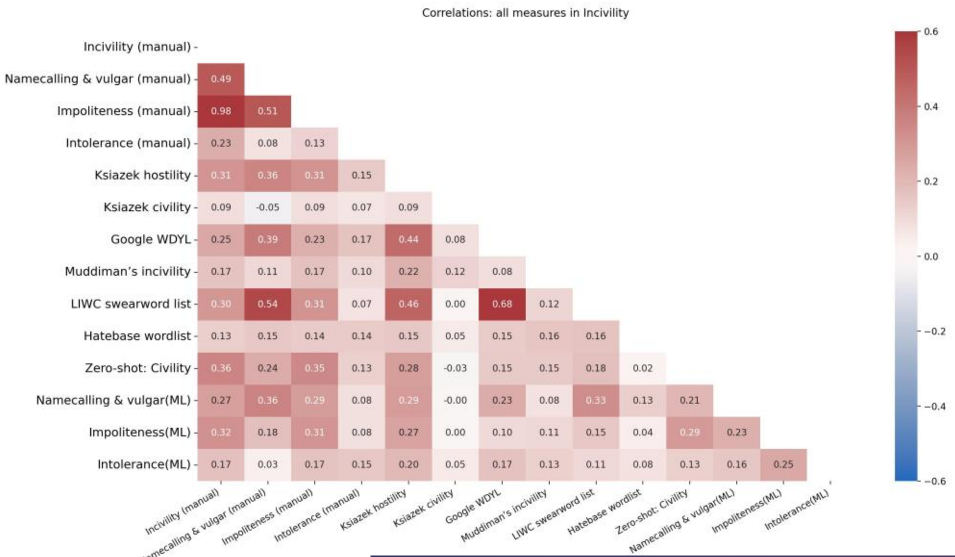
Working with codes in social science research: Web scraping and data analysis

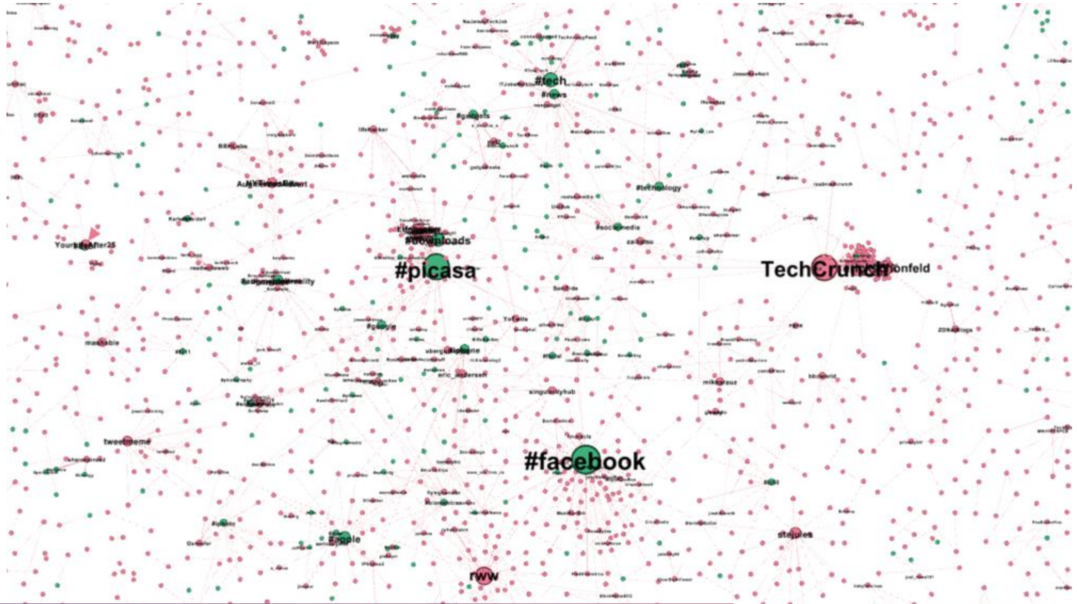
Justin Yeung|Digital Text and Data Analysis

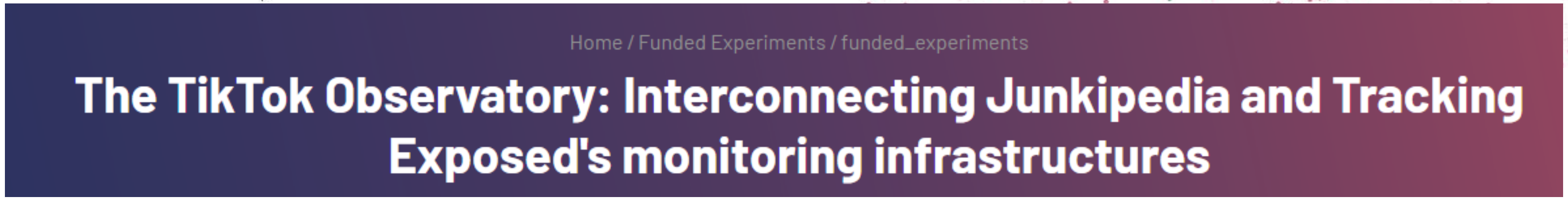
Universiteit Leiden
The Netherlands

A little bit about myself

- Background in communication science and computational social sciences at the University of Amsterdam
- Affiliate of the Edinburgh SDS hub (you can also join!) and incoming visiting researcher at the Max Planck Institute of Human Development (Berlin)
- Research topics including computational methods, algorithmic violence and news framing
- Research methods including network analysis, ACA, lab / online experiments



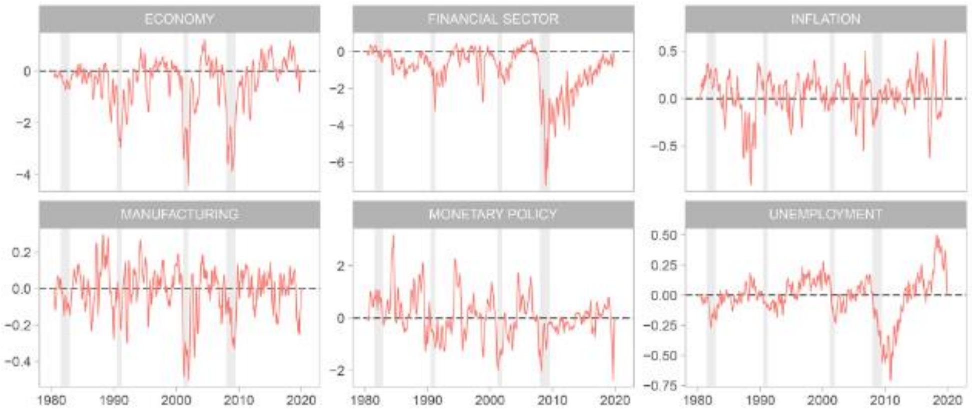



Home / Funded Experiments / funded_experiments

The TikTok Observatory: Interconnecting Junkipedia and Tracking Exposed's monitoring infrastructures

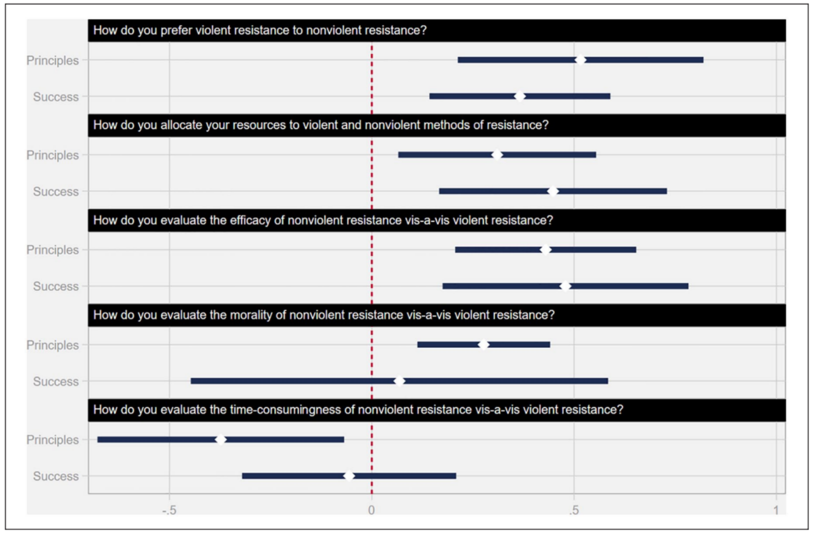
Asking questions as a social scientist

- Social scientists are often interested in questions about (1) why, (2) how and (3) when (4) what and (5) who happens / behaves in certain situations
- The common paradigms of social science research: Descriptive / exploratory, explanatory and predictive
- Different subfields of social sciences ask different questions



Barbagliaa et al., (2022)

How can economic news forecast economic performance?

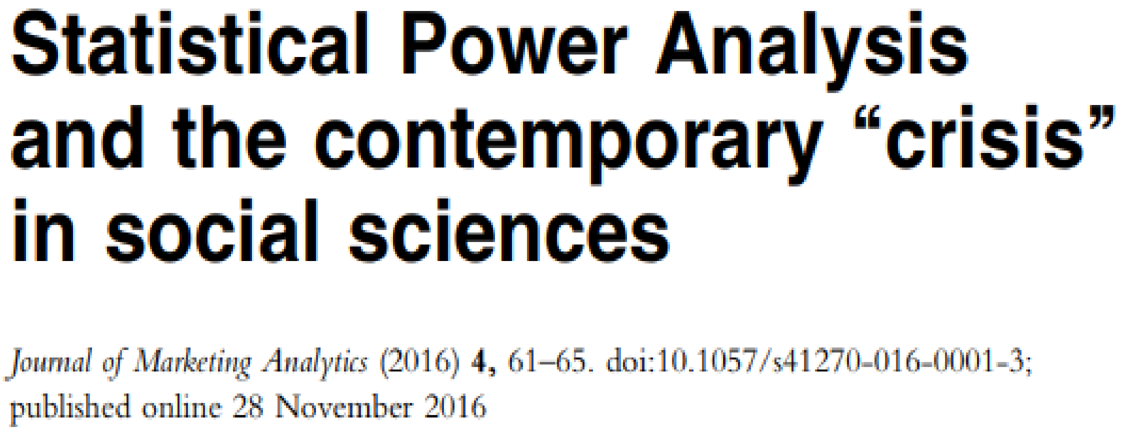


RezaeeDaryakenari & Asadzade (2020)

When do individuals tend to prefer nonviolent to violent resistance?

Era of ‘big’ data in the social sciences

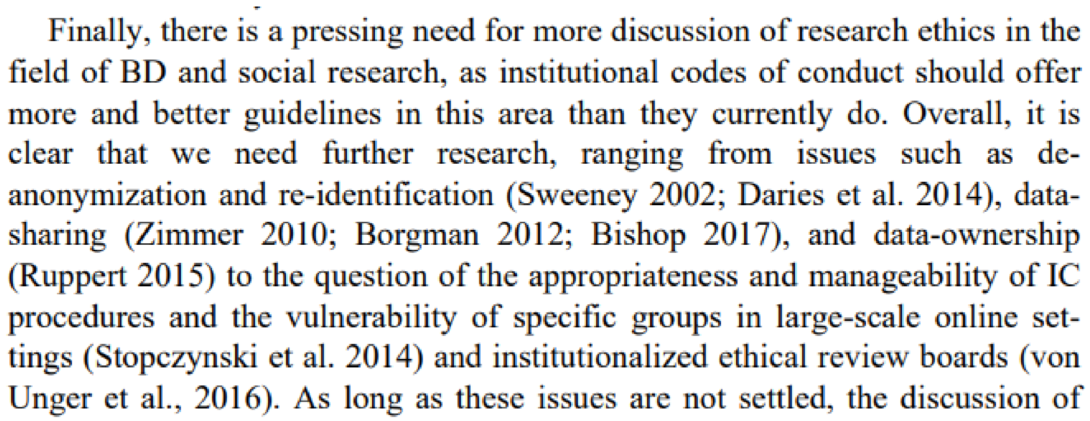
- We cannot merely collect data from surveys and (lab/online) experiments to understand our online world
- We are interested in settings where external validity can be maximised
- Great data come with great responsibility: research integrity and ethics



Statistical Power Analysis and the contemporary “crisis” in social sciences

Journal of Marketing Analytics (2016) **4,** 61–65. doi:10.1057/s41270-016-0001-3; published online 28 November 2016

(Breur, 2016)



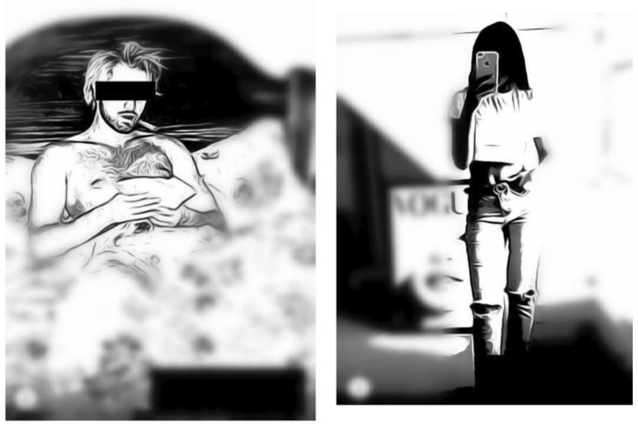
Finally, there is a pressing need for more discussion of research ethics in the field of BD and social research, as institutional codes of conduct should offer more and better guidelines in this area than they currently do. Overall, it is clear that we need further research, ranging from issues such as de-anonymization and re-identification (Sweeney 2002; Daries et al. 2014), data-sharing (Zimmer 2010; Borgman 2012; Bishop 2017), and data-ownership (Ruppert 2015) to the question of the appropriateness and manageability of IC procedures and the vulnerability of specific groups in large-scale online settings (Stopczynski et al. 2014) and institutionalized ethical review boards (von Unger et al., 2016). As long as these issues are not settled, the discussion of

(Weinhardt, 2020)

What and how to collect (digital) data?

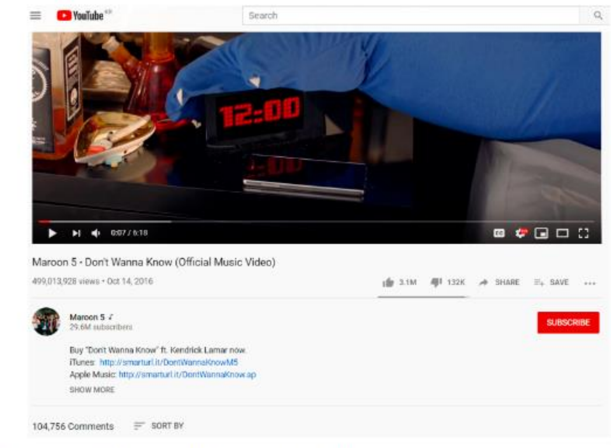
Visual

Tinder profile pictures
(e.g., Degen & Kleeberg-Niepage, 2021)



Aural

YouTube videos
(e.g., Oh & Choeh, 2021)



Maroon 5 - Don't Wanna Know (Official Music Video)

499,013,928 views • Oct 14, 2016

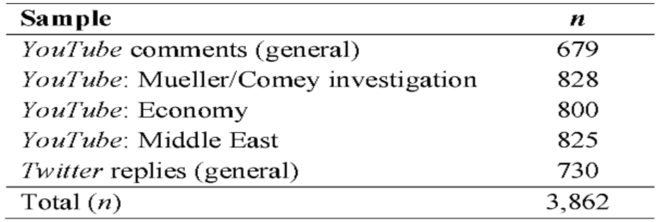
Maroon 5 ✓
29.6M subscribers

Buy "Don't Wanna Know" ft. Kendrick Lamar now.
iTunes: http://smarturl.it/DontWannaKnowM5
Apple Music: http://smarturl.it/DontWannaKnow.ap

104,756 Comments

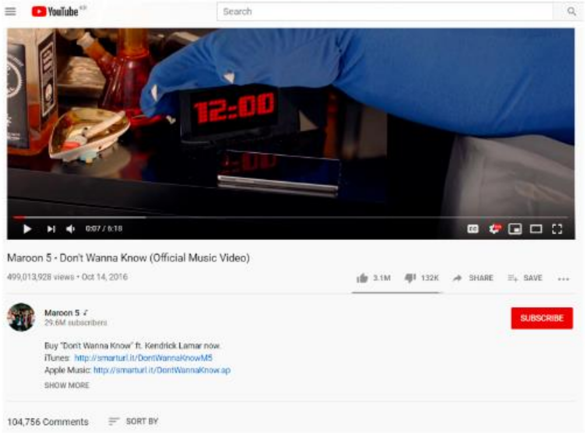
Textual

Social media comments
(e.g., Boukes et al., 2022)



Sample	***n***
YouTube comments (general)	679
YouTube: Mueller/Comey investigation	828
YouTube: Economy	800
YouTube: Middle East	825
Twitter replies (general)	730
Total (*n*)	3,862

What and how to collect (digital) data?



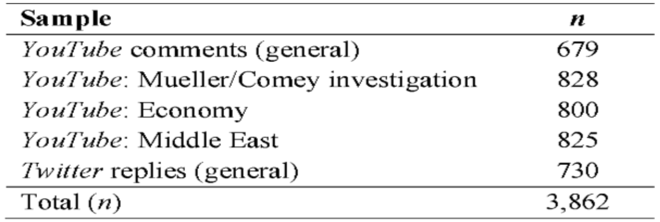
Maroon 5 - Don't Wanna Know (Official Music Video)

499,013,928 views • Oct 14, 2016

Maroon 5
29.6M subscribers

Buy "Don't Wanna Know" ft. Kendrick Lamar now.
iTunes: http://smarturl.it/DontWannaKnowM5
Apple Music: http://smarturl.it/DontWannaKnow.ap

104,756 Comments



Sample	*n*
YouTube comments (general)	679
YouTube: Mueller/Comey investigation	828
YouTube: Economy	800
YouTube: Middle East	825
Twitter replies (general)	730
Total (*n*)	3,862

Manual / semi-automated:
Agent-based data collection

Automated:
API and online database

Automated: API

What is / are the problem(s) ?

The fragile stairway to heaven: APIs for social media platforms

- Application Programming Interface (API) allows for easier and technically stable access to the data available on digital platforms
- Collectable data are determined by the platform's infrastructure
- Very vulnerable to the platform's policies or management's decisions
- What can we do if API is no more available if we want to collect data from these platforms?

Twitter Dev
@TwitterDev

Starting February 9, we will no longer support free access to the Twitter API, both v2 and v1.1. A paid basic tier will be available instead

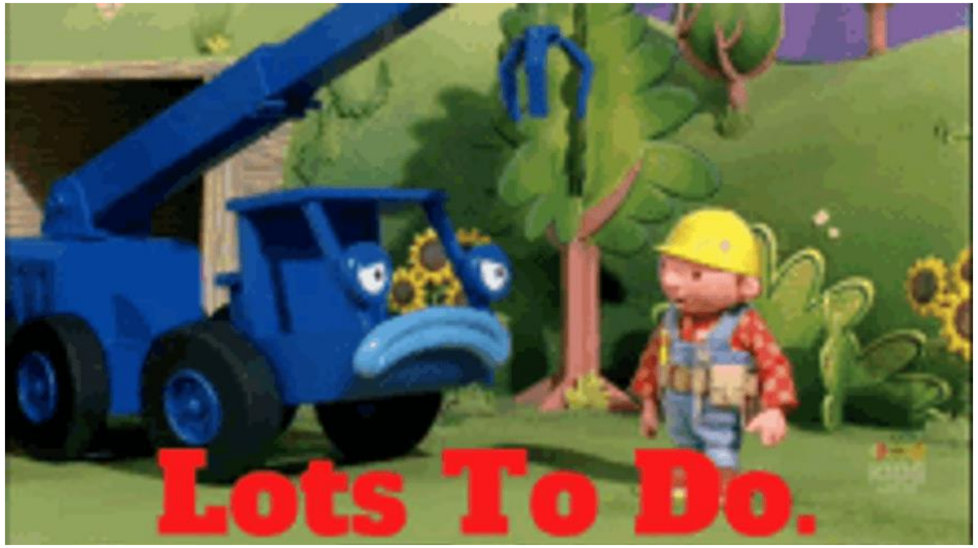
But not for Academic Track access right?
But not for Academic Track access right?

Source: @favstats on Twitter

Source: @danhett on Twitter

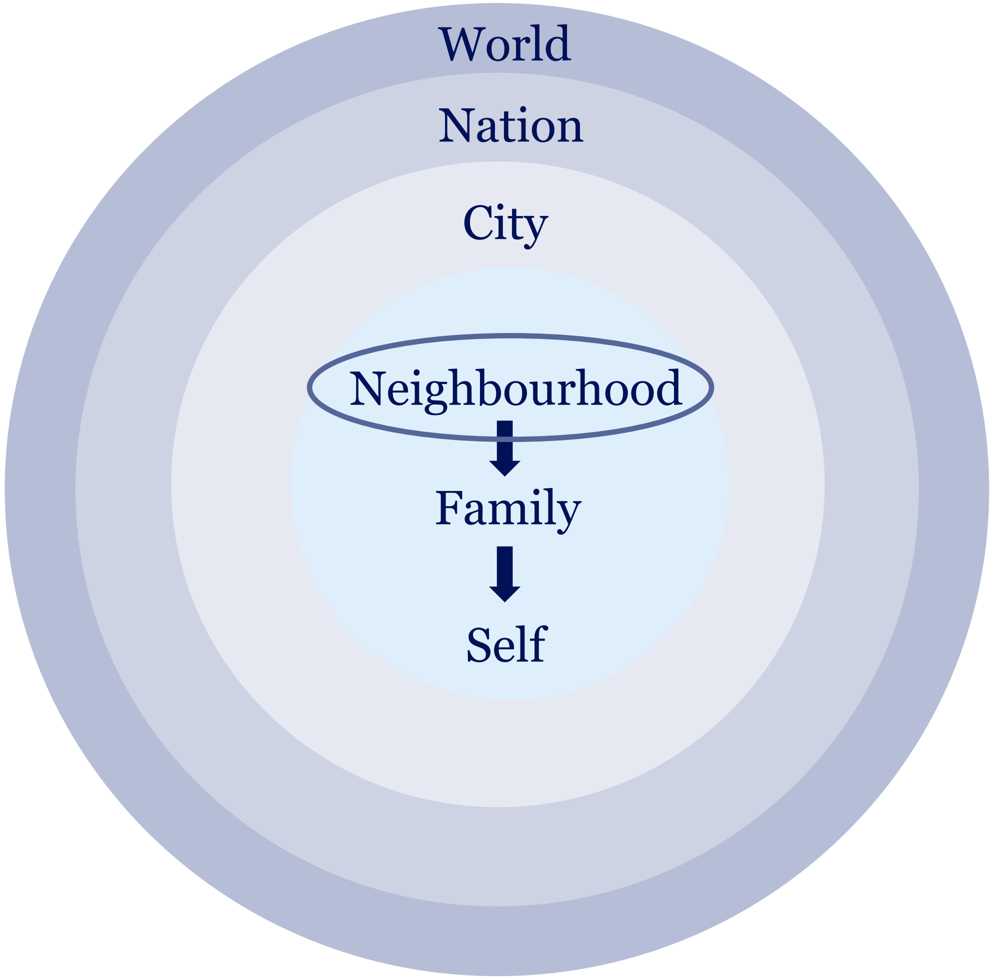
Building your own infrastructure!

- The lazy way: Use ready-made tools such as Octoparse and ParseHub
  - User-friendly
  - Commercial
  - Not tailored to your research project
- The good way: Create a tailored web scraper to collect your digital data with Scrapy, BeautifulSoup4 and Selenium
  - Difficult to set up
  - Open-source and free
  - Tailored to your research project
- The common problem: Subject to platform's changes



Communitarianism, diversity and social cohesion

- **Paradox of diversity:** (ethnic) diversity will inevitably pose threats to solidarity and social capital of a society in the short-medium run but nurture a better society in the long run (Putnam, 2007).
- **Value of community:** The values of community are necessary to balance a society too often tipped in the direction of self-centeredness, greed, and power seeking (Glass and Rud, 2012).
- **Community and exclusion:** A thick notion of belonging will end up in forms of exclusion, which are observable from some Muslim communities in the Netherlands (Duyvendak, 2011).
- **A solution?** Establishing an overarching communitarian blueprint – Diversity Within Unity – to uphold universal values without sacrificing diversity (Etzioni, 2002).



Computational social science in action

- An interdisciplinary project combining communication science, sociology and urban studies
- Aim
  - Uncover some realities in an ethno-racially mixed neighbourhood, as a manifestation of immigration and diversity
  - Examine why and how building a uniform and close community among various subgroups is difficult in a diverse neighbourhood.
  - Introducing Communication Infrastructure Theory (CIT) as a theoretical underpinning to achieve cohesion
  - Assessing and criticising existing platforms (i.e., NextDoor)
- Our results show
  - Nextdoor is an underdeveloped meso-level communication infrastructure. Out of 8865 residents who age 15 or above (Statline, 2021), only 326 are participating in the respective Nextdoor page of the neighbourhood, which accounts for merely 3.68% of the population.
- Our question now…
  - Why Nextdoor failed to establish the meso-level storytelling network in the neighbourhood?

Communication Infrastructure as an approach to a cohesive neighbourhood

- Previous literature has much focus on a **macro-level policies** and **legal aspects** to attain DWU (e.g., Baltic states: Barrington, 1999, Canada: Howard-Hassmann, 2000; Britain and France: Banton, 2001).
- **Communication Infrastructure Theory (CIT):** a triple-level storytelling network that facilitates the construction of objective and subjective sense of belonging to a certain community (Ball-Rokeach et al., 2001).
  - Establishing strong community through setting common goal.
  - Communication infrastructures as tools to "reconstitute a social world in which the **"I" and the "we" can survive**" (Ball-Rokeach et al., 2001, p.393).
- Example of CIs
  - National and local newspapers
  - Interpersonal networks

Macro

NOS

Meso

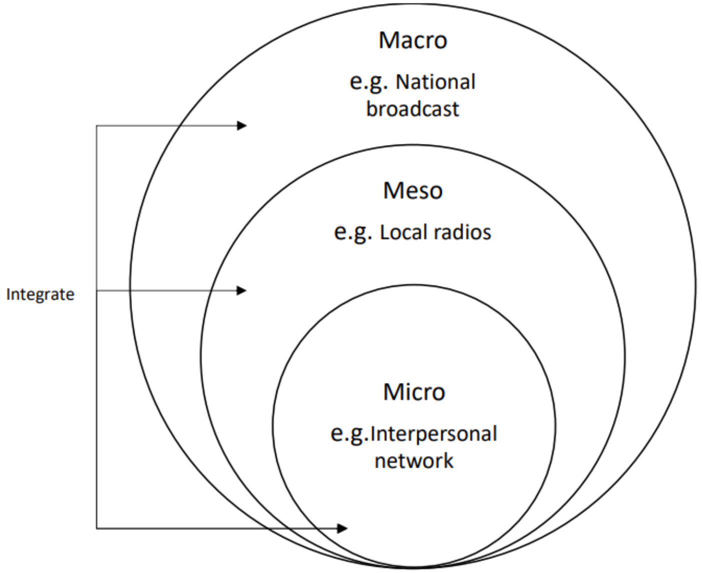
Leidsch Dagblad

Micro

Current research on neighbourhoods and belonging

- The presence and integration of CIs can strengthen **sense of belonging**, **civic engagement** and **community organization connection** (e.g., Choi et al., 2021; Nah et al., 2021)
- Meso-level networks work better in dominantly black neighbourhoods (only in the U.S.)
- Interesting moderators: commuting time, home identification and home ownership

Can the implementation of meso-level communication infrastructures encourage neighbourhood engagement?

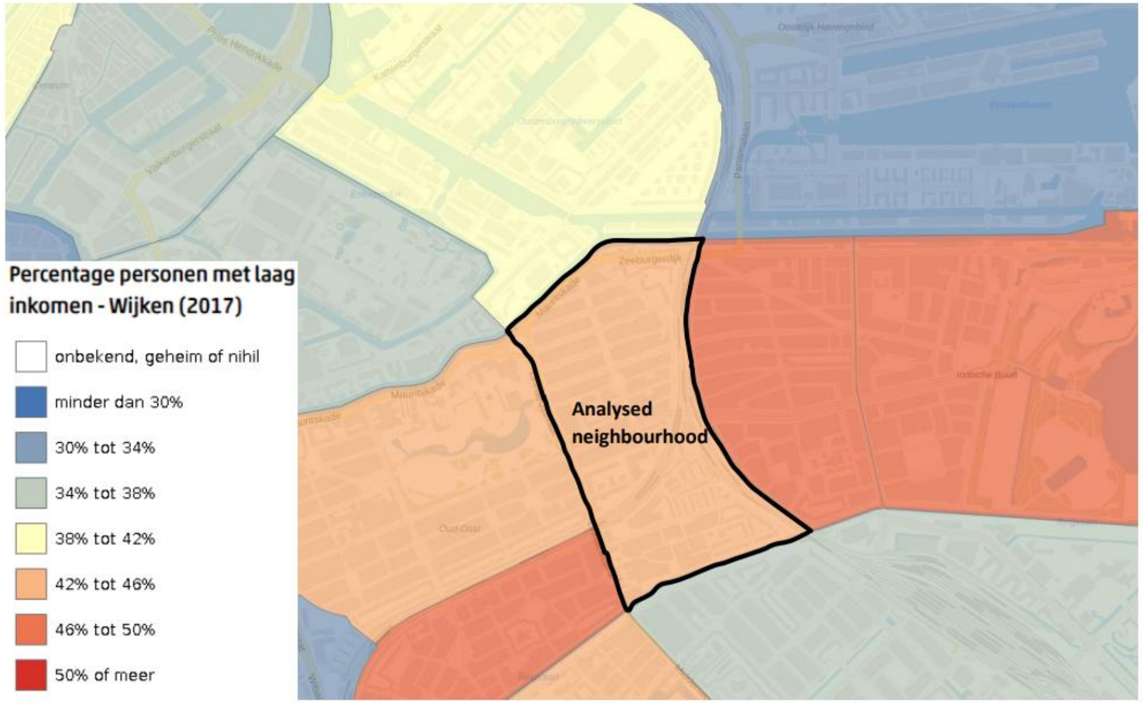


Studying Nextdoor as a communication infrastructure

- Nextdoor is a **hyperlocal social networking site** (SNS) that allows co-residents to share relevant information to a specific neighbourhood.
- Nextdoor is both a **meso- and micro-level infrastructure**.
  - Nextdoor allows exclusively relevant agencies (i.e., neighbours) and information to penetrate the network
  - Individuals are able to connect with each other (i.e., friending) and post updates within the certain (i.e., following and posting).
- What do people do / say on Nextdoor?
  - receive trusted information
  - give and get help
  - get things done
  - build real-world connections with those nearby — neighbors, businesses, and public services

Sampling and manual content analysis

- Agent-based data collection
  - Data collected in a neighbourhood located in Amsterdam Oost
- Collected 30 consecutive posts on the news feed
  - Included comments
- $N_{total} = 119$
- Thematic analysis (Shanahan, 2019)
  - Data familiarisation
  - Code creation
  - Alignment with RQ
  - Theme categorisation and reviewing
  - Theme naming



Yeung, 2021

Results (Thematic analysis)

Category	Example	Count
Self-introduction	"It's great to be here. I'm [name] … and I come from [place]. Me and my husband recently moved to the [neighbourhood] and love the area already! Looking forward to discover more and more about the neighbourhood.	9
Ask for help / Provide help	"Urgent Alert. Has anyone seen this kitten before? Send me a message ASAP!! It is urgent, a small and orange kitten."	15
Community events	"Neighbourhood tour through [neighbourhood] on Sunday 26 September!"	3
Businesses	"Dear neighbours! It is lucky, we are open! … Come thrift at [company] and spread the words!"	1

* Original comments and posts are in Dutch

Did Nextdoor failed? Some background of the neighbourhood

- Out of 8865 residents who age 15 or above (Statline, 2021), only 326 are participating in the respective Nextdoor page of the neighbourhood, which accounts for merely 3.68% of the population.
- Half of the households of the neighbours fall into the category of low-income and mid-to-low education household.
- 80% of the properties in the studied neighbourhood are rental
  - > Amsterdam grand mean of 68.5% (Amsterdamse Federatie van Woningcorporaties, 2020)
  - Some of the rental apartments are social housing.

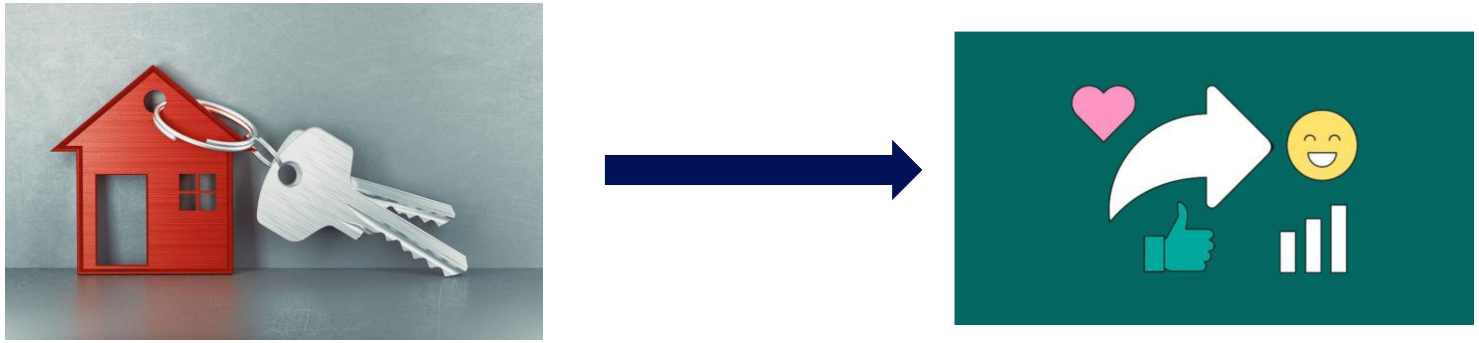
So what could have been the problem(s)?

- Reciprocal relationship between civic engagement and storytelling network (Liu et al., 2018).
  - Low % of home ownership lead to weak place attachment and sense of belonging – (La Grange & Yip, 2001; Elliott & Wadley, 2013; Leviten-Reid & Matthew, 2018).
- Digital divide in Amsterdam (Goedhart & Dedding, 2020)
  - Internet access is hugely determined by socio-economic factors such as age, education level and income (Friemel, 2014)
  - Complex digital divide in Amsterdam particularly for those who live under poverty and vulnerability

How do we test the hypotheses?

- First – we need to have a larger dataset for greater statistical power
  - We should collect more data and also a more diverse dataset
  - Data from other neighbourhoods to be able to compare means
- Next, what other variables can we think about?
  - Home ownership rate
  - Education level
  - Age (how should we code it?)
  - Income
  - Population
  - …
- Finally, is our study…
  - Quasi-experimental / nonexperimental?
  - Causal / Correlational?

To keep it simple, we assume that there are no omitted variables



But before we start with the analysis

- How do we collect the data?

NextDoorScraper

Universiteit Leiden
The Netherlands

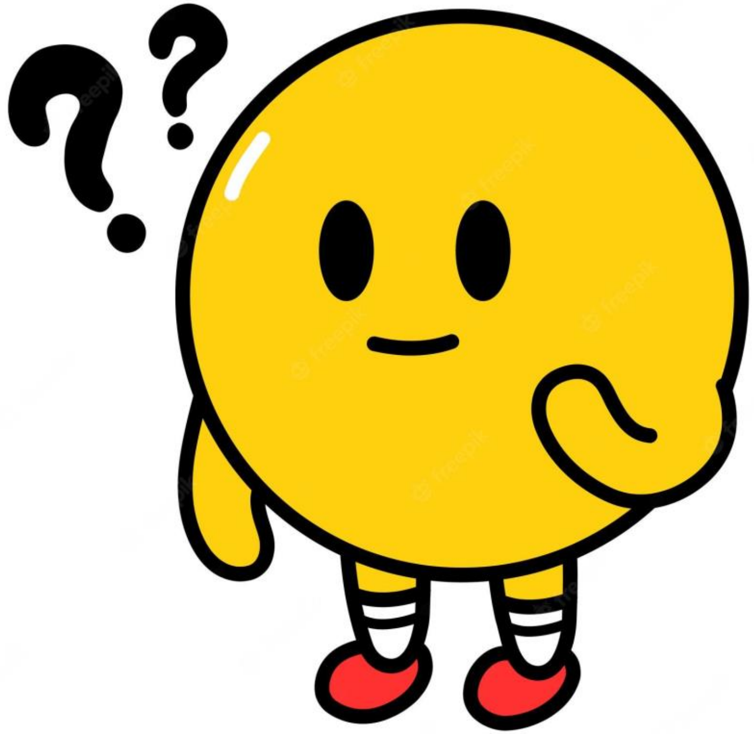
Three directions of study

1. Text (with VoyantTools or Python)
   - **Question:** What are the most frequently mentioned words / noun chunks on Nextdoor?
   - **Recommended visualisation:** Wordcloud / Frequency bar plot
2. Socio-economic factors and engagement (with Excel or Python)
   - **Question:** Do users in neighbourhoods with higher socio-economic status interact more than those in lower ones?
   - **Challenge:** Can neighbourhood average income, population and home ownership predict platform engagement?
   - **Recommended visualisation:** Regression or correlation plot / heatmap (confusion matrix)
   - **Tip:** Think about how you define platform engagement
3. Text + socio-economic (with Python / Excel, for more advanced students)
   - **Question:** Do neighbourhoods with a higher average income level talk more about X,Y,Z…?'
   - Free choice

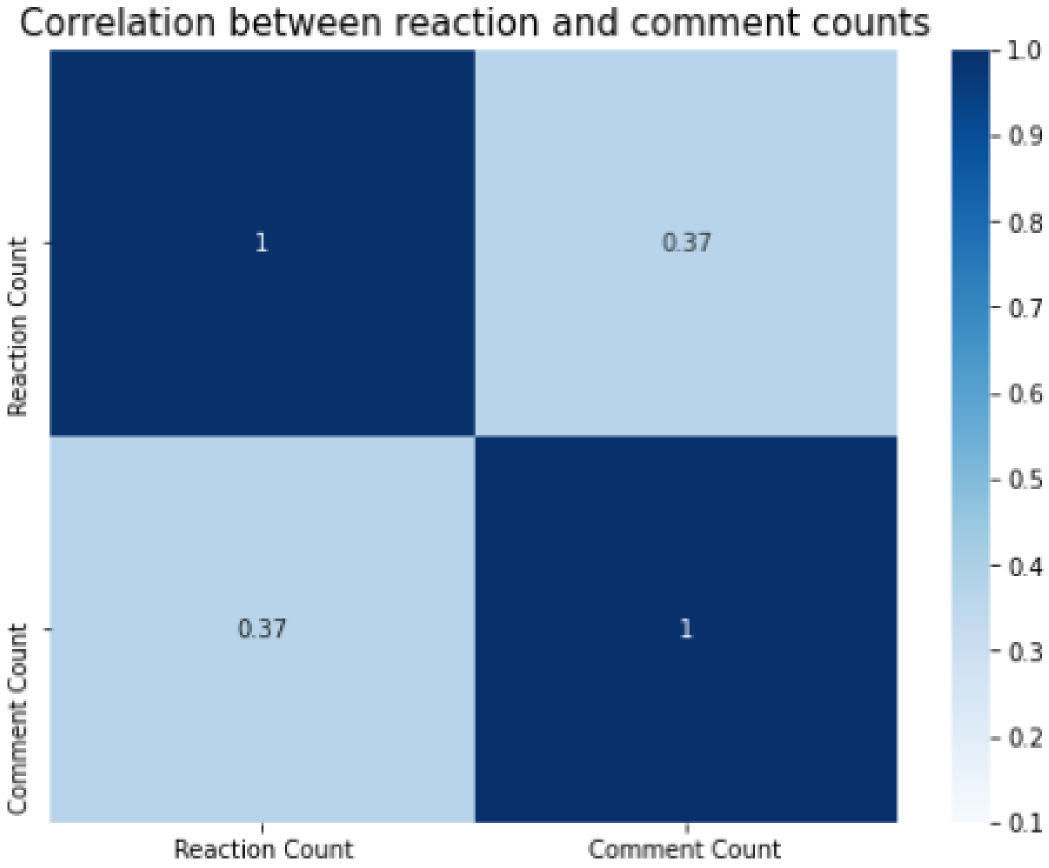
Datasets

- You do not have to do it yourself
- Datasets have been uploaded to the designated GitHub classroom
- **Not** all datasets have been cleaned and therefore you should check before you work on these datasets
  - **Tip:** Be careful while importing CSV files because there are ‘weird’ characters (e.g., emojis)
- If you want to play with some other datasets, it is also fine
  - Keep in mind that you are humanists and social scientists: there should be a theoretical basis why these datasets should be used

Your results and conclusions?



My results: Correlation between reaction and comment counts



Note. p < .001.

User activities

Figure 3. Reaction count per active user (outliers removed).

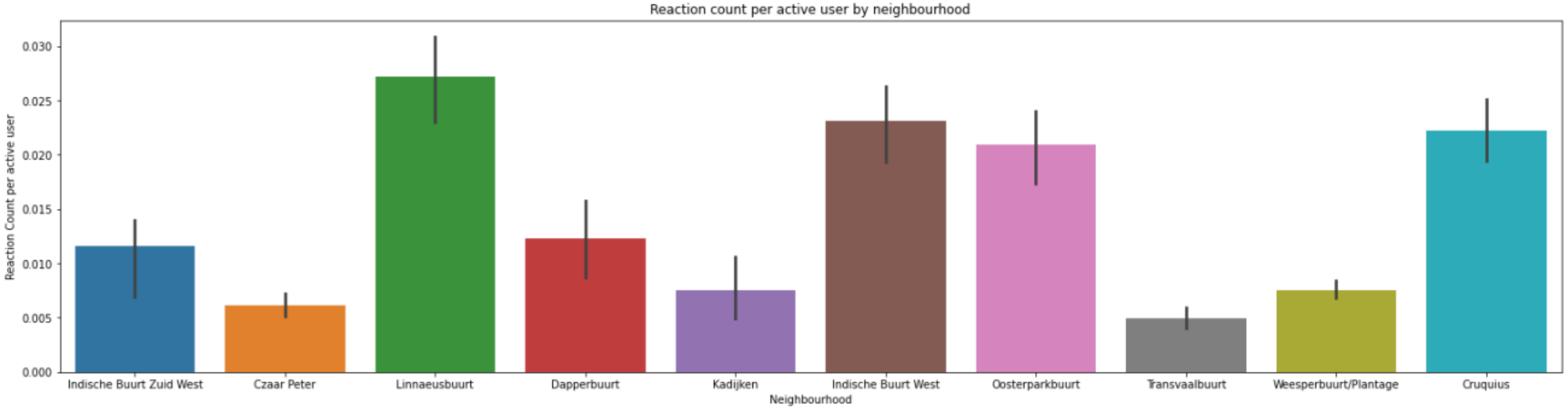
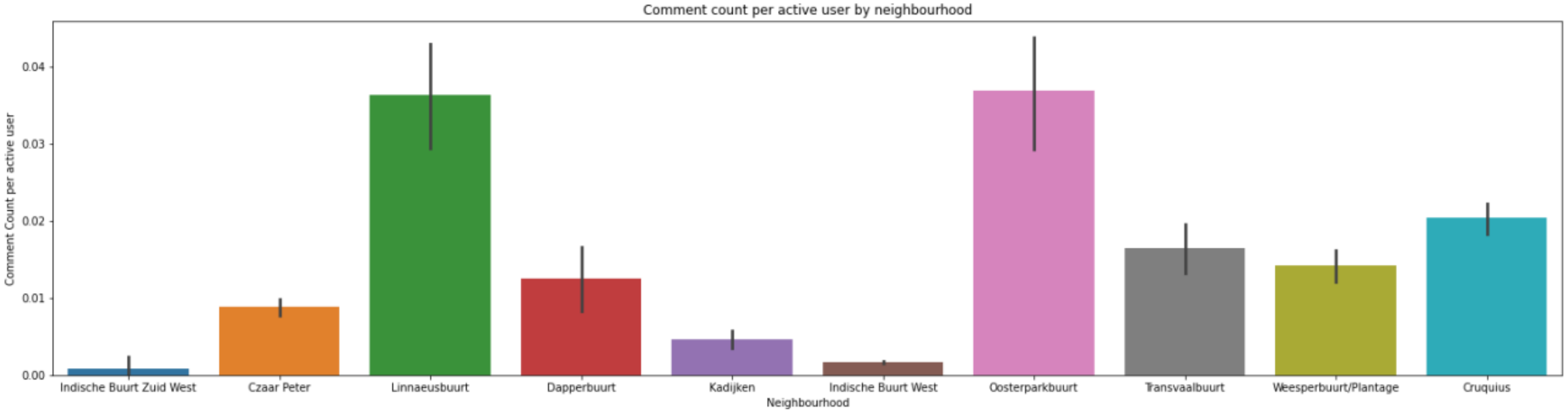
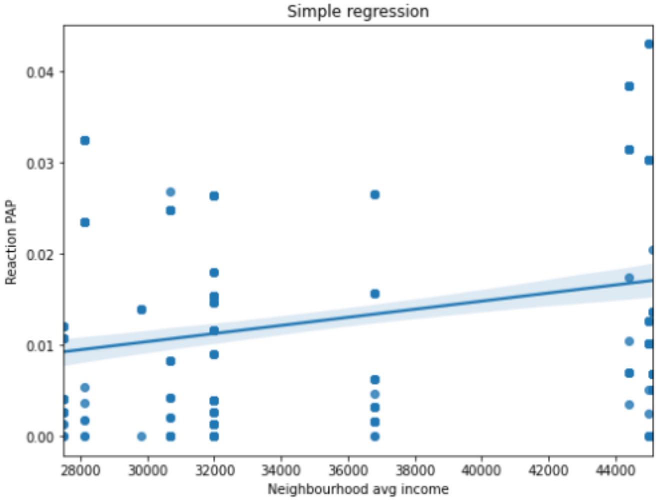
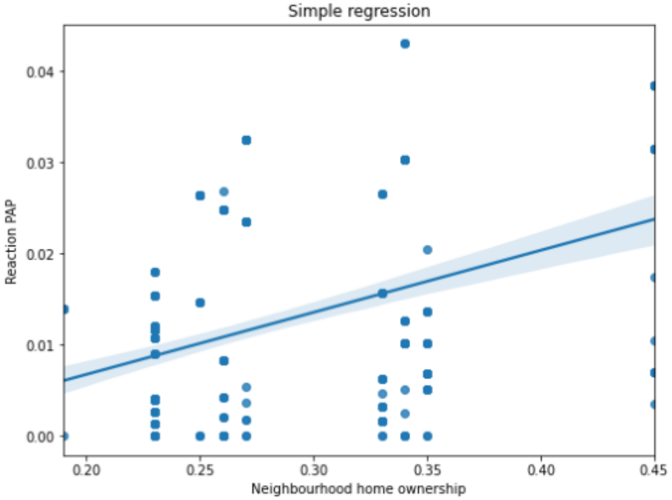


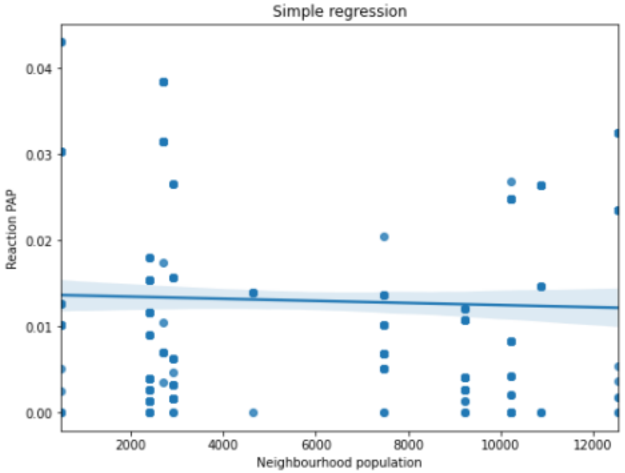
Figure 4. Comment count per active user (outliers removed).



Socio-economic factors and user activities


Simple regression


Simple regression


Simple regression

Altogether, neighbourhood average income, population and home ownership predicts 17.3% of the variance of the outcome variable reaction count per active user, p > .001, F (453, 3) = 31.50. Both home ownership and average income could significantly predict the number of reaction count per active user, whereas neighbourhood population could not significantly predict the number of reaction count.

Conclusion & Future steps?

- It appears that **home ownership** is consistently an important predictor of user engagement on NextDoor, while income can only predict reaction count but not comment count.
  - If home ownership plays an important role in encouraging social media engagement, can we rely on hyperlocal social media to promote cohesion (through enhancing civic engagement and trust) in socially disadvantaged neighbourhoods?
  - Interestingly, population does not play a role in influencing the number of user engagement
- We must be cautious in making a causal claim: This is not a strictly controlled and randomly assigned study
  - Multilevel model may be more useful in analysing nested, hierarchical data (neighbourhood -> user)
- Future steps: Gather more info from different neighbourhoods and combine the data with experimental study?

Thank You and Have Fun in the Digital World!

Universiteit Leiden
The Netherlands