

gesis

Leibniz-Institut  
für Sozialwissenschaften



# Digital traces of human behaviour in online platforms – Research design and error sources

## Meet the Experts

*Best practice methods in Survey Methodology and  
Computational Social Science*

*Fabian Flöck and Indira Sen, 01.07.2021*

# Logistics

- This talk will be recorded. We are not recording the Q&A session after the talk.
- Participants are muted during the session.
- For questions, please only use the private chat function and send questions exclusively to the „meetexperts“ account.
- If you send a message to the general chat, this message (incl. your name) will be visible to all participants.
- Questions will be collected and answered after the talk.

# Speakers



## Dr. Fabian Flöck

- Team lead „Data Science“ at the CSS department
- PhD in Computer / Web Science
- CSS methods and data quality
- Contact: [fabian.floeck@gesis.org](mailto:fabian.floeck@gesis.org)



## Indira Sen

- Doctoral Candidate in CSS
- Works on Natural Language Processing and Social Media Analysis
- Contact: [indira.sen@gesis.org](mailto:indira.sen@gesis.org)

gesis

Leibniz-Institut  
für Sozialwissenschaften



# Digital traces of human behaviour in online platforms – Research design and error sources

## Meet the Experts

*Best practice methods in Survey Methodology and  
Computational Social Science*

*Fabian Flöck and Indira Sen, 01.07.2021*

# Agenda

- The typical design steps in a study with digital traces of human behavior online
- Some typical error sources in such a study
- Exemplified via a use case

# Research with Digital Behavioural Data: More to come

08.07.2021: Combining survey data and digital behavioural data (Sebastian Stier/ Johannes Breuer)

New series with talks about CSS methods and data coming soon: September – December 2021

02-05.11.2021: Workshop on 'Introduction to Social Media as Research Data: Potentials and Pitfalls'

# Research with Digital Behavioural Data: GESIS Courses

## Fall Seminar in Computational Social Science

Date: 13.09 - 01.10.2021 ics-file

Location: Online via Zoom

### Week 1

#### Introduction to Computational Social Science with Python

13.09.2021 - 17.09.2021 - Dr. Orsolya Vásárhelyi, Luis Natera

#### Introduction to Computational Social Science with Applications in R

13.09.2021 - 17.09.2021 - Dr. Aleksandra Urman, Max Pellert

### Week 2

#### Web Data Collection and Natural Language Processing in Python

20.09.2021 - 24.09.2021 - Indira Sen, Dr. Arnim Bleier, Julian Kohne, Dr. Fabian Flöck

#### Automated Web Data Collection with R

20.09.2021 - 24.09.2021 - Dr. Theresa Gessler, Hauke Licht

### Week 3

#### A Practical Introduction to Machine Learning in Python

27.09.2021 - 01.10.2021 - Dr. Damian Trilling, Dr. Anne Kroon

#### Social Network Analysis with R

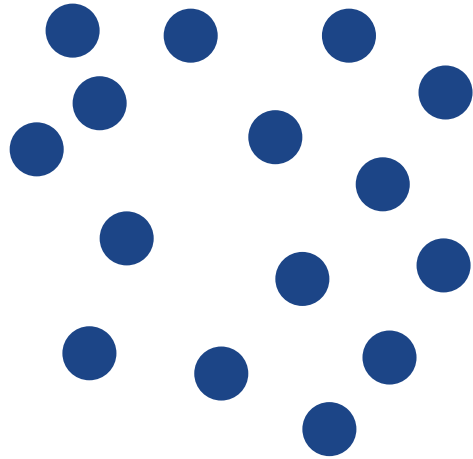
27.09.2021 - 01.10.2021 - Dr. Silvia Fierascu, Ianis Rusitoru

[training.gesis.org](https://training.gesis.org)

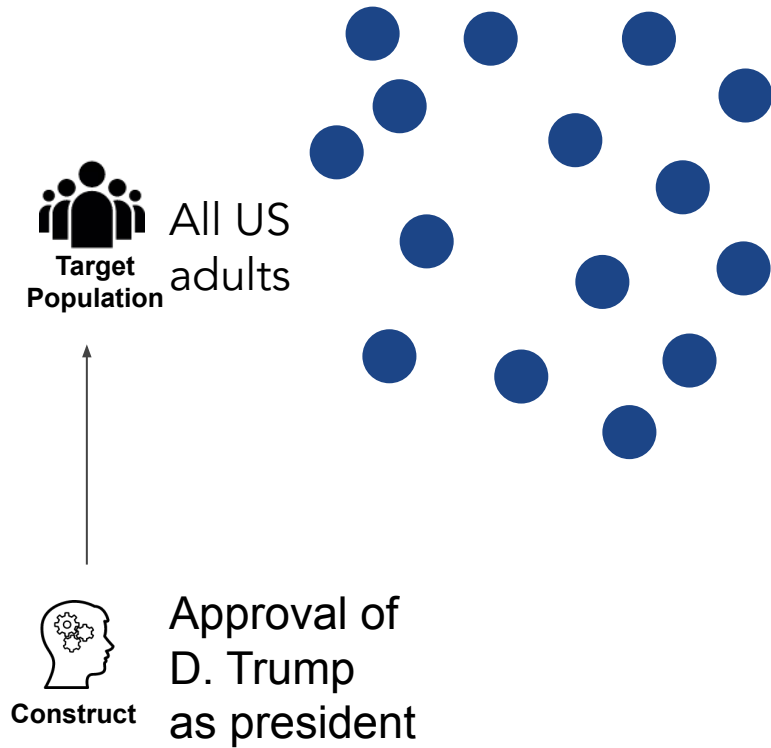
# A typical research pipeline with digital trace data



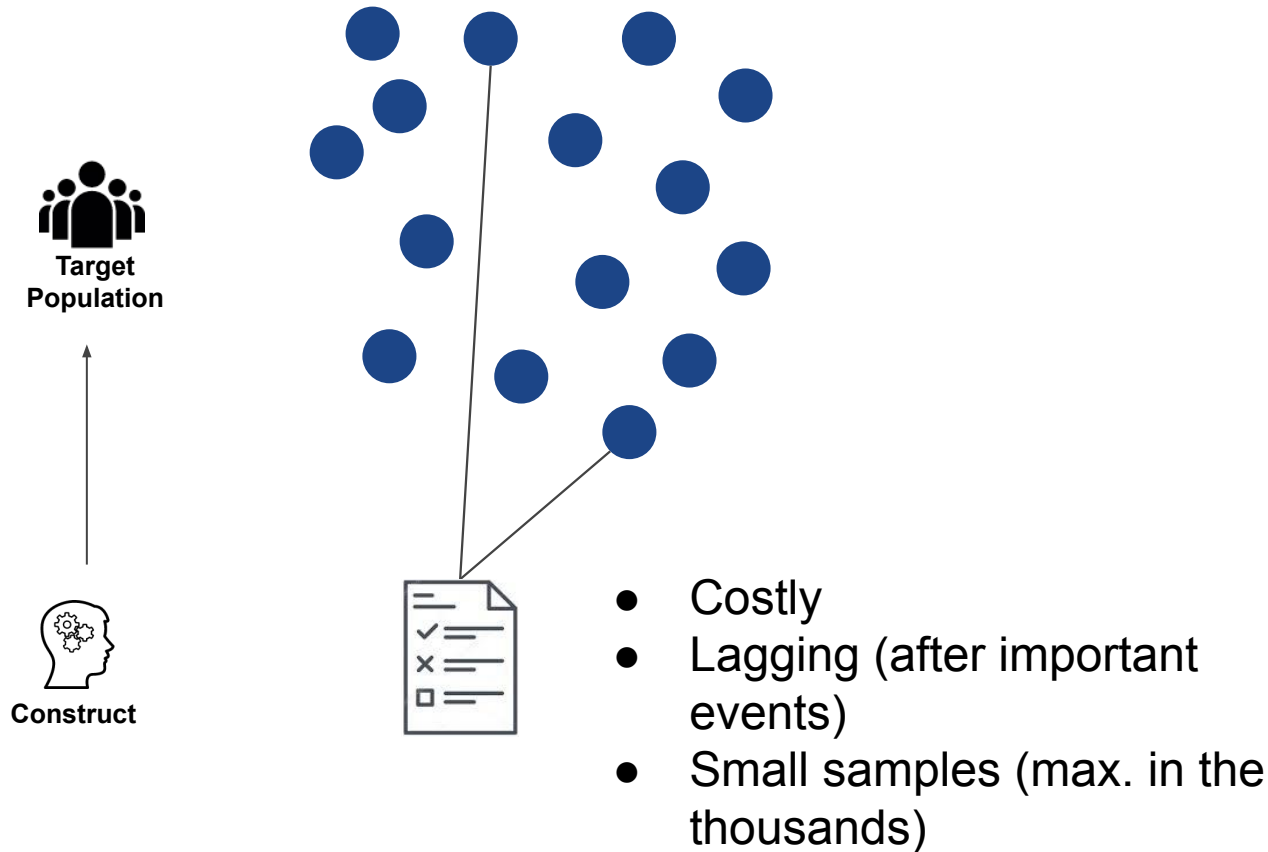
# A typical research design with digital traces



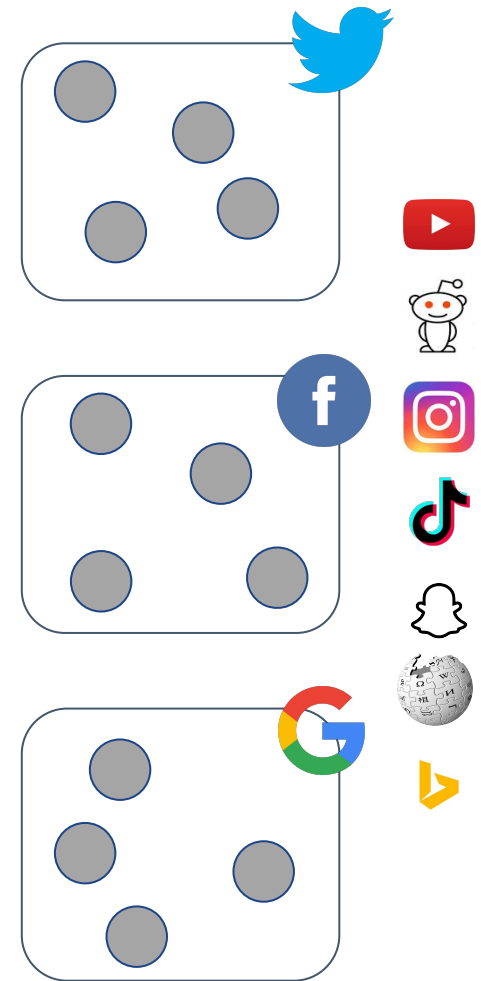
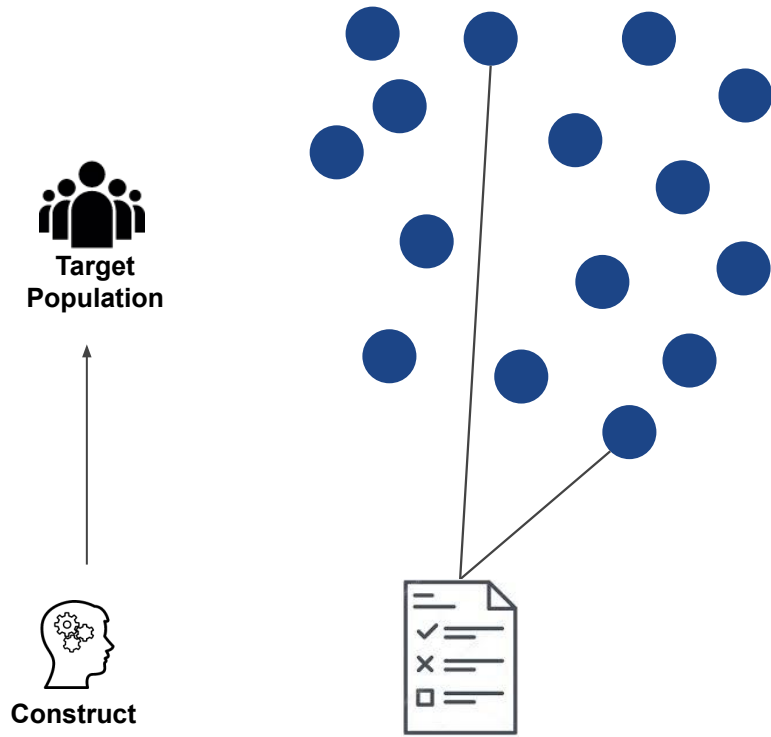
# A typical research design with digital traces



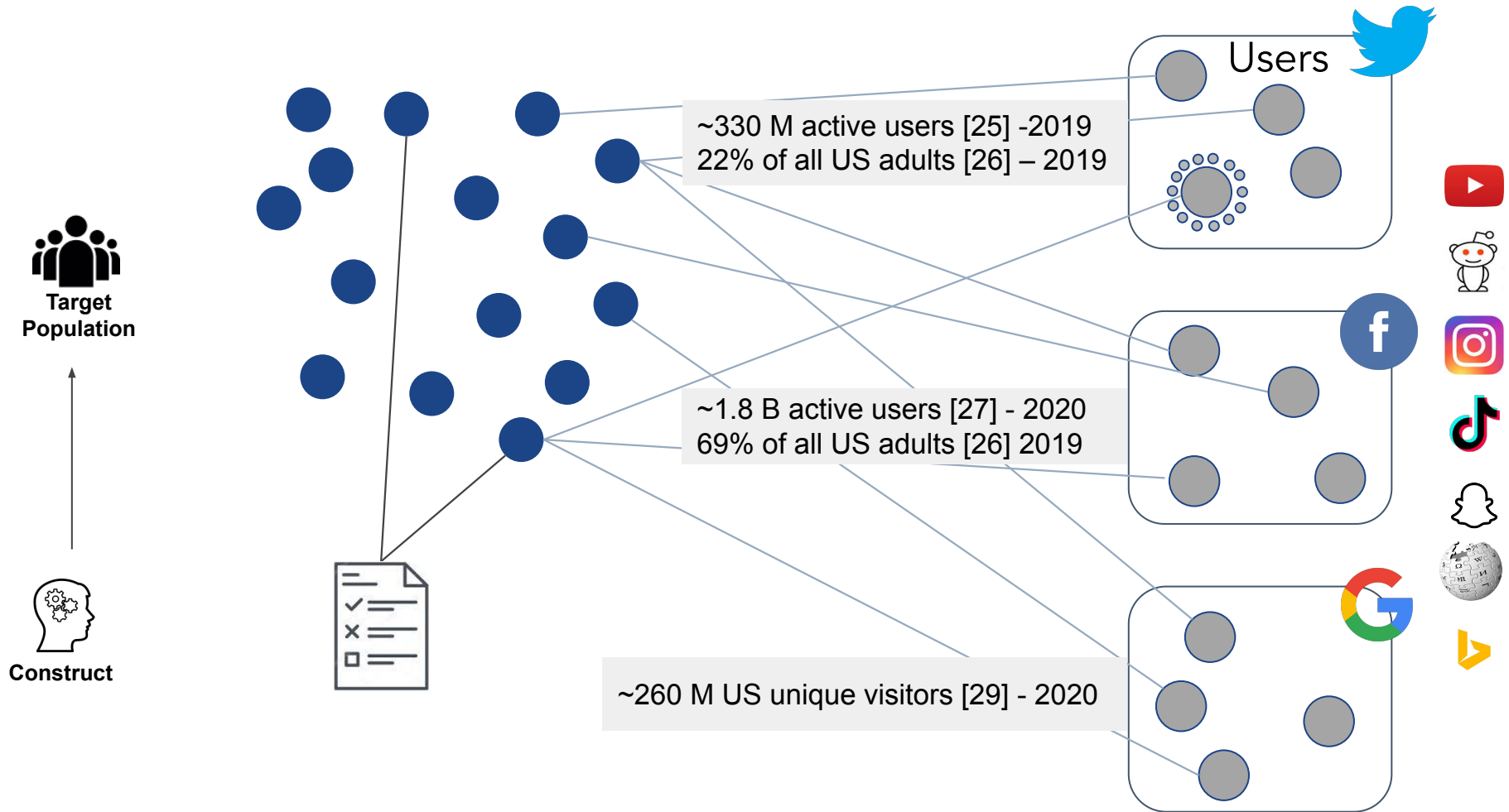
# A typical research design with digital traces



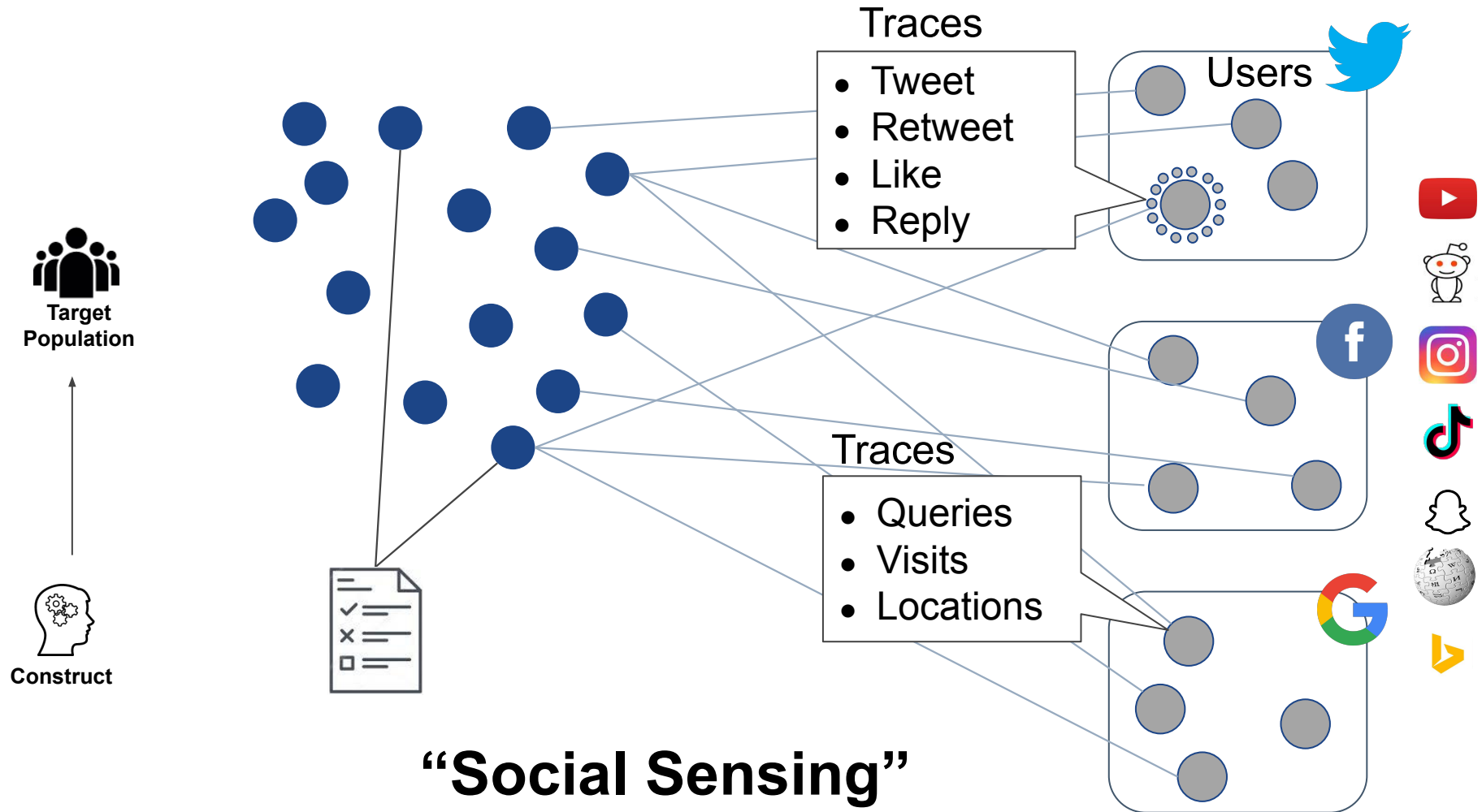
# A typical research design with digital traces



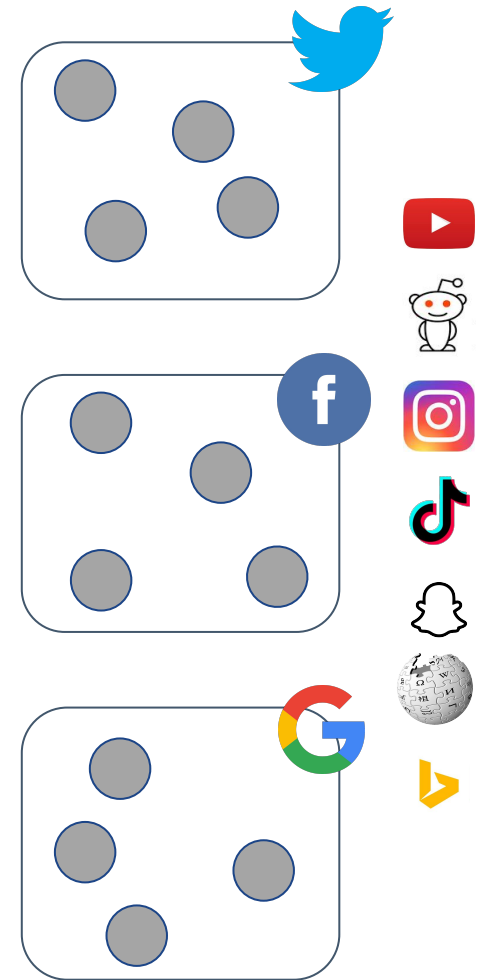
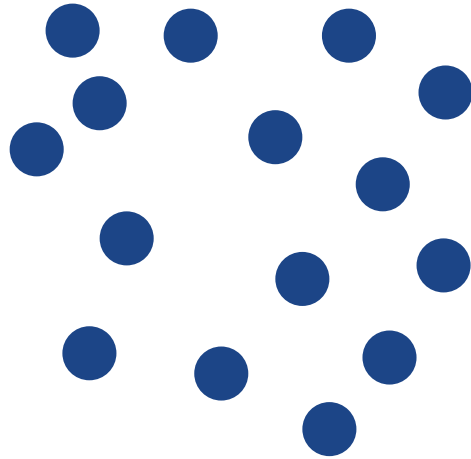
# A typical research design with digital traces



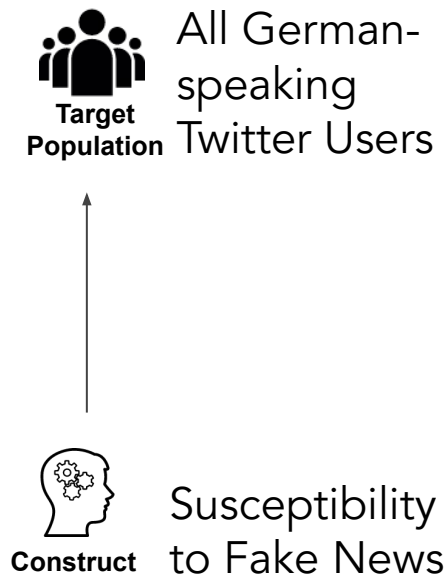
# A typical research design with digital traces



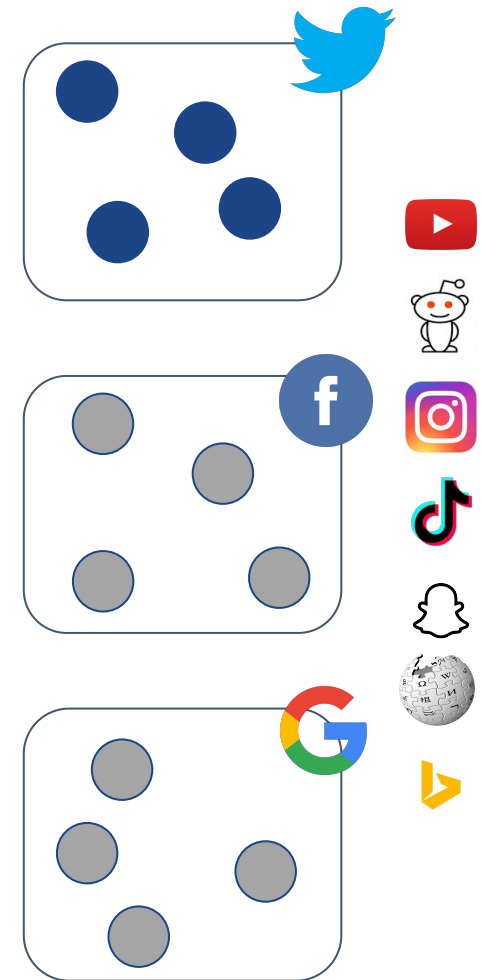
# A typical research design with digital traces



# A typical research design with digital traces

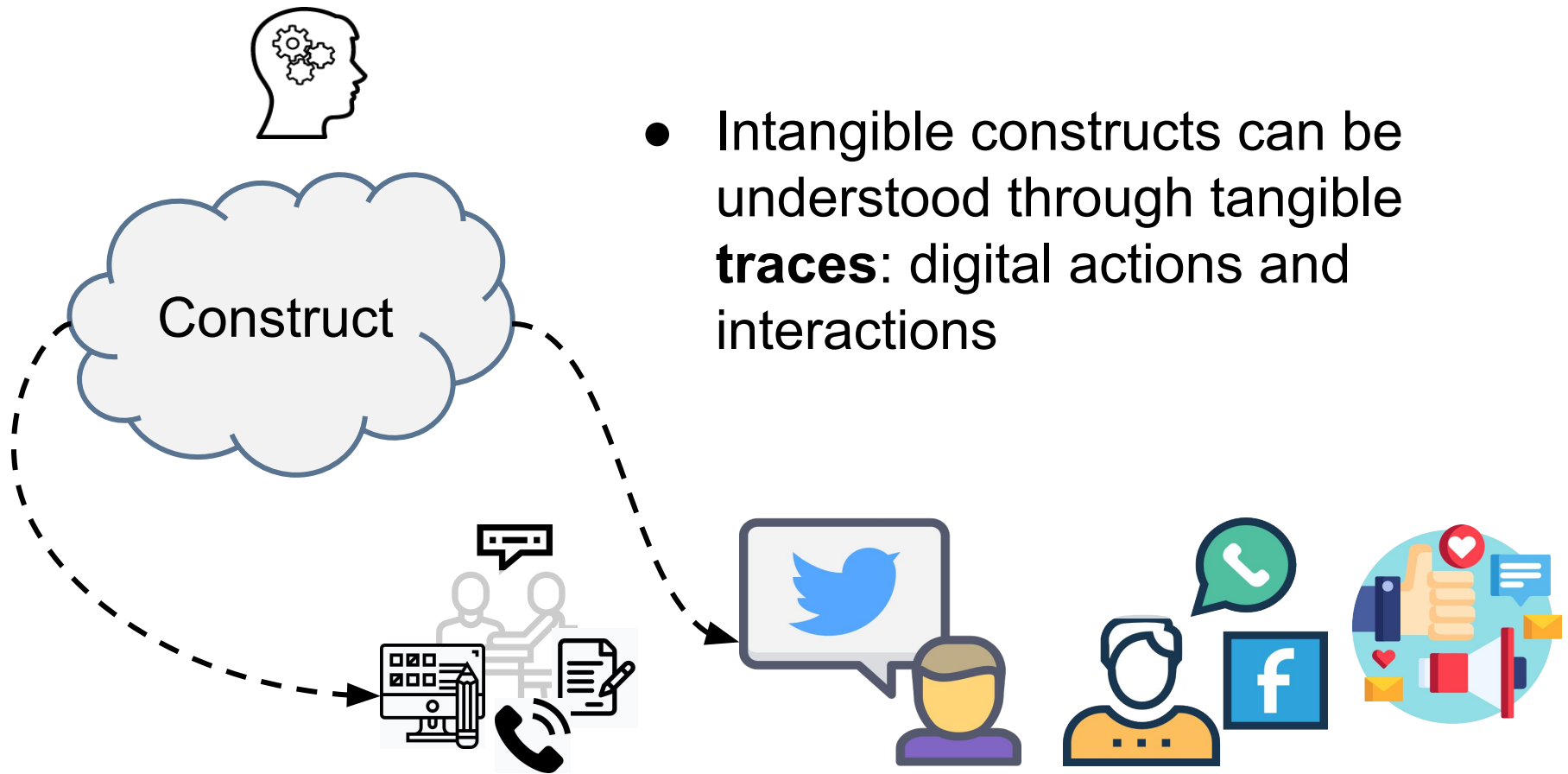


## “Platform Study”



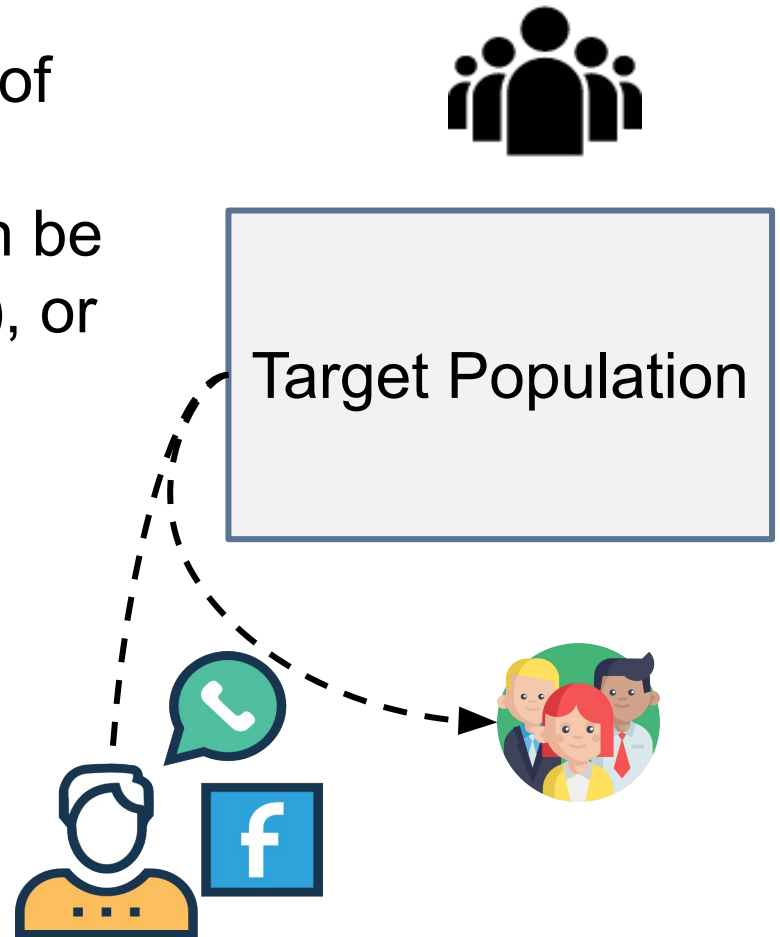


# Construct

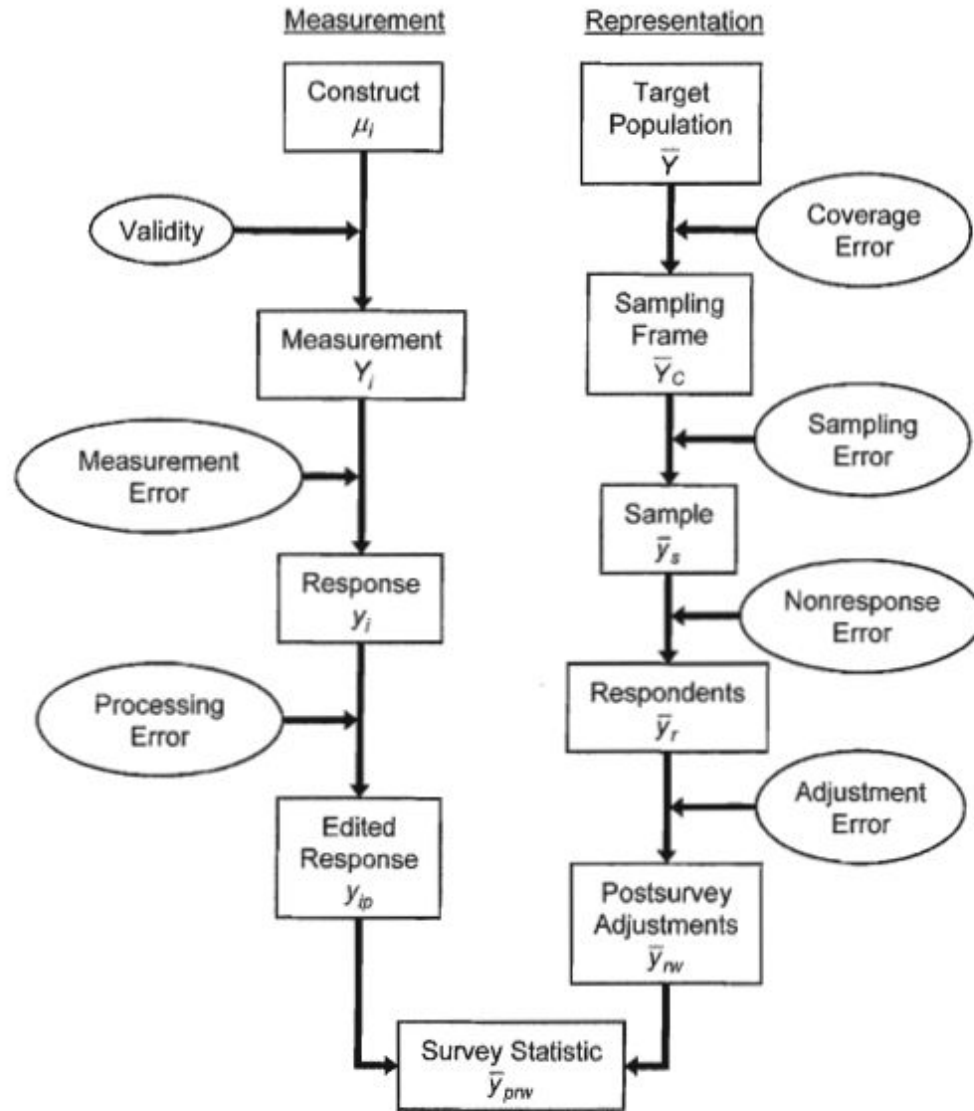


## Target population

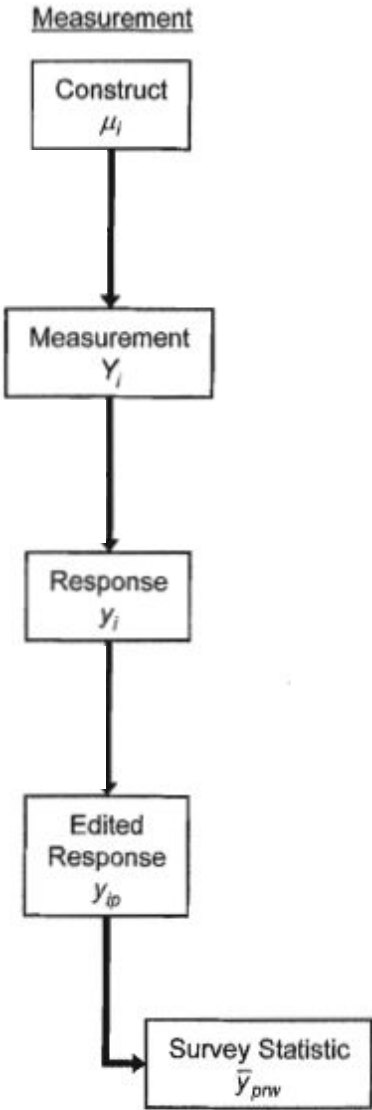
- **Users** are the representation of the target population (TP) elements in a record. This can be online platforms (profiles, IPs), or other digital forms (fitness trackers, cell numbers)



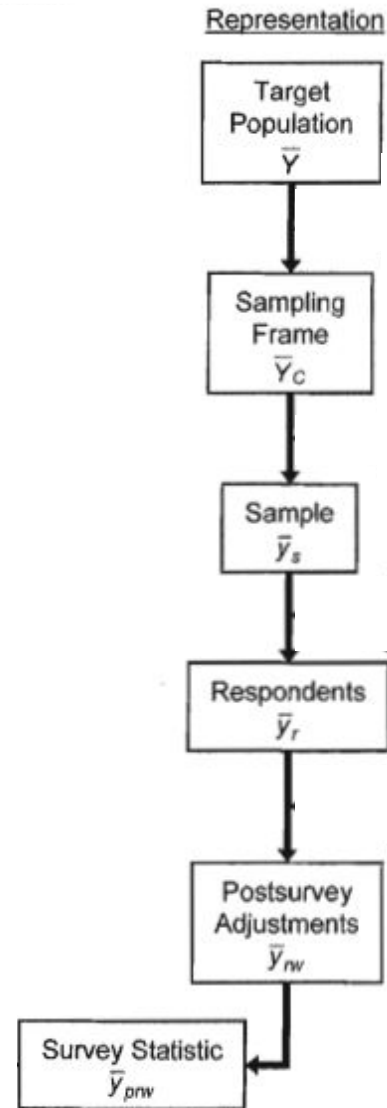
# The Total Survey Error Framework as a Starting Point



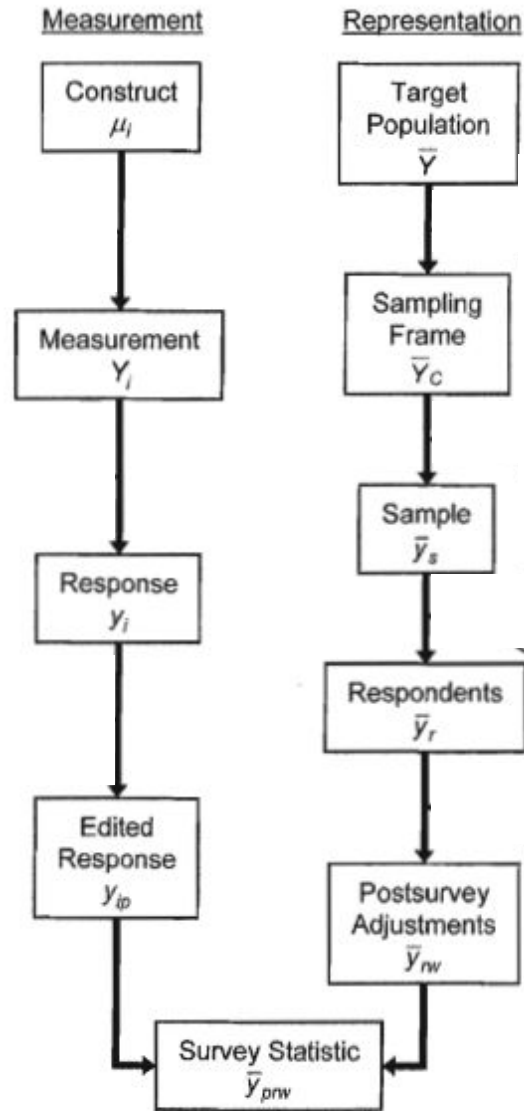
Survey Lifecycle from a Quality perspective. Adapted from Groves et al.(2011)



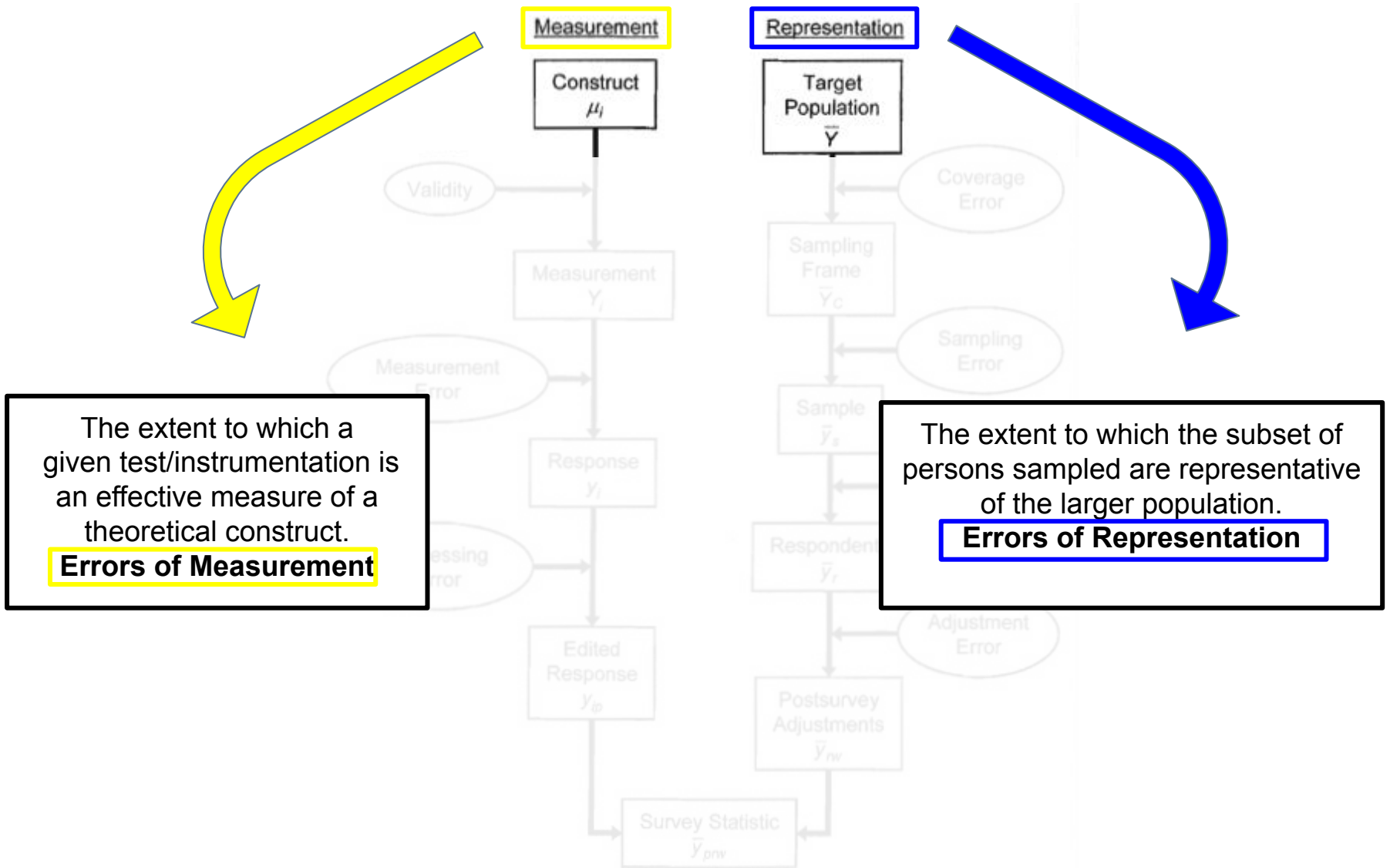
Survey Lifecycle from a Quality perspective. Adapted from Groves et al.(2011)



Survey Lifecycle from a Quality perspective. Adapted from Groves et al.(2011)



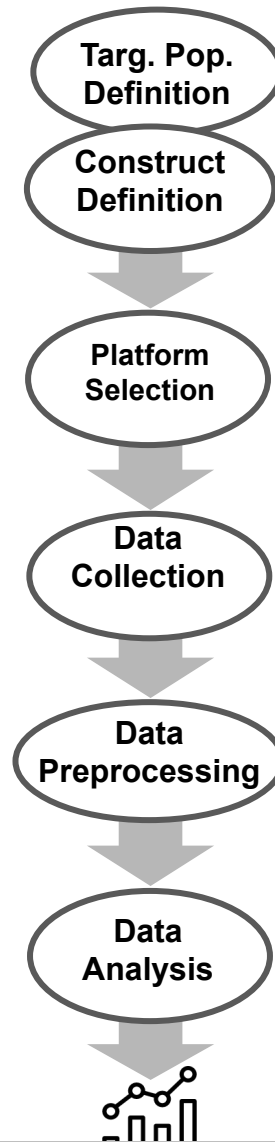
Survey Lifecycle from a Quality perspective. Adapted from Groves et al.(2011)



Survey Lifecycle from a Quality perspective. Adapted from Groves et al.(2011)



## A typical research pipeline with digital traces



“TED-On: A Total Error Framework for Digital Traces of Human Behavior on Online Platforms” Sen et al., 2021, Public Opinion Quarterly (to appear): <https://arxiv.org/pdf/1907.08228>

# A use case: Detecting the flu with digital traces

## Google Flu (trends)

nature

Vol 457 | 19 February 2009 | doi:10.1038/nature07634

# LETTERS

---

## **Detecting influenza epidemics using search engine query data**

Jeremy Ginsberg<sup>1</sup>, Matthew H. Mohebbi<sup>1</sup>, Rajan S. Patel<sup>1</sup>, Lynnette Brammer<sup>2</sup>, Mark S. Smolinski<sup>1</sup> & Larry Brilliant<sup>1</sup>

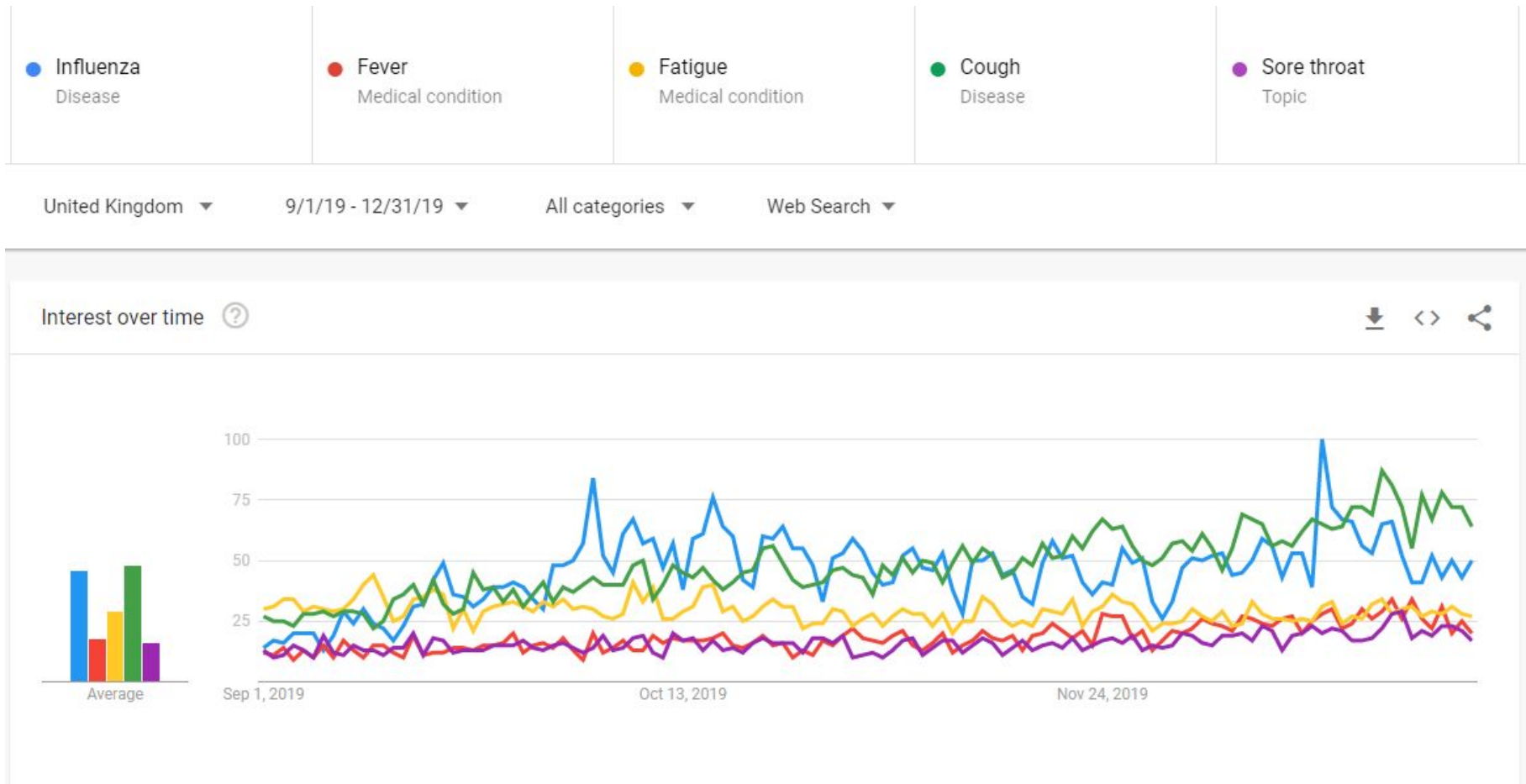
What proportion of US-Americans have the flu?

What is the approval rating of A. Merkel?

Are anti-immigration sentiments on the rise?

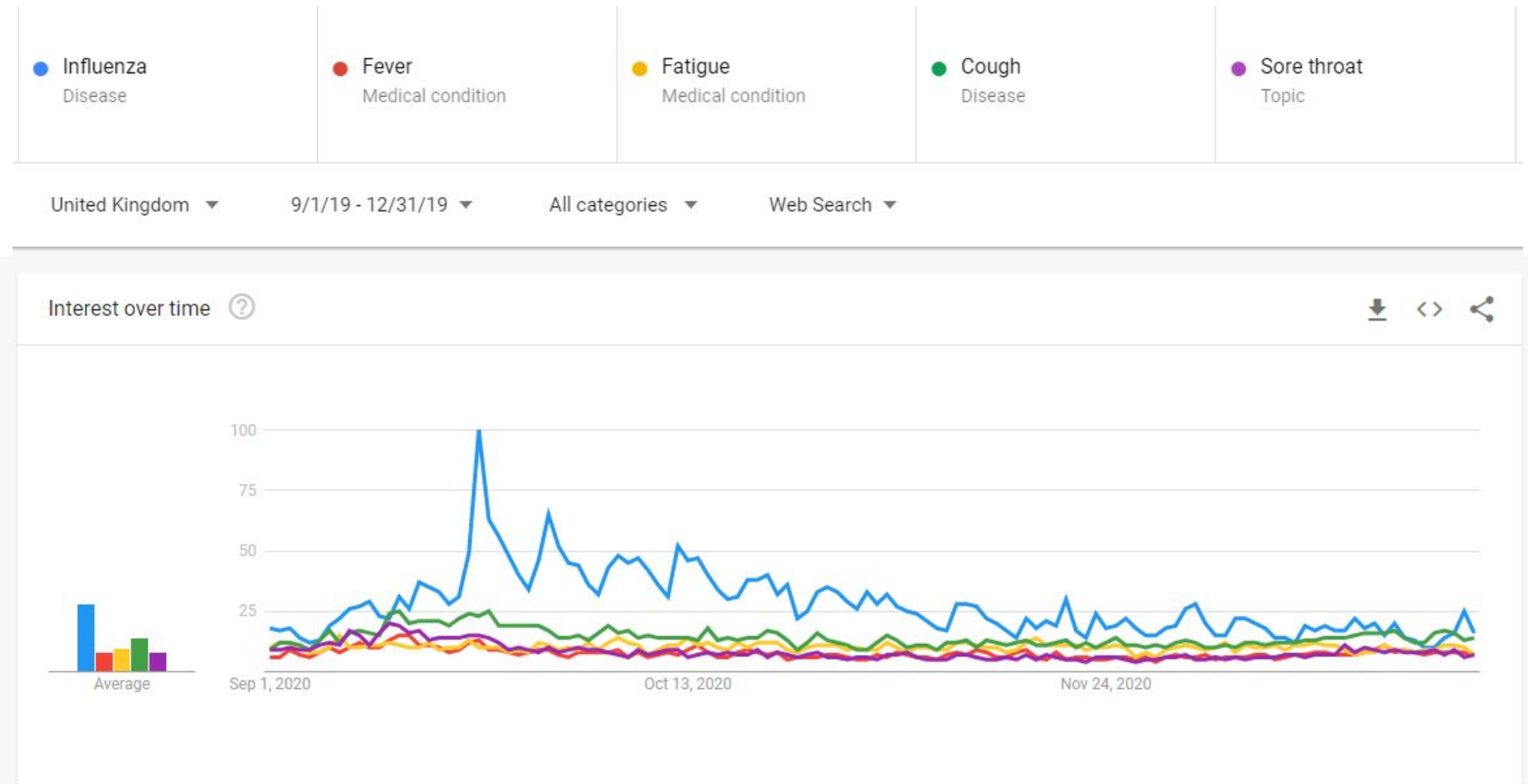
# Approach

## Trends in Google searches related to influenza-like illnesses



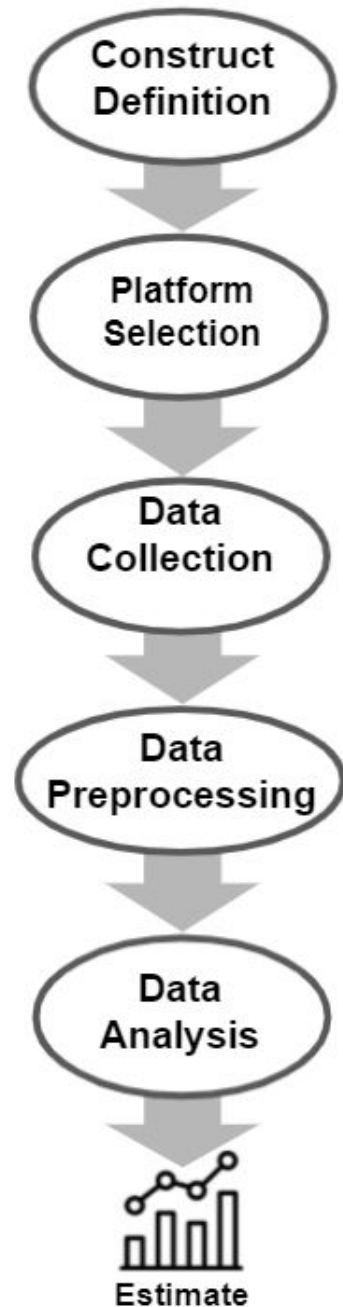
# Approach

## □ Trends in Google searches related to influenza-like illnesses



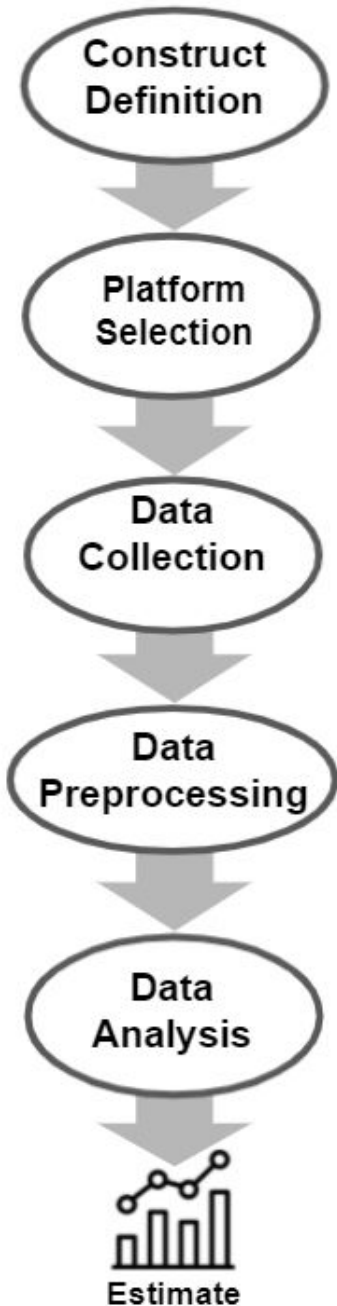
## Approach

- Construct: influenza-like illness (ILI)
- Target population(s): national and regional populations of the US
- Platform: Google web search logs 2003-2008, “hundreds of billions of individual searches”
- “Sample”: “About 90 million American adults are believed to search online for information about specific diseases or medical problems each year.” (in 2009)
- Data collection: “a time series of weekly counts for 50 million of the most common search queries in the United States” and per each state (concrete query terms not shared)
- Preprocessing: aggregated per each region, normalized by overall search activity, location via IP
- Analysis: Regression where DV= ILI doctor visits, IV= ILI-related search fractions, for 1152 data points



## Approach in a nutshell

- Construct: influenza-like illness (ILI)
- Target population(s): national and regional populations of the US
- Platform: Google web search
- Data collection: Logs 2003-2008, weekly, From 50 mio. search terms, select those that best predict.
- Preprocessing: aggregated per each region, normalized by overall search activity, location via IP.
- Analysis: Regression where DV= ILI doctor visits, IV= ILI-related search fractions, for 1152 data points



## Approach

### □ Result

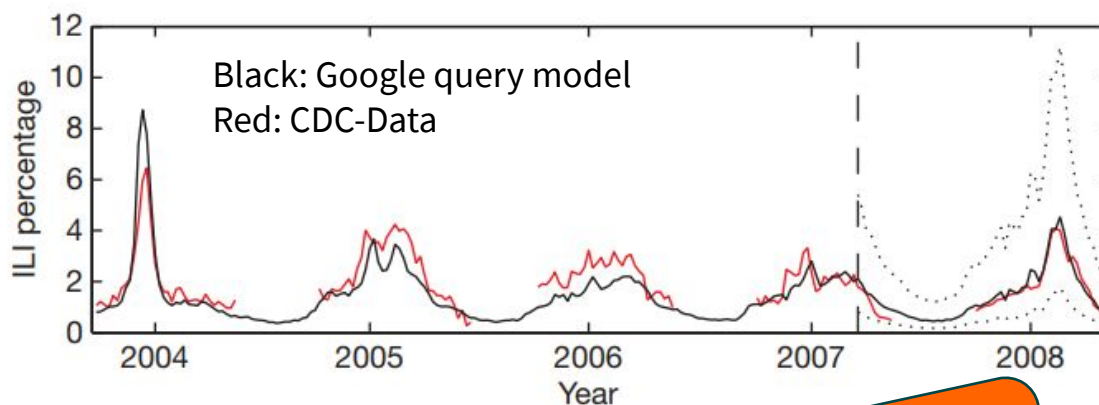


Figure 2 | A comparison of model estimates (black) against CDC-reported data (red) over 128 points. Dotted lines indicate 95% prediction intervals. The region comprises New York, New Jersey and Pennsylvania.

Good concurrent validity. Content validity of queries seems ok. + other validations.

□ Now possible: Now- or fore-casting, fine-grained local detection, other countries



# The fail



WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY SIGN IN SUBSCRIBE

CORONAVIRUS FAQs BEST FACE MASKS ESSENTIALS HOW TO DISINFECT EVERYTHING SYMPTOMS AND TESTING NEWSLETTER LATEST NEWS

DAVID LAZER RYAN KENNEDY OPINION 10.01.2015 07:00 AM

## What We Can Learn From the Epic Failure of Google Flu Trends

GFT seen searches

people's

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 494 > Issue 7436 > News > Article

NATURE | NEWS

عربي

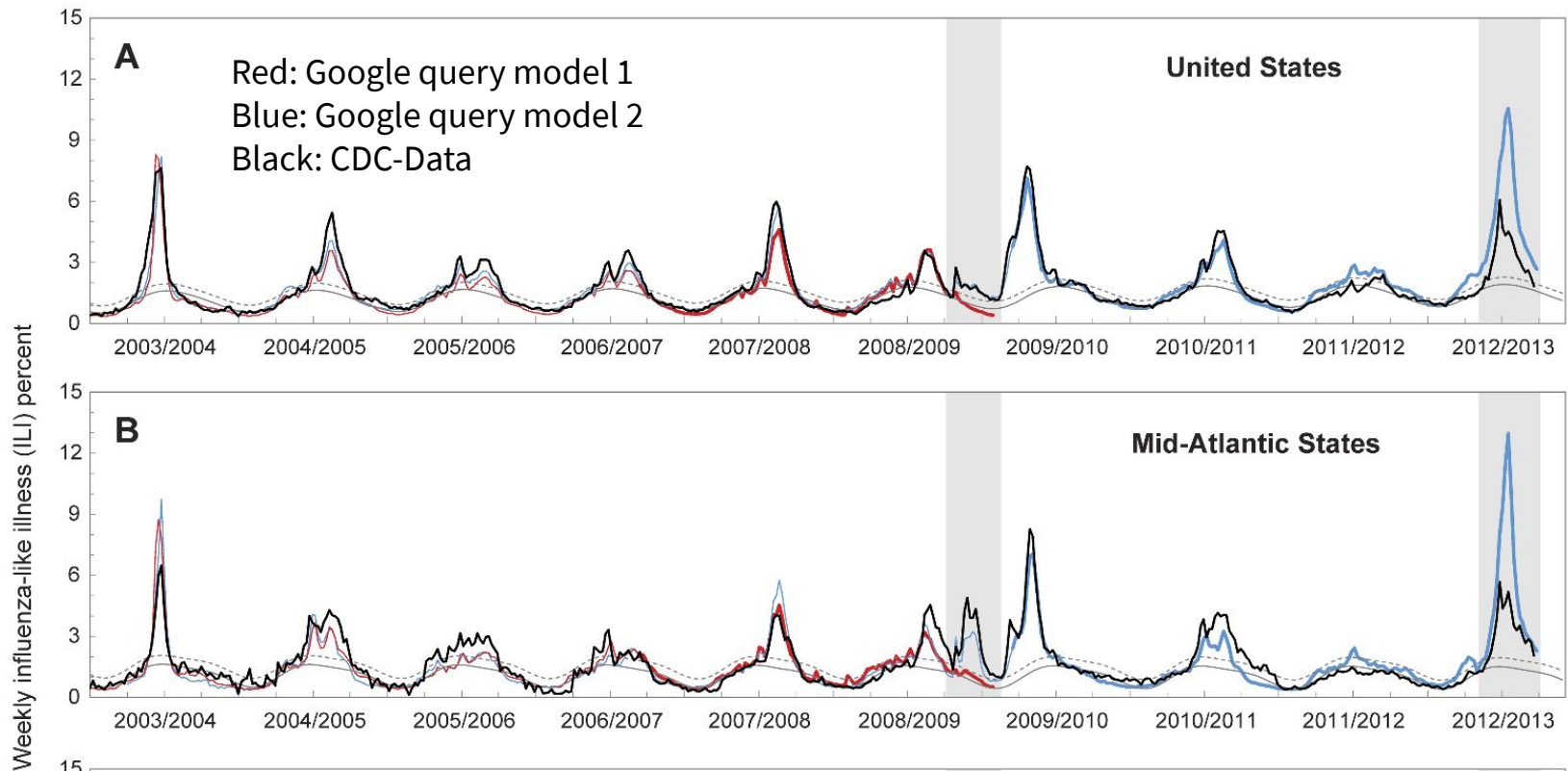
## When Google got flu wrong

US outbreak foxes a leading web-based method for tracking seasonal flu.

Declan Butler

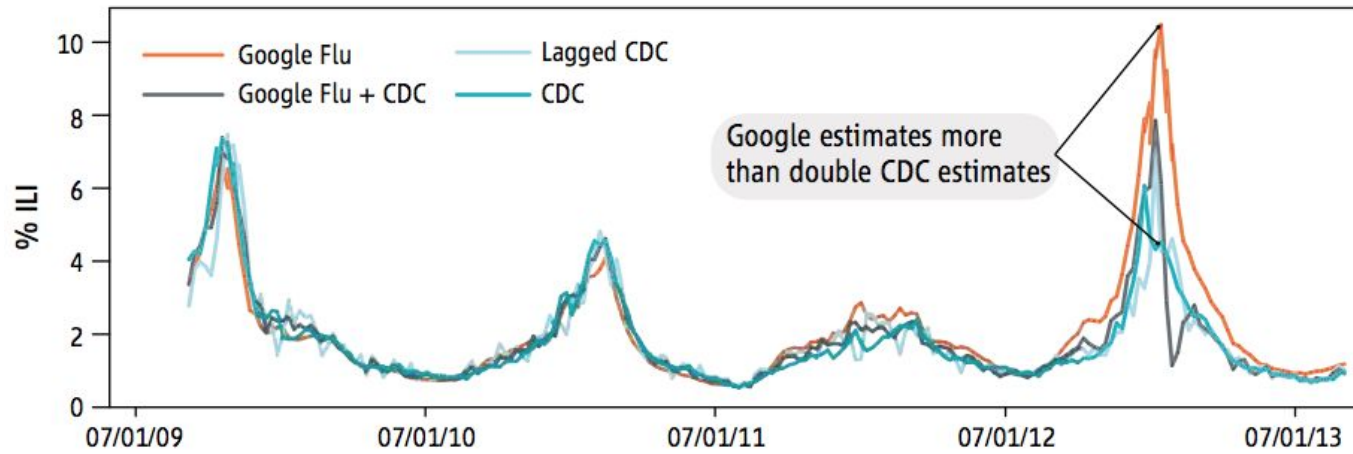
13 February 2013

# Overestimation



„part flue detector, part winter detector“ □ errors are auto-correlated & direction and magnitude change with seasons

# Overestimation



## The fail

- “**Big data hubris** is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis”

Lazer, David, et al. "The parable of Google Flu: traps in big data analysis." *Science* 343.6176 (2014)

## Where to be careful

“[...] there are enormous scientific possibilities in big data [but] the core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.”

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu, *Science*.

Let's dissect the process







Construct Definition

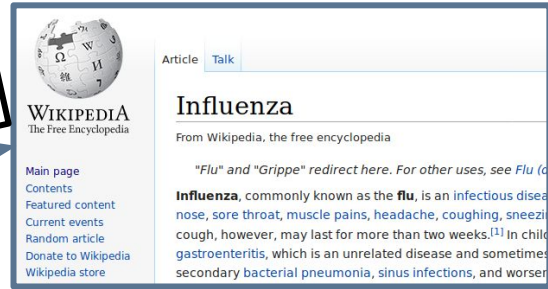
Platform Selection

Data Analysis

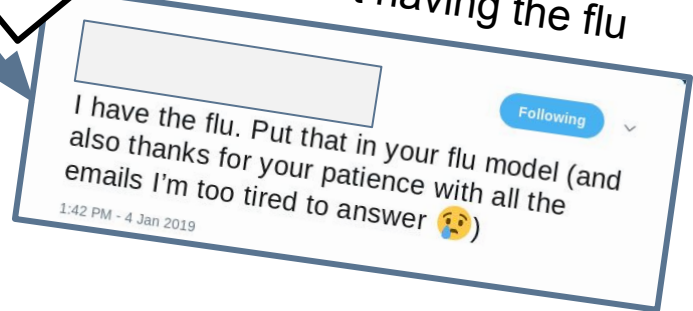
search queries related to flu



flu related information usage



posts about having the flu



TRACES

TRACES

TRACES



feeling flu-like symptoms



Construct Definition

Platform Selection

Data Analysis



information search

search queries related to flu



feeling flu-like symptoms

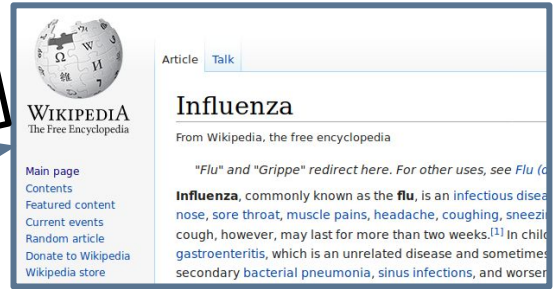


TRACES

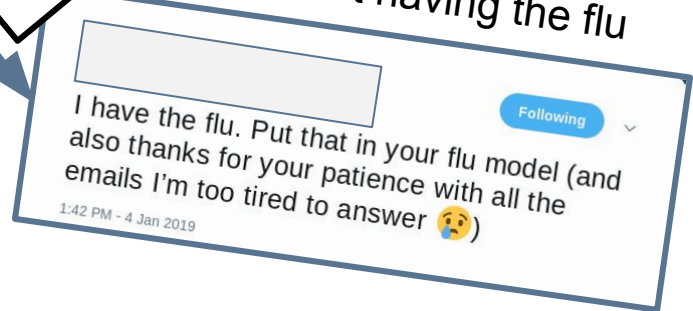
TRACES

TRACES

flu related information usage



posts about having the flu



Construct Definition

Platform Selection



information search

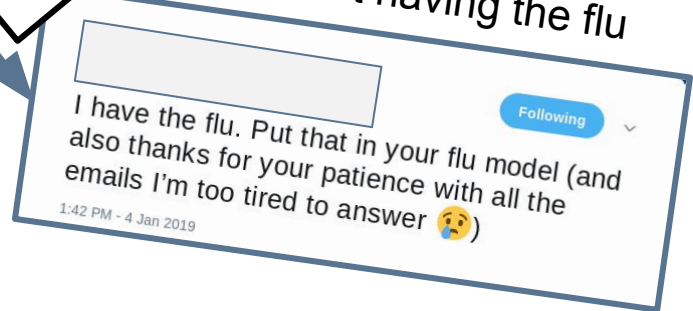
search queries related to flu



flu related information usage



posts about having the flu



feeling flu-like symptoms

Coronavirus confirmed as pandemic by World Health Organization

© 11 March 2020

43

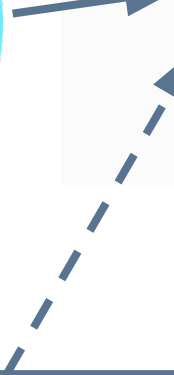
media coverage



TRACES

TRACES

TRACES



Construct Definition

Platform Selection

Construct Definition

Platform Selection

**MEASUREMENT**

**Validity**

**Construct  
Definition**

**What does searching for  
flu-related terms actually  
mean?  
Media coverage/hype?  
Infected peers?  
Similar  
illness/symptoms?**

Platform  
Selection

Data  
Collection

Data  
Preprocessing

Data  
Analysis

MEASUREMENT

Validity

Construct  
Definition

What does searching for  
flu-related terms actually  
mean?  
Media coverage/hype?  
Infected peers?  
Similar  
illness/symptoms?

Platform  
Selection

Data  
Collection

= The insufficient linkage between a clearly defined  
construct and an envisioned measurement

Data  
Preprocessing

Data  
Analysis

Construct  
Definition



**Platform  
Selection**



Data  
Collection



Data  
Analysis



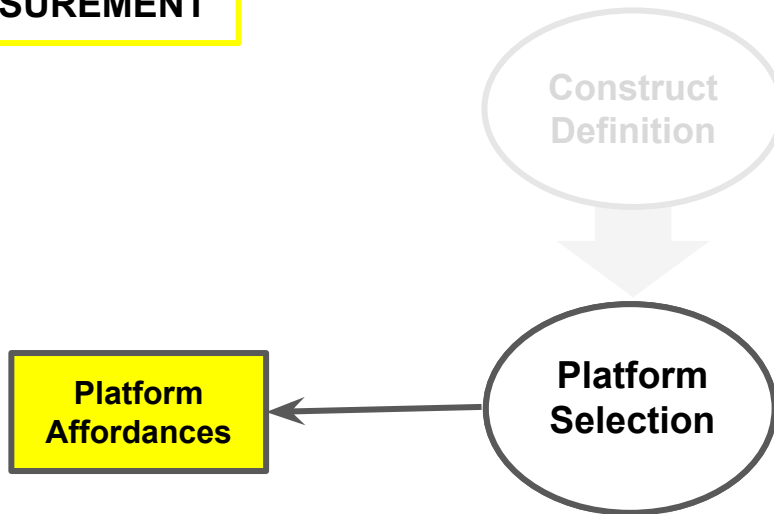


- Recommendations for search queries
- The algorithm is changing!
- Data not publicly available



- 280-character limit
- Trending tweets can stoke imitation behavior
- Free API is limited

## MEASUREMENT



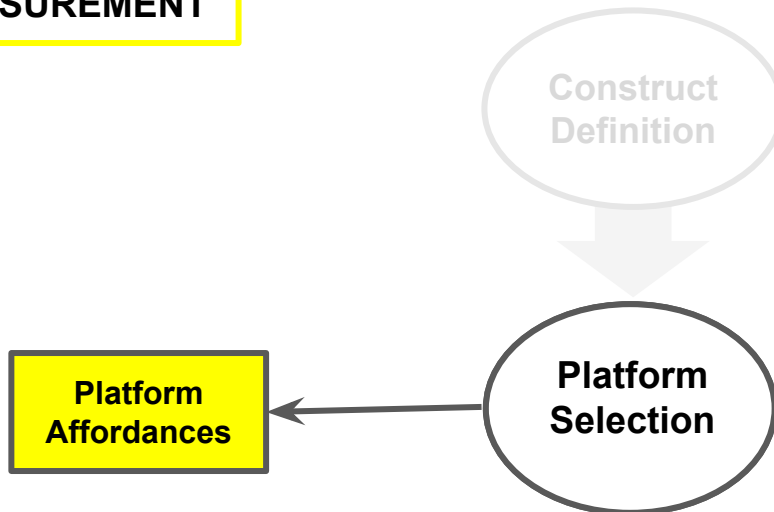
- Recommendations for search queries
- The algorithm is changing!
- Data not publicly available



- 280-character limit
- Trending tweets can stoke imitation behavior
- Free API is limited



## MEASUREMENT



= The gap between the 'true' traces and traces distorted by platform affordances  technical or community standards, terms & conditions



- Recommendations for search queries
- The algorithm is changing!
- Data not publicly available



- 280-character limit
- Trending tweets can stoke imitation behavior
- Free API is limited





Platform population = people who access Google (and search for health topics)



Platform population = people with Twitter accounts (and write about health topics)



younger, male,  
western, educated



Platform population = people who  
access Google  
(and search for health topics)



Platform population =  
people with Twitter  
accounts  
(and write about health  
topics)

**REPRESENTATION**

Construct Definition

younger, male,  
western, educated



Platform Selection

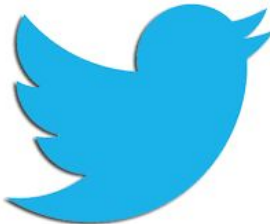
Platform Coverage

Data Collection



Platform population = people who access Google (and search for health topics)

Data Preprocessing

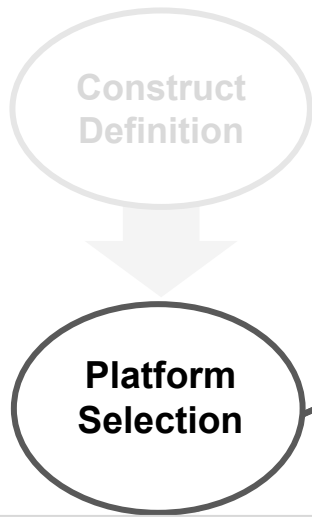
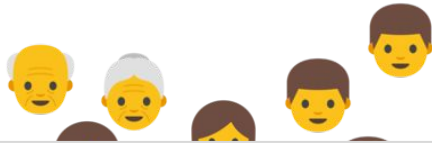


Platform population = people with Twitter accounts (and write about health topics)

Data Analysis

**REPRESENTATION**

younger, male,  
western, educated



**Platform Coverage**

= The gap between the target population (US Population) and the platform population



Platform population = people who access Google (and search for health topics)



Platform population = people with Twitter accounts (and write about health topics)





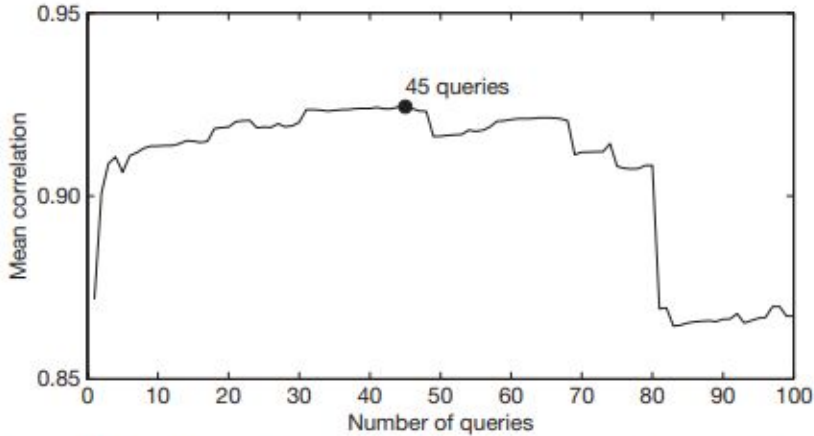


Google Flu: “Automated” + manual --- pick search queries correlated with flu rates





Construct  
Definition



Form  
ation



Google Flu: “Automated” +  
manual --- pick search queries  
correlated with flu rates

Data  
Collection

Data  
Preprocessing

Data  
Analysis



Manual:

Tweets with keywords related to Influenza: Avian influenza, Influenza Virus B, Centers for Disease Control and Prevention, Common Cold, Vaccine, Influenza

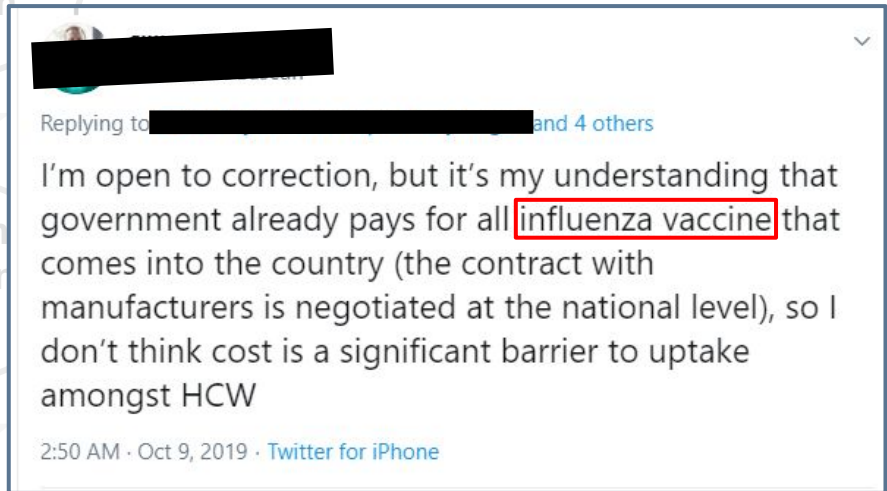
Construct Definition

Platform Selection

Data Collection

Data Preprocessing

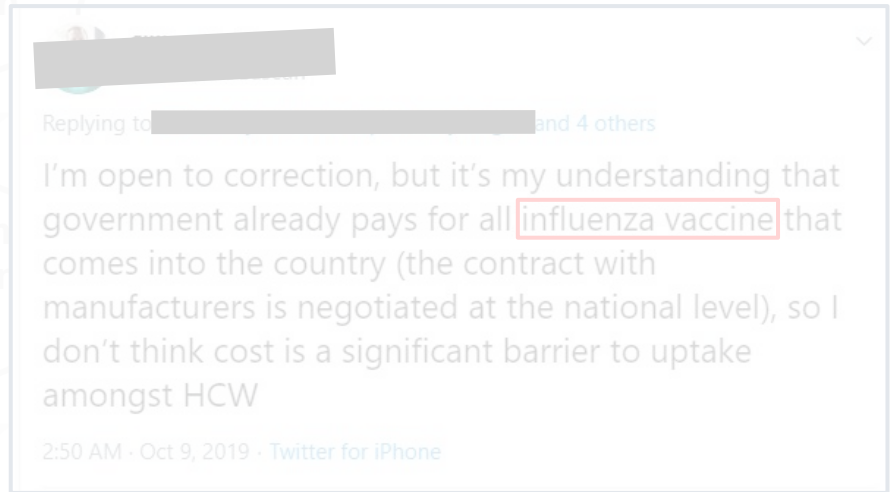
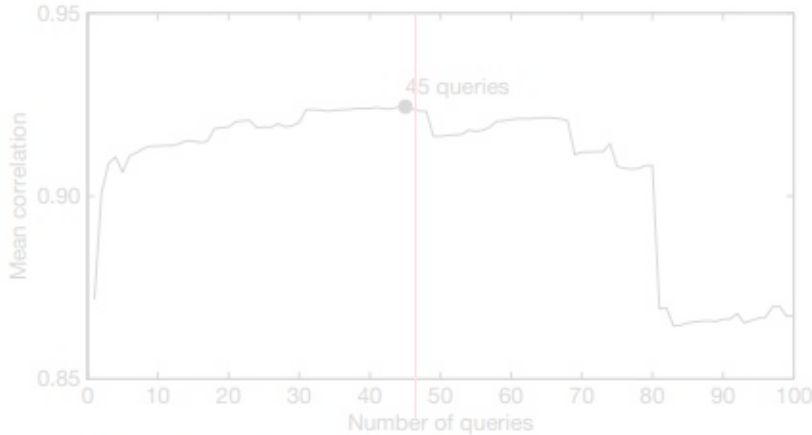
Data Analysis



Manual:

Tweets with keywords related to Influenza: Avian influenza, Influenza Virus B, Centers for Disease Control and Prevention, Common Cold, Vaccine, Influenza

# MEASUREMENT



**Trace Selection**

**Data Collection**



Google Flu: „Automated“ + manual

More common: Predefined, e.g. by keywords

Tweets with keywords related to Influenza: Avian influenza, Influenza Virus B, Centers for Disease Control and Prevention, Common Cold, Vaccine, Influenza

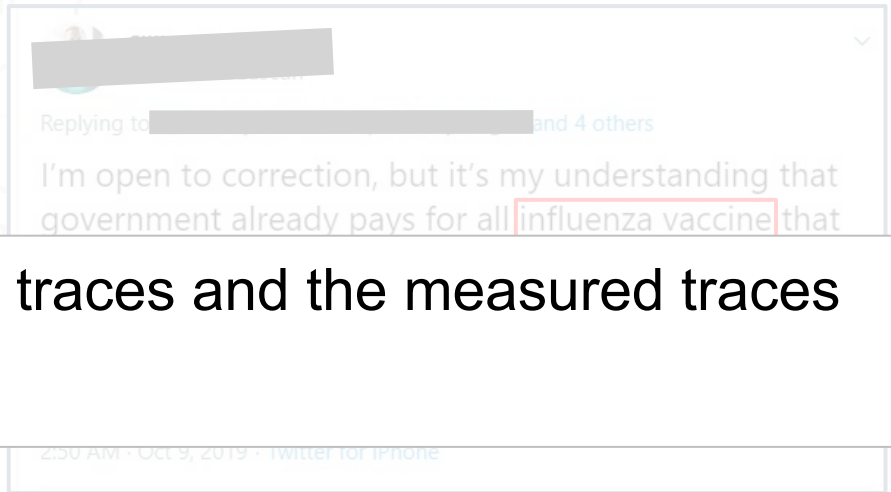
Data Preprocessing

Data Analysis

# MEASUREMENT



Construct Definition



= The gap between the relevant traces and the measured traces

**Trace Selection**

**Data Collection**



Google Flu: „Automated“ + manual

More common: Predefined, e.g. by keywords

Tweets with keywords related to Influenza: Avian influenza, Influenza Virus B, Centers for Disease Control and Prevention, Common Cold, Vaccine, Flu(the Band), Influenza

Data Preprocessing

Data Analysis



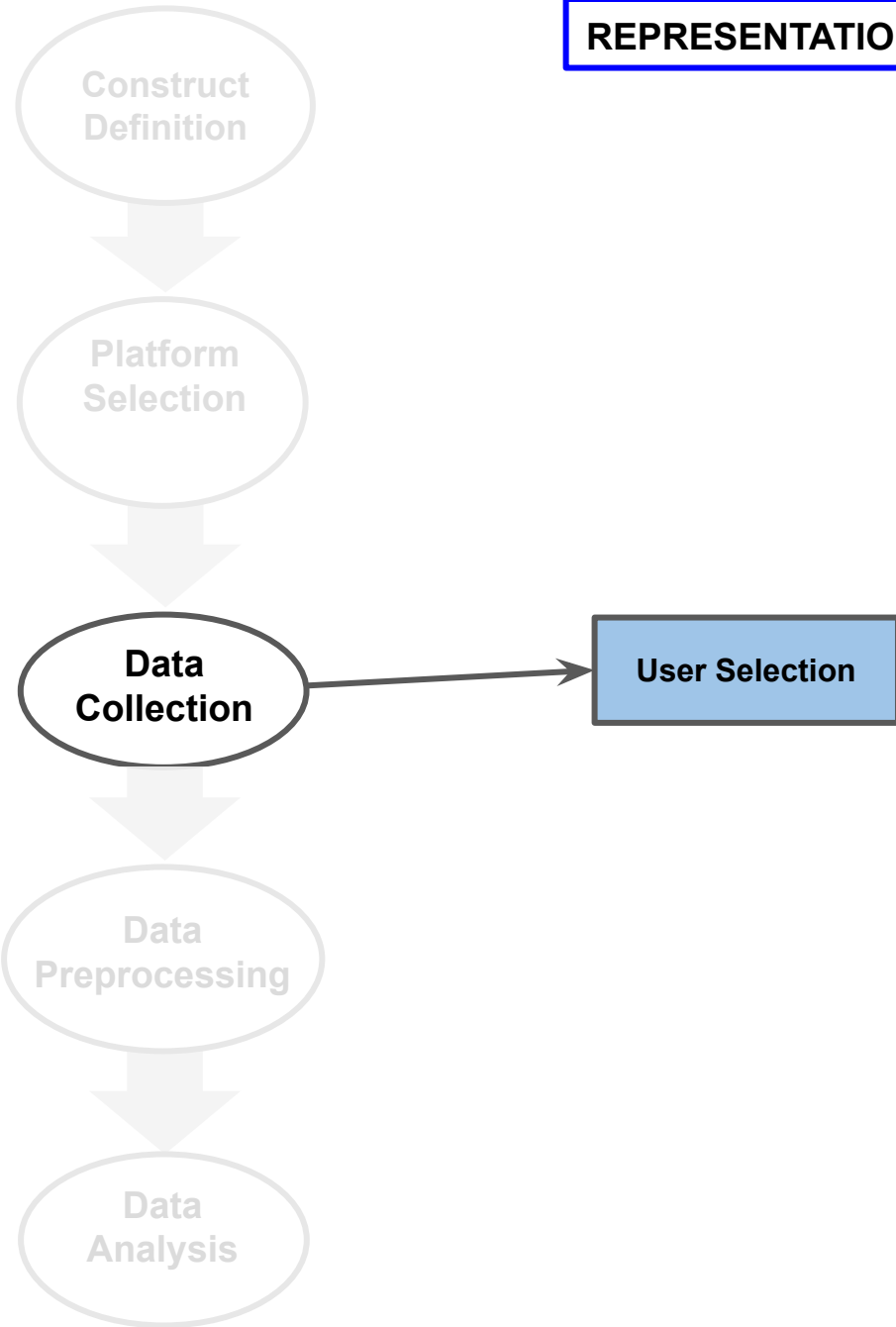
1. „Fever remedy“
2. „Throat hurts“
3. „Symptoms flu“
4. „What to do against flu“
5. „Healing flu“
- ...
61. „fiebre que hacer“
62. „síntomas influenza“

Google



**REPRESENTATION**

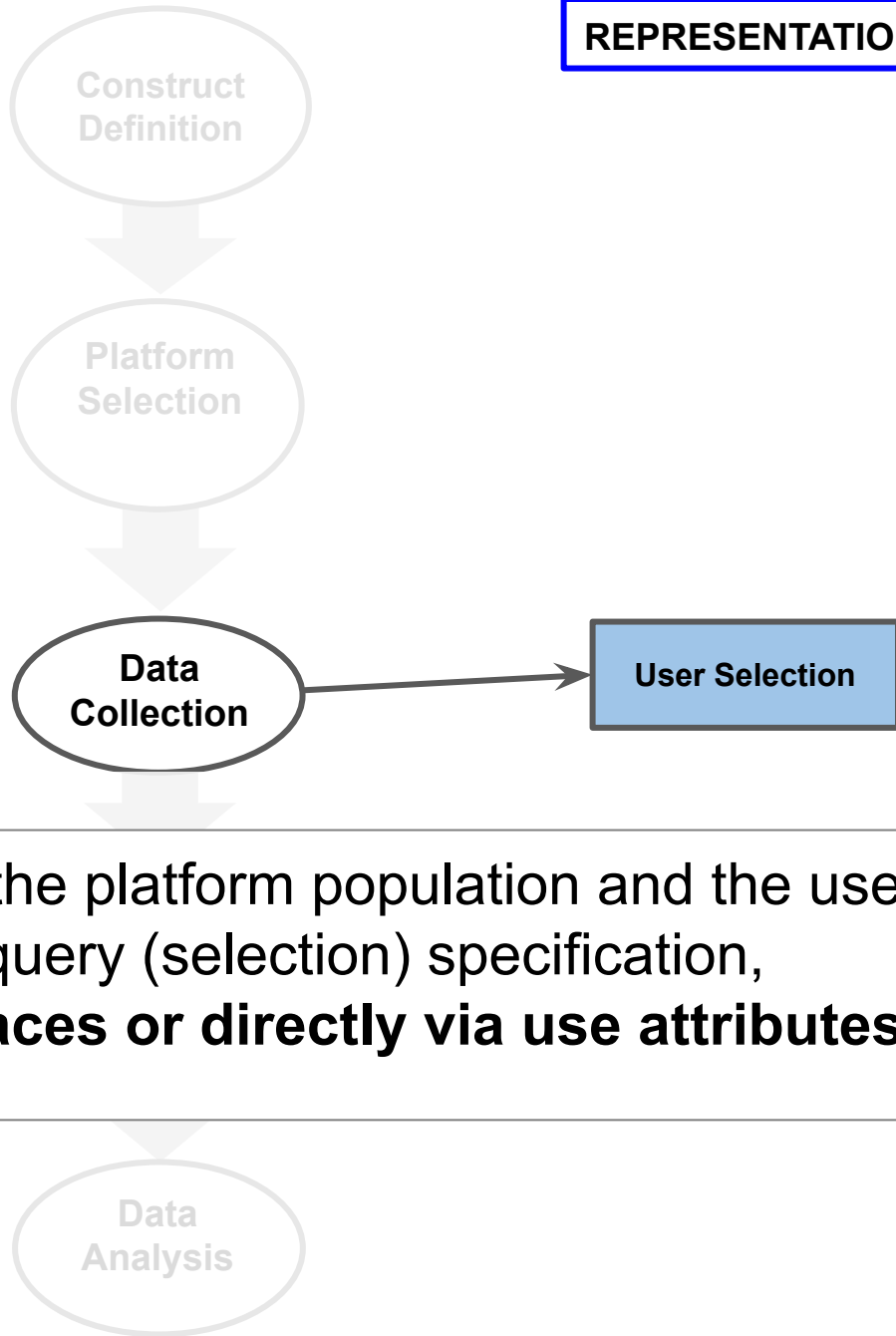
- 1. „Fever remedy“
- 2. „Throat hurts“
- 3. „Symptoms flu“
- 4. „What to do against flu“
- 5. „Healing flu“
- ...
- 61. „fiebre que hacer“
- 62. „síntomas influenza“





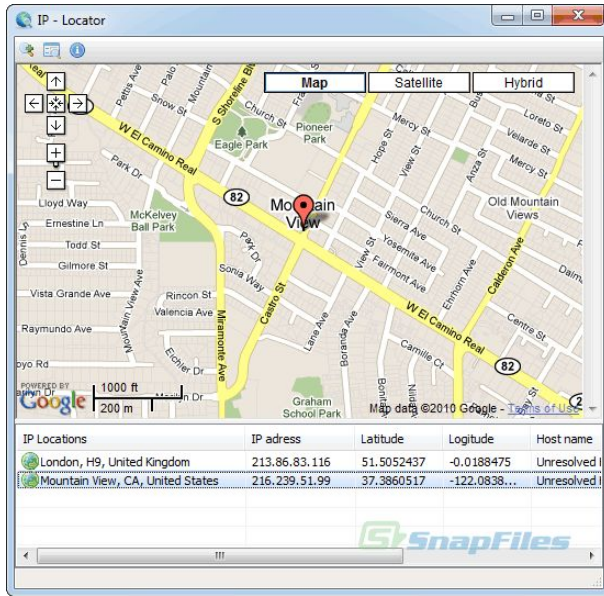
**REPRESENTATION**

- 1. „Fever remedy“
- 2. „Throat hurts“
- 3. „Symptoms flu“
- 4. „What to do against flu“
- 5. „Healing flu“
- ...
- 61. „fiebre que hacer“
- 62. „síntomas influenza“



= The difference between the platform population and the users chosen due to the query (selection) specification, **indirectly via traces or directly via use attributes**





Construct

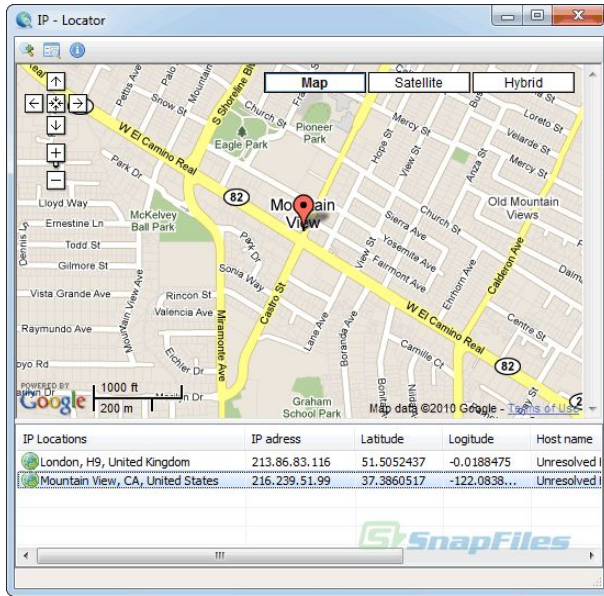
A screenshot of a Twitter profile. The profile picture is a circular image of a person with dark curly hair wearing a yellow shirt. To the right of the profile picture are three icons: a three-dot menu, an envelope icon, and a 'Follow' button. Below the profile picture, the name is redacted with a black bar, and the pronouns 'she/her' are circled in red. Below the name, the bio reads: 'Data Scientist 🧑‍💻 Speaker 🎤 Mentor 🧡 Founder @ [redacted] 🌴 I'll help you with your career, but these hands are still rated E for Everybody'. Below the bio, the location 'San Diego, CA' is circled in red, followed by a link icon, a redacted URL, a calendar icon, and the text 'Joined June 2019'.

Data  
Collection

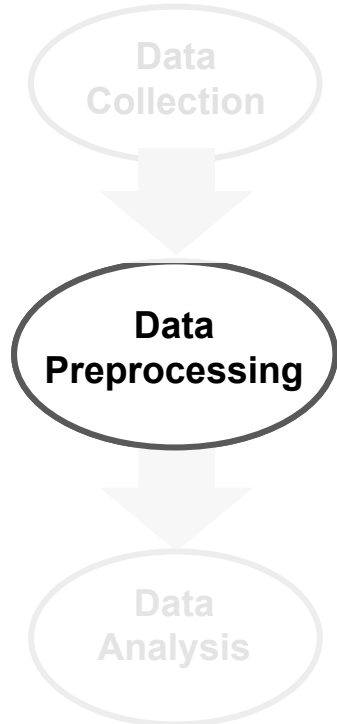
Data  
Preprocessing

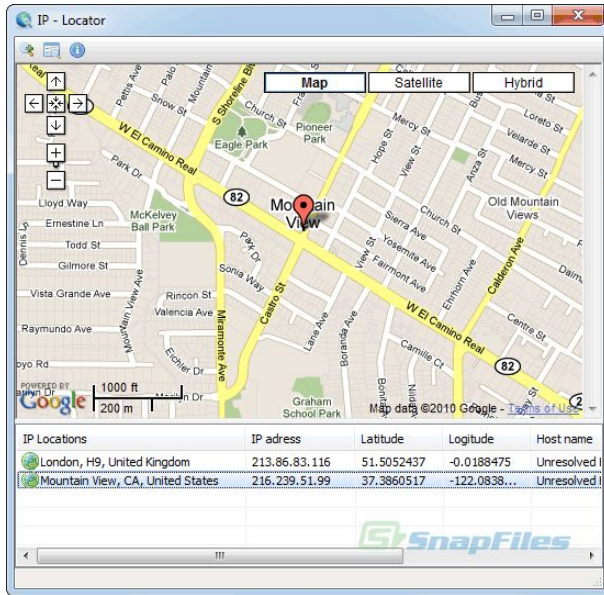
Data  
Analysis



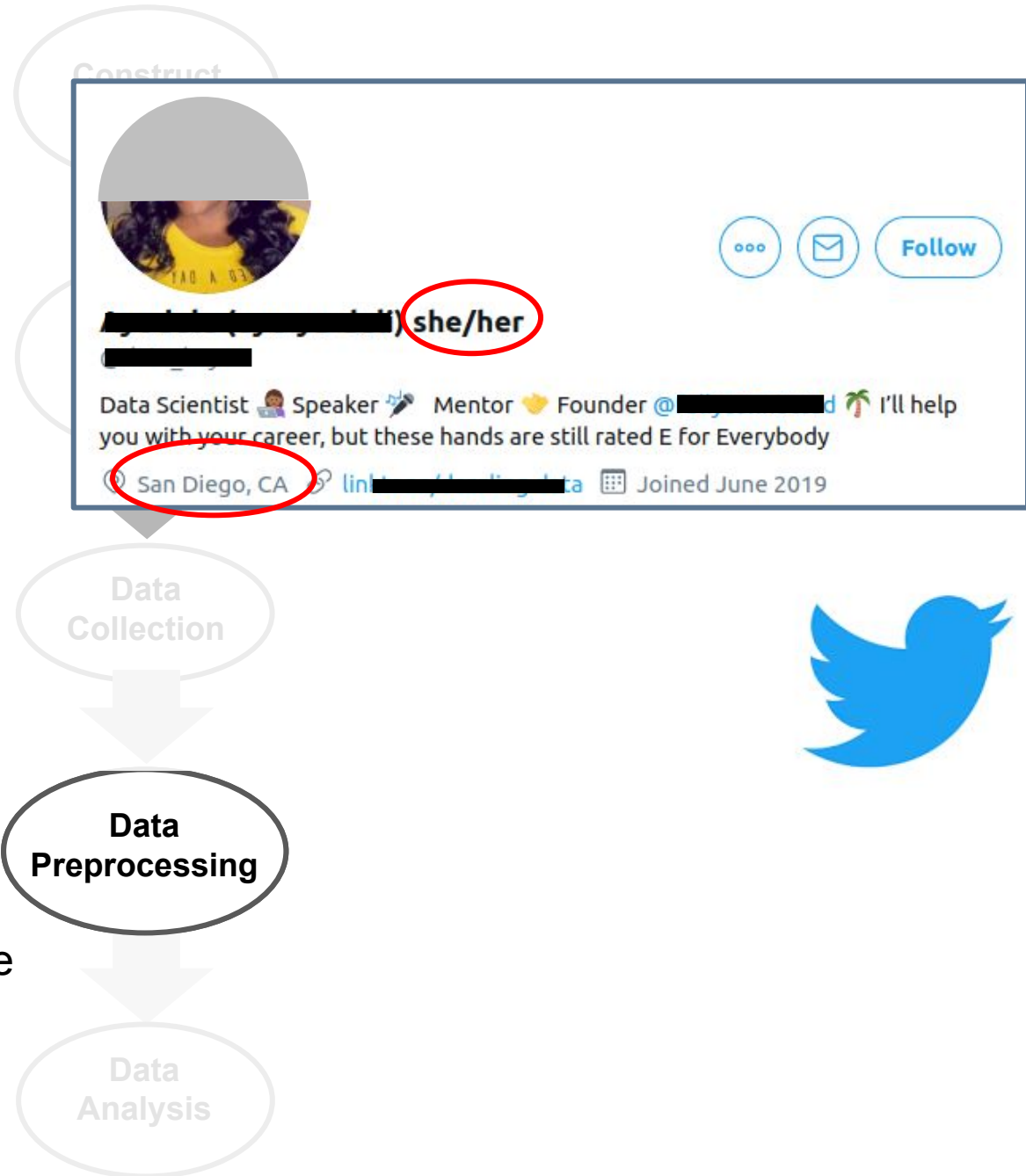


Construct

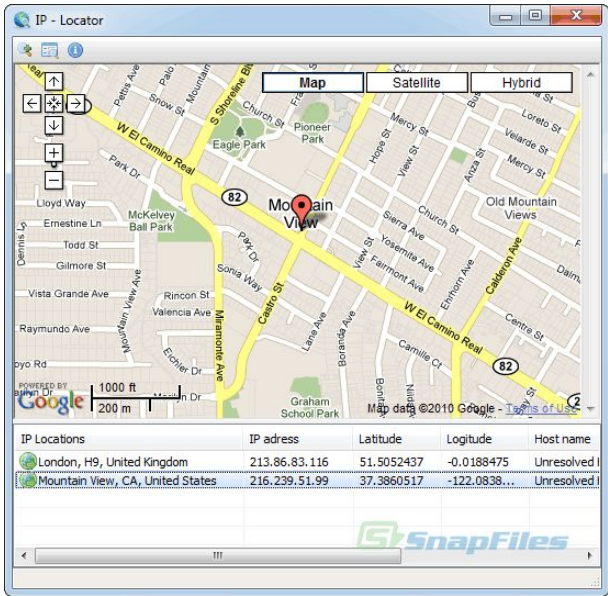




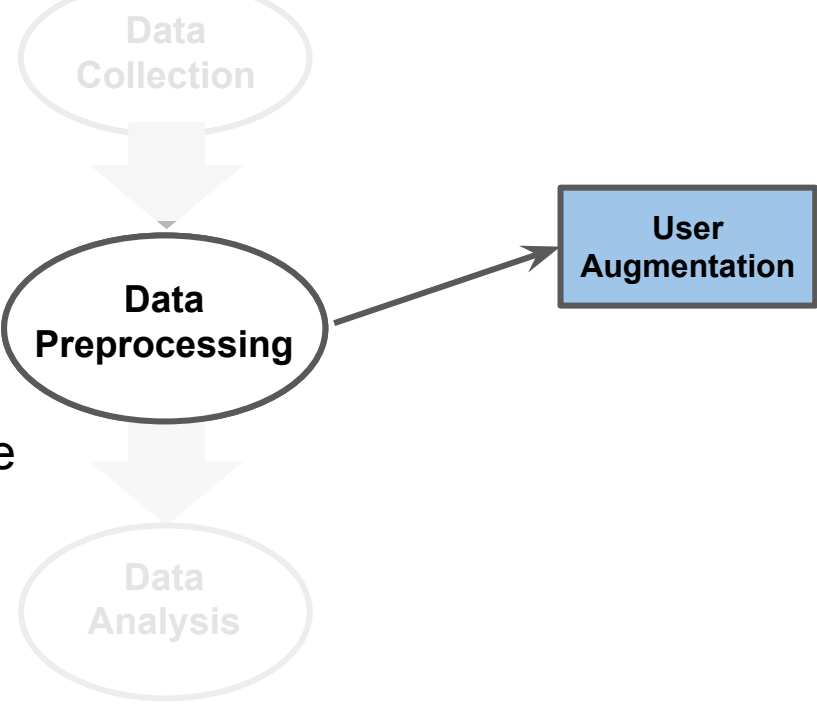
We augment user characteristics, often, automatically, through machine learning methods



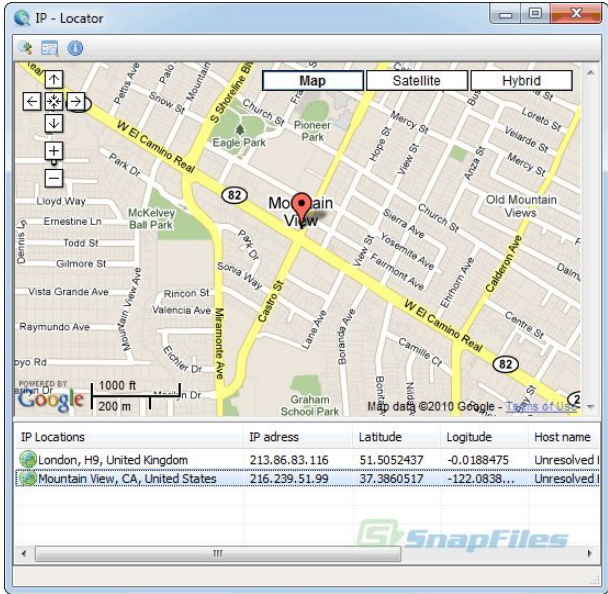
# REPRESENTATION



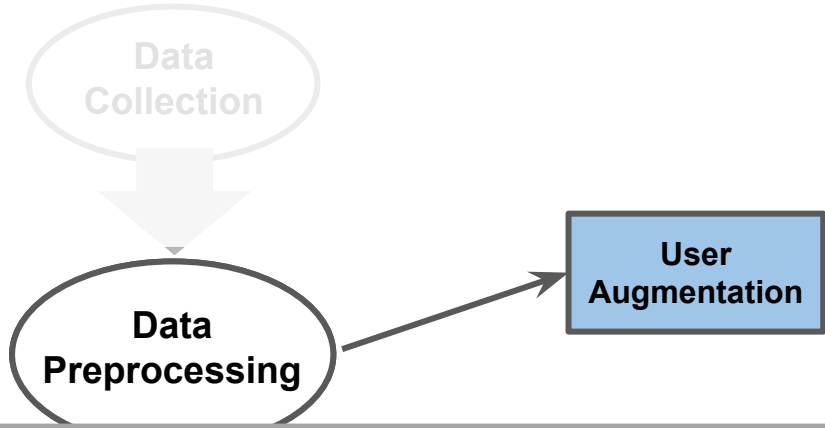
We augment user characteristics, often, automatically, through machine learning methods



# REPRESENTATION



We augment user characteristics often



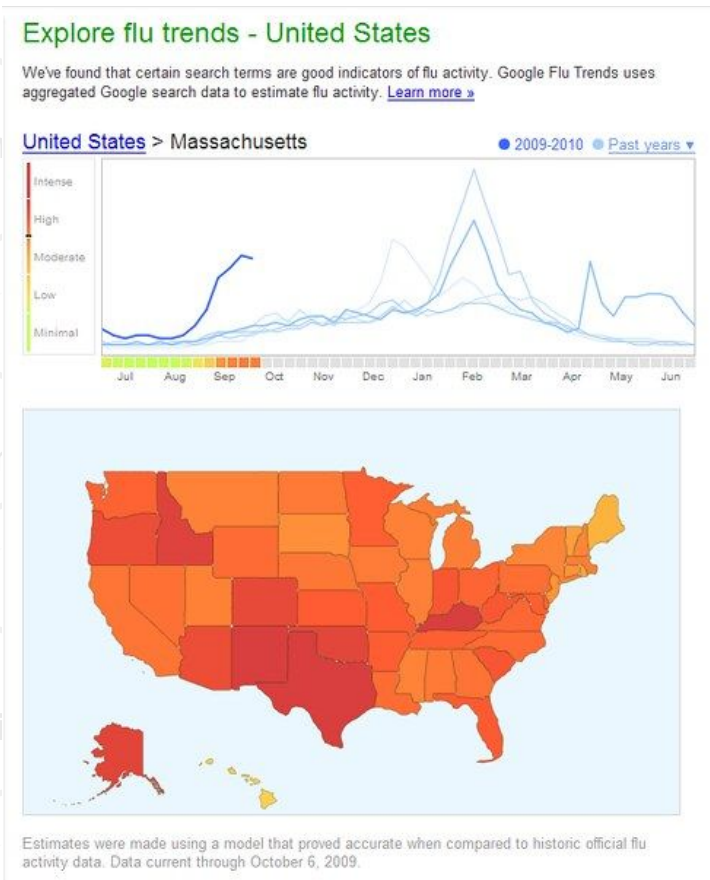
= The potential issues in the study population due to wrongly inferring user characteristics







Use demographic data, location to correct for coverage errors using post-stratification



**REPRESENTATION**

Construct Definition

Platform Selection

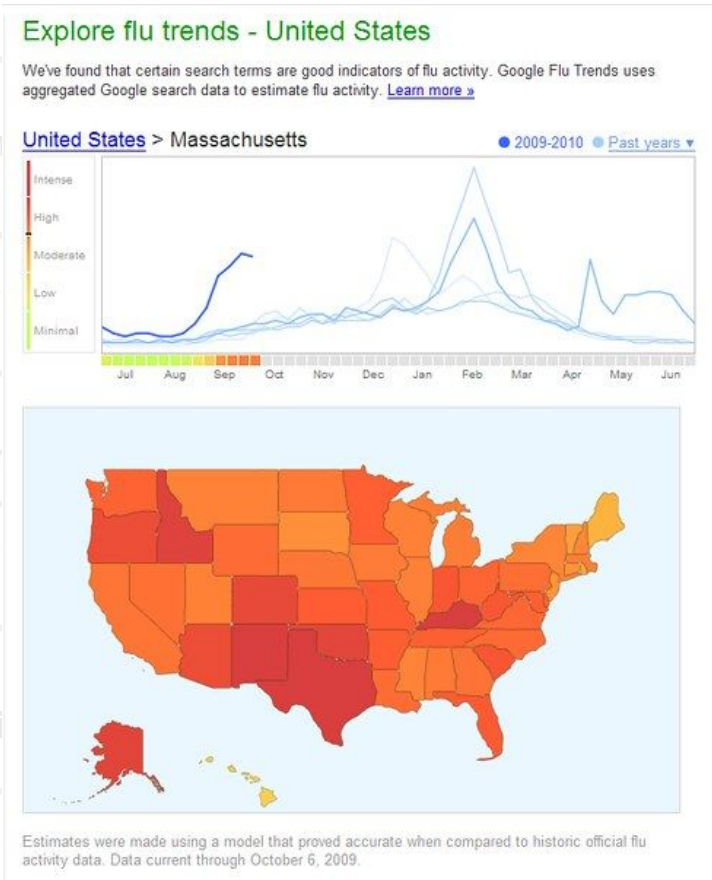
Data Collection

Data Preprocessing

**Data Analysis**

**Adjustment**

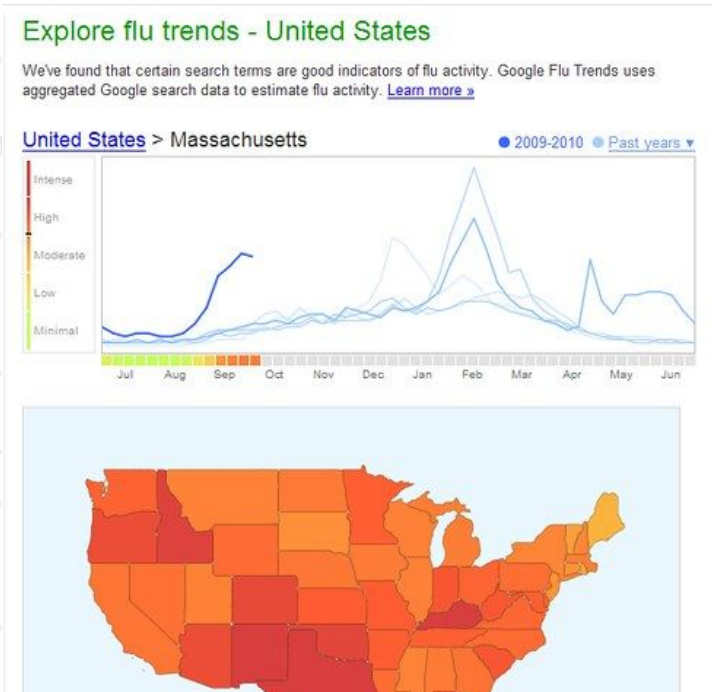
Use demographic data, location to correct for coverage errors using post-stratification



**REPRESENTATION**



Use demographic data, location to correct for coverage errors using post-stratification



= Errors due to correcting for representation errors through reweighting

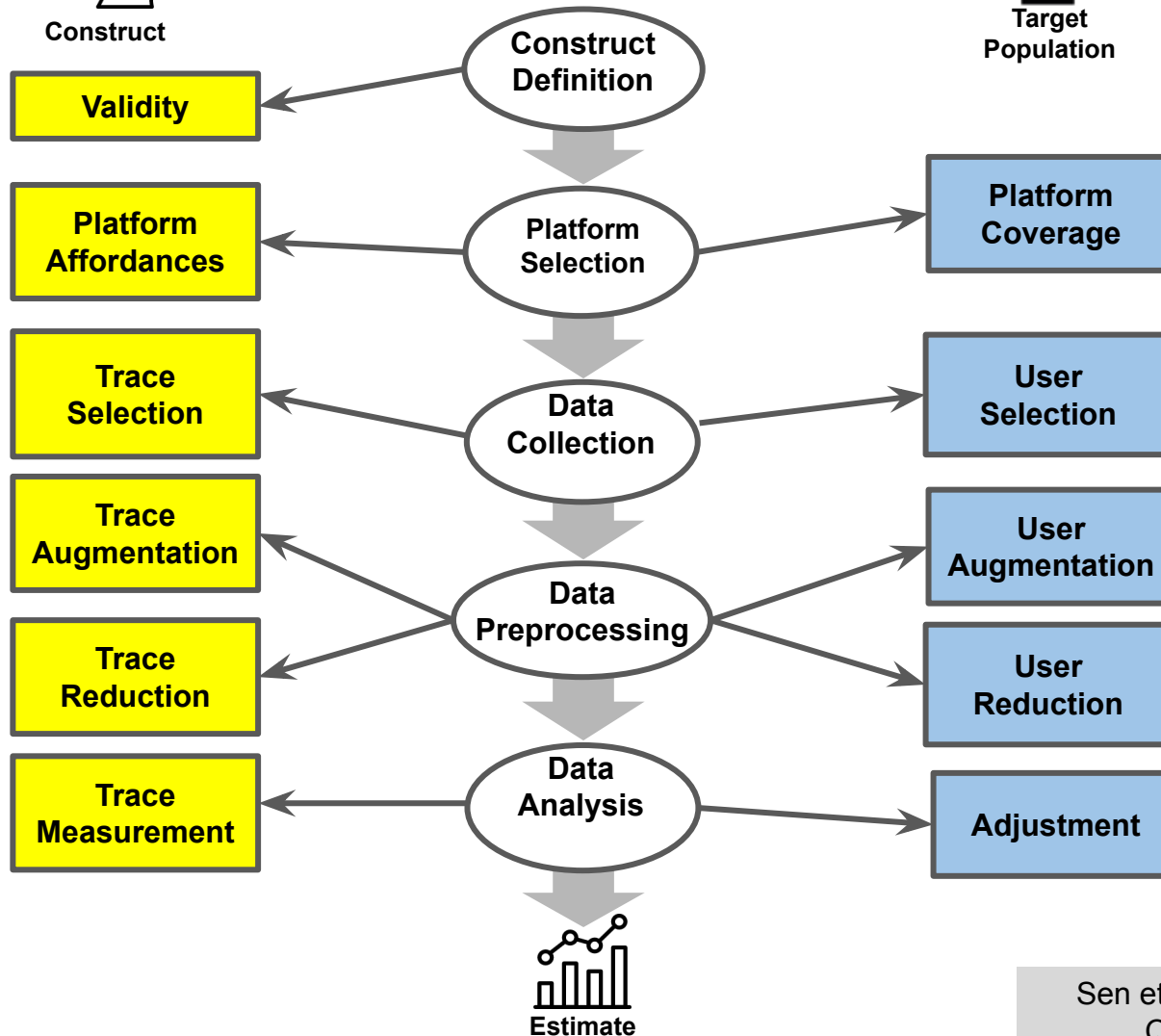


# TED: A Total Error Framework for Digital Traces of Humans

Errors of measurement



Errors of representation



Thank you for your attention!

gesis

Leibniz-Institut  
für Sozialwissenschaften

Leibniz  
Gemeinschaft

## GESIS offers Consulting

GESIS offers individual consulting in a number of areas – including survey design & methodology, data archiving, digital behavioral data & computational social science – and across the research data cycle. Please visit our website [www.gesis.org](http://www.gesis.org) for more [detailed information](#).

GESIS consulting is *free of charge* for researchers who conduct

- scientific projects – financed institutionally or by third-party-funds – at universities or publicly funded research institutions, or
- scientific projects at institutions of the Federal Government or the *Länder* or other publicly funded institutions.


For other projects consulting is *subject to a charge* and to available resources.



Expert contact: fabian.floeck | indira.sen @gesis.org

Please find on the GESIS website consulting contacts for:  
[Planning Studies](#), [Accessing Data](#), [Analyzing Data](#), [Archiving Data](#)

## More Services from GESIS

- [GESIS Survey Guidelines](#) provide short and hands-on explanations to frequent challenges in survey design and methodology.
- [GESIS Training](#) offers a wide range of seminars, workshops, and other courses.
- Use GESIS data services for [finding data](#) for secondary analysis and [sharing your own data](#).
- Get materials for [capacity building](#) in computational social science and take advantage of our expanding expertise and resources in [digital behavioral data](#).
- Check out the [GESIS blog](#) "Growing Knowledge in the Social Sciences" for topics, methods and discussions from the GESIS cosmos – and beyond.
- Keep up with GESIS activities and subscribe to our monthly [newsletter](#).
-  for publications, tools &



And now: your questions!

gesis

Leibniz-Institut  
für Sozialwissenschaften

Leibniz  
Gemeinschaft

# Backup

gesis

Leibniz-Institut  
für Sozialwissenschaften

Leibniz  
Leibniz  
Gemeinschaft

## MEASUREMENT

Construct  
Definition



I'm surprised I managed to draw that since I feel sick  
as heck x'D Probably gonna go sleep now.

Augment traces with syntactic relations indicating  
the speaker being afflicted by the flu  
(~feature engineering)



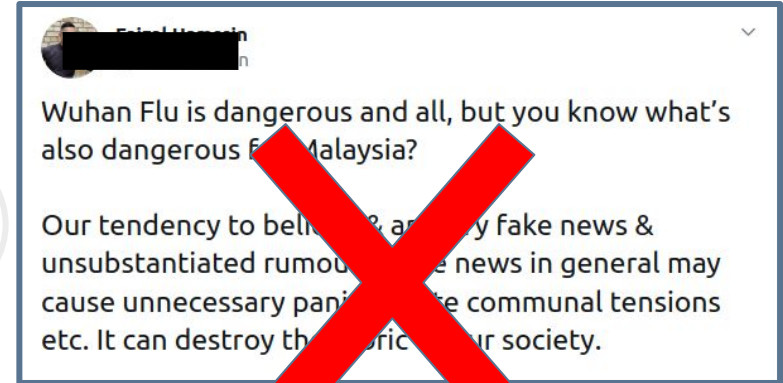
Data  
Collection

**Data  
Preprocessing**

Data  
Analysis

# MEASUREMENT

Remove non 'first-person' tweets



# MEASUREMENT

Construct  
Definition

Platform  
Selection


Data  
Collection

Data  
Access

Data  
Analysis

First\_person\_account: True  
Overall sentiment: +1.3  
Flu\_relevance: 0.8

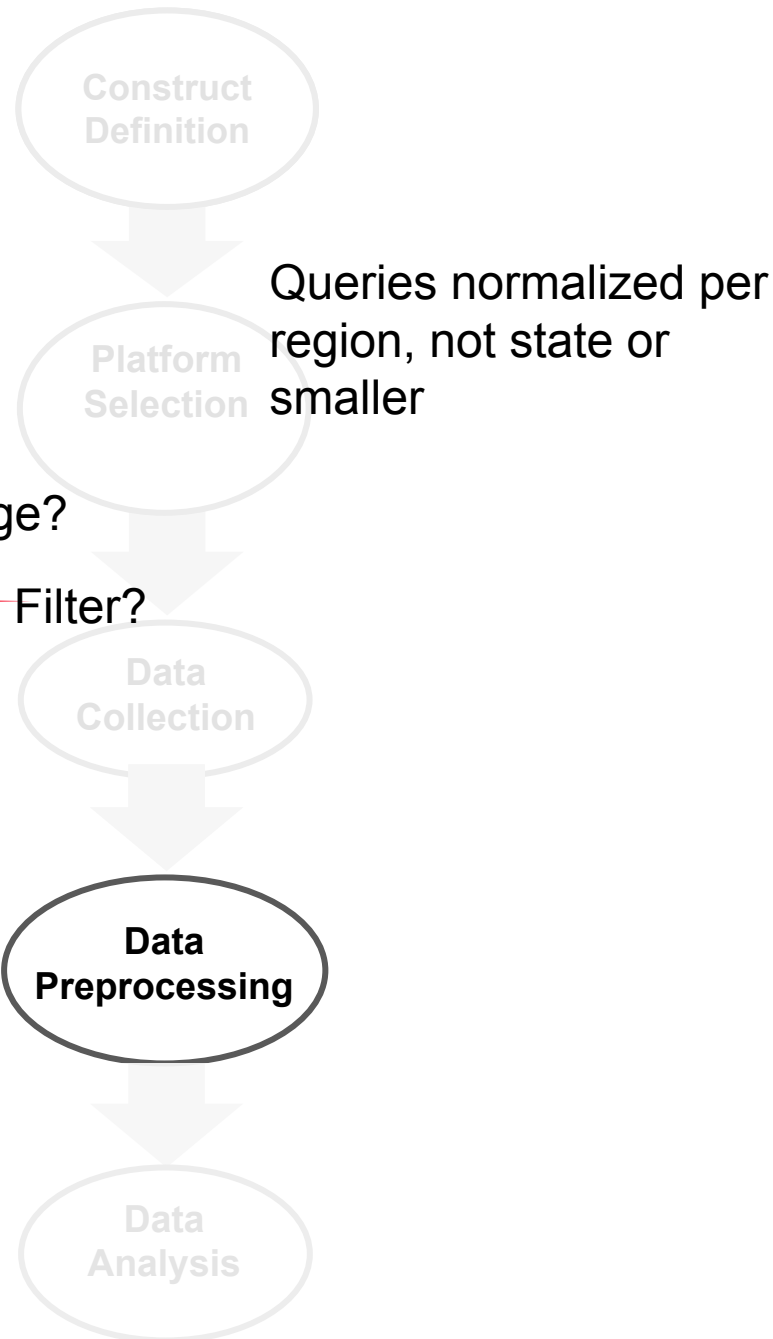
Text\_PoS: [PP, V, PP, PRE,  
ADJ, V, KONJ, PART, V,  
ADJ, ...]  
Text\_Sent: [0, +1, 0, 0, +1,  
0, -1, 0, 0, 0, ...]  
Embedding: [0.15, -2.03,  
0.93, -0.16, 0.98, 0.06]

 )  
I'm surprised I managed to draw that since I feel sick  
as heck x'D Probably gonna go sleep now.

Username: "Guy Dudeson"  
Timestamp: "18.06.18"  
TweetID: "6315127"  
Text: [I, surprised, I, managed,  
draw, I, feel, sick, heck,  
probably, going to, sleep,  
now]  
Likes: 725  
Retweets: 51

## MEASUREMENT

1. „Fever remedy“
  2. „Throat hurts“
  3. „Symptoms flu“
  4. „What to do against flu“
  5. „Healing flu“
  - ...
  61. „flu in Trucktown“
  62. „flu in Oklahoma“
  63. „flu infection numbers“
- } Merge?
- } Filter?



## MEASUREMENT

1. „Fever remedy“
2. „Throat hurts“
3. „Symptoms flu“
4. „What to do against flu“
5. „Healing flu“
- ...
61. „flu in Trucktown“
62. „flu in Oklahoma“
63. „flu infection numbers“

Queries normalized per region, not state or smaller

Google

Trace Augmentation Error

Trace Reduction Error

Wrong normalization

Data Preprocessing

- Misleading queries remain
- Connected queries not aggregated

= Errors due to augmenting or filtering out traces



# MEASUREMENT

Construct Definition

Develop model to measure rates of flu based on preprocessed/augmented traces

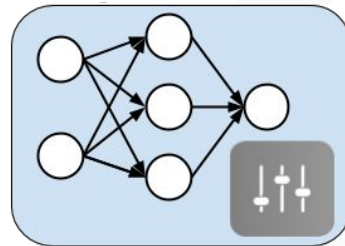


Sentiment score: -3.2, First person: yes, Embedding vector:  $\vec{w}$ , Location: Michigan

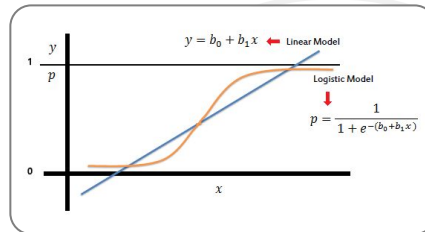
I feel sick 🤒

Aight twitter I still have the flu. Someone @ me with some entertainment 🙄😏💜

Alright, I probably have a fever. I'm going under the blankets again and hope to see the replies to this tweet filled with wholesome fan art of good waifus.



Trained Model

$$y = f(x)$$


Data Analysis



## MEASUREMENT

Construct  
Definition

Develop model to measure  
rates of flu based on  
preprocessed/augmented  
traces

$$\text{logit}(I(t)) = \alpha \cdot \text{logit}(Q(t)) + \varepsilon$$

Where

$I(t)$ : percentage of ILI visits at time ( $t$ )

$Q(t)$ : ILI-related query fraction at time ( $t$ )

$\alpha$ : coefficient

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

Google



Collection

Data  
Preprocessing

Data  
Analysis

## MEASUREMENT

Construct  
Definition

Develop model to measure  
rates of flu based on  
preprocessed/augmented  
traces



$$\text{logit}(I(t)) = \alpha \cdot \text{logit}(Q(t)) + \varepsilon$$

$$\text{logit}(I(t)) = \sum_{i=1}^n \alpha_i(t) \cdot \text{logit}(Q_i(t)) + \varepsilon$$

### Research Practice and Methods

What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends?\*

Mauricio Santillana, Ph.D., M.S.<sup>1</sup>; D. Wendong Zhang M.A.<sup>1</sup>; Benjamin M. Althouse, Ph.D., Sc.M.<sup>2</sup> John W. Ayers, Ph.D. M.A.<sup>3</sup>

Data  
Collection

Individual query proportions  $Q_i$ , with own coefficients  $\alpha_i$ , top 100 most correlating queries & no cut-off, LASSO regularization, weekly refitting on last two weeks



Data  
Preprocessing

Data  
Analysis

# MEASUREMENT

Construct  
Definition

Develop model to measure  
rates of flu based on  
preprocessed/augmented  
traces



$$\text{logit}(I(t)) = \alpha \cdot \text{logit}(Q(t)) + \varepsilon$$

$$\text{logit}(I(t)) = \sum_{i=1}^n \alpha_i(t) \cdot \text{logit}(Q_i(t)) + \varepsilon$$

## Research Practice and Methods

What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends?\*

Mauricio Santillana, Ph.D., M.S.<sup>1</sup>; D. Wendong Zhang M.A.<sup>1</sup>; Benjamin M. Althouse, Ph.D., Sc.M.<sup>2</sup> John W. Ayers, Ph.D. M.A.<sup>3</sup>

Data  
Collection

Individual query proportions  $Q_i$ , with own coefficients  $\alpha_i$ , top 100 most correlating queries & no cut-off, LASSO regularization, weekly refitting on last two weeks



Data  
Preprocessing

= Errors due to the choice of modeling or aggregation

Trace  
Measurement  
Error

Data  
Analysis

# MEASUREMENT

Construct  
Definition

Develop model to measure rates of flu based on preprocessed/augmented traces



$$\text{logit}(I(t)) = \alpha \cdot \text{logit}(Q(t)) + \varepsilon$$

$$\text{logit}(I(t)) = \sum_{i=1}^n \alpha_i(t) \cdot \text{logit}(Q_i(t)) + \varepsilon$$

## Research Practice and Methods

What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends?<sup>\*</sup>

Mauricio Santillana, Ph.D., M.S.<sup>1</sup>; D. Wendong Zhang M.A.<sup>1</sup>; Benjamin M. Althouse, Ph.D., Sc.M.<sup>2</sup> John W. Ayers, Ph.D. M.A.<sup>3</sup>

Data  
Collection

Individual query proportions  $Q_i$ , with own coefficients  $\alpha_i$ , top 100 most correlating queries & no cut-off, LASSO regularization, weekly refitting on last two weeks

Data  
Preprocessing

**Undue weight to certain query terms**  
**No nuance in term weighting**  
**Overfitting to time period**  
**Linearity assumed**

to the choice of modeling or aggregation

Data  
Analysis

Trace  
Measurement  
Error