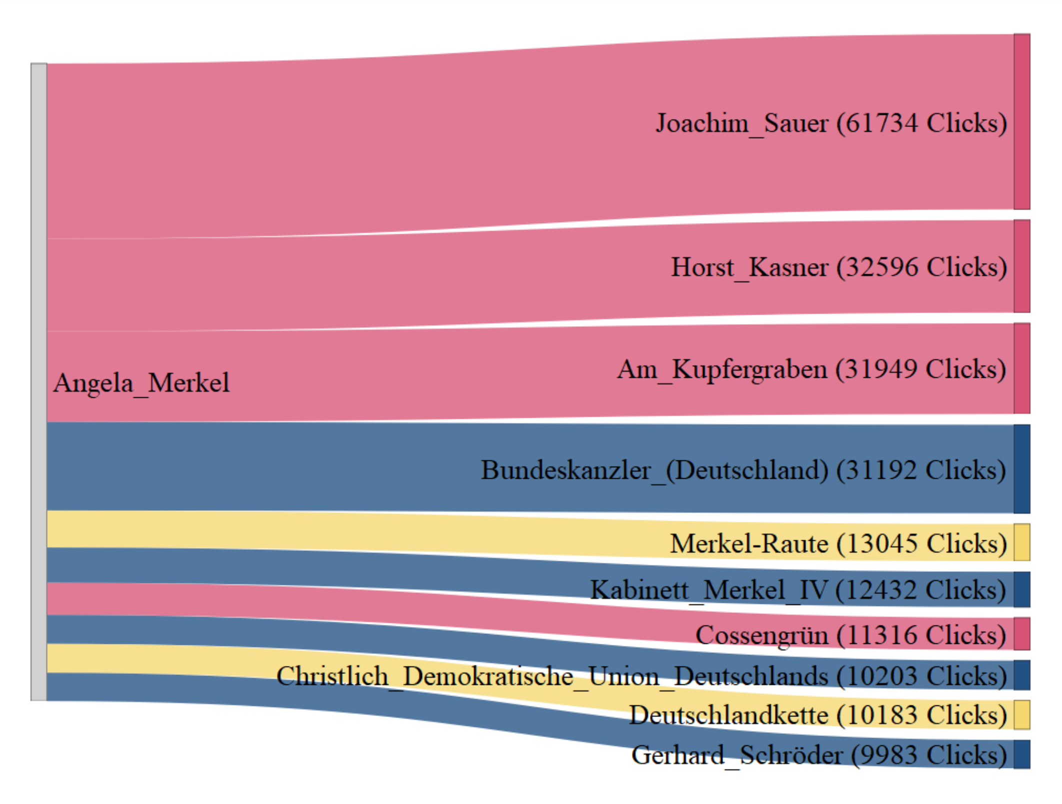
- conceptualizing **attention and interest**
  - attention to events, issues, people
  - event-based and long-term

→ **conceptual clarification** - e.g. attention vs. support (Yasseri and Bright, 2016; Salem and Stephany, 2021; Smith and Gustafson, 2017)

- **traces**: selection of scope (articles)
  - selection of articles
  - theoretical or empirical combination
- **measurement**
  - combining multiple time series
  - disentangle aspects of a concept

Measurement

Theoretical constructs & their measurement



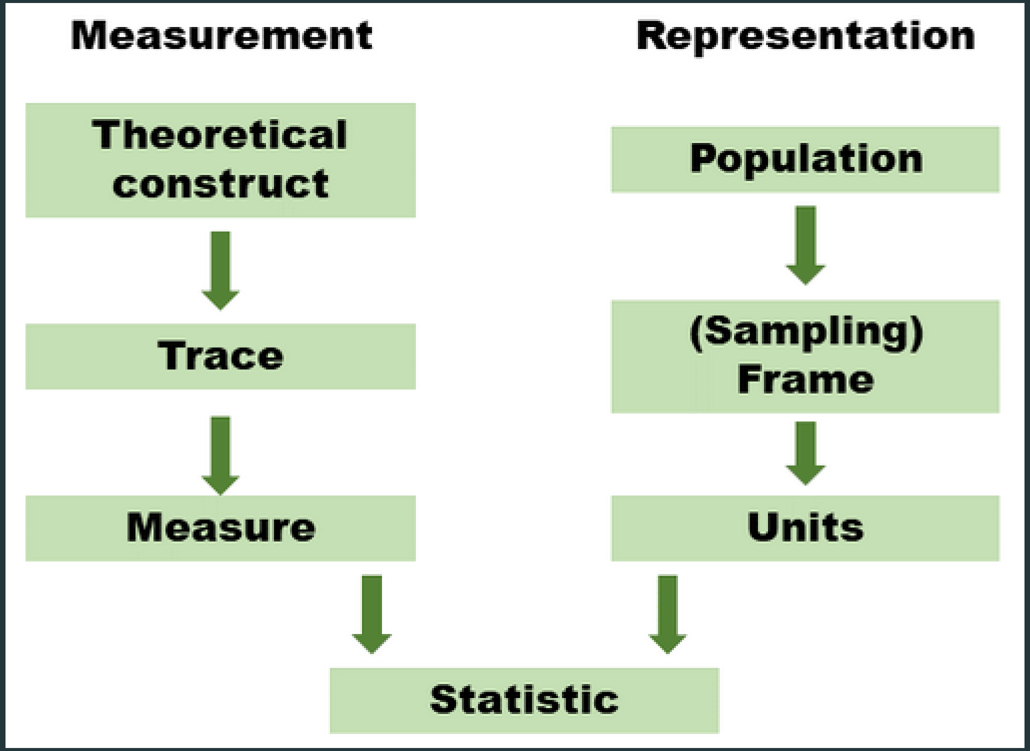
- conceptualizing **framing and stereotypes**
  - how framing shapes link-following
  - how stereotypes shape navigation
  - → Focus on one source of variation
- **traces**: parse and classify clickstreams
- **measurement**: calculate property of interest
  - e.g. link clicks per page view

Measurement

Traces

	Representation and Bias	**Attention and Interest**	**Framing and Stereotypes**
Mechanism	Wikipedia content as reflection	Event → Wikipedia outcome	Wikipedia framing / stereotypes → Wikipedia outcome
Typical Outcome	Page content, edits	Page views, edits	content, link clicks
Processing	text as data methods that account for platform affordances	indicator building and accounting for cyclical effects	calculation of measures such as clicks-per-view

Representation



Representation

Population

- **Wikipedia as a context**
  - populations of users (e.g. readers, editors, ...)
  - populations of subjects (e.g. politicians, political activists, historical figures)

→ outline the relevance of this context

- **Wikipedia as a sensor**
  - large user base but little demographic information
  - Wikipedia user surveys (Singer Lemmerich, et al., 2017; Lemmerich Saez-Trumper, et al., 2019)

→ discuss representativity of Wikipedia

→ **Wikipedia content and edits represent Wikipedia editors**

→ **Wikipedia content and edits represent article subjects**

→ **Wikipedia page views and clickstreams may represent the wider online population**

...all conditional on a set of pages

Representation

(Sampling) Frame

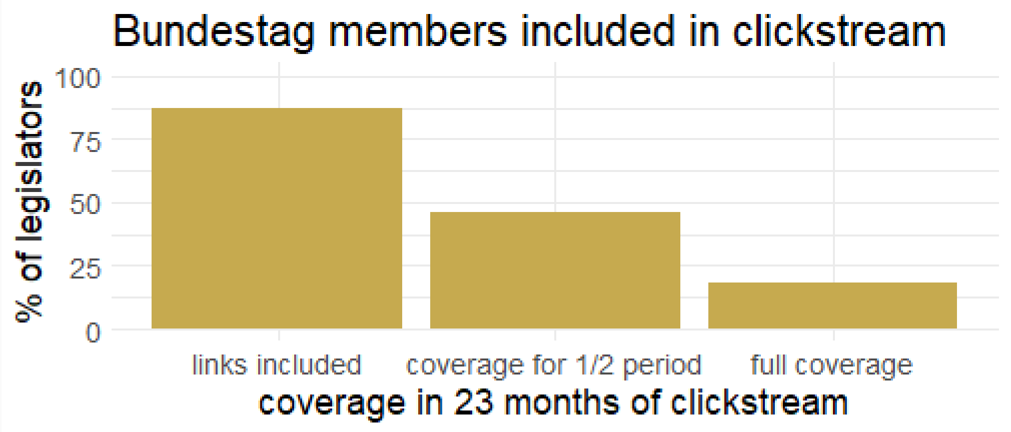
- **for well-defined populations: measuring coverage**
    - variation in article availability
- **quantification of coverage for weakly defined populations**
    - inductive sampling strategies
        - Wikipedia categories
        - linked pages
        - textual similarity
    - deductive sampling strategies
        - keywords
        - operationalizations

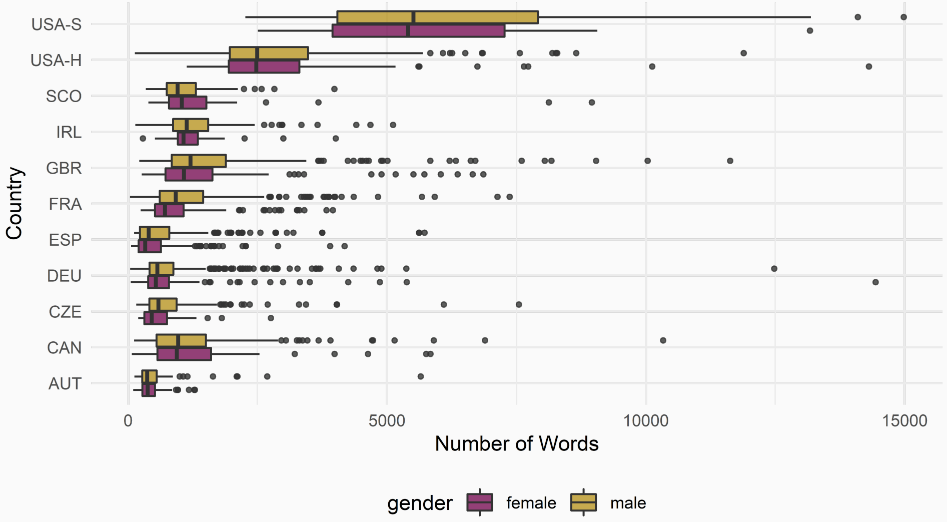
→ *only deductive strategies allow a thorough assessment of coverage*

Representation

Units

- **assess ability to generate measurements for each unit**
    - articles: length
    - page views & edits: history of page
    - clickstreams: page views as ceiling
- explain biases by internal and external metrics
    - e.g. demographic features, google search frequencies, ...





Conclusion

Measurement & Representation

- **core concepts**
    - representation and bias
    - interest and attention
    - framing and stereotypes
- **two types of inference**
    - measurement: generating statistics that reflect data
    - representation: generalizing to a population of interest
- **comprehensive perspective on subject matter**
    - content, content creation, content consumption

Potential for future research

- **Wikipedia as one type of evidence in applied research**
    - first exploratory evidence
    - generalization from experimental data
    - combination with individual-level browsing history data
- **Wikipedia as a platform to generate data**
    - randomized control trials (Thompson and Hanley, 2018; Hinnosaar Hinnosaar, et al., 2021)
- **use of platform content in researcher-run experiments**
    - high quality and adaptable content as stimulus
- **focus on underexplored data types**
    - edits, clickstreams
    - other datasources from the Wikipedia family (Wikidata, DBpedia)

Conclusion

- Wikipedia has **limitations**
    - data as a by-product of processes of interest
    - unknown user base / lack of demographic information
    - aggregate data
    - changing features

- Wikipedia provides **opportunities**
    - **advantages of digital trace data**
        - always-on
        - non-reactive nature
        - time series data at massive scale
        - global reach
    - **unique advantages of Wikipedia**
        - encyclopedic approach
        - relevance for political behavior
        - comparison groups and structure
        - comprehensive data sharing

Thanks for your attention!

gessler@ipz.uzh.ch | @th_ges

Literature

Amaya, A., P. P. Biemer, et al. (2020). "Total Error in a Big Data World: Adapting the TSE Framework to Big Data". In: *Journal of Survey Statistics and Methodology* 8.1, pp. 89-119. ISSN: 2325-0984. DOI: 10.1093/jssam/smz056.

Atkinson, M. D. and D. DeWitt (2019). "Does Celebrity Issue Advocacy Mobilize Issue Publics?" In: *Political Studies* 67.1, pp. 83-99. ISSN: 0032-3217. DOI: 10.1177/0032321717751294.

Brown, A. R. (2011). "Wikipedia as a Data Source for Political Scientists: Accuracy and Completeness of Coverage". In: *PS: Political Science & Politics* 44.2, pp. 339-343. ISSN: 1537-5935, 1049-0965. DOI: 10.1017/S1049096511000199.

Dimitrov, D., F. Lemmerich, et al. (2018). "Query for Architecture, Click through Military: Comparing the Roles of Search and Navigation on Wikipedia". In: *Proceedings of the 10th ACM Conference on Web Science*. WebSci '18. New York, NY, USA: Association for Computing Machinery, pp. 371-380. ISBN: 978-1-4503-5563-6. DOI: 10.1145/3201064.3201092.

Göbel, S. and S. Munzert (2018). "Political Advertising on the Wikipedia Marketplace of Information". In: *Social Science Computer Review* 36.2, pp. 157-175. ISSN: 0894-4393. DOI: 10.1177/0894439317703579.

Göbel, S. and S. Munzert (2021). "The Comparative Legislators Database". In: *British Journal of Political Science* FirstView, pp. 1-11. ISSN: 0007-1234, 1469-2112. DOI: 10.1017/S0007123420000897.

Literature (cont.)

Groves, R. M., F. J. Fowler, et al., ed. (2009). *Survey Methodology*. 2nd ed. Wiley Series in Survey Methodology. Hoboken, N.J: Wiley. ISBN: 978-0-470-46546-2.

Herrmann, M. and H. Döring (2021). "Party Positions from Wikipedia Classifications of Party Ideology". In: *Political Analysis*. DOI: 10.31235/osf.io/5fg8n.

Hinnosaar, M., T. Hinnosaar, et al. (2021). "Wikipedia Matters". In: *Journal of Economics & Management Strategy* FirstView. ISSN: 1556-5068. DOI: 10.1111/jems.12421.

Hobbs, W. R. and M. E. Roberts (2018). "How Sudden Censorship Can Increase Access to Information". In: *American Political Science Review* 112.3, pp. 621-636. ISSN: 0003-0554, 1537-5943. DOI: 10.1017/S0003055418000084.

Howison, J., A. Wiggins, et al. (2011). "Validity Issues in the Use of Social Network Analysis with Digital Trace Data". In: *Journal of the Association for Information Systems* 12.12, pp. 767-797. ISSN:

1. DOI: 10.17705/1jais.00282.

Langrock, I. and S. Gonzalez-Bailon (2020). *The Gender Divide in Wikipedia: Quantifying and Assessing the Impact of Two Feminist Interventions*. SSRN Scholarly Paper ID 3739176. Rochester, NY: Social Science Research Network. DOI: 10.2139/ssrn.3739176.

Literature (cont.)

Lemmerich, F., D. Saez-Trumper, et al. (2019). "Why the World Reads Wikipedia: Beyond English Speakers". In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. WSDM '19. Melbourne VIC, Australia: Association for Computing Machinery, pp. 618-626. ISBN: 978-1-4503-5940-5. DOI: 10.1145/3289600.3291021.

Margolin, D. B., S. Goodman, et al. (2016). "Wiki-Worthy: Collective Judgment of Candidate Notability". In: *Information, Communication & Society* 19.8, pp. 1029-1045. ISSN: 1369-118X. DOI: 10.1080/1369118X.2015.1069871.

Menghini, C., A. Anagnostopoulos, et al. (2019). "Wikipedia Polarization and Its Effects on Navigation Paths". In: *2019 IEEE International Conference on Big Data (Big Data)*. , pp. 6154-6156. DOI: 10.1109/BigData47090.2019.9005566.

Neff, J. J., D. Laniado, et al. (2013). "Jointly They Edit: Examining the Impact of Community Identification on Political Interaction in Wikipedia". In: *PLOS ONE* 8.4, p. e60584. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0060584.

Okoli, C., M. Mehdi, et al. (2012). *The People's Encyclopedia Under the Gaze of the Sages: A Systematic Review of Scholarly Research on Wikipedia*. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network. DOI: 10.2139/ssrn.2021326.

Pan, J. and M. E. Roberts (2020). "Censorship's Effect on Incidental Exposure to Information: Evidence From Wikipedia". In: *SAGE Open* 10.1. ISSN: 2158-2440, 2158-2440. DOI: 10.1177/2158244019894068.

Literature (cont.)

Pradel, F. (2020). "Biased Representation of Politicians in Google and Wikipedia Search? The Joint Effect of Party Identity, Gender Identity and Elections". In: *Political Communication* 0.0, pp. 1-32. DOI: 10.1080/10584609.2020.1793846.

Salem, H. and F. Stephany (2021). "Wikipedia: A Challenger's Best Friend? Utilizing Information-Seeking Behaviour Patterns to Predict US Congressional Elections". In: *Information, Communication & Society* 0.0, pp. 1-27. DOI: 10.1080/1369118X.2021.1942953.

Schroeder, R. and L. Taylor (2015). "Big Data and Wikipedia Research: Social Science Knowledge across Disciplinary Divides". In: *Information, Communication & Society* 18.9, pp. 1039-1056. DOI: 10.1080/1369118X.2015.1008538.

Sen, I., F. Flöck, et al. (2021). "A Total Error Framework for Digital Traces of Human Behavior on Online Platforms". In: *Public Opinion Quarterly* 85.S1, pp. 399-422. ISSN: 0033-362X. DOI: 10.1093/poq/nfab018.

Shi, F., M. Teplitskiy, et al. (2019). "The Wisdom of Polarized Crowds". In: *Nature Human Behaviour* 3.4, pp. 329-336. ISSN: 2397-3374. DOI: 10.1038/s41562-019-0541-6.

Singer, P., F. Lemmerich, et al. (2017). "Why We Read Wikipedia". In: *Proceedings of the 26th International Conference on World Wide Web - WWW '17*, pp. 1591-1600. DOI: 10.1145/3038912.3052716. arXiv: 1702.05379.

Literature (cont.)

Smith, B. K. and A. Gustafson (2017). "Using Wikipedia to Predict Election Outcomes: Online Behavior as a Predictor of Voting". In: *Public Opinion Quarterly* 81.3, pp. 714-735. DOI: 10.1093/poq/nfx007.

Thompson, N. and D. Hanley (2018). *Science Is Shaped by Wikipedia: Evidence From a Randomized Control Trial*. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network. DOI: 10.2139/ssrn.3039505.

Yasseri, T. and J. Bright (2016). "Wikipedia Traffic Data and Electoral Prediction: Towards Theoretically Informed Models". In: *EPJ Data Science* 5.1, pp. 1-15. ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-016-0083-3.

Yasseri, T., R. Sumi, et al. (2012). "Dynamics of Conflicts in Wikipedia". In: *PLOS ONE* 7.6. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0038869.