gesis Leibniz Institute for the Social Sciences

Potentials and Pitfalls of Social Media Data

Indira Sen & Katrin Weller

GESIS workshop - December 2022

Leibniz Association

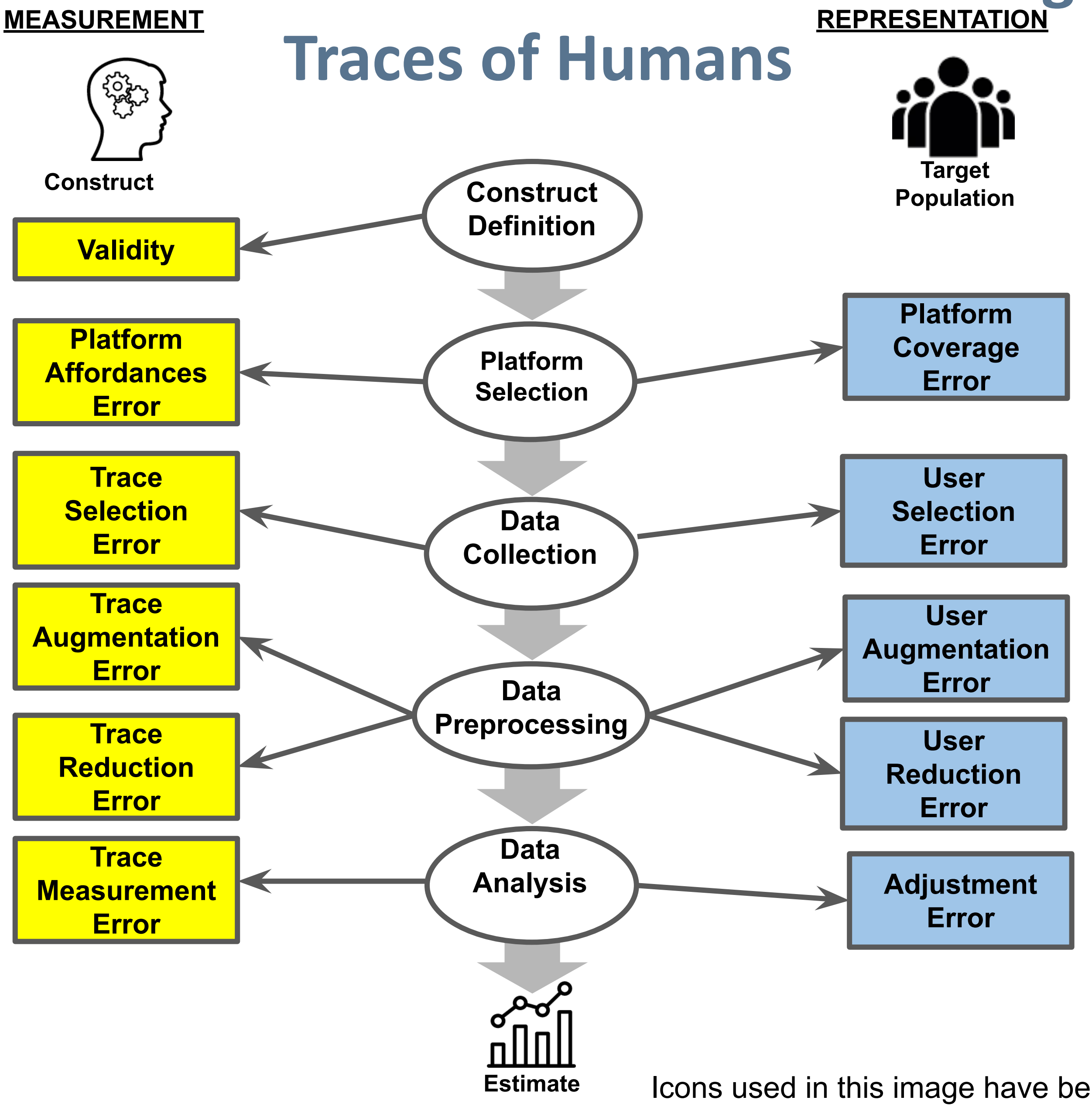
Today’s Schedule

Monday, 12.12.	
9:30-11:00	Recap of hands-on examples from last week and exemplary research scenarios
11:00-11:15	*Break*
11:15-12:00	Potential error sources in social media research
12:00-12:30	*Break*
12:30-14:00	Hands-on error identification + Case Study: Mitigating potential error sources in social media research

Agenda

- Session 1: Introduction to Research with Social Media Data (SMD)
- Session 2: SM Data Collection
- Session 3: SMD Preprocessing and Analysis
- Session 4: Potential Pitfalls of SMD
- Session 5: Identifying Pitfalls with help from surveys
- Session 6: Identifying Pitfalls in SMD
- Session 7: Mitigating Pitfalls
- Session 8: Documenting Pitfalls
- Session 9: Recap and Conclusions

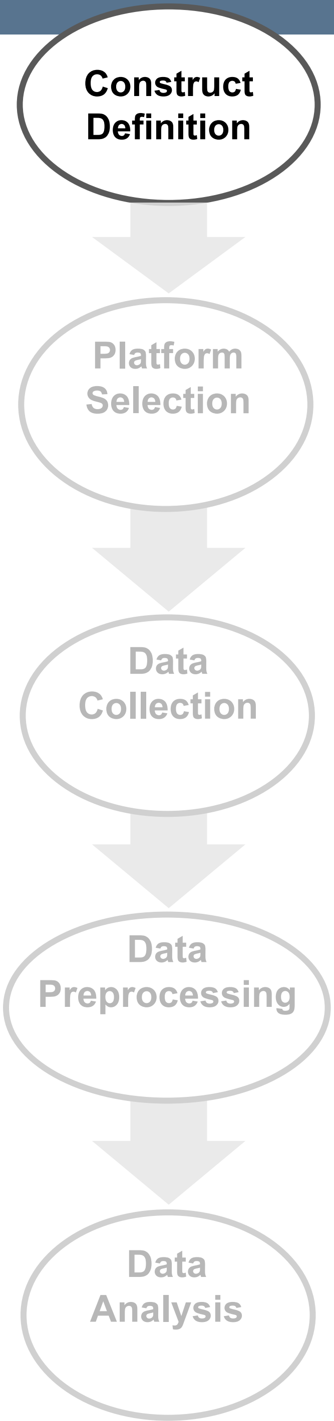
TED-On: A Total Error Framework for Digital Traces of Humans



Icons used in this image have been designed by Becris, EliasBikbulatov and Pixel perfect from www.flaticon.com

Example study

- How would a researcher study influenza prevalence in a national population using digital traces?


Construct Definition
Platform Selection
Data Collection
Data Preprocessing
Data Analysis

MEASUREMENT

Construct Definition

Platform Selection

Preprocessing

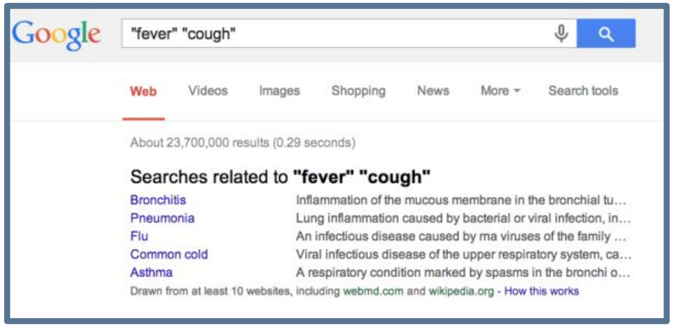
Data Analysis

TRACES

TRACES

TRACES

search queries related to flu



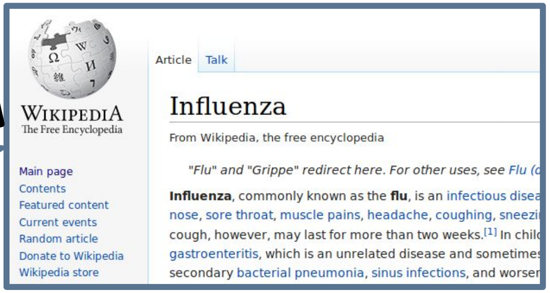
About 23,700,000 results (0.29 seconds)

Searches related to "fever" "cough"

Bronchitis	Inflammation of the mucous membrane in the bronchial tu…
Pneumonia	Lung inflammation caused by bacterial or viral infection, in…
Flu	An infectious disease caused by rna viruses of the family …
Common cold	Viral infectious disease of the upper respiratory system, ca…
Asthma	A respiratory condition marked by spasms in the bronchi o…

Drawn from at least 10 websites, including webmd.com and wikipedia.org - How this works

flu related information usage



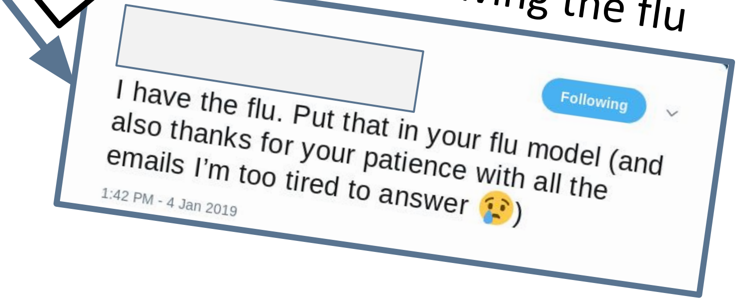
Influenza

From Wikipedia, the free encyclopedia

"Flu" and "Grippe" redirect here. For other uses, see Flu (c

Influenza, commonly known as the **flu**, is an infectious disea
nose, sore throat, muscle pains, headache, coughing, sneezi
cough, however, may last for more than two weeks.[1] In child
gastroenteritis, which is an unrelated disease and sometimes
secondary bacterial pneumonia, sinus infections, and worser

posts about having the flu



I have the flu. Put that in your flu model (and also thanks for your patience with all the emails I'm too tired to answer 😢)

1:42 PM - 4 Jan 2019

MEASUREMENT

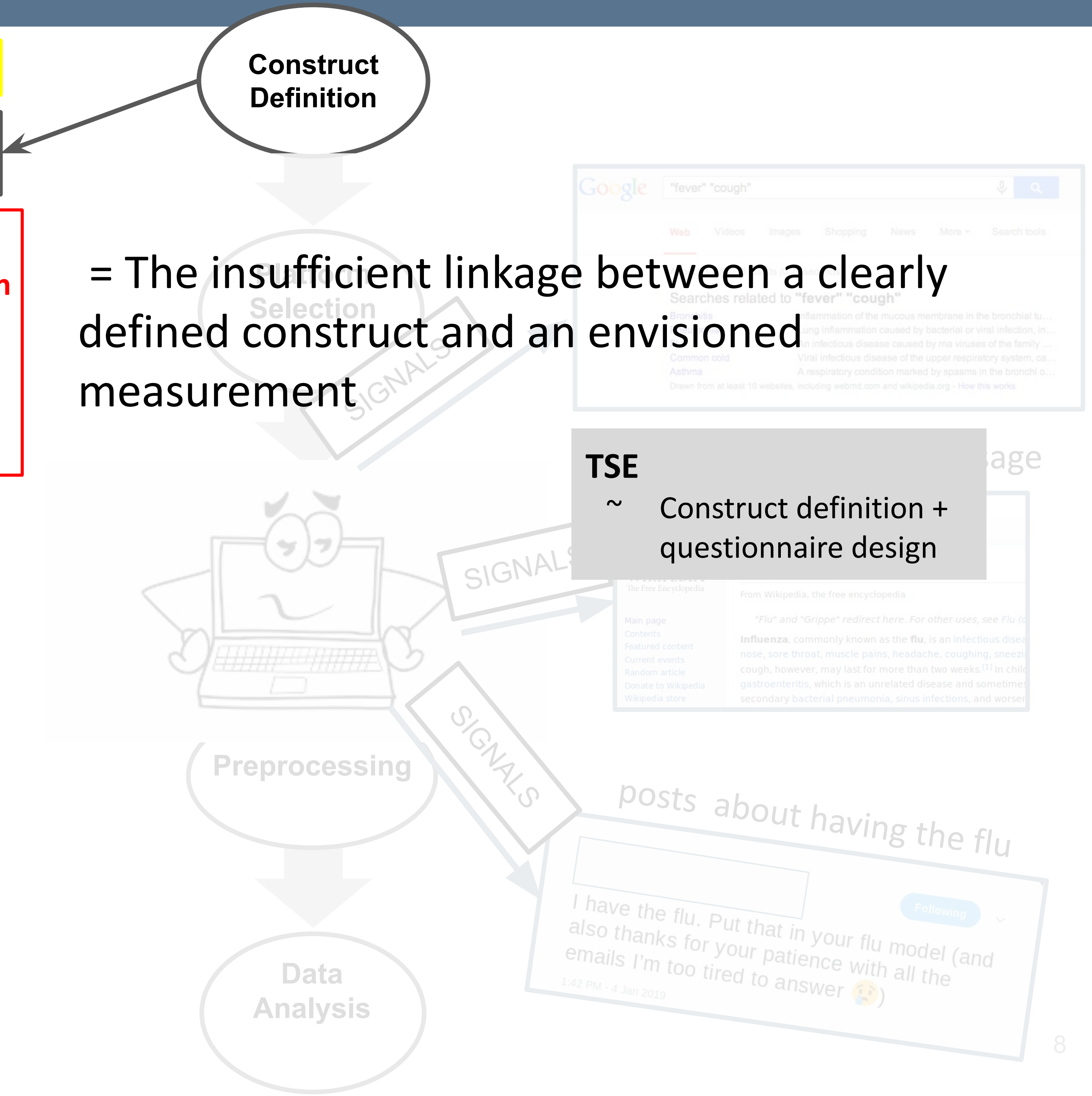
Validity

searching for flu related information or tweeting about flu may misrepresent incidences of flu

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

TSE

~ Construct definition + questionnaire design

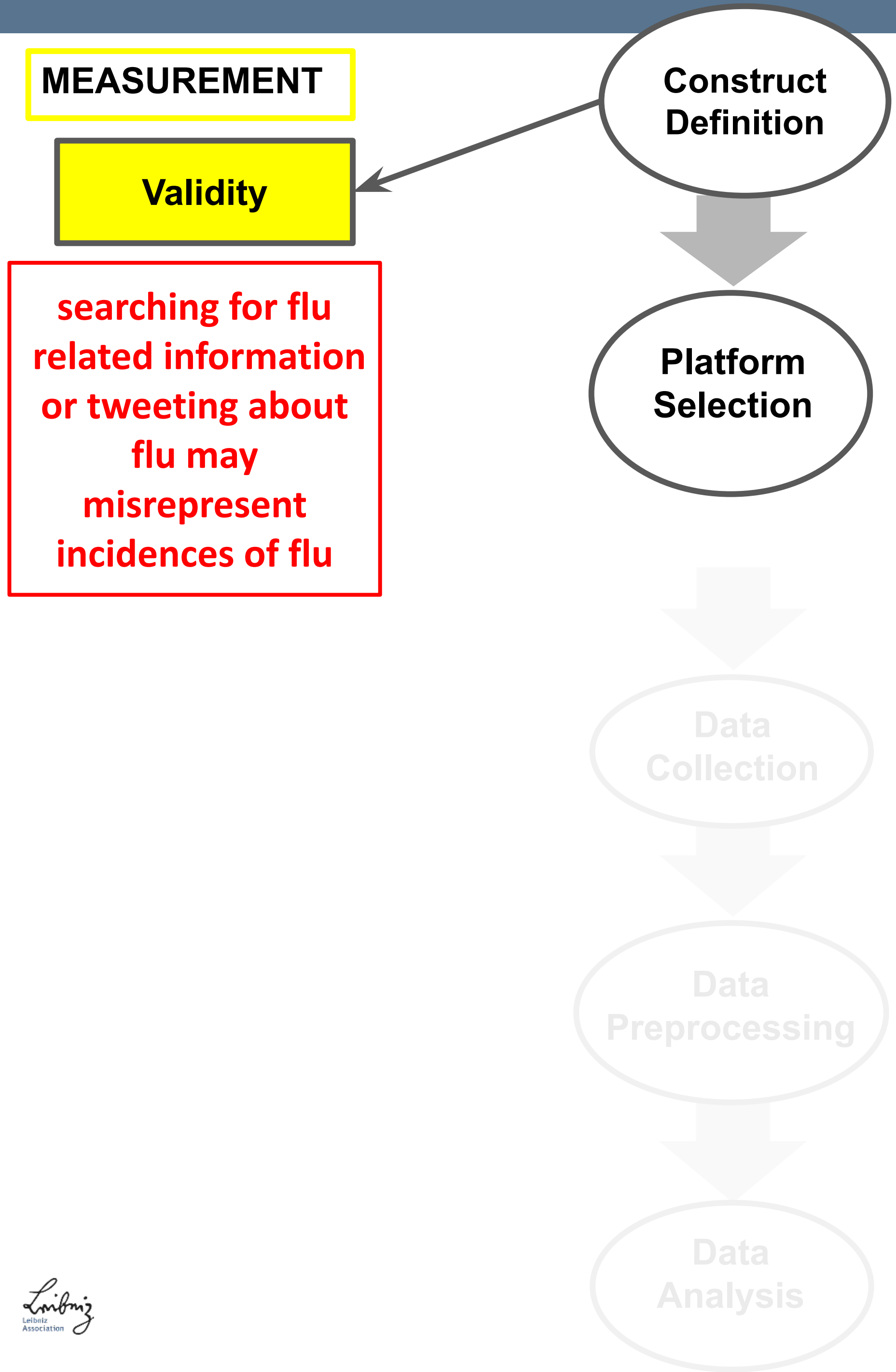


Inflammation of the mucous membrane in the bronchial tu…
Lung inflammation caused by bacterial or viral infection, in…
An infectious disease caused by rna viruses of the family …
Viral infectious disease of the upper respiratory system, ca…
A respiratory condition marked by spasms in the bronchi o…
Drawn from at least 10 websites, including webmd.com and wikipedia.org - How this works

"Flu" and "Grippe" redirect here. For other uses, see Flu (…

Influenza, commonly known as the **flu**, is an infectious disea… nose, sore throat, muscle pains, headache, coughing, sneezi… cough, however, may last for more than two weeks.[1] In child… gastroenteritis, which is an unrelated disease and sometimes… secondary bacterial pneumonia, sinus infections, and worse…

I have the flu. Put that in your flu model (and also thanks for your patience with all the emails I'm too tired to answer 😢)



MEASUREMENT

Validity

searching for flu related information or tweeting about flu may misrepresent incidences of flu

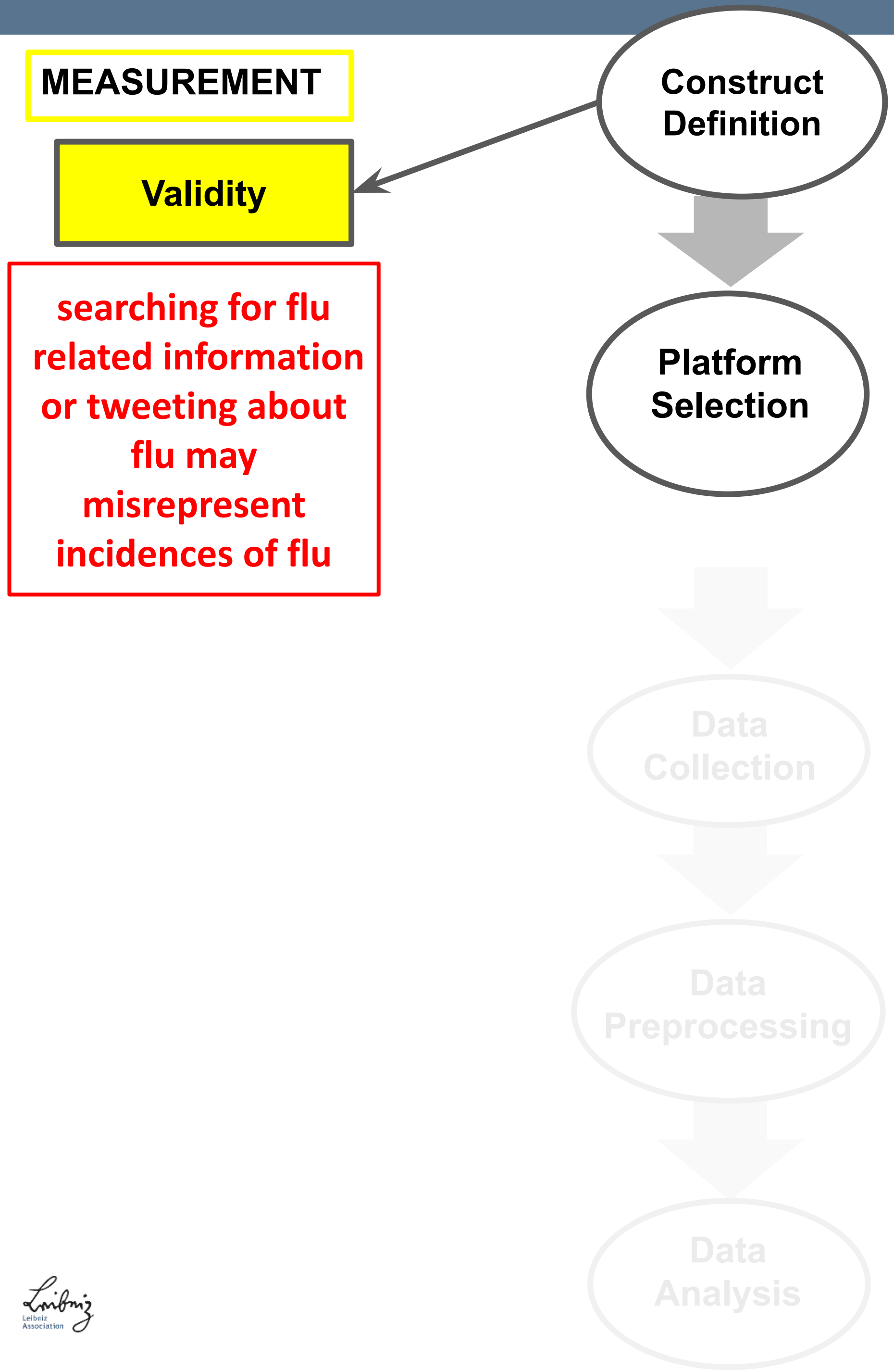
Construct Definition

Platform Selection

Data Collection

Data Preprocessing

Data Analysis



MEASUREMENT

Validity

searching for flu related information or tweeting about flu may misrepresent incidences of flu

Construct Definition

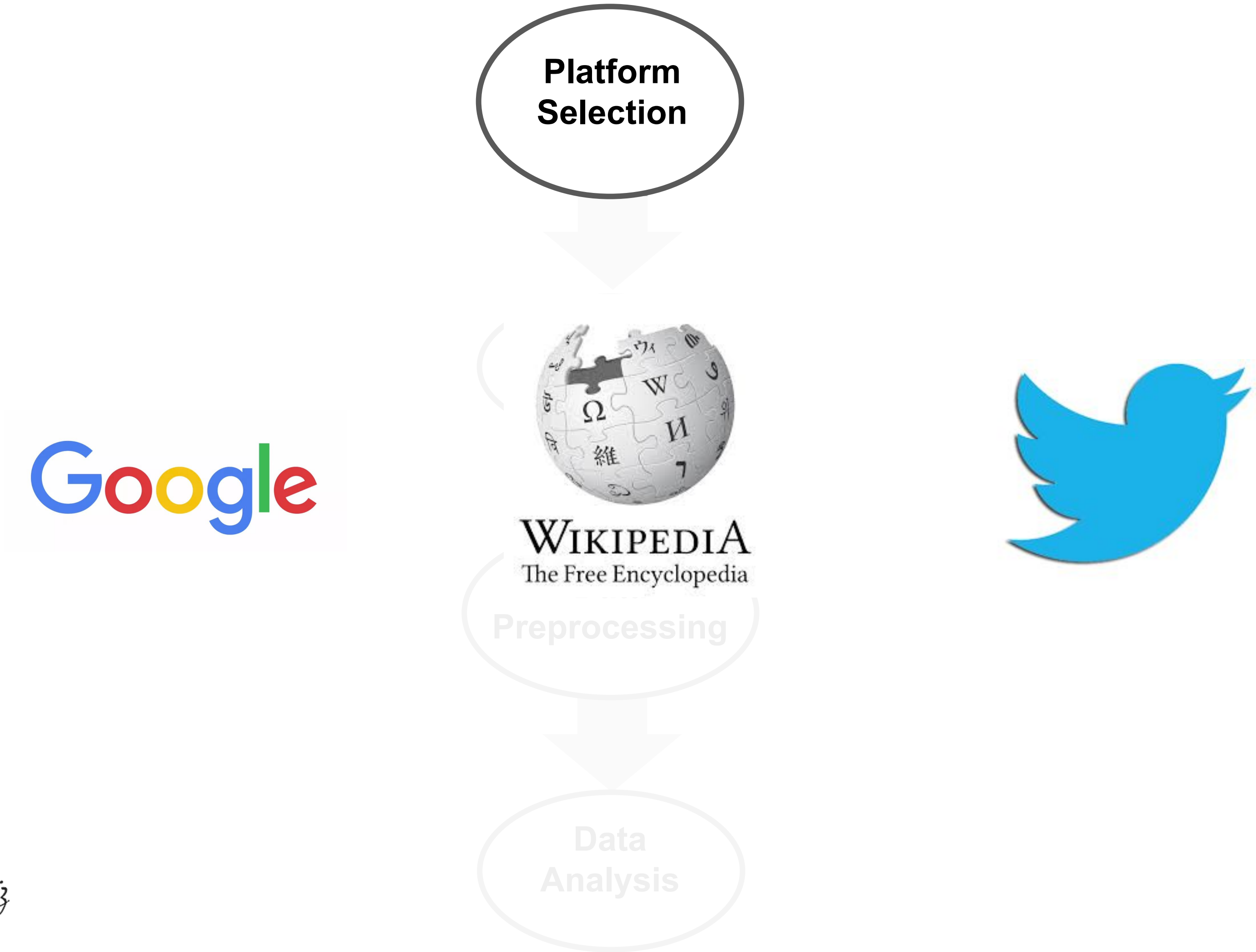
Platform Selection

Data Collection

Data Preprocessing

Data Analysis

MEASUREMENT

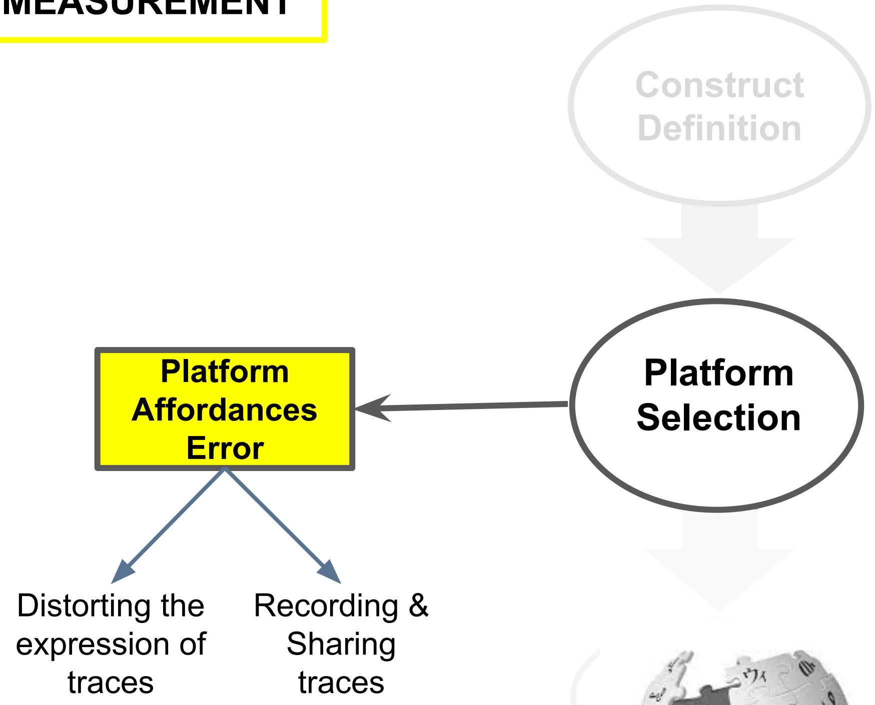


Platform Selection

Preprocessing

Data Analysis

MEASUREMENT



Construct Definition

Platform Selection

Platform Affordances Error

Distorting the expression of traces

Recording & Sharing traces

Preprocessing

Data Analysis

Recommendations for search queries

Only aggregate queries

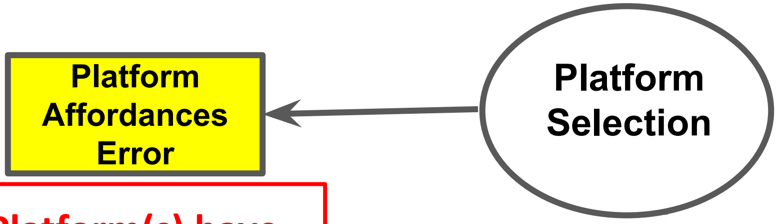
Session length for readers

280-character limit

Trending tweets

MEASUREMENT

= The gap between the ‘true’ traces and traces distorted by platform affordances - technical/community standards, terms & conditions



Platform(s) have affordances which distort traces

TSE

~ Reponse collection (collection mode): Response Error

Recommendations for search queries

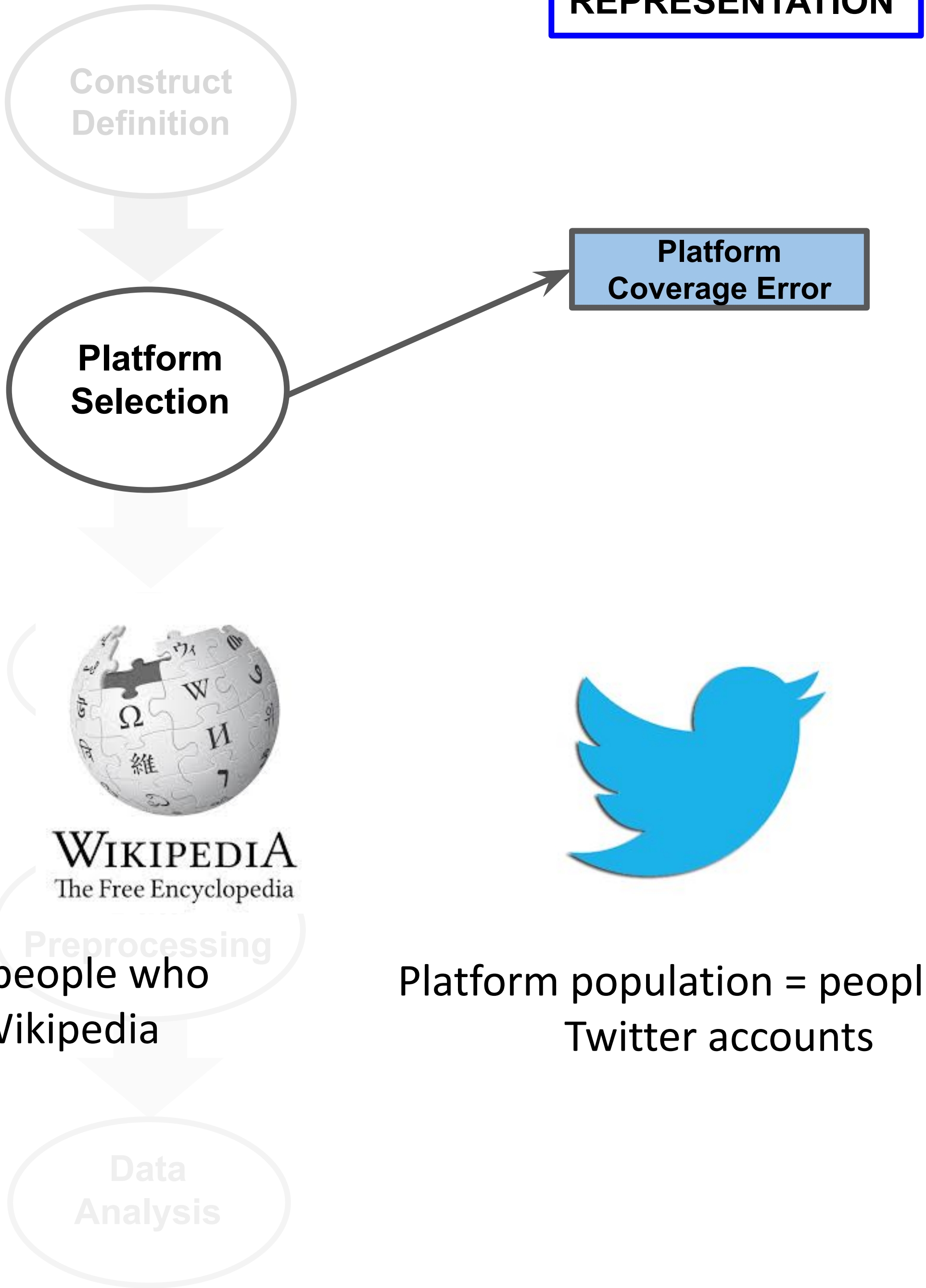
Only aggregate queries

Session length for readers

280-character limit

Trending tweets

REPRESENTATION

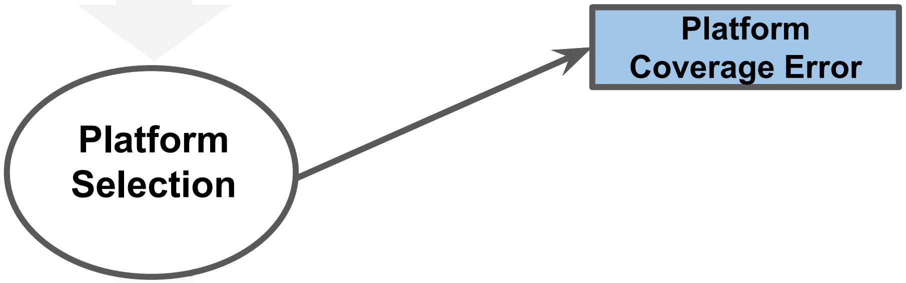


Platform population = people who access Google and Wikipedia

Platform population = people with Twitter accounts

REPRESENTATION

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

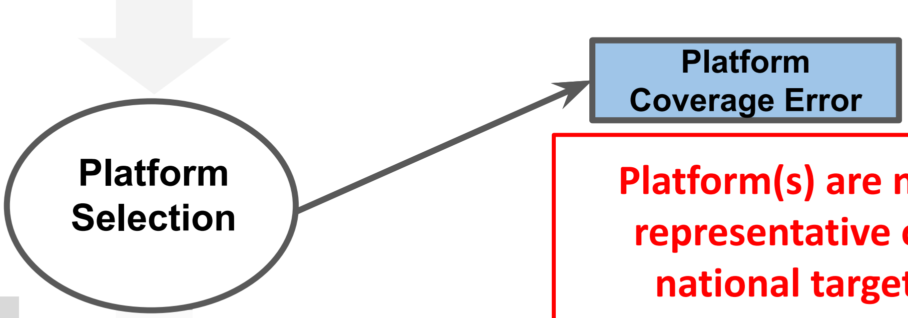


Platform population = people who access Google and Wikipedia

Platform population = people with Twitter accounts

REPRESENTATION

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



Platform(s) are not representative of national target population

TSE

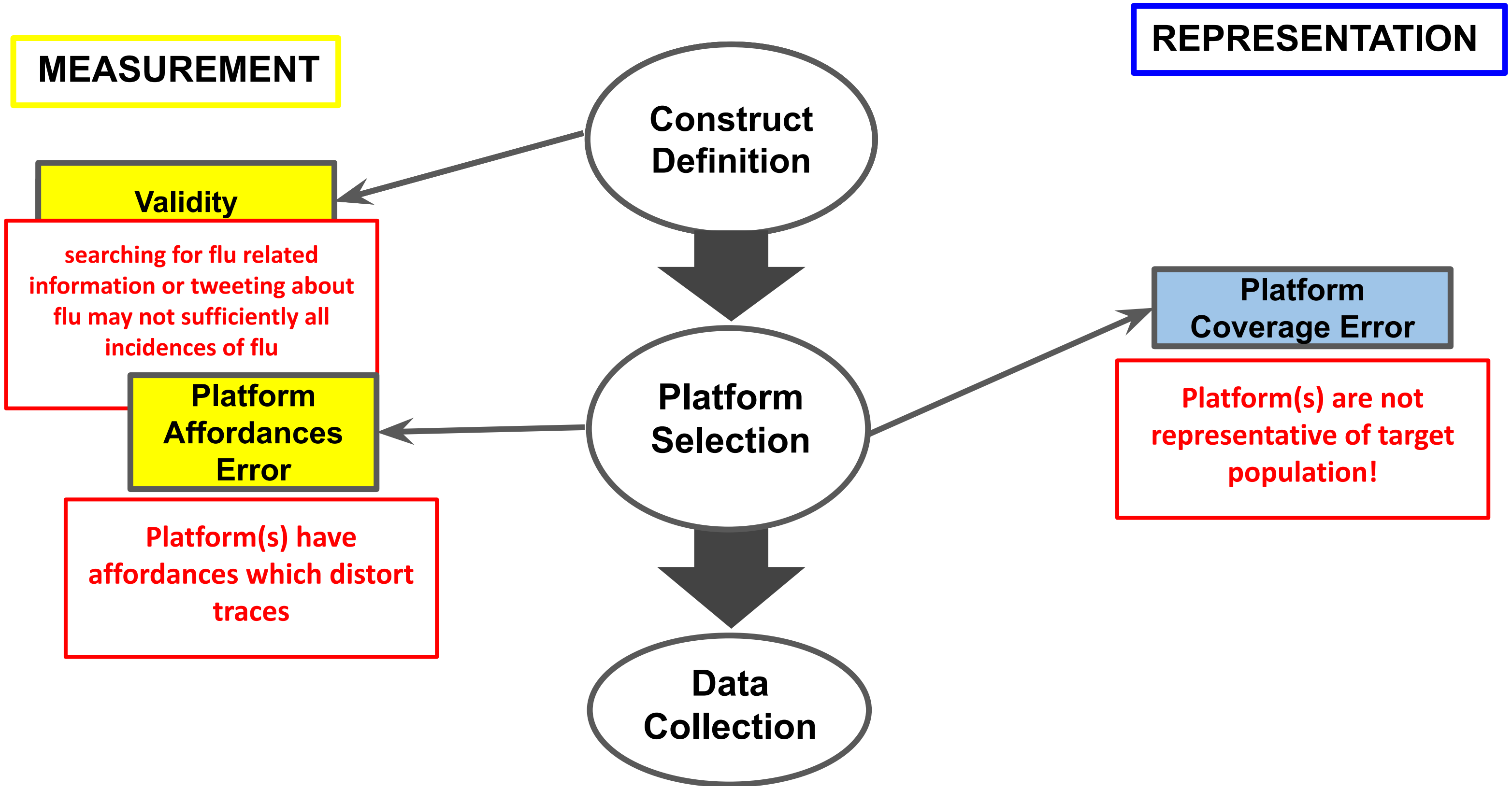
- Sampling Frame Selection: Coverage Error

Google

WIKIPEDIA The Free Encyclopedia

Platform population = people who access Google and Wikipedia

Platform population = people with Twitter accounts



MEASUREMENT

REPRESENTATION

Construct Definition

Validity

searching for flu related information or tweeting about flu may not sufficiently all incidences of flu

Platform Selection

Platform Affordances Error

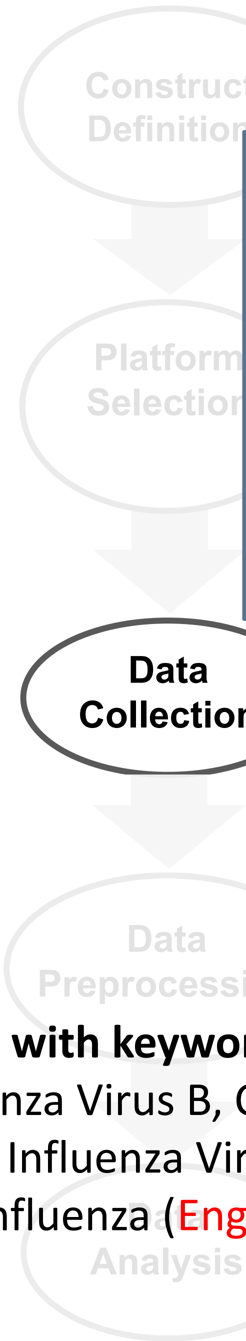
Platform(s) have affordances which distort traces

Platform Coverage Error

Platform(s) are not representative of target population!

Data Collection

MEASUREMENT



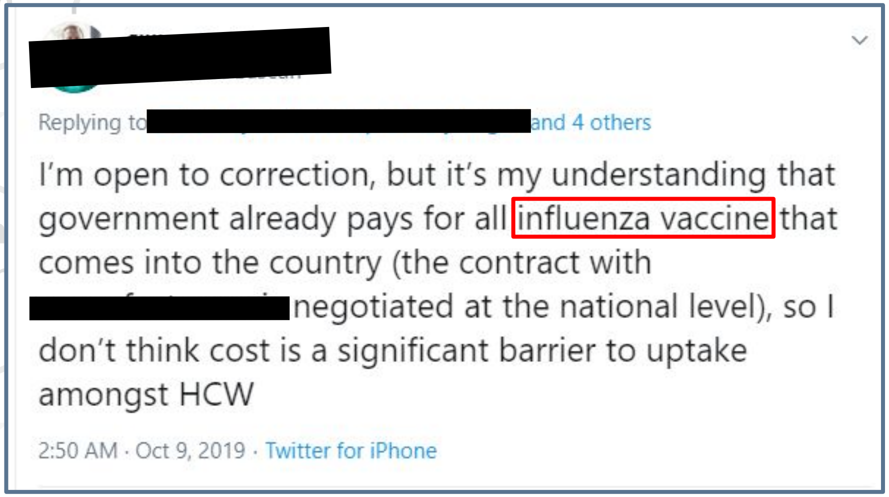
Construct Definition

Platform Selection

Data Collection

Data Preprocessing

Analysis



[redacted]

Replying to [redacted] and 4 others

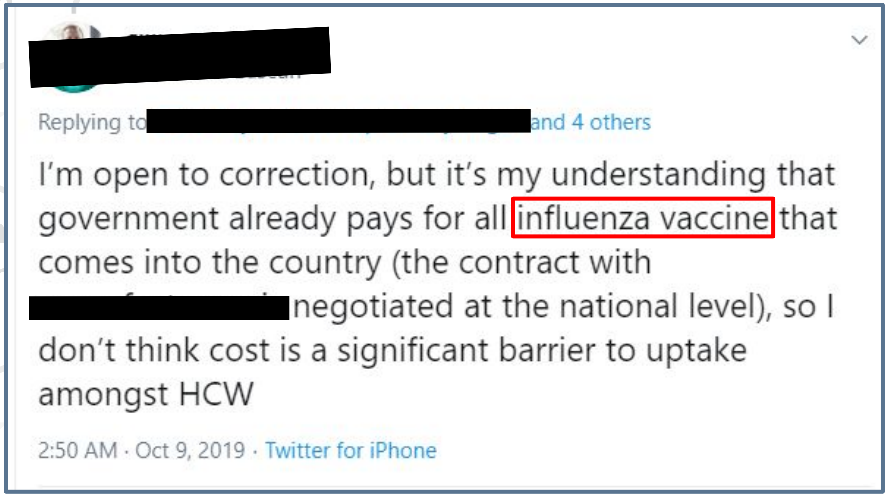
I'm open to correction, but it's my understanding that government already pays for all influenza vaccine that comes into the country (the contract with [redacted] negotiated at the national level), so I don't think cost is a significant barrier to uptake amongst HCW

2:50 AM · Oct 9, 2019 · Twitter for iPhone

Retrieve tweets with keywords related to Influenza: Avian influenza, Influenza Virus B, Centers for Disease Control and Prevention, Influenza Virus C, Common Cold, Vaccine, Flu(the Band), Influenza (English tweets only!)

MEASUREMENT

= The gap between the ideal response and the measured trace

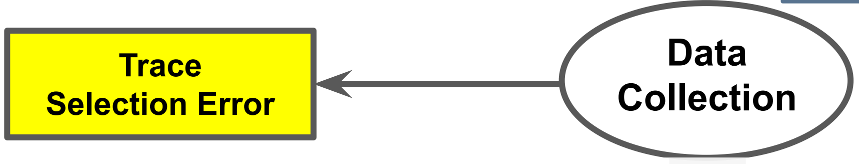


[redacted]

Replying to [redacted] and 4 others

I'm open to correction, but it's my understanding that government already pays for all influenza vaccine that comes into the country (the contract with [redacted] negotiated at the national level), so I don't think cost is a significant barrier to uptake amongst HCW

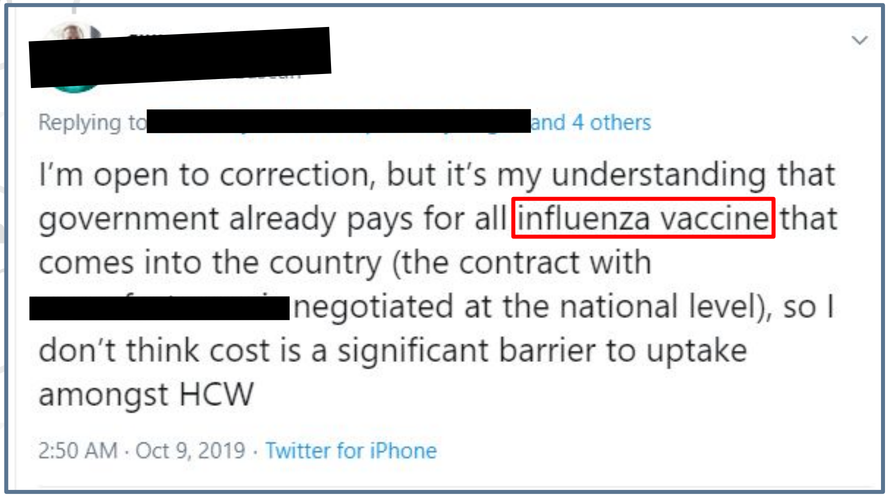
2:50 AM · Oct 9, 2019 · Twitter for iPhone



Retrieve tweets with keywords related to Influenza: Avian influenza, Influenza Virus B, Centers for Disease Control and Prevention, Influenza Virus C, Common Cold, Vaccine, Flu(the Band), Influenza (English tweets only!)

MEASUREMENT

= The gap between the ideal response and the measured trace

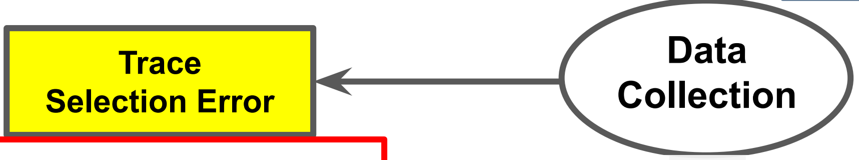


[redacted]

Replying to [redacted] and 4 others

I'm open to correction, but it's my understanding that government already pays for all influenza vaccine that comes into the country (the contract with [redacted] negotiated at the national level), so I don't think cost is a significant barrier to uptake amongst HCW

2:50 AM · Oct 9, 2019 · Twitter for iPhone



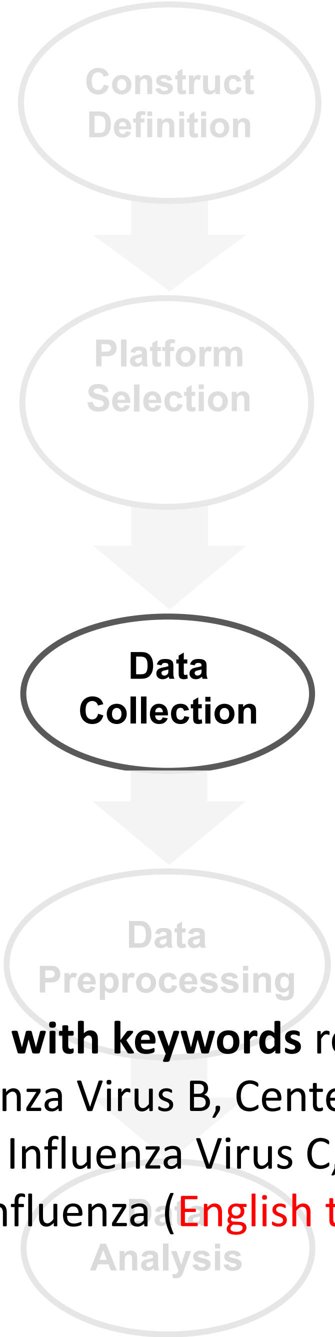
Some tweets chosen may not be relevant for influenza like illnesses

TSE

~ Response collection: response error

Retrieve tweets with keywords related to Influenza: Avian influenza, Influenza Virus B, Centers for Disease Control and Prevention, Influenza Virus C, Common Cold, Vaccine, Flu(the Band), Influenza (English tweets only!)

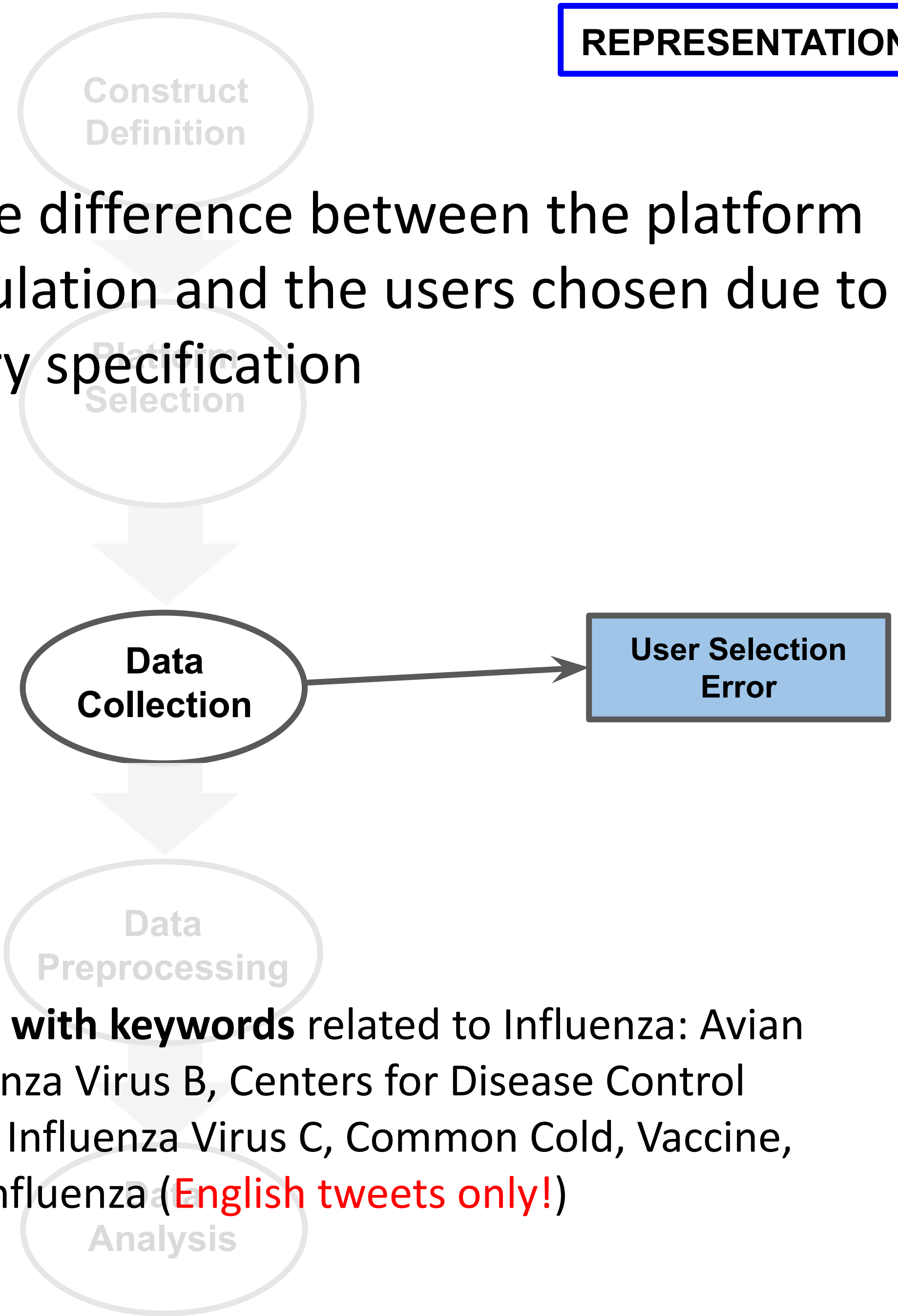
REPRESENTATION



Retrieve tweets with keywords related to Influenza: Avian influenza, Influenza Virus B, Centers for Disease Control and Prevention, Influenza Virus C, Common Cold, Vaccine, Flu(the Band), Influenza (English tweets only!)

REPRESENTATION

= The difference between the platform population and the users chosen due to the query specification



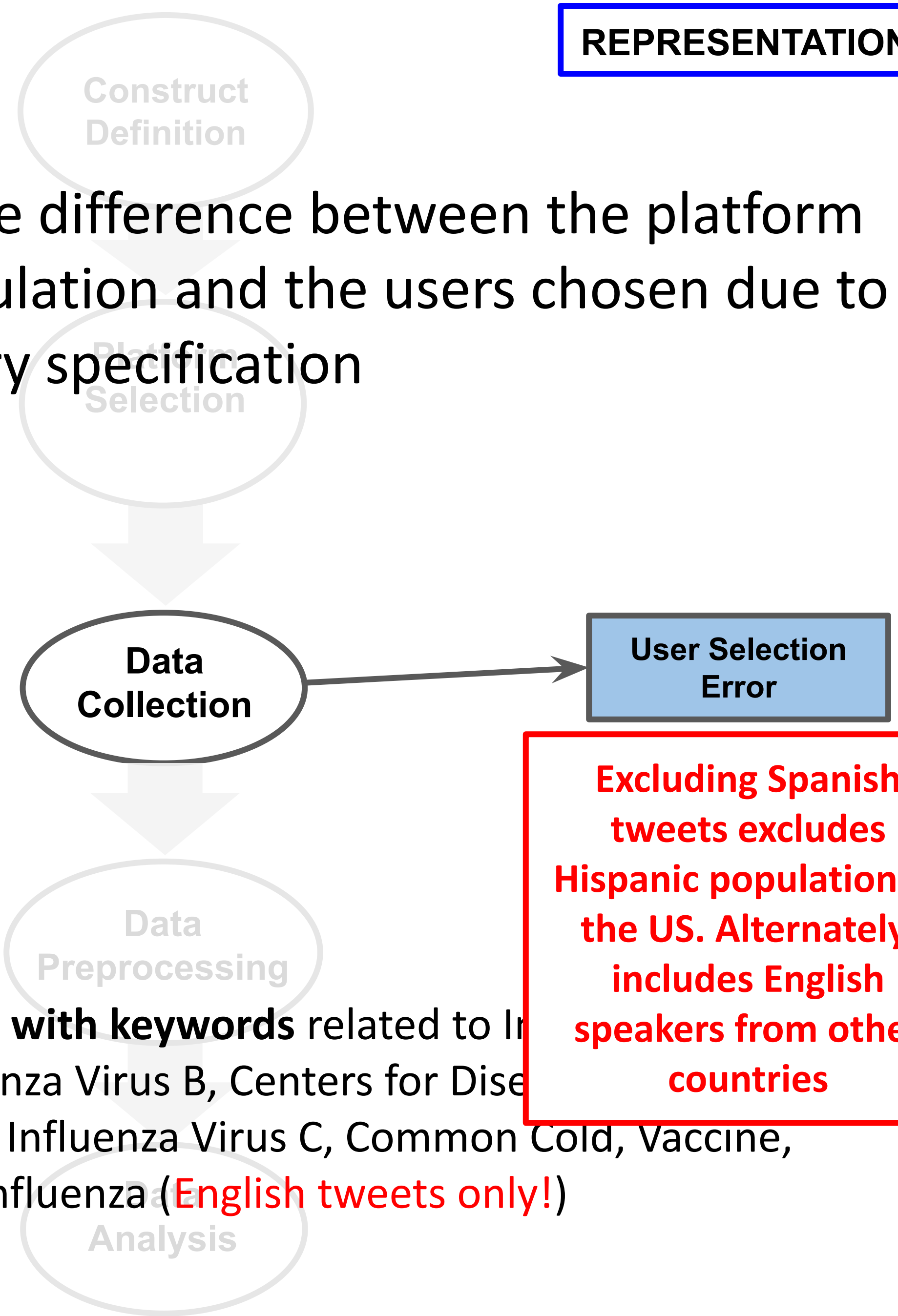
Retrieve tweets with keywords related to Influenza: Avian influenza, Influenza Virus B, Centers for Disease Control and Prevention, Influenza Virus C, Common Cold, Vaccine, Flu(the Band), Influenza (English tweets only!)

REPRESENTATION

= The difference between the platform population and the users chosen due to the query specification

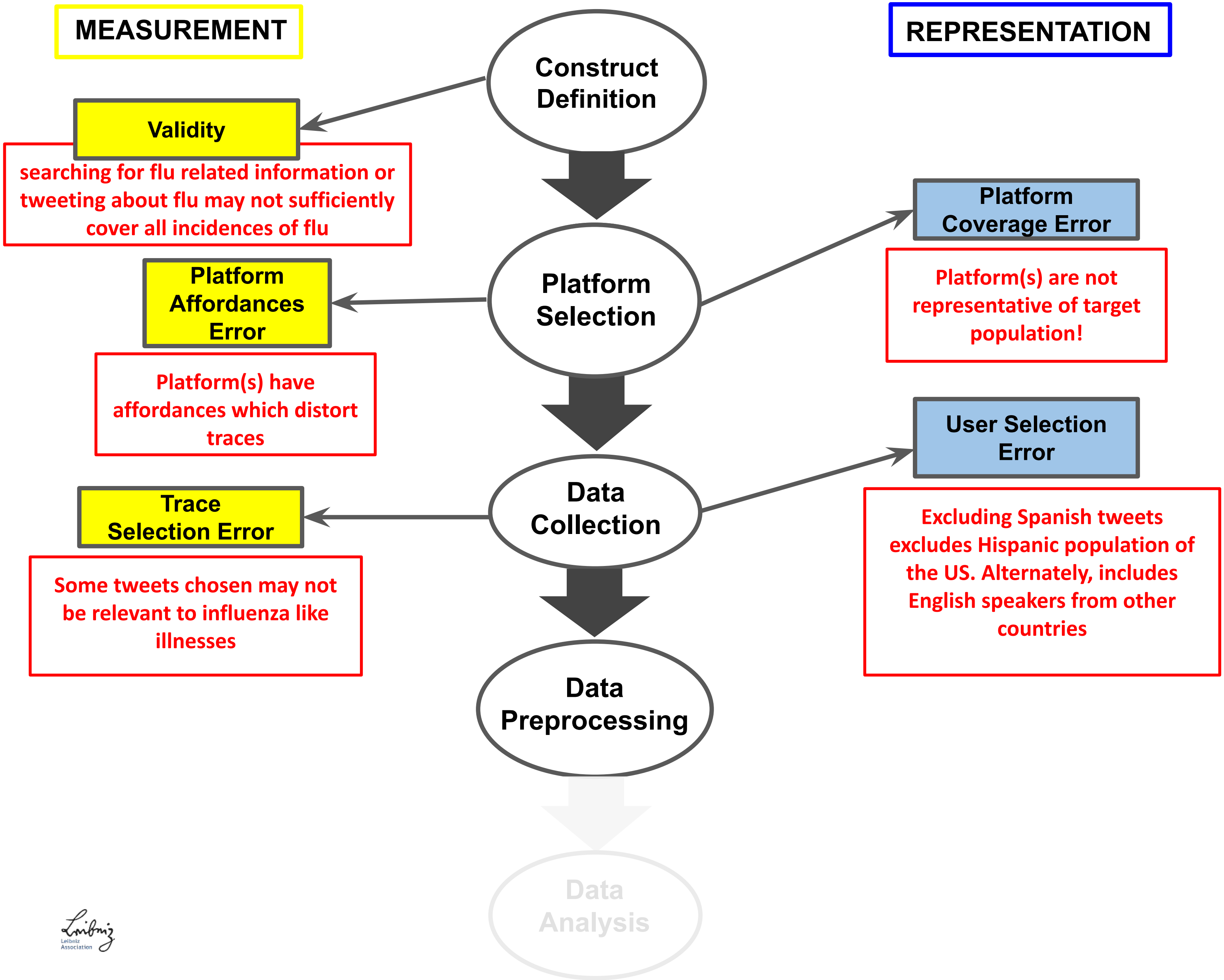
TSE

- ~ Sampling: Sampling error
- ~ Coverage error



Excluding Spanish tweets excludes Hispanic population of the US. Alternately, includes English speakers from other countries

Retrieve tweets with keywords related to Ir influenza, Influenza Virus B, Centers for Dise and Prevention, Influenza Virus C, Common Cold, Vaccine, Flu(the Band), Influenza (English tweets only!)



MEASUREMENT

REPRESENTATION

Construct Definition

Platform Selection

Data Collection

Data Preprocessing

Data Analysis

Validity

searching for flu related information or tweeting about flu may not sufficiently cover all incidences of flu

Platform Affordances Error

Platform(s) have affordances which distort traces

Trace Selection Error

Some tweets chosen may not be relevant to influenza like illnesses

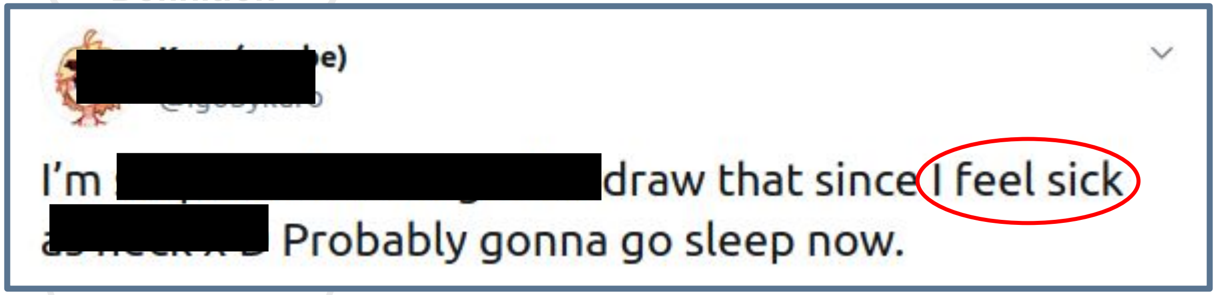
Platform Coverage Error

Platform(s) are not representative of target population!

User Selection Error

Excluding Spanish tweets excludes Hispanic population of the US. Alternately, includes English speakers from other countries

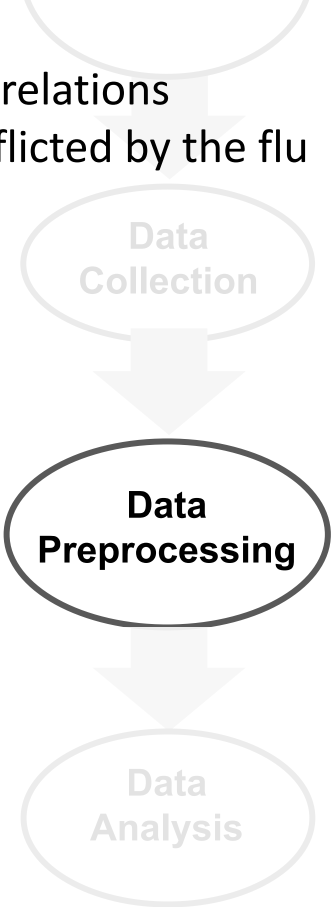
MEASUREMENT



[redacted]

I’m [redacted] draw that since I feel sick [redacted] Probably gonna go sleep now.

Augment traces with syntactic relations indicating the speaker being afflicted by the flu



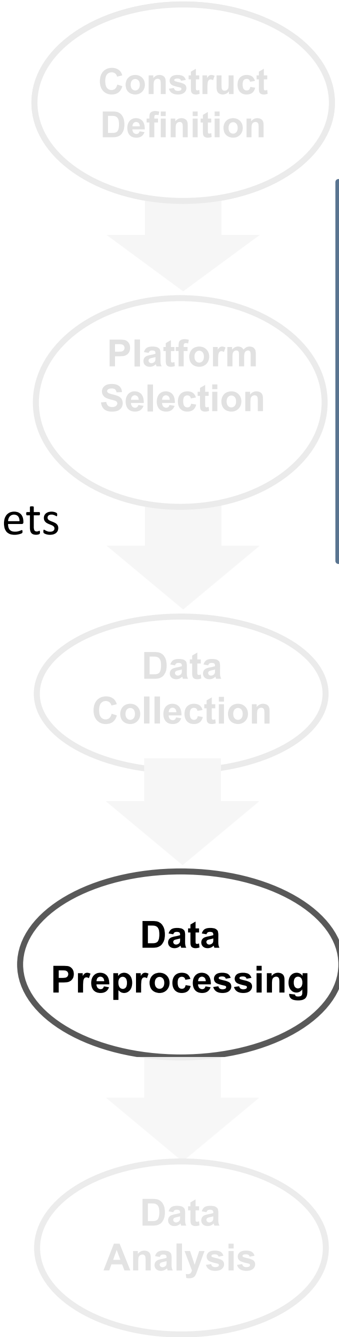
Construct Definition

Data Collection

Data Preprocessing

Data Analysis

MEASUREMENT


Construct Definition

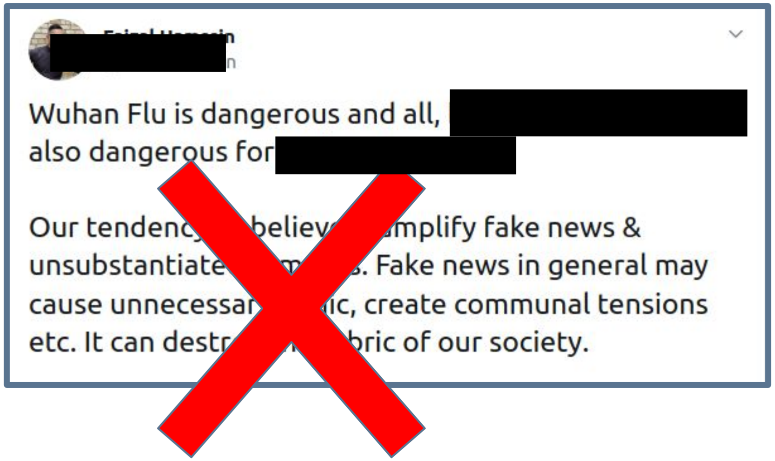
Platform Selection

Data Collection

Data Preprocessing

Data Analysis

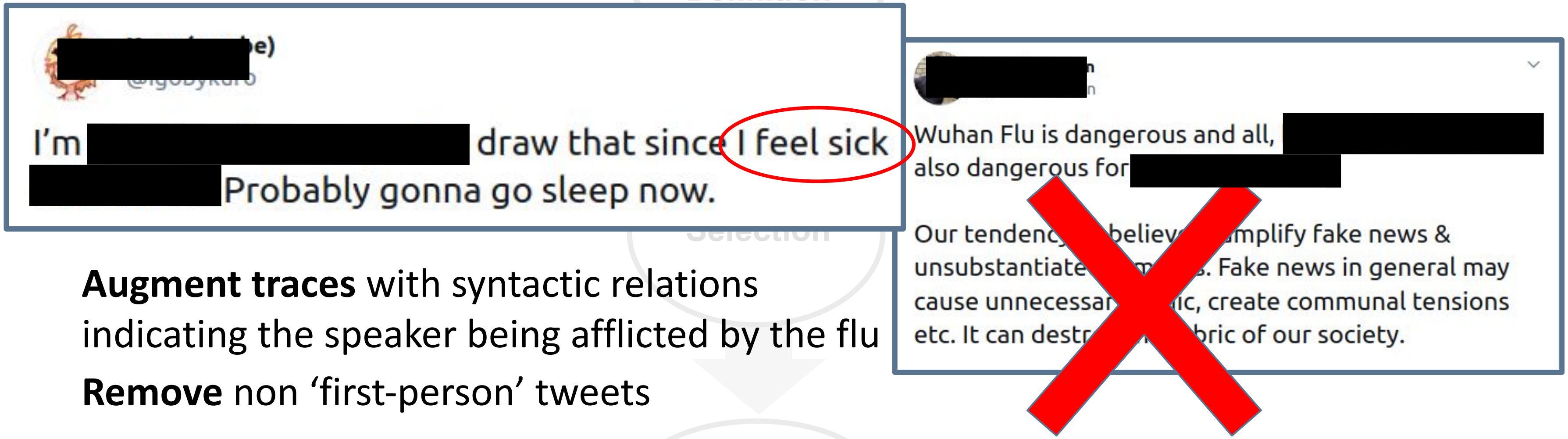
Remove non ‘first-person’ tweets


[redacted]

Wuhan Flu is dangerous and all, [redacted] also dangerous for [redacted]

Our tendenc believ mplify fake news & unsubstantiate m s. Fake news in general may cause unnecessar ic, create communal tensions etc. It can destr bric of our society.

MEASUREMENT



[redacted]

I'm [redacted] draw that since I feel sick [redacted] Probably gonna go sleep now.

[redacted]

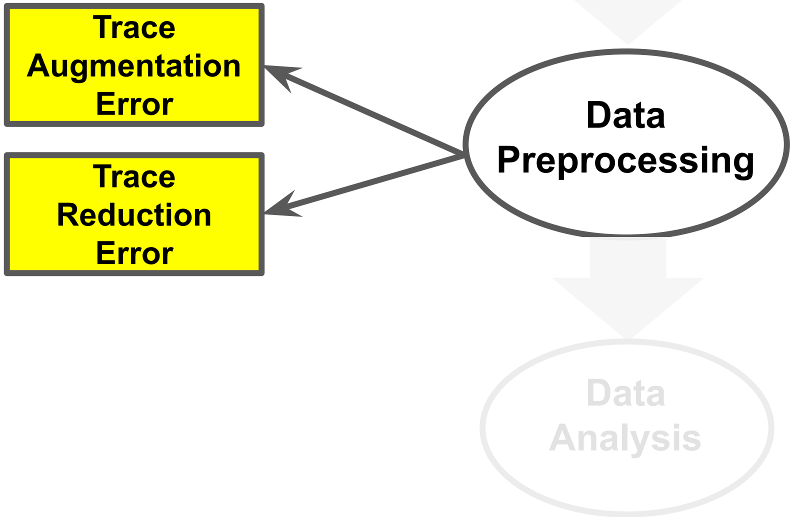
Wuhan Flu is dangerous and all, [redacted] also dangerous for [redacted]

Our tendenc believ mplify fake news & unsubstantiate m s. Fake news in general may cause unnecessar ic, create communal tensions etc. It can destr bric of our society.

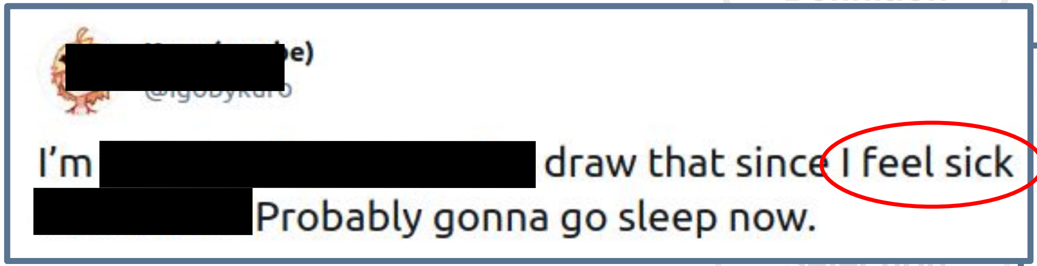
Augment traces with syntactic relations indicating the speaker being afflicted by the flu

Remove non 'first-person' tweets

= Errors due to **augmenting** or **filtering** out traces

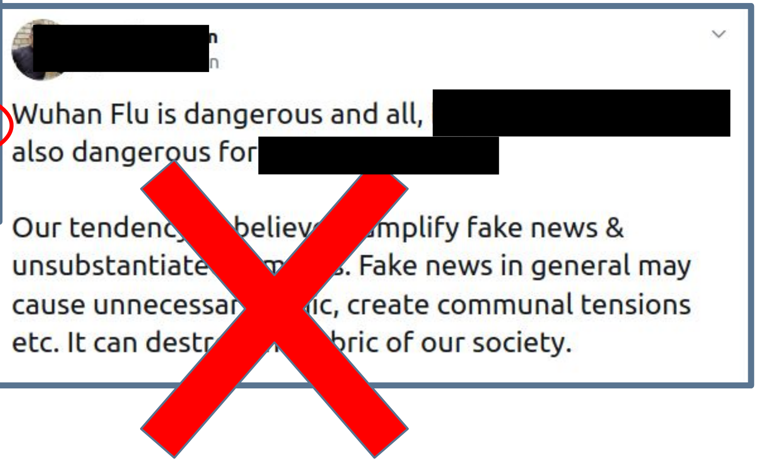


MEASUREMENT



[redacted] e)

I'm [redacted] draw that since I feel sick [redacted] Probably gonna go sleep now.



[redacted]

Wuhan Flu is dangerous and all, [redacted] also dangerous for [redacted]

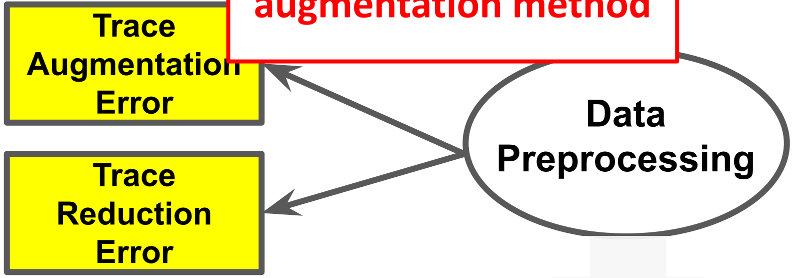
Our tendenc believ mplify fake news & unsubstantiate m s. Fake news in general may cause unnecessar ic, create communal tensions etc. It can destr bric of our society.

Augment traces with syntactic relations indicating the speaker being afflicted by the flu

Remove non 'first-person' tweets

= Errors due to **augmenting** or **filtering** out tra

error rate of augmentation method

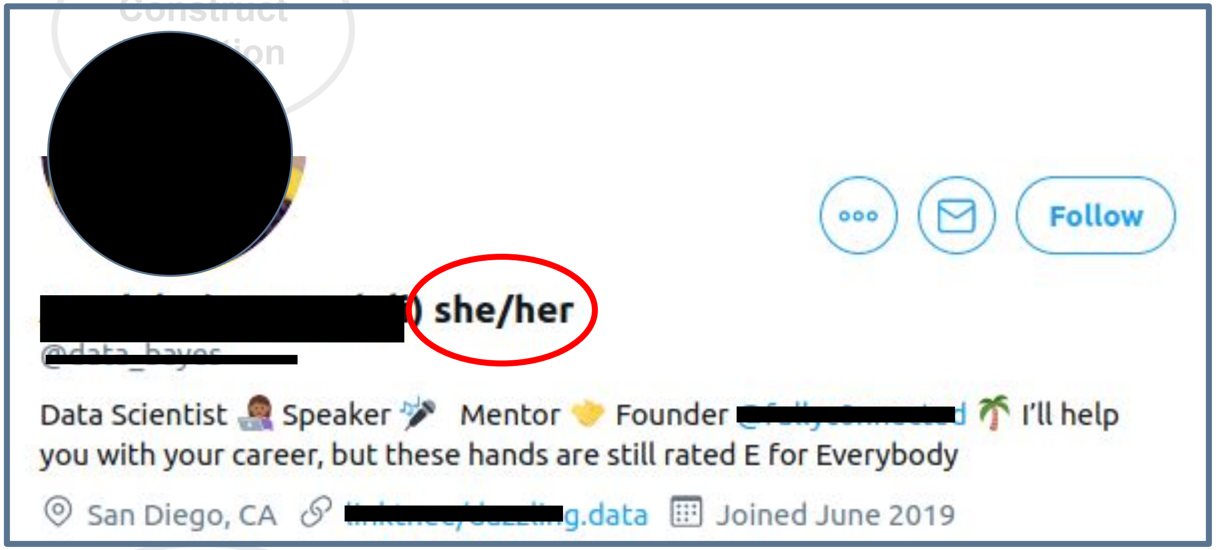


criteria used for filtering tweets may be error-prone

TSE

- Preprocessing: Processing Error

REPRESENTATION



[redacted]

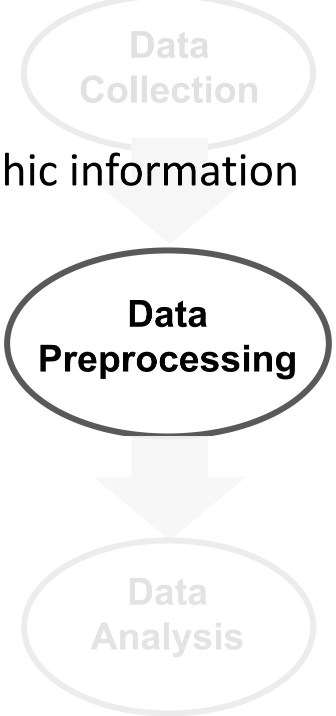
[redacted] she/her

[redacted]

Data Scientist 👩🏾‍💻 Speaker 🎤 Mentor 👏 Founder [redacted] 🌴 I'll help you with your career, but these hands are still rated E for Everybody

San Diego, CA [redacted]g.data Joined June 2019

Augment users with demographic information



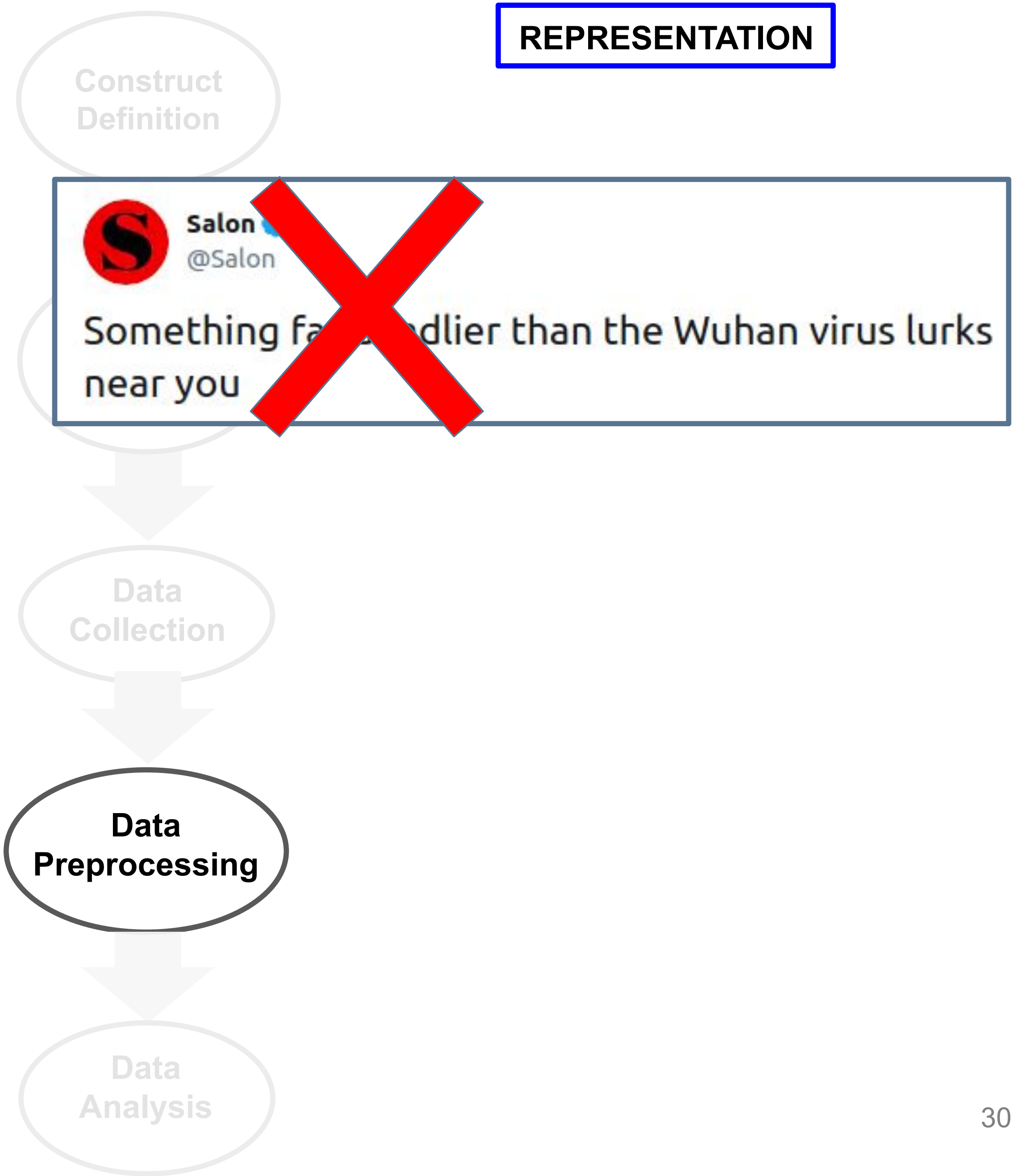
Construct

Data Collection

Data Preprocessing

Data Analysis

REPRESENTATION

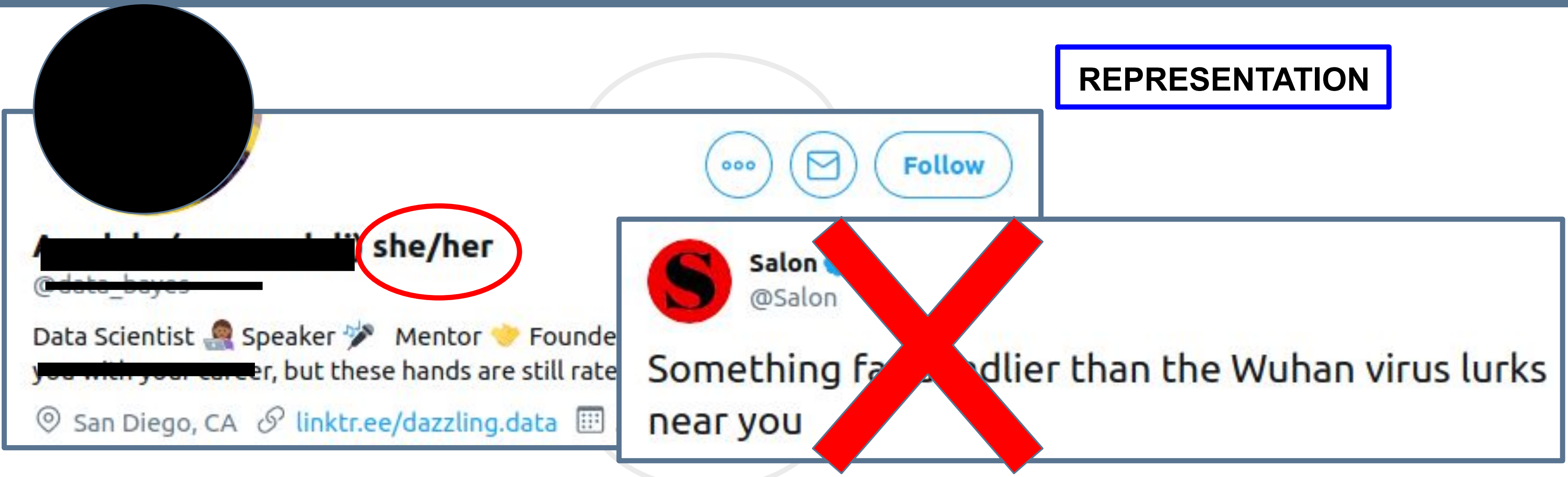


Salon
@Salon

Something fa dlier than the Wuhan virus lurks near you

Remove bots or organizations

REPRESENTATION



[redacted]

[redacted] she/her

[redacted]

Data Scientist 👩🏾‍💻 Speaker 🎤 Mentor 🤝 Founde

[redacted]r, but these hands are still rate

San Diego, CA linktr.ee/dazzling.data

Salon

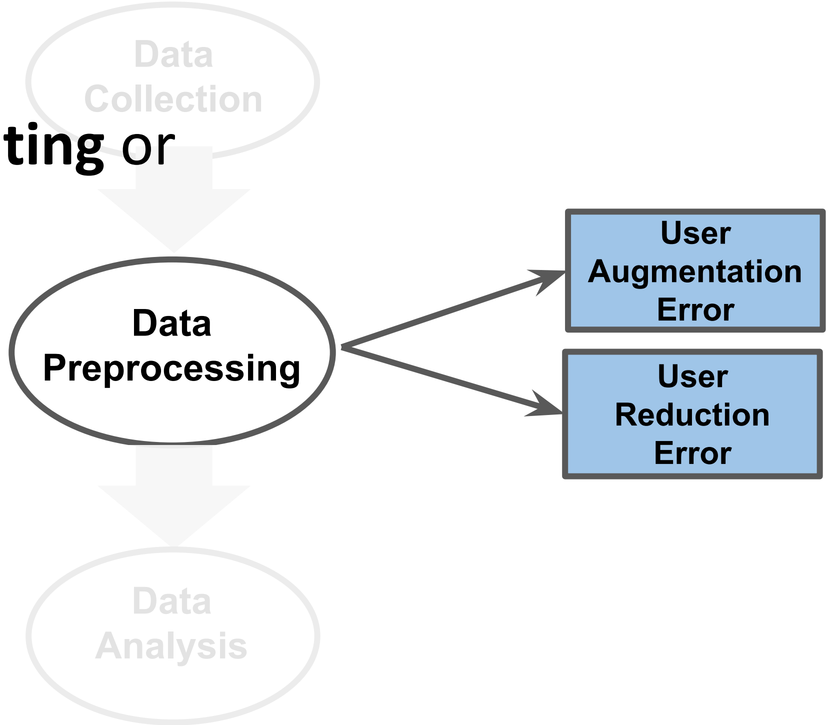
@Salon

Something fa dlier than the Wuhan virus lurks near you

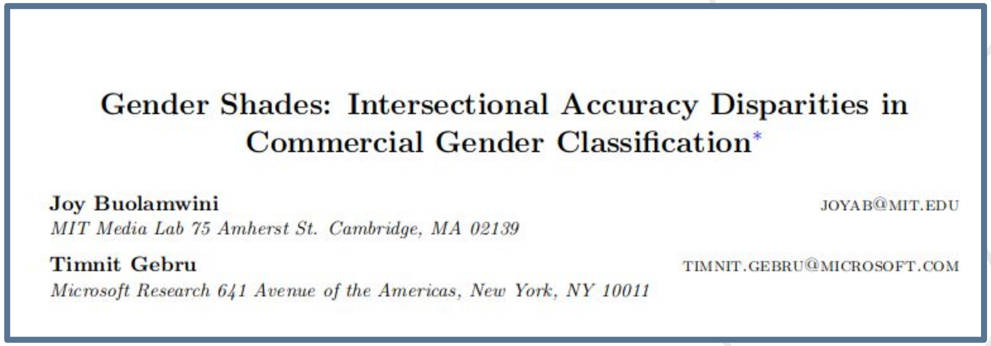
Augment users with demographic information

Remove bots or organizations

= Errors due to **augmenting** or **filtering** out users



REPRESENTATION



Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

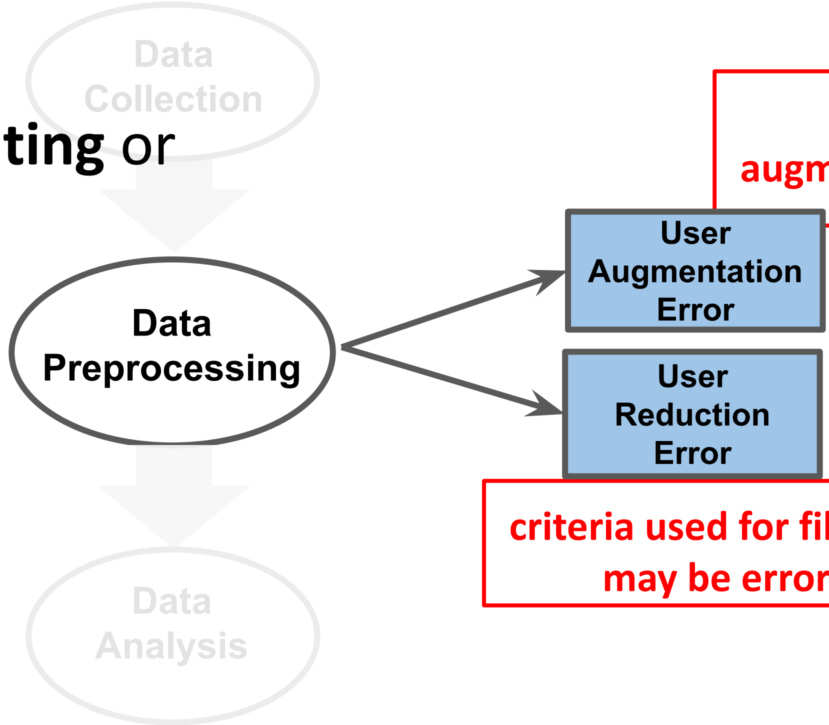
Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

Augment users with demographic information
Remove bots or organizations

= Errors due to **augmenting** or **filtering** out users

TSE
~ (Preprocessing: Processing Error)



Construct

Data Collection

Data Preprocessing

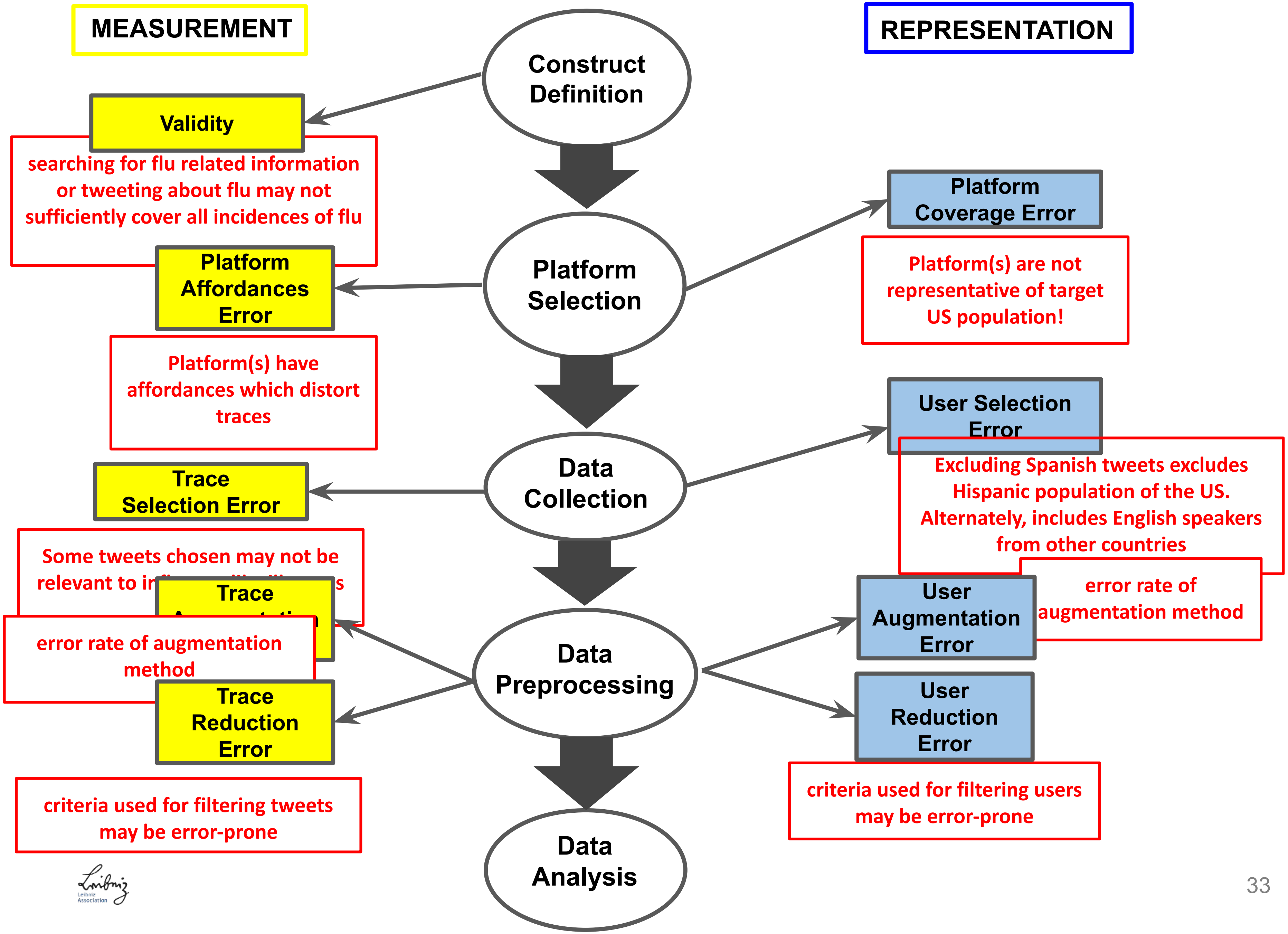
User Augmentation Error

User Reduction Error

Data Analysis

error rate of augmentation method

criteria used for filtering users may be error-prone



MEASUREMENT

REPRESENTATION

Construct Definition

Validity

searching for flu related information or tweeting about flu may not sufficiently cover all incidences of flu

Platform Selection

Platform Affordances Error

Platform(s) have affordances which distort traces

Platform Coverage Error

Platform(s) are not representative of target US population!

Data Collection

Trace Selection Error

Some tweets chosen may not be relevant to in[illegible]s

User Selection Error

Excluding Spanish tweets excludes Hispanic population of the US. Alternately, includes English speakers from other countries

Data Preprocessing

Trace A[illegible]n

error rate of augmentation method

Trace Reduction Error

criteria used for filtering tweets may be error-prone

User Augmentation Error

error rate of augmentation method

User Reduction Error

criteria used for filtering users may be error-prone

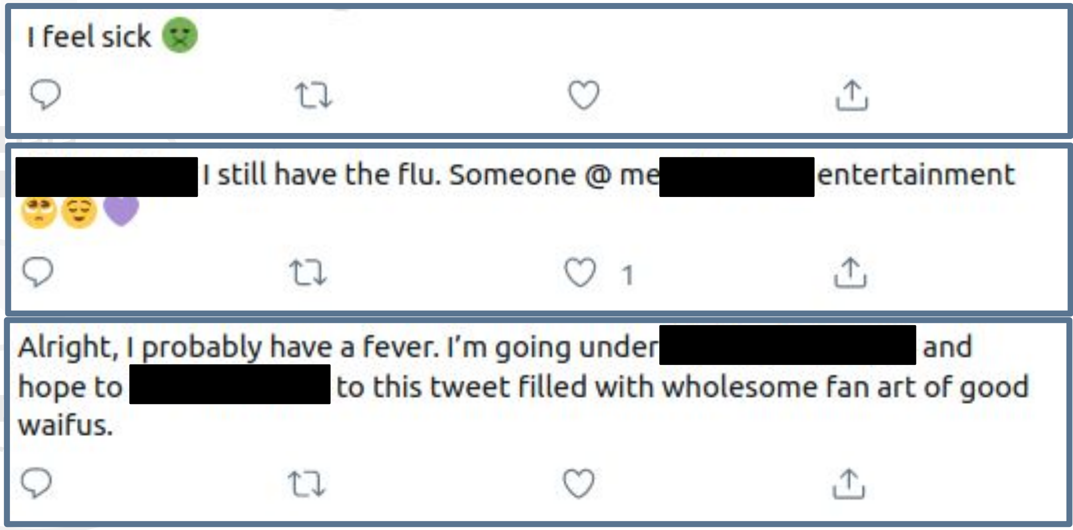
Data Analysis

MEASUREMENT

Construct Definition

Platform Selection

Develop machine learning model to **measure rates of flu** based on tweeting activity



I feel sick 🤢

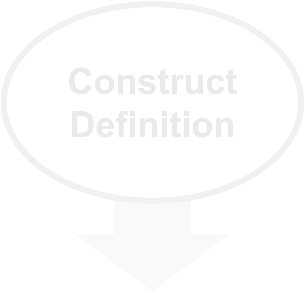
[redacted] I still have the flu. Someone @ me [redacted] entertainment 🙄😌💜

1

Alright, I probably have a fever. I'm going under [redacted] and hope to [redacted] to this tweet filled with wholesome fan art of good waifus.

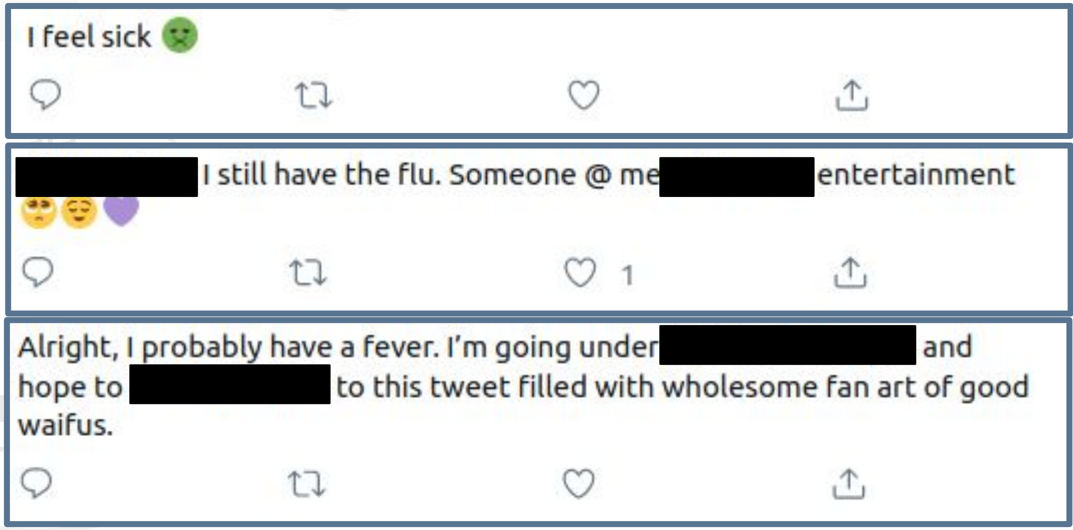
Data Analysis

MEASUREMENT



Develop machine learning model to **measure rates of flu** based on tweeting activity

= Errors due to the choice of modeling or aggregation used by the researcher as well as how the traces are mapped to the users

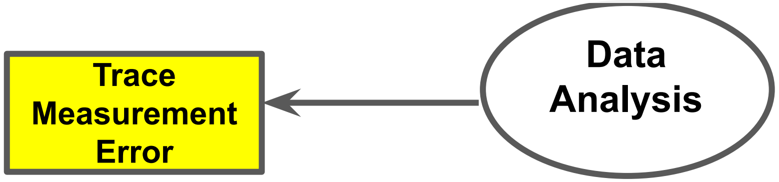


I feel sick 🤢

[redacted] I still have the flu. Someone @ me [redacted] entertainment 🙄😌💜

1

Alright, I probably have a fever. I'm going under [redacted] and hope to [redacted] to this tweet filled with wholesome fan art of good waifus.



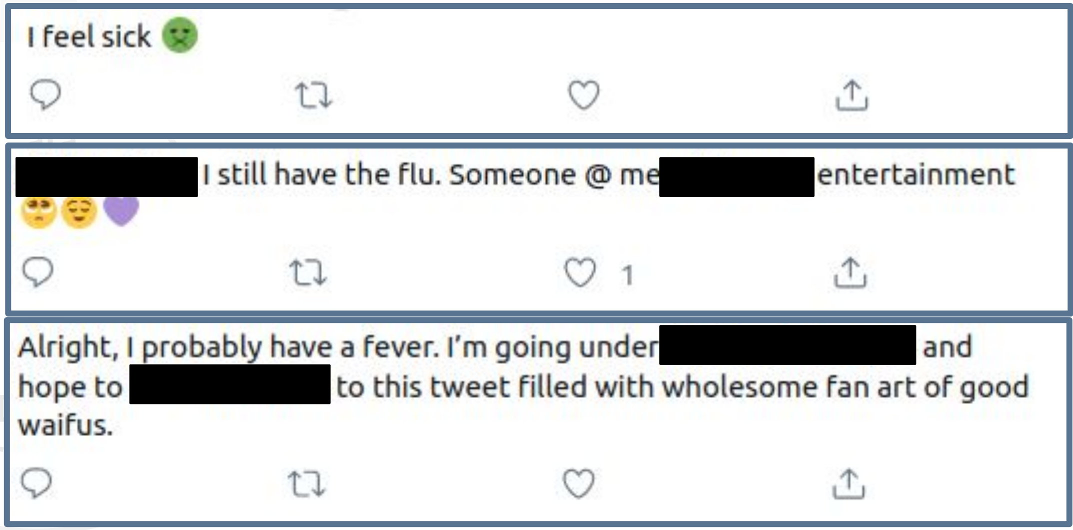
MEASUREMENT

Construct Definition

Platform Selection

Develop machine learning model to **measure rates of flu** based on tweeting activity

= Errors due to the choice of modeling or aggregation used by the researcher as well as how the traces are mapped to the users

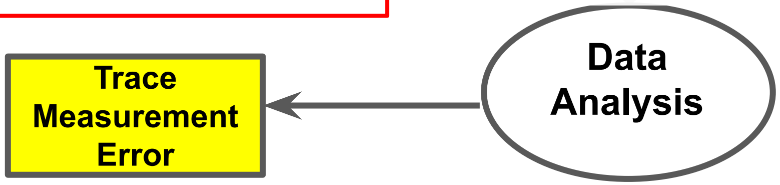

I feel sick 🤢

[redacted] I still have the flu. Someone @ me [redacted] entertainment 😩😌💜

1

Alright, I probably have a fever. I'm going under [redacted] and hope to [redacted] to this tweet filled with wholesome fan art of good waifus.

error due to how traces are included in the model



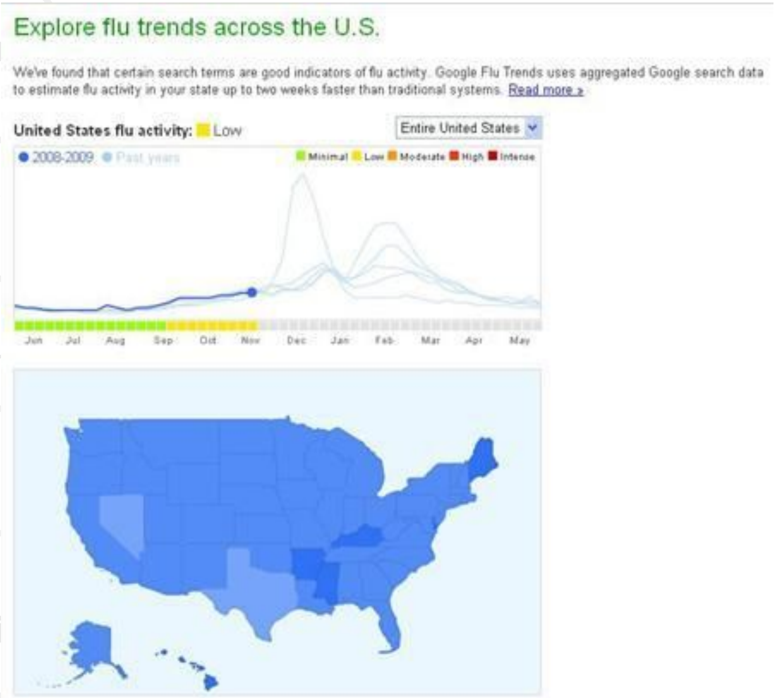
TSE
~ Analysis
~ Questionnaire design: validity

REPRESENTATION

Construct Definition

Platform Selection

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



Explore flu trends across the U.S.

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity in your state up to two weeks faster than traditional systems. Read more »

United States flu activity: 🟨 Low

Entire United States

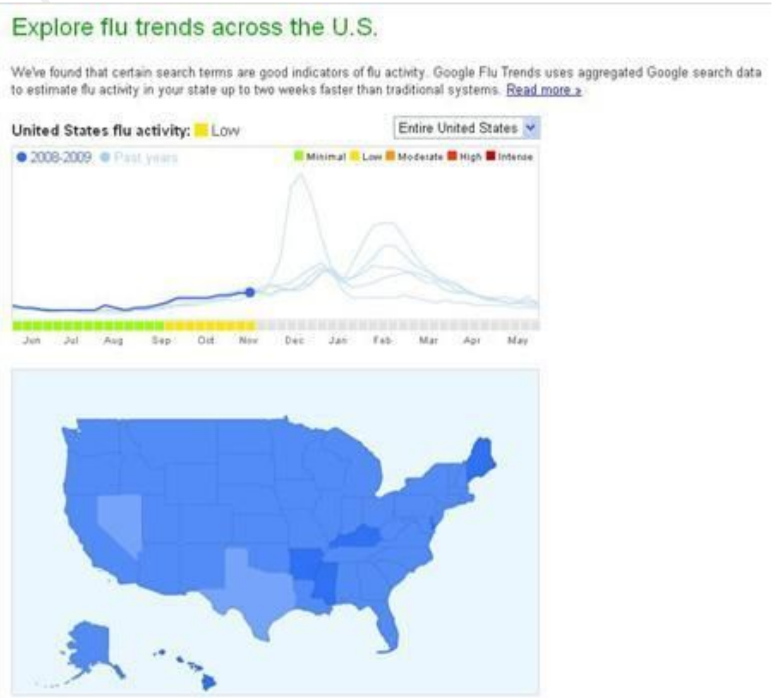
Data Collection

Data Preprocessing

Data Analysis

REPRESENTATION

Use demographic data (say, location) to **correct for coverage errors** using post-stratification



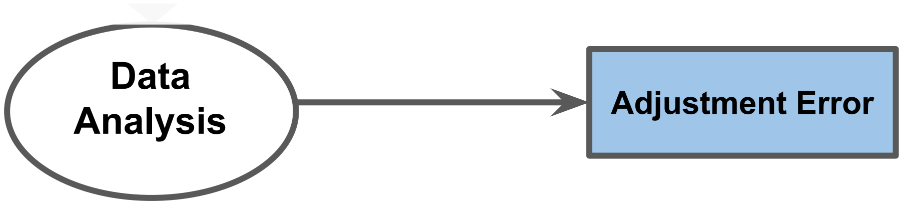
Explore flu trends across the U.S.

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity in your state up to two weeks faster than traditional systems. Read more »

United States flu activity: 🟨 Low

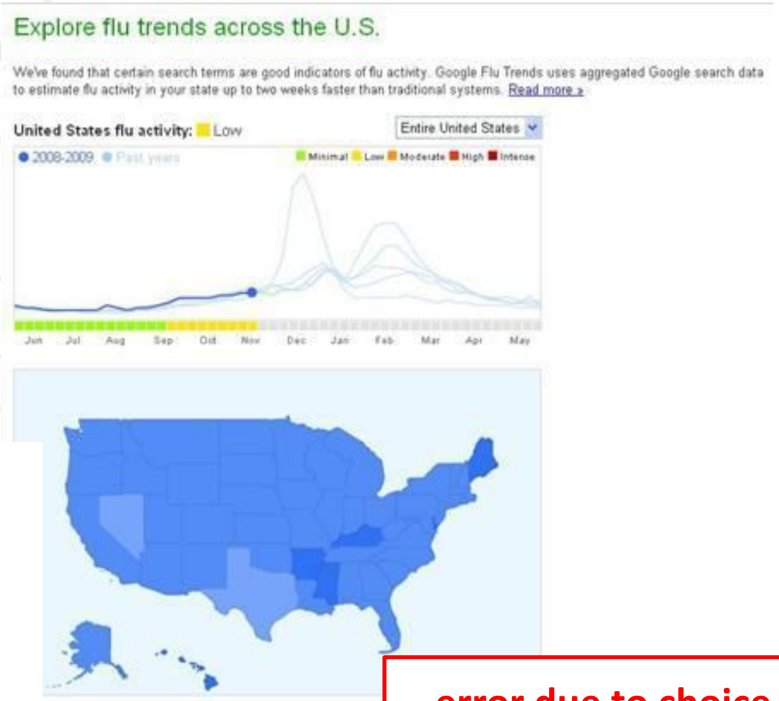
Entire United States

= Errors due to correcting for representation errors through reweighting



REPRESENTATION

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



Explore flu trends across the U.S.

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity in your state up to two weeks faster than traditional systems. Read more »

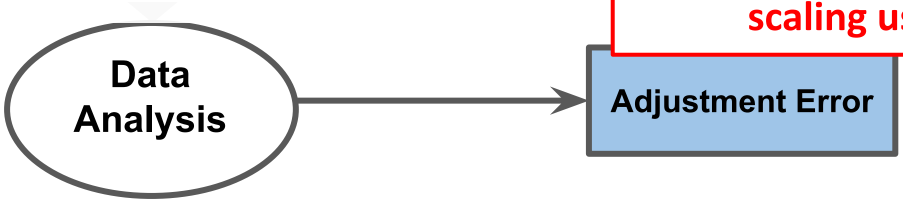
United States flu activity: 🟨 Low

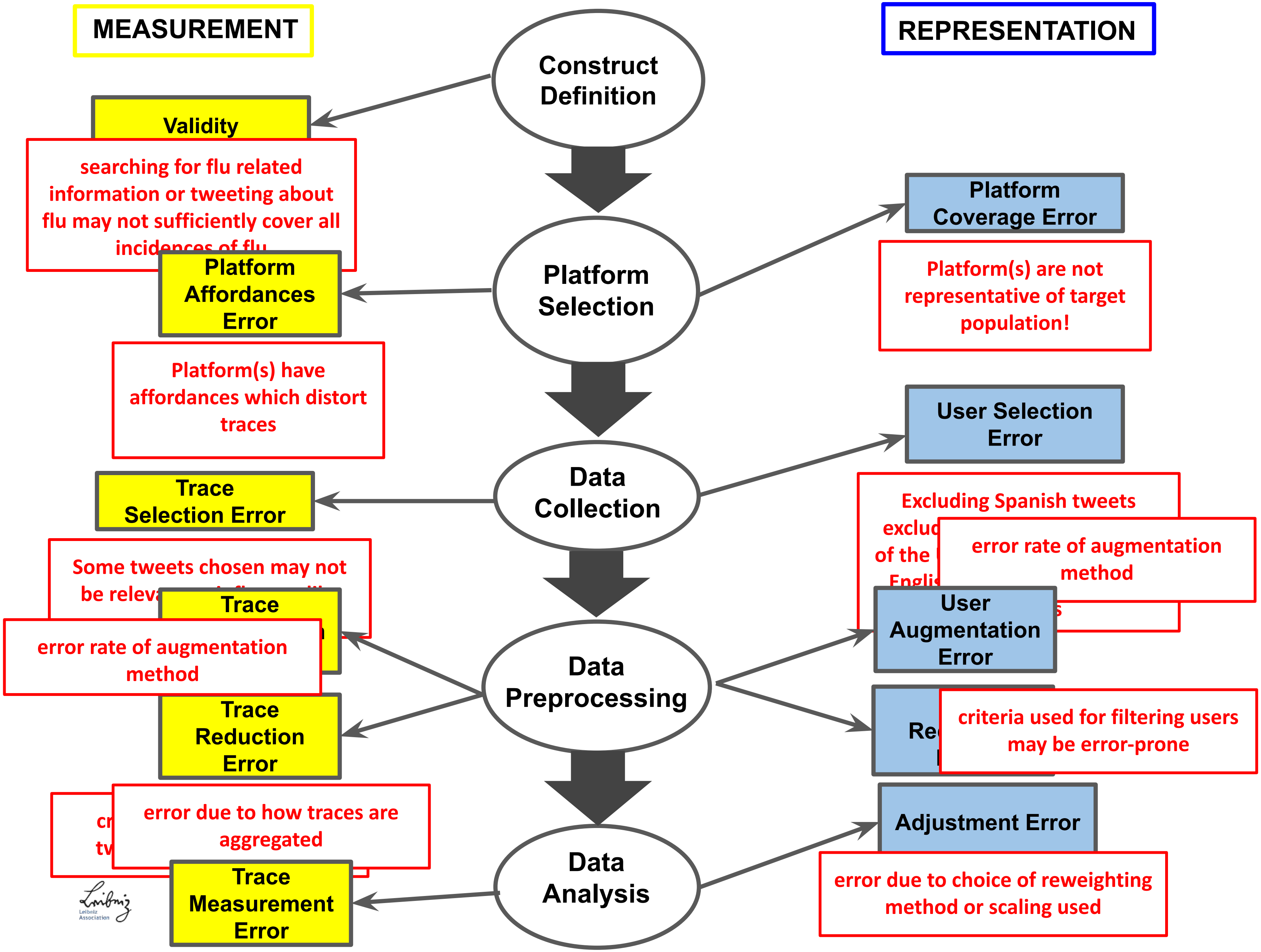
TSE

- Adjustment error

= Errors due to correcting for representation errors through reweighting

error due to choice of reweighting method or scaling used





MEASUREMENT

Validity

searching for flu related information or tweeting about flu may not sufficiently cover all incidences of flu

Platform Affordances Error

Platform(s) have affordances which distort traces

Trace Selection Error

Some tweets chosen may not be releva

Trace

n

error rate of augmentation method

Trace Reduction Error

cr

tw

error due to how traces are aggregated

Trace Measurement Error

Construct Definition

Platform Selection

Data Collection

Data Preprocessing

Data Analysis

REPRESENTATION

Platform Coverage Error

Platform(s) are not representative of target population!

User Selection Error

Excluding Spanish tweets

excluc

of the

Englis

s

error rate of augmentation method

User Augmentation Error

Re

criteria used for filtering users may be error-prone

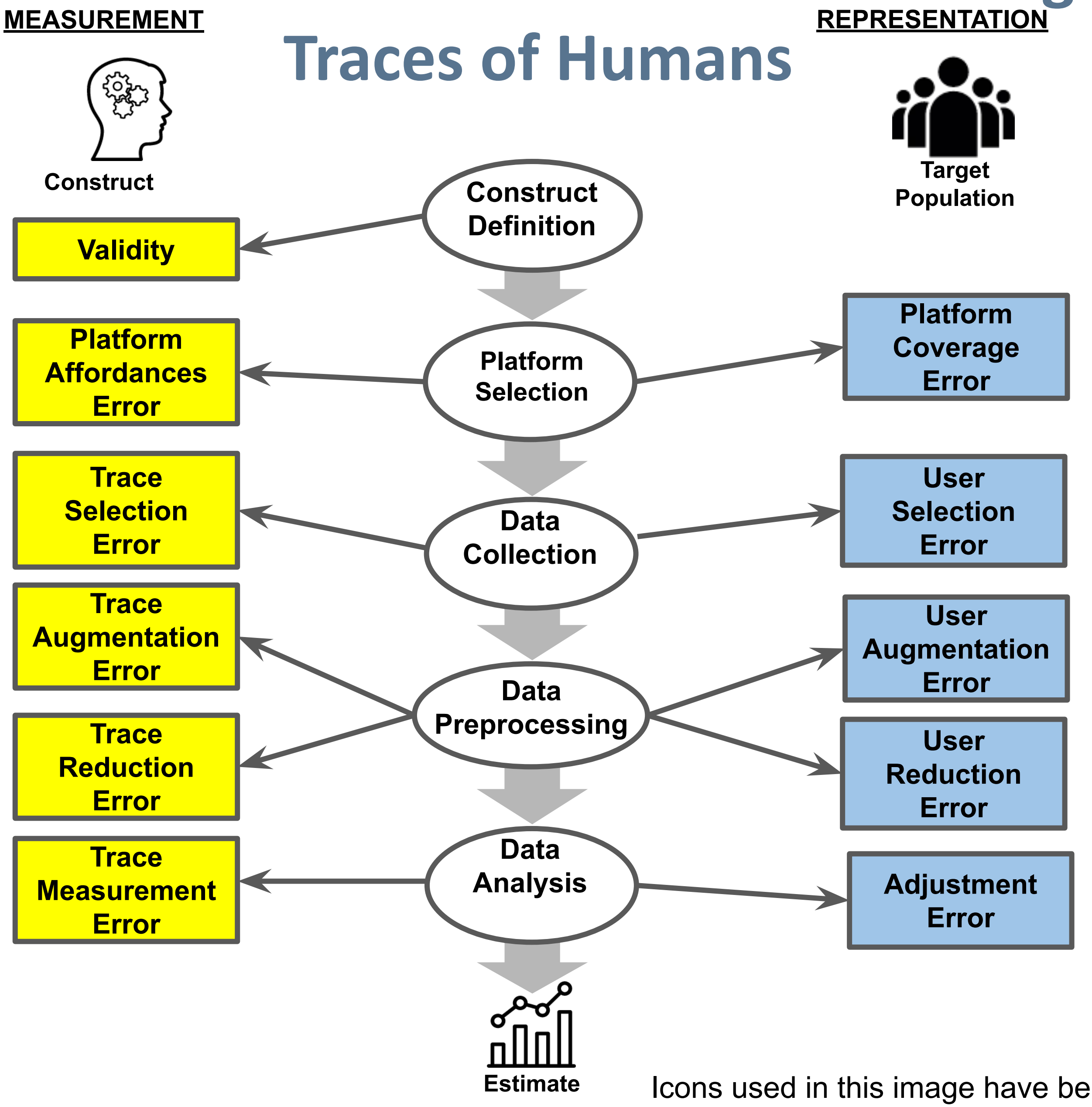
Adjustment Error

error due to choice of reweighting method or scaling used

Time as factor

- Every error has a time component, e.g.:
  - System drift → Platform affordances
  - Population drift → Coverage
  - Behavioral drift → Trace selection
  - Behavioral drift, System drift → User augmentation
  - ...

TED-On: A Total Error Framework for Digital Traces of Humans



MEASUREMENT

Construct

REPRESENTATION

Target Population

Construct Definition

Validity

Platform Selection

Platform Affordances Error

Platform Coverage Error

Data Collection

Trace Selection Error

User Selection Error

Data Preprocessing

Trace Augmentation Error

Trace Reduction Error

User Augmentation Error

User Reduction Error

Data Analysis

Trace Measurement Error

Adjustment Error

Estimate

Icons used in this image have been designed by Becris, EliasBikbulatov and Pixel perfect from www.flaticon.com