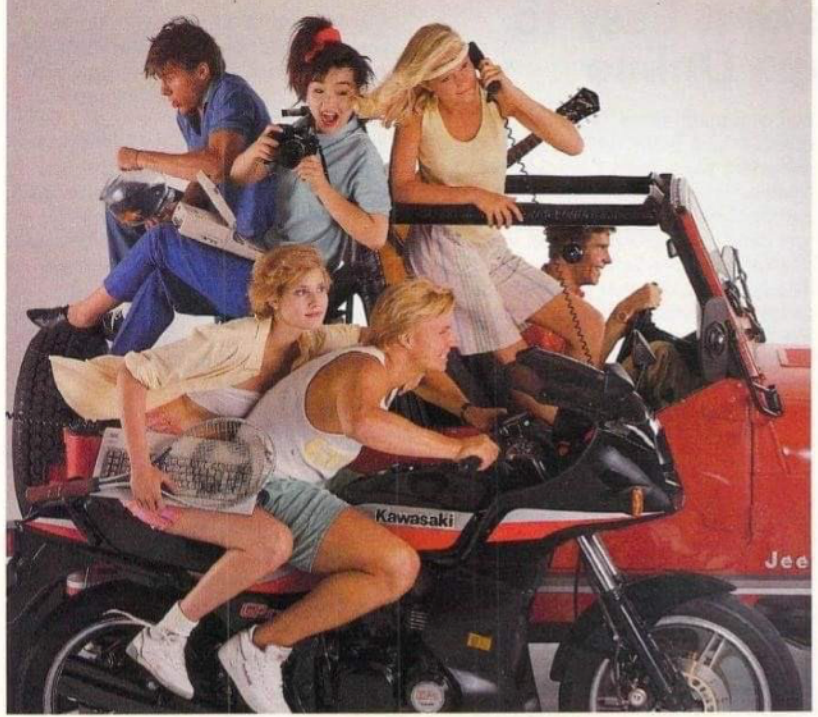
UNIVERSITÄT LEIPZIG

Forschungsseminar CSS – Digital Trace Data

NSR 423, 22.10.2024

Felix Lennert, M.Sc.



TO BE YOUNG AND ONLINE

~~Youth~~ Explore Databases, Find Friends ~~in Forums~~

CLARIFICATION – WEEKLY CHECK-IN

These emails can have any format

Purpose: giving me a “feel” for you students and what works and doesn’t work and what to expect

⟹ Some examples

CLARIFICATION – WEEKLY CHECK-IN

Hi Felix,

I found last week very inspirational. It really got me thinking about my final project. For now my plan is to acquire data from this platform called x.com (it's not porn despite the name). My goal is to study the diffusion of ideas, perhaps combining it with a little ABM and maybe some geospatial analyses. How does that sound?

Kind regards,
Xilef Trennel

Hi Xilef,

Great, thanks for your feedback :)

x.com has been used quite a lot for research. You can find some articles attached. It is a bit finicky to scrape though, since this billionaire has taken over. But where there's a headless browser there is a way, as we nerds say hehe.

Anyways, find attached some papers you might deem interesting. Please keep me posted.

Best,
Felix

> On 21 Oct 2024, at 14:20, LENNERT Felix <Felix.LENNERT@ensae.fr> wrote:
>
> Hi Felix,
>
> I found last week very inspirational. It really got me thinking about my final project. For now my plan is to acquire data from this platform called x.com (it's not porn despite the name). My goal is to study the diffusion of ideas, perhaps combining it with a little ABM and maybe some geospatial analyses. How does that sound?
>
> Kind regards,
> Xilef Trennel

CLARIFICATION – WEEKLY CHECK-IN

Hi Felix,

I got hung up on reading some papers last week and really went down a deep rabbit hole. They deal with how people have measured the bundling of lifestyles around political camps.

Thought that this could be right up your alley, and maybe they would be cool additional readings for the session on ABMs. You can find them attached.

Best,
Lexif Nenlert

Hi Lexif,

Wow, they look amazing. Thank you very much for sending this! Give me a couple of days to read them and maybe we can discuss them later this week? But don't worry if not!

Excited to hear your thoughts!

Best,
Felix

> On 21 Oct 2024, at 14:23, LENNERT Felix <Felix.LENNERT@ensae.fr> wrote:
>
> Hi Felix,
>
> I got hung up on reading some papers last week and really went down a deep rabbit hole. They deal with how people have measured the bundling of lifestyles around political camps.
>
> Thought that this could be right up your alley, and maybe they would be cool additional readings for the session on ABMs. You can find them attached.
>
> Best,
> Lexif Nenlert

CLARIFICATION – WEEKLY CHECK-IN

Hi Felix,

Last week I was sort of overwhelmed with the material. Do you think you could give me some pointers on where I could find some additional material?
Also maybe if you have 30 minutes to spare on perhaps Wednesday, I would like to look over some code with you.

Thank you!
Lefix Nellert

Hi Lefix,

This is completely understandable. If you have gone through the material provided in both syllabus and R script, I can recommend these additional resources:

- YouTube video on topic A
- Blog entry on topic A
- Hands-on tutorial on topic B
- YouTube video on topic C

I will be free all Wednesday, just suggest a time, and we can have a look.

Best,
Felix

> On 21 Oct 2024, at 14:14, LENNERT Felix <Felix.LENNERT@ensae.fr> wrote:
>
> Hi Felix,
>
> Last week I was sort of overwhelmed with the material. Do you think you could give me some pointers on where I could find some additional material?
> Also maybe if you have 30 minutes to spare on perhaps Wednesday, I would like to look over some code with you.
>
> Thank you!
> Lefix Nellert

CLARIFICATION – WEEKLY CHECK

Hi,

Just wanted to do a quick check-in:

Last week my horse got sick and the doctor gave me some Ketamine.
The horse is still sick but I had a wonderful week. No work has been done though.

Kind regards,
Lelix Fennert

Dear Lelix,

This is very unfortunate. I recommend electrolytes and perhaps some cuddles for the horse.

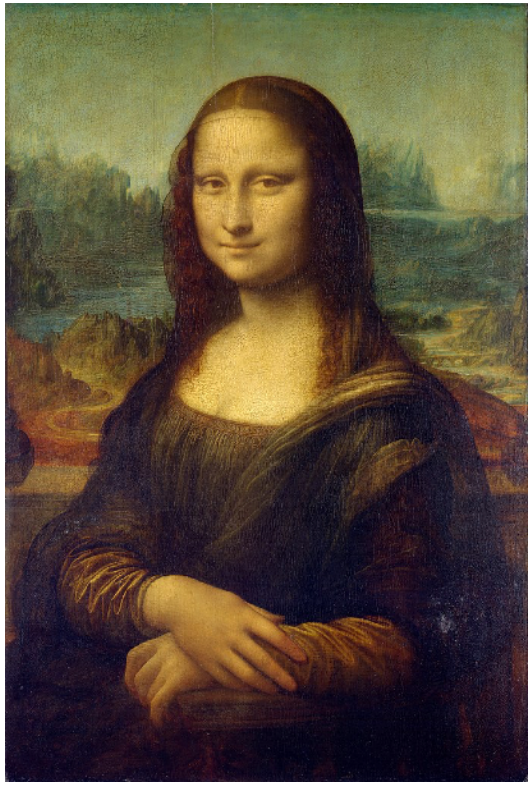
Peaceful wishes,
Felix

> On 21 Oct 2024, at 14:12, LENNERT Felix <Felix.LENNERT@ensae.fr> wrote:
>
> Hi,
>
> Just wanted to do a quick check-in:
>
> Last week my horse got sick and the doctor gave me some Ketamine.
> The horse is still sick but I had a wonderful week. No work has been done though.
>
> Kind regards,
> Lelix Fennert

OUTLINE

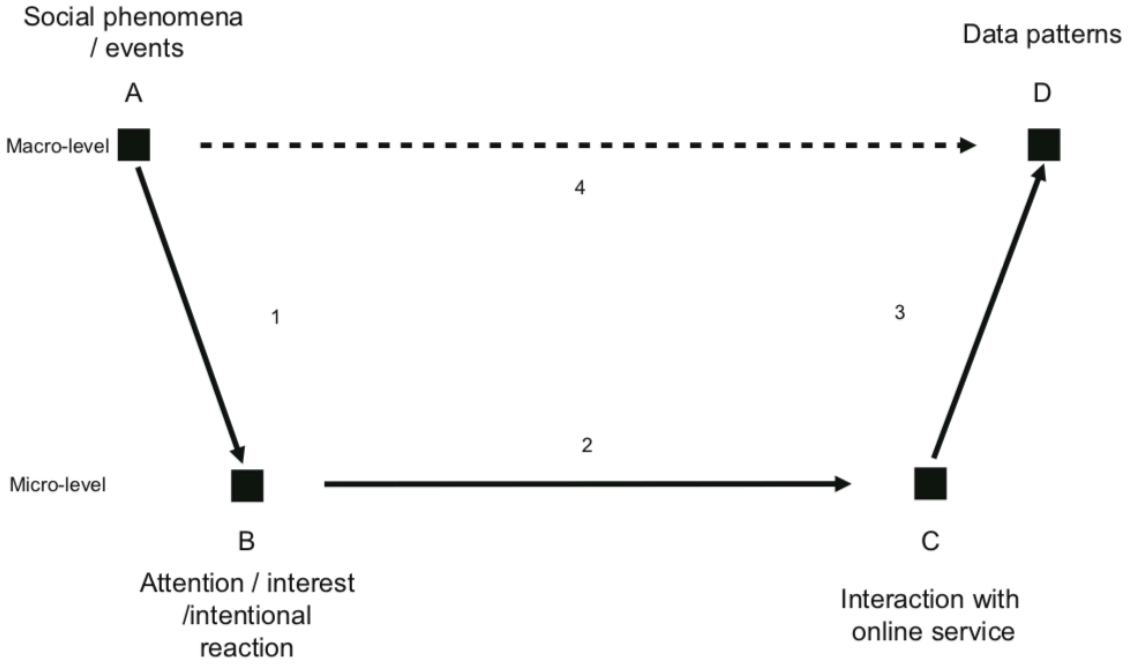
- defining “digital trace data”
- characteristics of “big data” (including some “infotainment”)
  - Garg et al. 2018
  - Phan & Airoldi 2015
- conclusion: what can we do?
- R – string manipulation and regular expressions

MoMA / Marcel Duchamp



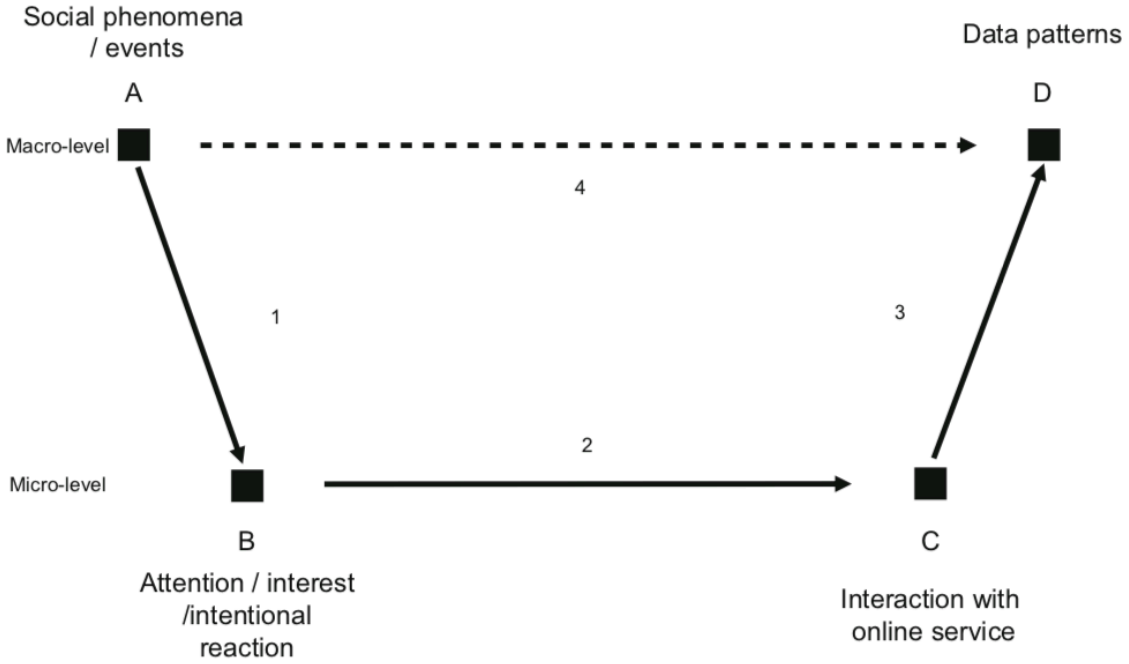
Wikimedia / Leonardo Da Vinci

ANALOG	DIGITAL
"Over the past century, there has been no shortage of social theory, but there have been severe constraints on access to data. The reason is simple: Social life is very hard to observe." (Golder/Macy 2014: 130)	"Now, in the digital age, the behaviors of billions of people are recorded, stored, and analyzable." (Salganik 2017: 13)
– physical traces – "self-reported" data → Interviews, questionnaires, participant observation, etc. – lab experiments	– human behavior is constantly recorded – digital devices as sensors of human behavior
→ data are generated in order to answer a priorly defined question → **"designed/custom-made data"**	→ data are inadvertently generated (or at least not with social-science research goals in mind → **"found/ready-made data"**



Jungherr 2015: 38

traces of collective human behavior



Jungherr 2015: 38

International Men's Day
Holiday

+ Compare

Germany | Past 12 months | All categories | Web Search

Interest over time

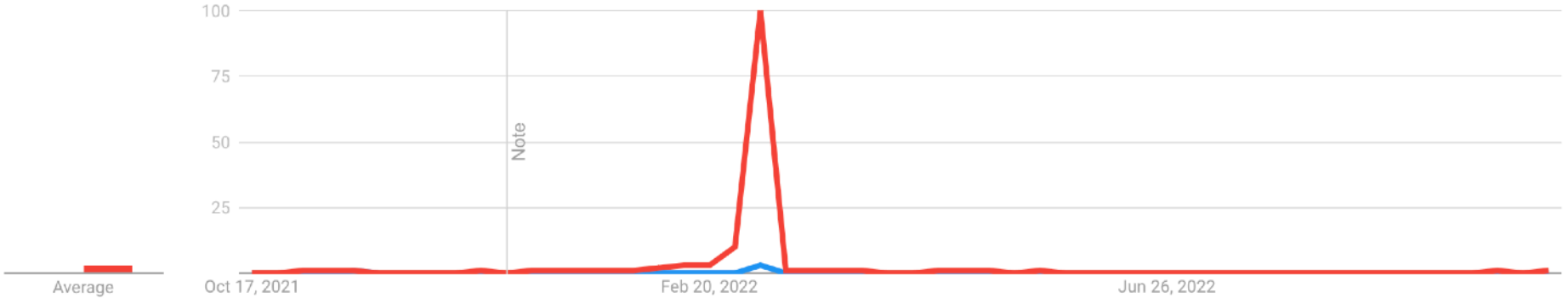
International Men's Day
Holiday

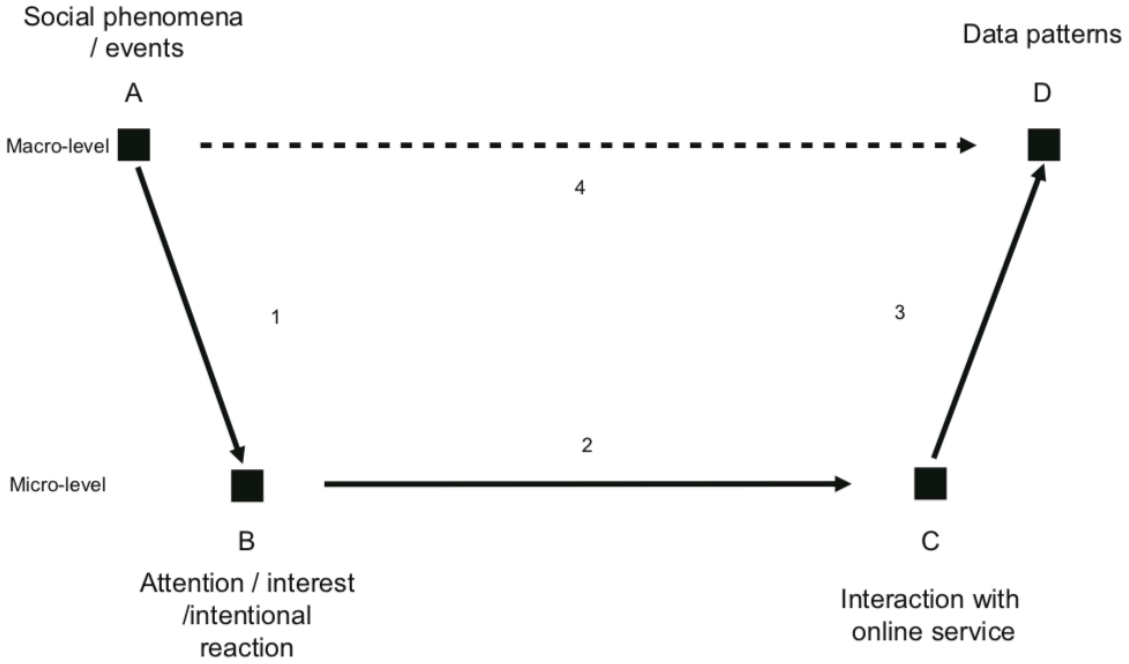
International Women's Day
Holiday

+ Add comparison

Germany | Past 12 months | All categories | Web Search

Interest over time





Jungherr 2015: 38

Salganik (2017) defines 10 characteristics of *Big Data*:

- Big
- Always-on
- Nonreactive
- Incomplete
- Inaccessible
- Nonrepresentative
- Drifting
- Algorithmically confounded
- Dirty
- Sensitive

BIG

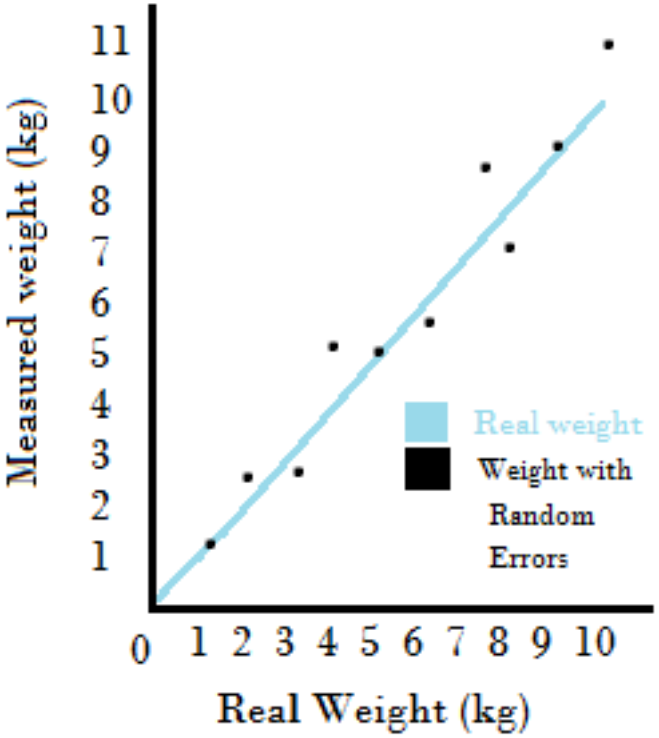
„Large datasets are a means to an end; they are not an end in themselves.” (Salganik 2017: 17)

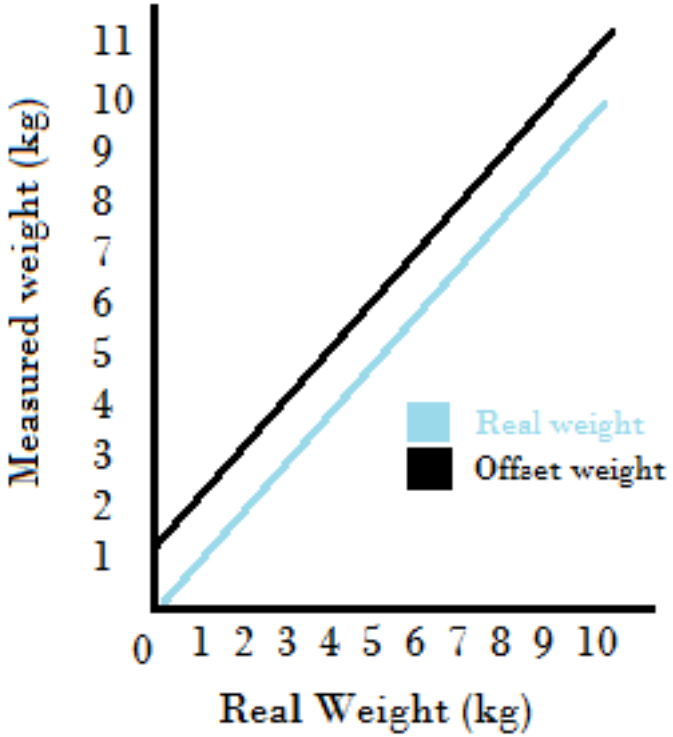
Enables:

- Looking at rare events
- “zooming in”
- Finding small differences

But:

They eliminate *random error* – but: we cannot control how the data are generated, hence we cannot fully rule out *systematic errors/biases*



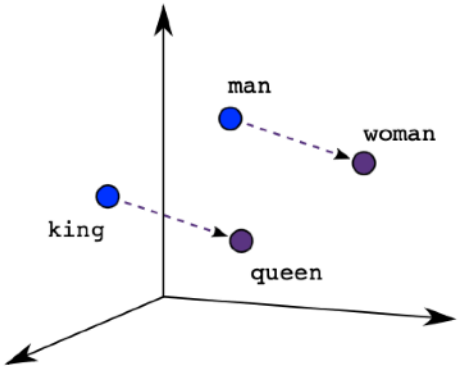


https://www.statisticshowto.com/systematic-error-random-error/

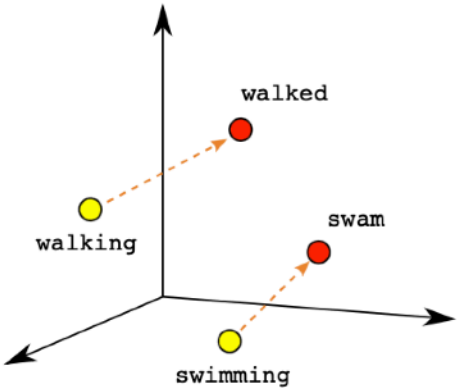
BIG – GARG ET AL. 2018

- the hot new trend in Natural Language Processing (NLP): learning huge models of human language
  - words become “embedded” into a semantic vector space
- basic logic: the computer tries to learn a numerical representation of human language
  - big problem: our language is always biased to some extent – this can lead to big problems in downstream prediction tasks
  - however, it’s great for us as social science researchers – we can search for and quantify these biases
- this is what Garg et al. (2018) are doing: gender stereotypes and how they have evolved

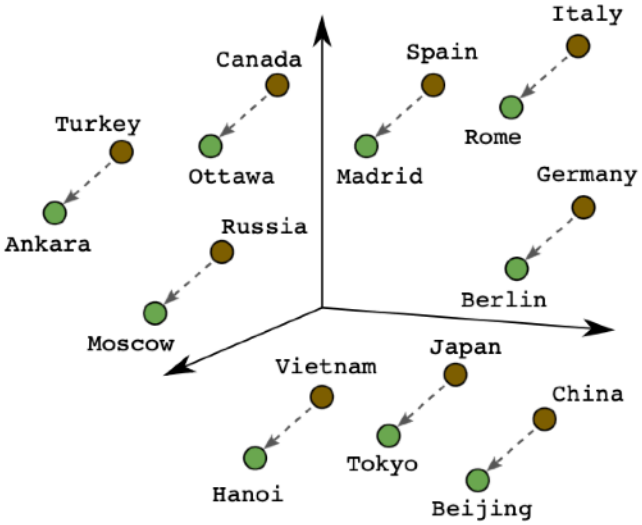
  “We use the word embeddings as a quantitative lens through which to study historical trends—specifically trends in the gender and ethnic stereotypes in the 20th and 21st centuries in the United States.” (Garg et al. 2018)



Male-Female



Verb Tense



Country-Capital

https://developers.google.com/machine-learning/crash-course/embeddings/translating-to-a-lower-dimensional-space

BIG – GARG ET AL. 2018

- data source: “Google Books/Corpus of Historical American English (COHA) embeddings, which are a set of nine embeddings, each trained on a decade in the 1900s, using the COHA and Google Books”
- constructed axes based on gendered words and pronouns (e.g., man/woman, he/she, his/her)
- projected adjectives and occupations on the resulting axis
  - idea: if an adjective or occupation co-occurs more often with male words, they are closer to male in the vector space
- validated using “quantifiable demographic trends in the occupation participation” and “historical surveys of stereotypes”

Fig. 1. Women's occupation relative percentage vs. embedding bias in Google News vectors. More positive indicates more associated with women on both axes. $P < 10^{-10}$, $r^2 = 0.499$. The shaded region is the 95% bootstrapped confidence interval of the regression line. In this single embedding, then, the association in the embedding effectively captures the percentage of women in an occupation.

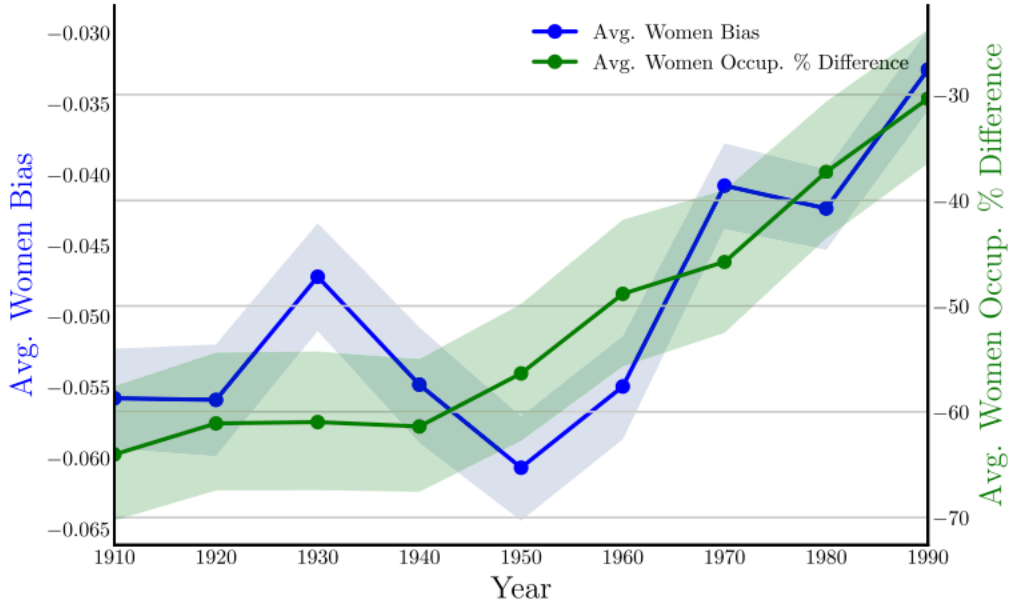


Fig. 2. Average gender bias score over time in COHA embeddings in occupations vs. the average percentage of difference. More positive means a stronger association with women. In blue is relative bias toward women in the embeddings, and in green is the average percentage of difference of women in the same occupations. Each shaded region is the bootstrap SE interval.

Table 2. Top adjectives associated with women in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

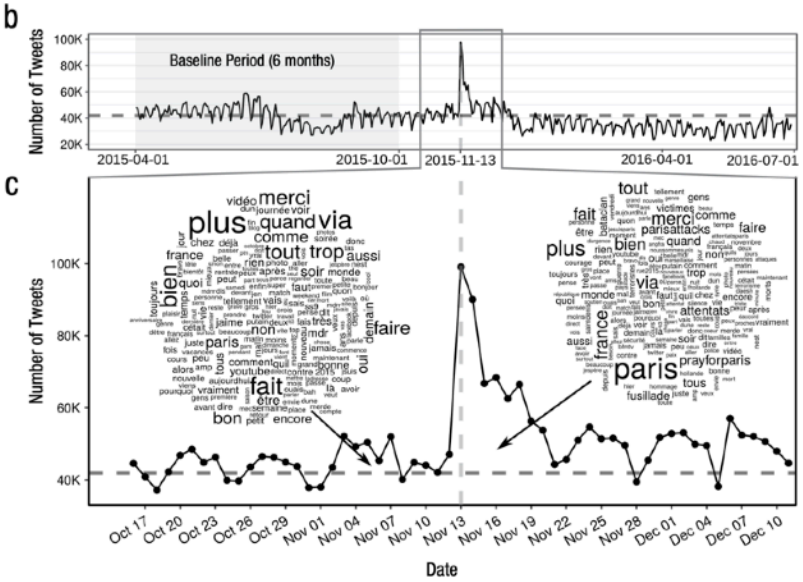
1910	1950	1990
Charming	Delicate	Maternal
Placid	Sweet	Morbid
Delicate	Charming	Artificial
Passionate	Transparent	Physical
Sweet	Placid	Caring
Dreamy	Childish	Emotional
Indulgent	Soft	Protective
Playful	Colorless	Attractive
Mellow	Tasteless	Soft
Sentimental	Agreeable	Tidy

ALWAYS-ON

„Many big data systems are *always-on*; they are constantly collecting data.”
(Salganik 2017: 21)

facilitates:

- studying unexpected phenomena
- *nowcasting*



Garcia/Rimé 2019: 620

„In social sciences, an experiment is a research strategy used by a social scientist to establish **causal relationships** between one or more independent variables and one or more dependent variables. “ (Peng 2004: 349)

„We can approximate experiments that we can’t do. Two approaches that especially benefit from the digital age are natural experiments and matching.”
(Salganik 2017: 51)

STRATEGIES

- **natural experiments**: are there events where people were randomly affected ("treatment group") or not ("control group")?

→ "random … variation + always-on data = natural experiment" (Salganik 2017: 52)

- **matching**: simulation of experiments based on data that can be statistically manipulated

→ „In matching, the researcher looks through non-experimental data to create pairs of people who are similar except that one has received the treatment and one has not … [R]esearchers are actually also *pruning*, that is, discarding cases where there is no obvious match." (Salganik 2017: 55)

DATA

Motivation:

- social networks matter a lot for people
- however, we do not know how they evolve over time
- new results: networks appear unstable over time, volatility to short-term changes in hobbies and interests
- natural disasters are a sort of psychological shock
- literature a bit unsure about how communities react – do people get closer, intensify their ties, or try to expand (maybe also some sort of Durkheimian “effervescence”)

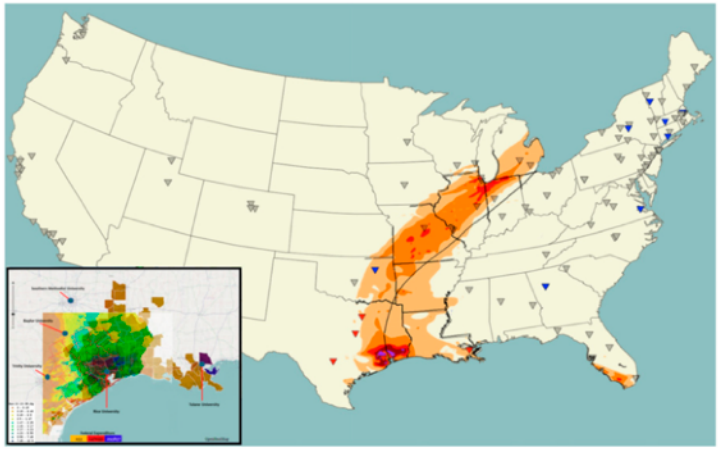
Question:

what were the short-term and long-term effects of hurricane Ike (09/2008) on the social networks of 1.5 million students in the US?

DATA

- **630,000,000 privat messages and 590,000,000 postings by 1,500,000 students**
- students from130 universities (criteria: born 1985–90, university mentioned in FB profile)
- period of investigation: 09/2007–05/2011

- treatment group: students from five universities that are located in affected regions (red triangles)
- control group: students from ten universities that are located in non-affected regions (blue triangles)



Phan und Airoldi 2015: 6596

Universities	No. of Users
Affected universities	
Baylor University	8,462
Rice University	2,355
Southern Methodist University	4,324
Trinity University	1,882
Tulane University	4,505
Unaffected universities	
Colgate University	2,359
The College of William and Mary	4,446
Georgia Institute of Technology	8,703
Middlebury College	2,374
Smith University	1,874
Tufts University	4,337
University of Pennsylvania	8,644
University of Tulsa	1,877
University of Utah	4,296
Yale University	4,519

Phan & Airoldi 2015: 6597

ANALYSIS

- search for „comparable pairs“ among universities via *propensity score matching*
  - criteria: size, socio-economic factors, regional and university-specific differences
- five measures from SNA – five dimensions of comparison:
  - number of friends (*average degree*)
  - who are they/how close is the circle (*transitivity*)
  - intensity of communication (*posts-per-week* and *messages-per-week*)
  - range of communication (*unique recipients* – measured as *recipients-per-week*)
  - *preferential attachment* – relationships with people who have central roles in the network
- Different periods:
  - four weeks before Ike (09/2008) ↔ four weeks after Ike
  - longitudinal 09/2007–05/2011

RESULTS

average degree:
- minor differences

transitivity:
- treatment group: more transitivity – become friends with friends of friends; keep “circle small”
- control group: extend circle

communication
- number of messages increases after event
- treatment group with less postings
- results stable over longer period (i.e., 52 weeks)

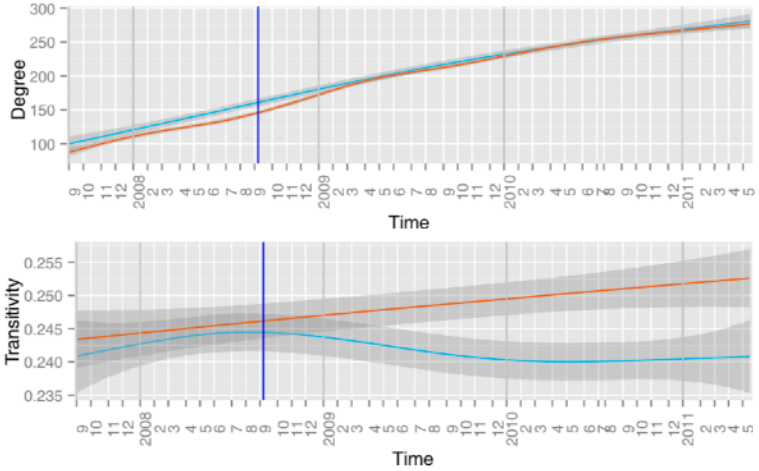


Table 2. Paired *t* tests using a 4-wk window before and after Hurricane Ike

Quantity	Treatment	SE	Control	SE	*P* value	*t* Statistic
Messaging	0.2654	15.5189	0.3470	32.1653	0.3628	−0.9100
Posting	−0.0871	8.6791	0.1209	9.3832	0.0000	−5.8176
No. of recipients	0.0196	3.4011	0.0597	3.7235	0.0044	−2.8498

Phan & Airoldi 2015: 6597-8

RESULTS

unique recipients
- increase in both groups (larger increase in control group)

preferential attachment
- behavior seems to be less common in treatment group
- hypothesis from the authors: people in treatment group might have better things to do

Table 2. Paired *t* tests using a 4-wk window before and after Hurricane Ike

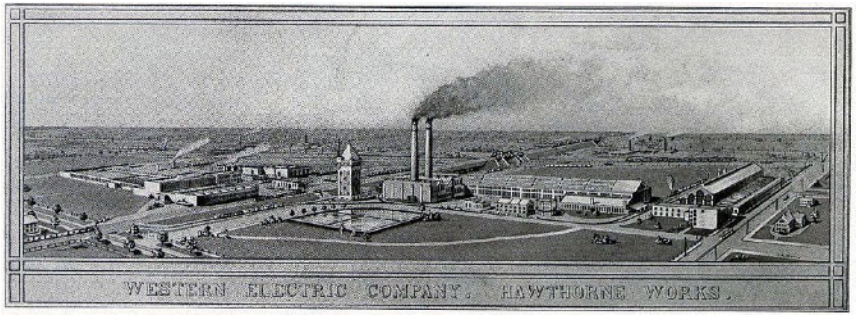
Quantity	Treatment	SE	Control	SE	*P* value	*t* Statistic
Messaging	0.2654	15.5189	0.3470	32.1653	0.3628	−0.9100
Posting	−0.0871	8.6791	0.1209	9.3832	0.0000	−5.8176
No. of recipients	0.0196	3.4011	0.0597	3.7235	0.0044	−2.8498

Table 3. Paired *t* tests using a 52-wk window before and after Hurricane Ike

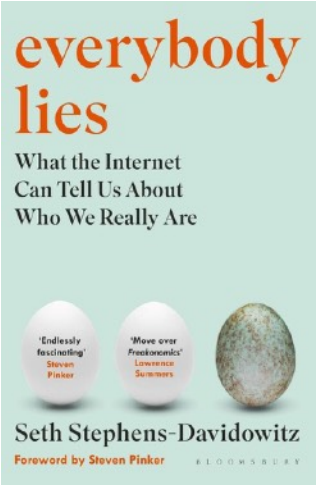
Quantity	Treatment	SE	Control	SE	*P* value	*t* Statistic
Messaging	0.0204	5.4548	0.0267	10.0864	0.4393	−0.7733
Posting	−0.0067	3.2523	0.0093	3.7682	0.0000	−4.1947
No. of recipients	0.0015	1.1383	0.0046	1.3107	0.0205	−2.3168

Phan & Airoldi 2015: 6598

NONREACTIVE



- Hawthorne effect: humans behave differently once they know that they are participants
- Advantage (D)TD: we analyze traces of human behavior – they are not aware of the fact that they are participating in our study
- BUT: online data are not perfect either – have their own biases (Ruths & Pfeffer 2014)



Stephens-Davidowitz 2016

INCOMPLETE

“No matter how big your big data, it probably doesn’t have the data you want.” (Salganik 2017: 24)

they usually lack:
- sociodemographics
- behavior on different parts of the internet
- the data you need to test your theory (e.g., geo data)

INACCESSIBLE

Utah Data Center: NSA collects incredible amounts of data such as citizens' communication data, but also digital pocket litter" (digital parking tickets, train tickets, digital invoices, etc.) – would be incredibly useful for us researcher – but of course we will never see them

Domestic Surveillance Center
https://nsa.gov1.info/utah-data-center/

"These data are inaccessible not because people at companies and governments are stupid, lazy, or uncaring. Rather, there are serious legal, business, and ethical barriers that prevent data access." (Salganik 2018: 27)

What could go wrong – an example:

AOL releases 657,000 users' search queries – did not anonymize the data properly – NYT reporters were able to link queries to people

A Face Is Exposed for AOL Searcher No. 4417749

By Michael Barbaro and Tom Zeller Jr.

"And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."
It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs." (Barbaro/Zeller 2006)

NONREPRESENTATIVE

Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape

Andranik Tumasjan[1], Timm O. Sprenger[1], Philipp G. Sandner[1], and Isabell M. Welpe[1]

Abstract

This study investigates whether microblogging messages on Twitter validly mirror the political landscape off-line and can be used to predict election results. In the context of the 2009 German federal election, we conducted a sentiment analysis of over 100,000 messages containing a reference to either a political party or a politician. Our results show that Twitter is used extensively for political deliberation and that the mere number of party mentions accurately reflects the election result. The tweets' sentiment (e.g., positive and negative emotions associated with a politician) corresponds closely to voters' political preferences. In addition, party sentiment profiles reflect the similarity of political positions between parties. We derive suggestions for further research and discuss the use of microblogging services to aggregate dispersed information.

Tumasjan et al. 2011

Table 5. Share of Tweets and Election Results

Party	All Mentions		Election	
	Number of Tweets	Share of Twitter Traffic	Election Result	Prediction Error
CDU	30,886	30.1%	29.0%	1.0%
CSU	5,748	5.6%	6.9%	1.3%
SPD	27,356	26.6%	24.5%	2.2%
FDP	17,737	17.3%	15.5%	1.7%
Die Linke	12,689	12.4%	12.7%	0.3%
Grüne	8,250	8.0%	11.4%	3.3%
			MAE:	1.65%

Note. CDU = Christian Democrats, CSU = Christian Social Union, SPD = Social Democrats, FDP = Liberals, Die Linke = Socialists, Grüne = Green Party; MAE = mean absolute error.

Tumasjan et al. 2011: 412

Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment"

Andreas Jungherr[1], Pascal Jürgens[2], and Harald Schoen[1]

Abstract

In their article "Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment," the authors Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe (TSSW) the authors claim that it would be possible to predict election outcomes in Germany by examining the relative frequency of the mentions of political parties in Twitter messages posted during the election campaign. In this response we show that the results of TSSW are contingent on arbitrary choices of the authors. We demonstrate that as of yet the relative frequency of mentions of German political parties in Twitter message allows no prediction of election results.

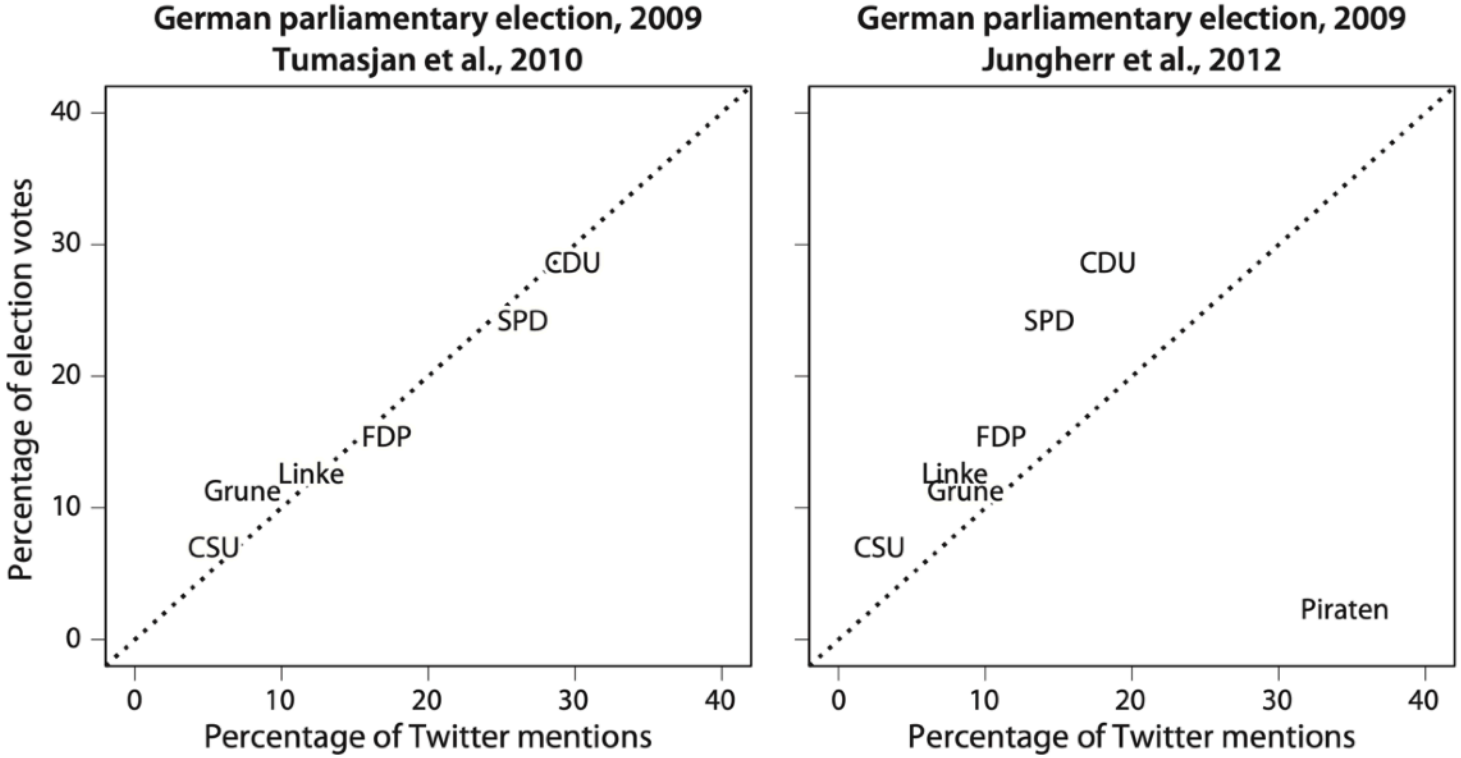
Jungherr/Jürgens/Schoen 2012

Table 2. Parties' Vote Shares and Proportions of Twitter Mentions Including the Pirate Party

Party	Election Results	Share of Twitter Messages (Replication)
CDU	28.4	18.6
CSU	6.8	3.0
SPD	24.0	14.7
FDP	15.2	11.2
Linke	12.4	8.3
Grüne	11.1	9.3
Piraten	2.1	34.8

Note. Following TSSW, when calculating vote shares, we included only the votes cast for the seven parties under scrutiny.

Jungherr/Jürgens/Schoen 2012: 231



Salganik 2017: 32

DRIFT

“If you want to measure change, don’t change the measure.” (as cited in Salganik 2018: 33)

Most big data sources collect data over time – longitudinal data. This enables us to measure changes in, for instance, people’s attitudes. Drift refers to the problem that the sources themselves do not remain stable either. Three types of drift:

Population drift – old people become familiar with Facebook

Behavioral drift – hashtags change, trends die

System drift – 280 characters on Twitter

ALGORITHMICALLY CONFOUNDED

- The platforms affect how people use them. We need to bear this in mind when we analyze their behavior.

- Some weird findings from Facebook…
  - A lot of Facebook users have exact 20 friends
  - Transitivity in friend networks on Facebook is fairly high – is this due to “normal” behavior?

DIRTY

- DTD are rarely in an analyzable format
- therefore, not only skills in terms of acquisition (e.g., scraping) required, but also w*rangling skills* and *cleaning techniques* (e.g., *regular expressions*, *tidyverse* in general)

SENSITIVE

- DTD are oftentimes sensitive
  - most of the time, this is straightforward:
    - health data
    - movement data
  - sometimes it's not: Netflix prize (Narayanan & Shmatikov 2008)

How To Break Anonymity of the Netflix Prize Dataset

Arvind Narayanan, Vitaly Shmatikov

They had information on a user's movie ratings on Netflix and the date they rated the movie

- they matched it with the IMDb, looking for users who rated the same movie within a range of two weeks from the rating on Netflix in quite the same manner
- political orientation, religious views, etc. can then be extracted from their view on related movies

SENSITIVE

How To Break Anonymity of the Netflix Prize Dataset

Arvind Narayanan, Vitaly Shmatikov

“First, we can immediately find his political orientation based on his strong opinions about ‘Power and Terror: Noam Chomsky in Our Times’ and 'Fahrenheit 9/11.’ Strong guesses about his religious views can be made based on his ratings on ‘Jesus of Nazareth’ and ‘The Gospel of John.’ He did not like ‘Super Size Me’ at all; perhaps this implies something about his physical size? Both items that we found with predominantly gay themes, ‘Bent’ and ‘Queer as folk’ were rated one star out of five. He is a cultish follower of ‘Mystery Science Theater 3000.’ This is far from all we found about this one person, but having made our point, we will spare the reader further lurid details.” (Narayanan/ Shmatikov 2008: 16)

⟹ they also “uncovered” an in-the-closet lesbian mother (WIRED article)

TO SUM IT UP

- Digital Trace Data immensely promising for a couple of reasons (big, always collecting, “honest”)
- But there are some pitfalls as well

When brainstorming the data you need for your questions, think about each of the criteria and how the data may affect your research/analyses in a negative way

⟹ if the data is not perfect: no big deal

⟹ it may be enriching to study a phenomenon just with a limited sample

⟹ but you need to discuss that in your paper

⟹ this strongly affects the **generalizability** of your findings

TO SUM IT UP

Mental crutch: the 5 Ws (Salganik 2017)

- Who (produced)
- what (kind of data)
- where
- when
- why

REFERENCES

- Garg, Nikhil/Schiebinger, Londa/Jurafsky, Dan/Zou, James 2018: Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences* 115(16), pp. 3635–44.
- Golder, Scott/Macy, Michael 2014: Digital Footprints: Opportunities and Challenges for Online Social Research. In: *Annual Review of Sociology* 40, pp. 129–152.
- Jungherr, Andreas 2015: Analyzing Political Communication with Digital Trace Data. Cham: Springer.
- Narayanan, Arvind/Shmatikov, Vitaly 2008: Robust De-Anonymization of Large Sparse Datasets. In: *Proceedings of the 2008 IEEE Symposium on Security and Privacy*. Washington, DC: IEEE Computer Society, pp. 111–125.
- Peng, Chao-Ying Joanne 2004: Experiment. In: Lewis-Beck, Michael S., Alan Bryman und Tim Futing Liao (Eds.) *The Sage Encyclopedia of Social Science Research Methods*. Volume 1. Thousand Oaks: Sage, pp. 349–354
- Phan, Tuan/Airoldi, Edoardo 2015. A Natural Experiment of Social Network Formation and Dynamics. In: *Proceedings of the National Academy of Sciences of the USA* 112 (21), pp. 6595–6600.
- Salganik, Matthew 2017: *Bit By Bit. Social Research in the Digital Age*. Princeton and Oxford: Princeton University Press.
- Tumasjan, Andranik/Sprenger, Timm/ Sander, Philipp/Welpe, Isabell 2010: Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment. In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pp. 178–185.

UNIVERSITÄT LEIPZIG

MERCI!

Felix Lennert
Institut für Soziologie

felix.lennert@uni-leipzig.de
www.uni-leipzig.de