

# Introduction to Digital Trace Data: Quality, ethics, and analysis

Final recap and Q&A

**Javier Garcia-Bernardo**

Assistant Professor

Department of Methodology and Statistics

# Exam and presentations

Grade of the presentation: Thijs is sick at the moment.

What to focus on:

- Goals of the course: understand and identify problems in data collection and analysis.
- Anything we covered in the lectures and the labs (and related material from readings).

## **19 multiple choice questions**

- Answer all questions, even if you have to answer at random

## **6 open text questions reflecting scenarios**

- Reflect on the advantages and disadvantages of different approaches
- Reflect on type of errors (either representation side or measurement side)
- Reflect on ethical principles

We won't ask you to calculate numbers.

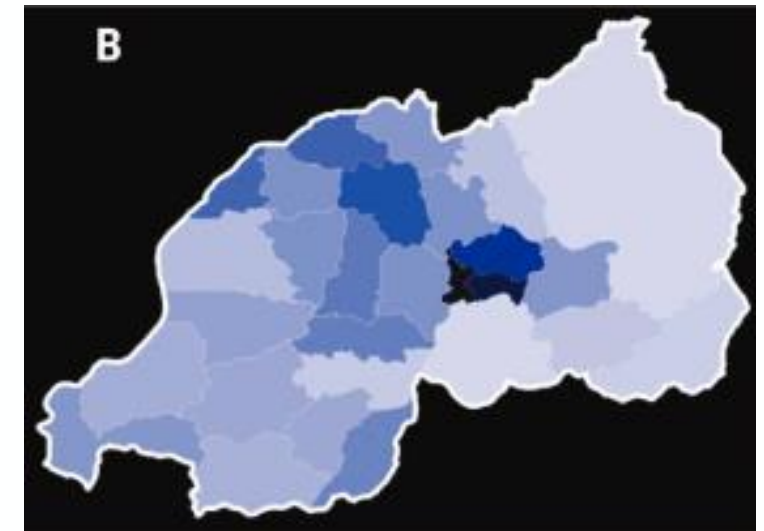
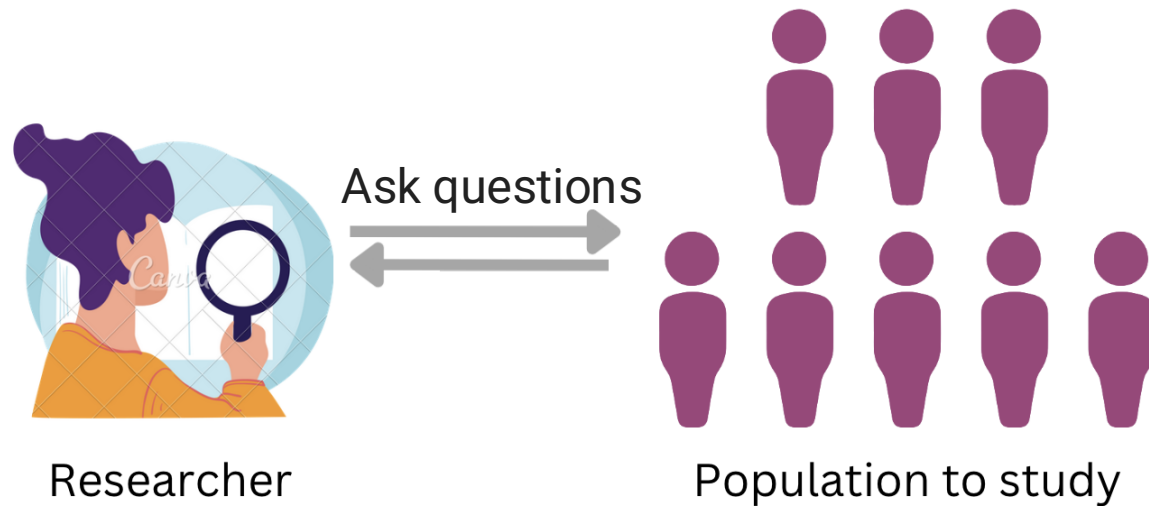
Read the email I sent last week, and bring your ID!

# How do we understand human behavior/societies?

e.g. determining poverty in Rwanda



Our traditional approach:

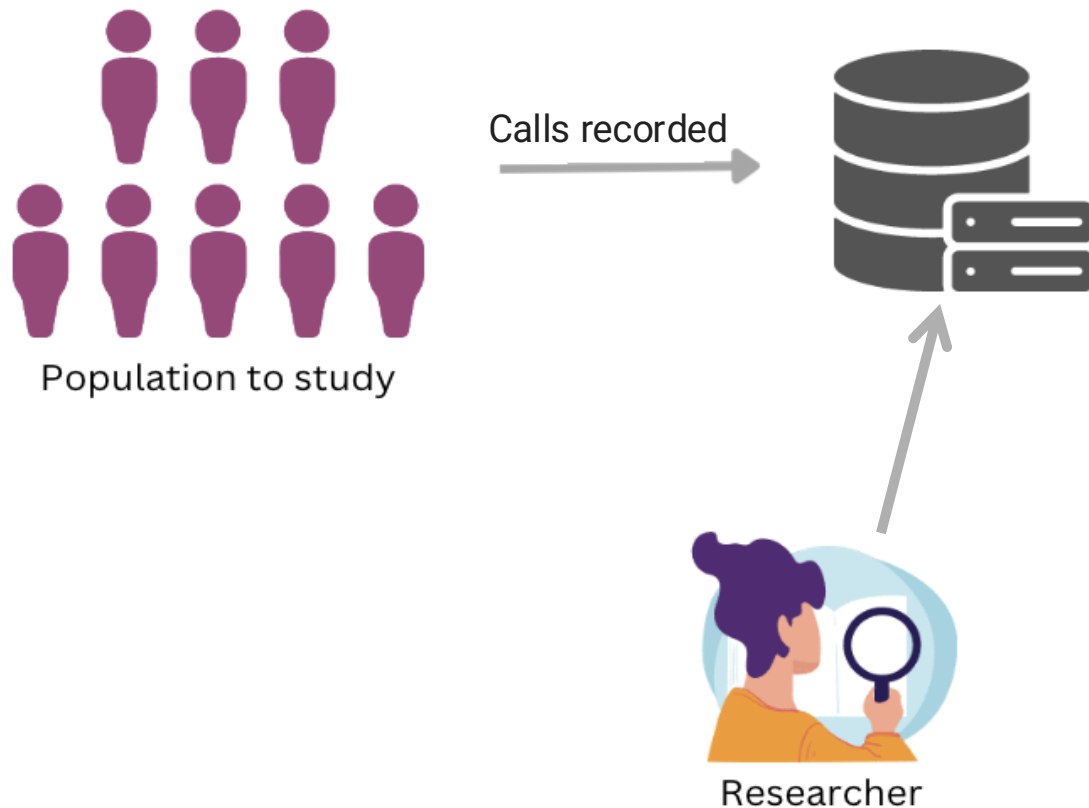


(Blumenstock et al., 2015)

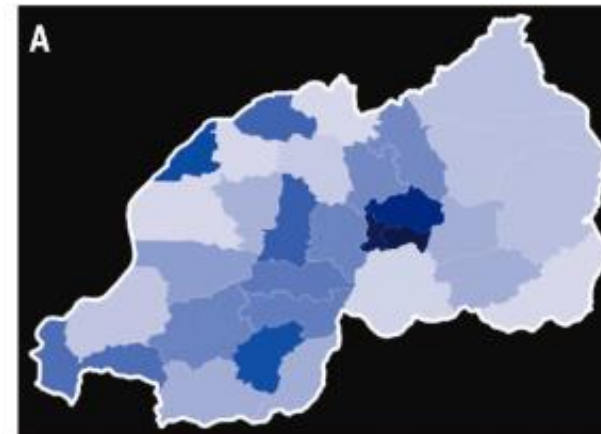
# How do we understand human behavior/societies?

But we could also use the records of individuals' digital activities, such as phone call records.

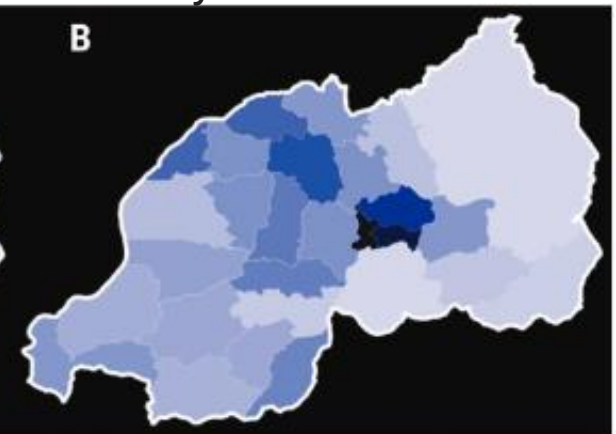
## Using Digital Trace Data:



Predicted



Survey



(Blumenstock et al., 2015)

# Advantages

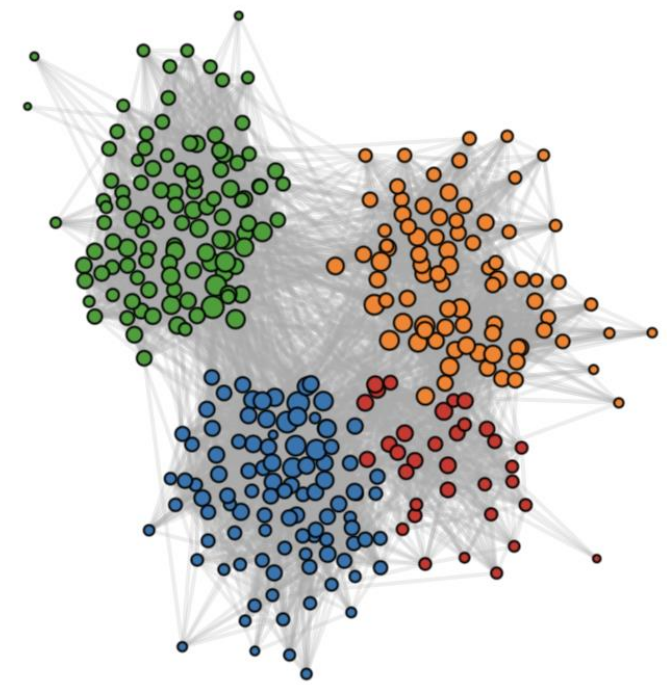
Unprecedented level of granularity: study small groups

Always on: Longitudinal data (dynamics! historical!)

It is non-reactive: it allows to study people “in-the-wild” (self-reported and real behaviour differ).

Cheaper than surveys.

New research possible: e.g. social interactions



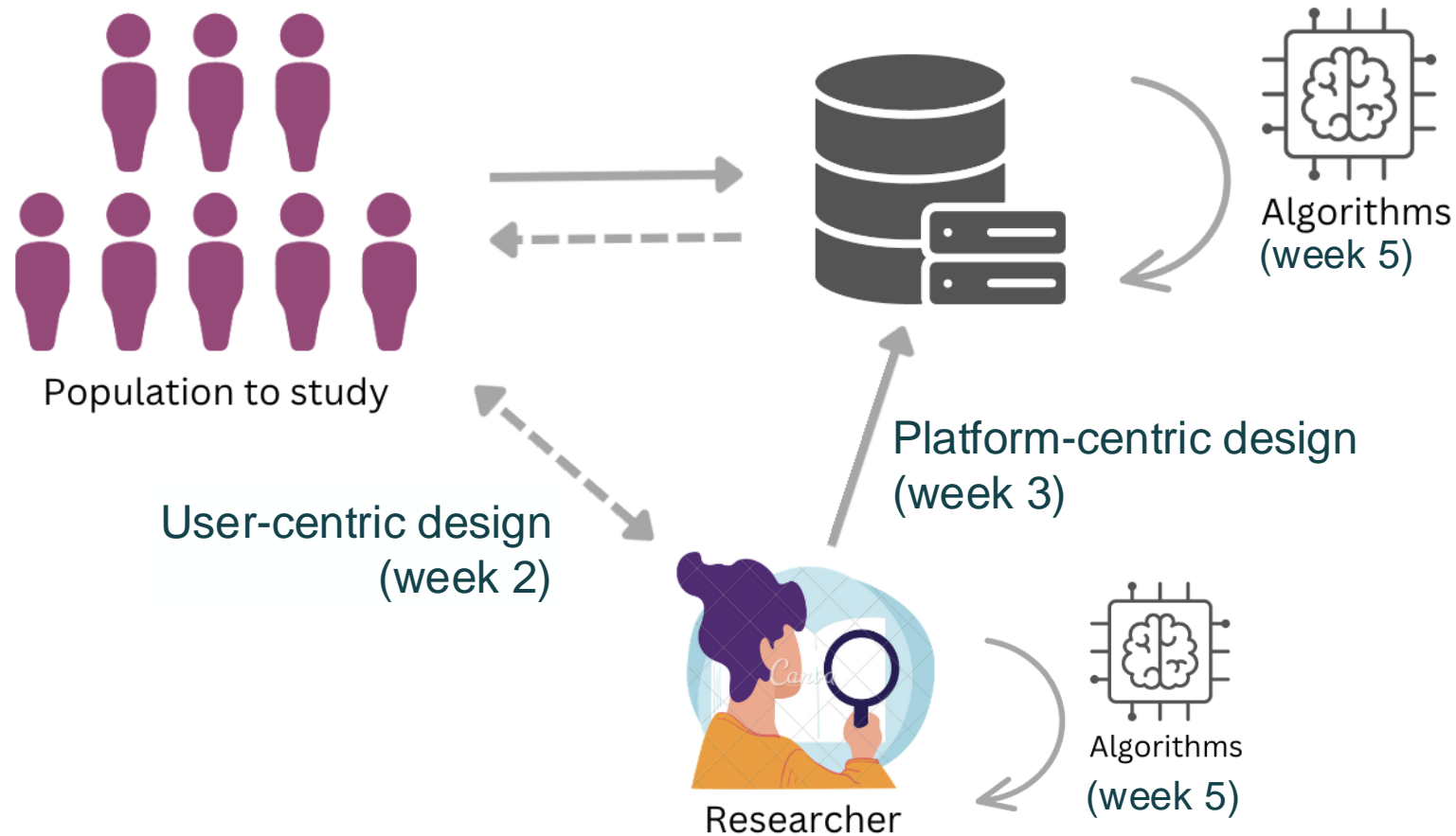
# Disadvantages

Prone to errors: measurement/representation

Confounded by algorithms; Incomplete; Drifting; Dirty; Sensitive; Inaccessible

# Collecting data (weeks 2 and 3)

Differences between user-centric and platform-centric approaches.  
Advantages/disadvantages of each approach.



# Collecting data (weeks 2 and 3)

**User-centric:** Tracking (sensors/apps) and digital data packages.

**Platform-centric:** Advantages of API and Web scraping

## **GDPR:**

- Enables data donation approaches
- And creates restrictions:
  - Collect data with a legal basis (informed consent *OR* legitimate interest)
  - Purpose/data/storage limitation
  - Accuracy/security/accountability

**[app.wooclap.com/DTD24](https://app.wooclap.com/DTD24)**



© 2024 Wooclap



# Errors in DTD (week 4)

## Two sides:

- **Representation:** is your *data* representative of the *target population*?

↓  
Twitter users using a hashtag

↓  
Dutch population

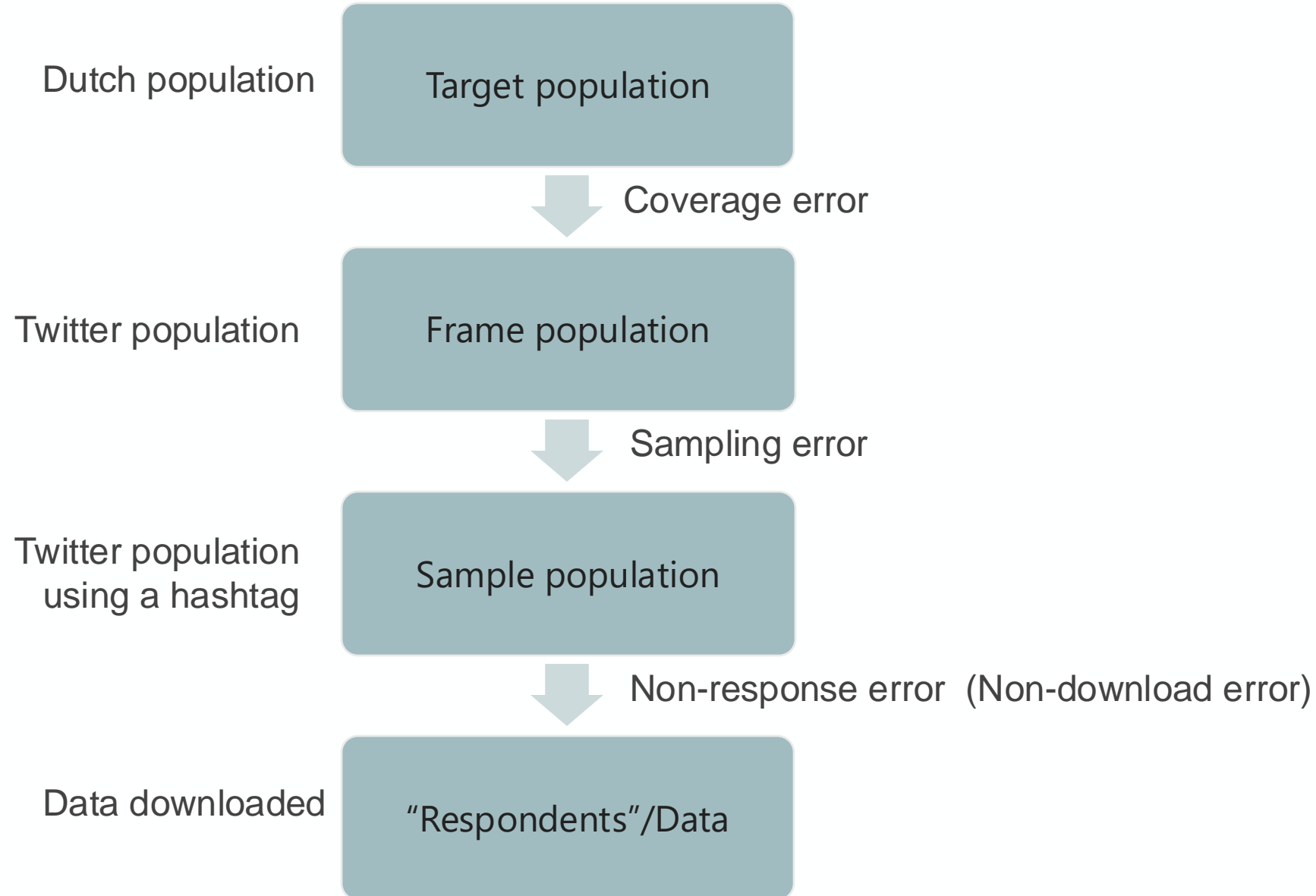
- **Measurement:** do your *variables* measure *what you are interested in*?

↓  
Number of likes from left-wing  
and right-wing users

↓  
Affective polarization

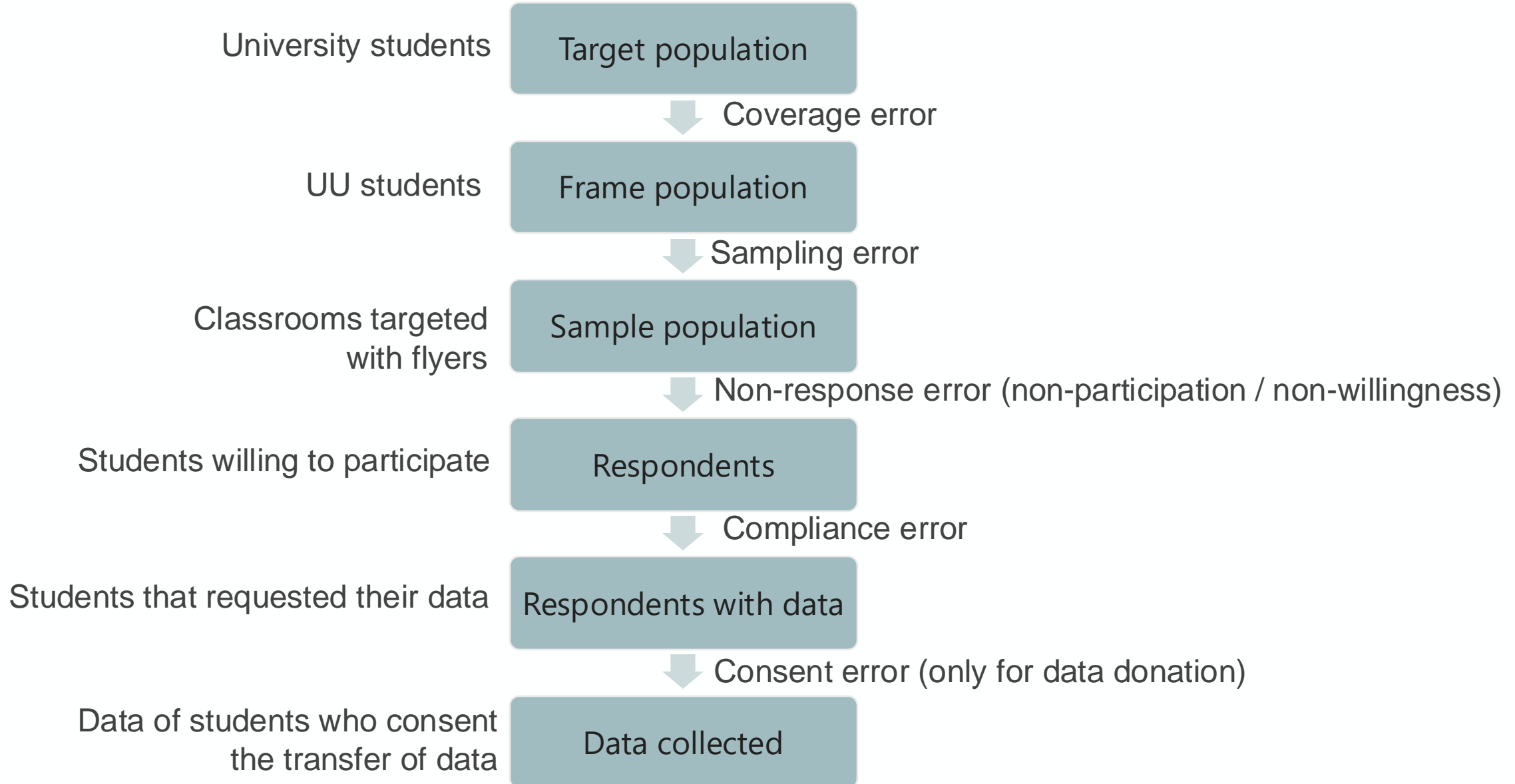
Example: You are interested in measuring affective polarization in Dutch society.

# Errors (week 4): representation, platform-centric



# Errors (week 4): representation, user-centric

Research: Music consumption during exams by university students



**[app.wooclap.com/DTD24](https://app.wooclap.com/DTD24)**



© 2024 Wooclap

# Errors in DTD (week 4): measurement

## Measurement:

### *Validity:*

- are likes on Twitter a good measurement of affective polarization?
- are Facebook friends a valid measurement of real social networks?

*Measurement error:* are the measurement correct? (maybe only the first 10 friends can be extracted)

*Processing error* (extraction, integration, labeling, etc)

→ Validation is key!

# Examine power (week 4)

**What actors are involved?**

**Examine power:**

Whose goals are prioritized and whose goals not?

Whose goals are going underserved?

Who is in charge of the institutions?

Who benefits the most from the status quo?

**How was Facebook users' data misused?**

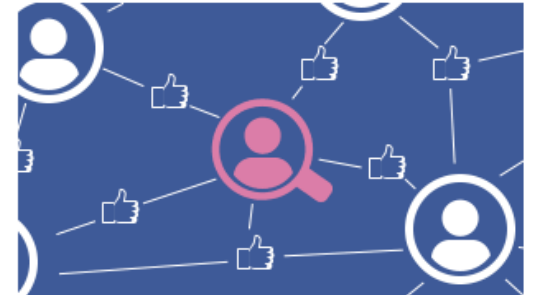
**1**

In 2014 a Facebook quiz invited users to find out their personality type



**2**

The app collected the data of those taking the quiz, but also recorded the public data of their friends



**3**

About 305,000 people installed the app, but it gathered information on up to 87 million people, according to Facebook



**4**

It is claimed the data was sold to Cambridge Analytica (CA), which used it to psychologically profile voters in the US



Cambridge analytica

# The role of AI (week 5)

AI is used:

- By companies (e.g. recommendation systems) → Validity problems (measurement)
- By research in the labeling process → Processing error (measurement)

Are these measurement or representation errors? Why?

# The role of AI (week 5)

ML models often yield errors different for different subpopulations (i.e., lack of fairness).

We can measure the quality of the model (for different subpopulations) using the confusion matrix.

In general:

- Assistive interventions: False negatives should be avoided (but also false positives if the budget is limited).
- Punitive interventions: False positives should be avoided.

Assistive		Predicted in need	Predicted not in need
In need	True positive 10	False negative 10	
Not in need	False positive 1	True negative 100	

Punitive		Predicted criminal	Predicted not criminal
Criminal	True positive 10	False negative 10	
Not criminal	False positive 1	True negative 100	



# The role of AI (week 5): sources of bias

Main idea: If the model was trained on biased data → the model will be biased when you use it

e.g. A company uses an ML model to predict whether a job candidate will be a successful employee based on their employment history, education level, and criminal record.

**Sample Bias:** Certain groups are (vastly) underrepresented in the training data.

Example: The training data has limited examples from international candidates

Impact: The model may perform poorly on these groups, leading to biased hiring recommendations.

**Feature Bias:** Certain features have different meaning for different subpopulation

Example: "Educational background" may favour candidates from prestigious universities, which are less accessible to lower-income groups.

Impact: The model may unfairly favour candidates from privileged backgrounds, reinforcing inequalities.

**Outcome Bias:** The outcome has different meaning for different subpopulation

Example: Success is defined by promotion history, which may be biased against historically marginalized groups due to prior discrimination in the workplace.

Impact: The model perpetuates past inequalities, overlooking qualified candidates from diverse backgrounds.

**Pipeline Bias:** errors in data processing and model training (e.g. removing candidates with international work experience)

# The role of AI (week 5)

Main idea: If the model was trained on biased data → the model will be biased when you use it

e.g. A researcher uses a machine learning model to label a large dataset of social media posts to understand public sentiment on a new public policy. The model is trained text from a right-wing forum labeled by UU students.

**Sample Bias:** Certain groups are (vastly) underrepresented in the training data.

Example: The model is trained on right-wing forum

Impact: May be biased for other groups.

**Feature Bias:** Certain features (the words in case of text) have different meaning for different subpopulation

Example: "woke" may be seen as something negative for right-wing people and positive for left-wing people.

Impact: The model may classify posts containing "woke" as negative.

**Outcome Bias:** The outcome has different meaning for different subpopulation

Example: UU students may be more left-wing than average and label as negative posts that were not intended to be negative.

Impact: The model reflects the views of the labellers.

**Pipeline Bias:** errors in data processing and model training (e.g. removing slang)

**[app.wooclap.com/DTD24](https://app.wooclap.com/DTD24)**



© 2024 DTD

# Ethics (week 6)

## Four principles:

- Respect for persons
- Beneficence
- Justice
- Respect for law and public interest

## How was Facebook users' data misused?

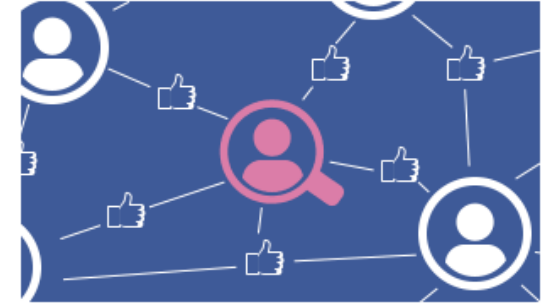
1

In 2014 a Facebook quiz invited users to find out their personality type



2

The app collected the data of those taking the quiz, but also recorded the public data of their friends



3

About 305,000 people installed the app, but it gathered information on up to 87 million people, according to Facebook



4

It is claimed the data was sold to Cambridge Analytica (CA), which used it to psychologically profile voters in the US



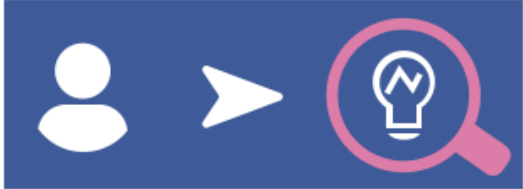
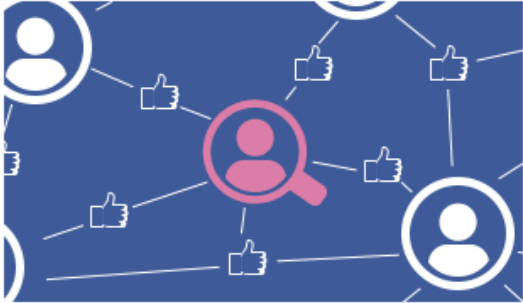

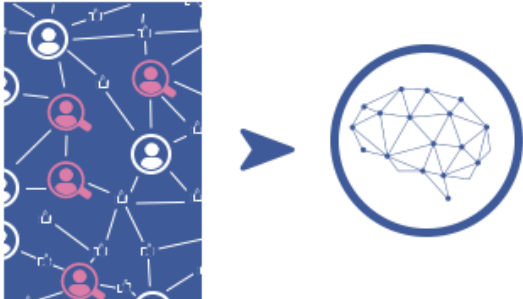
# Ethics (week 6)

## Four challenging areas:

- Informed consent
- Managing informational risk
- Privacy: appropriate flow of personal information.
  - Actors, attributes and transmission principles of contextual integrity
- Making decisions in the face of uncertainty
  - Precautionary principle: "Better safe than sorry"

The GDPR principles provide guidance!

## How was Facebook users' data misused?

- 1** In 2014 a Facebook quiz invited users to find out their personality type 
- 2** The app collected the data of those taking the quiz, but also recorded the public data of their friends 
- 3** About 305,000 people installed the app, but it gathered information on up to 87 million people, according to Facebook 
- 4** It is claimed the data was sold to Cambridge Analytica (CA), which used it to psychologically profile voters in the US 

# Ethics (week 6)

## Ethical frameworks:

- Deontology: the *means* matter most
- Consequentialism: the *ends* matter most
- Virtue ethics: be *good/virtuous*

**[app.wooclap.com/DTD24](https://app.wooclap.com/DTD24)**



© 2024 Wooclap

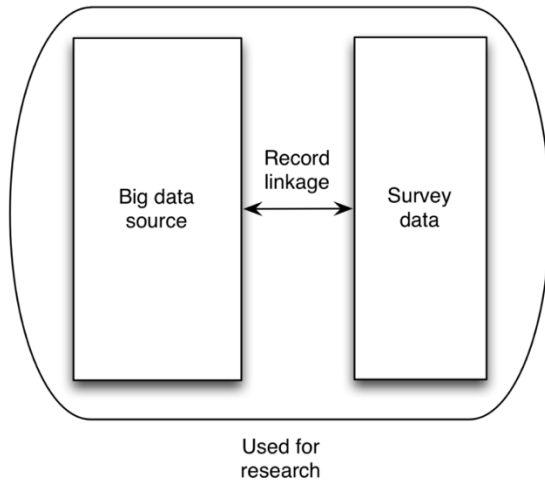
# Designed big data (week 7)

Merging survey and DTD

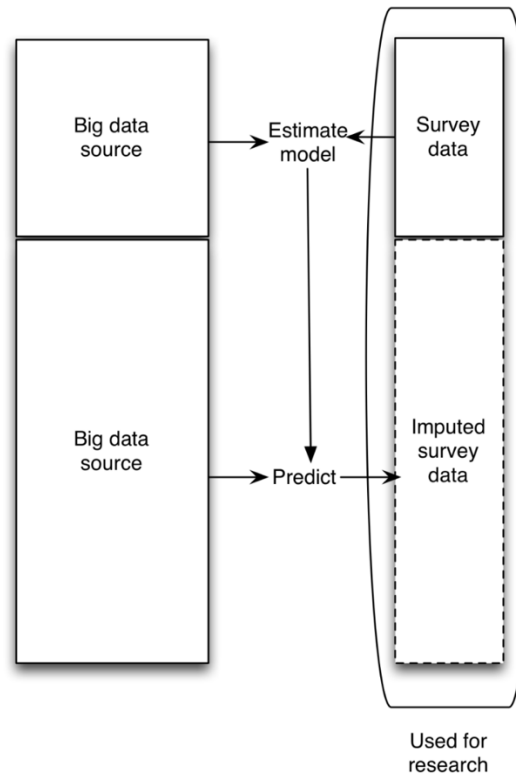
- Answering questions that cannot be answered with only one type of data (enriched asking)
- Creating predictions for a much larger population (amplified asking)

It can also help correct for errors of representation.

## Enriched asking



## Amplified asking





# Designed big data (week 7)

Record linkage: Find matching records between two data sets.

**Preprocessing:** Standardize text

**Blocking:** What comparisons to make?

**Matching:** Compare each record of dataset 1 with each record of dataset 2.

**Outcome:** Record pairs that correspond to the same entity.

**Evaluate:** Estimate the resulting error rates

**[app.wooclap.com/DTD24](https://app.wooclap.com/DTD24)**



© 2024 Wooclap

# Example open question

[app.wooclap.com/DTD24](https://app.wooclap.com/DTD24)



Copyrighted material

# Outlook

- DTD open **vast possibilities** for understanding societal trends, with the power to address global challenges if used ethically. But a lot can go wrong if the data is used carelessly.
- **Now you know how to ethically collect data and properly understand the errors!** You also know that data and models are not neutral but can consolidate or break down power relations.
- Next steps: Continue being critical. Learn how to analyze data (e.g. Applied Data Science and Visualization course, Text Mining course).

**Other questions?**



# Job opportunity

---

We are looking for **two** student assistants to work on a project involving political attitudes and Reddit data (~4 hours week/4 months).

## Requirements:

- Understanding of US attitudes
- EU passport

## Tasks:

- Manual coding of attitudes in Reddit comments
- Other tasks possible (e.g. prompt engineering, developing ML models, literature review, etc).

# Evaluation



[entry.caracal.uu.nl/39291](https://entry.caracal.uu.nl/39291)

**Good luck in the exam!**