

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

Final recap and Q&A

Laura Boeschoten

Assistant Professor

Department of Methodology and Statistics

Exam and presentations

Grade of the presentation: Individual grades are added to OSIRIS (hopefully) this afternoon

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

3 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

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

Student questions

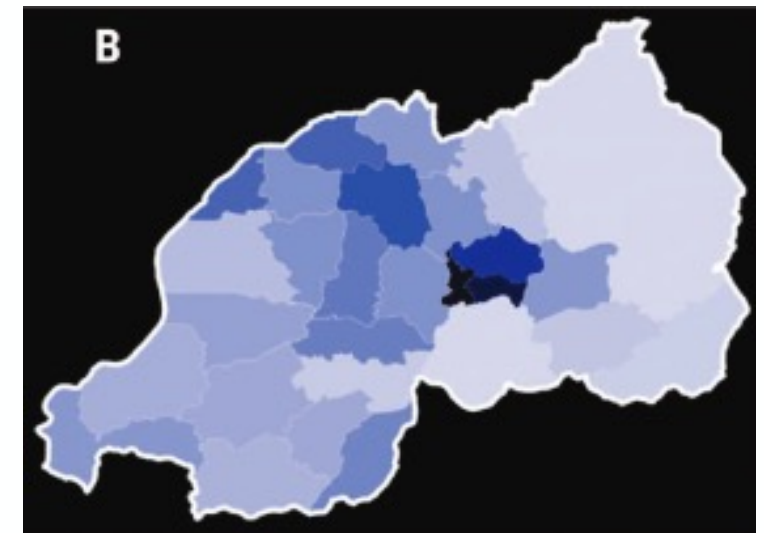
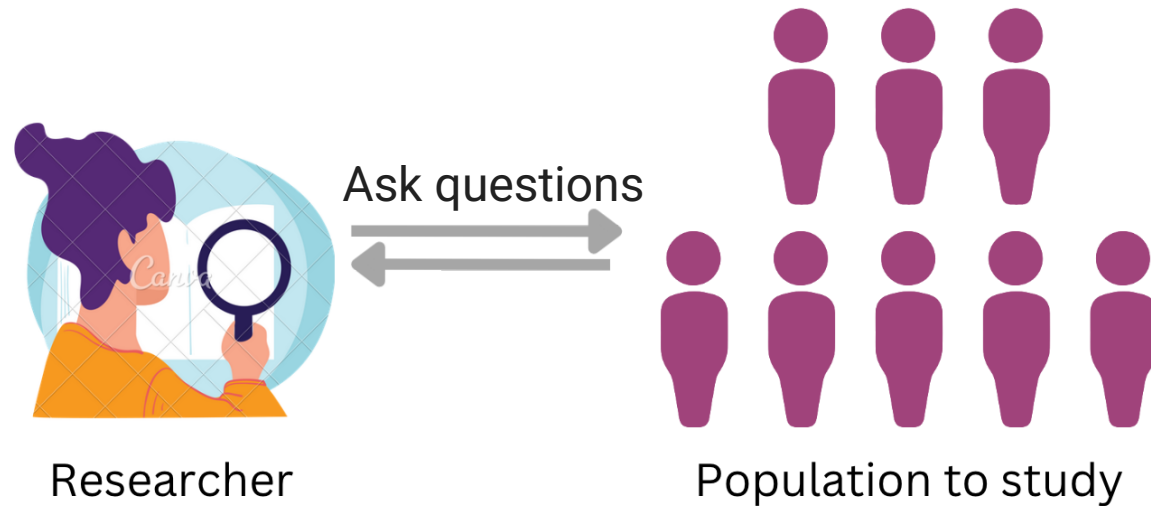
1. Does it require close reading of the literature? → in some cases yes, but be critical with details. Are they relevant for the course?
2. What kindsof questions will be there? → you will see some examples today
3. Will there be questions about coding? → no

How do we understand human behavior/societies?

e.g. determining poverty in Rwanda



Our traditional approach:

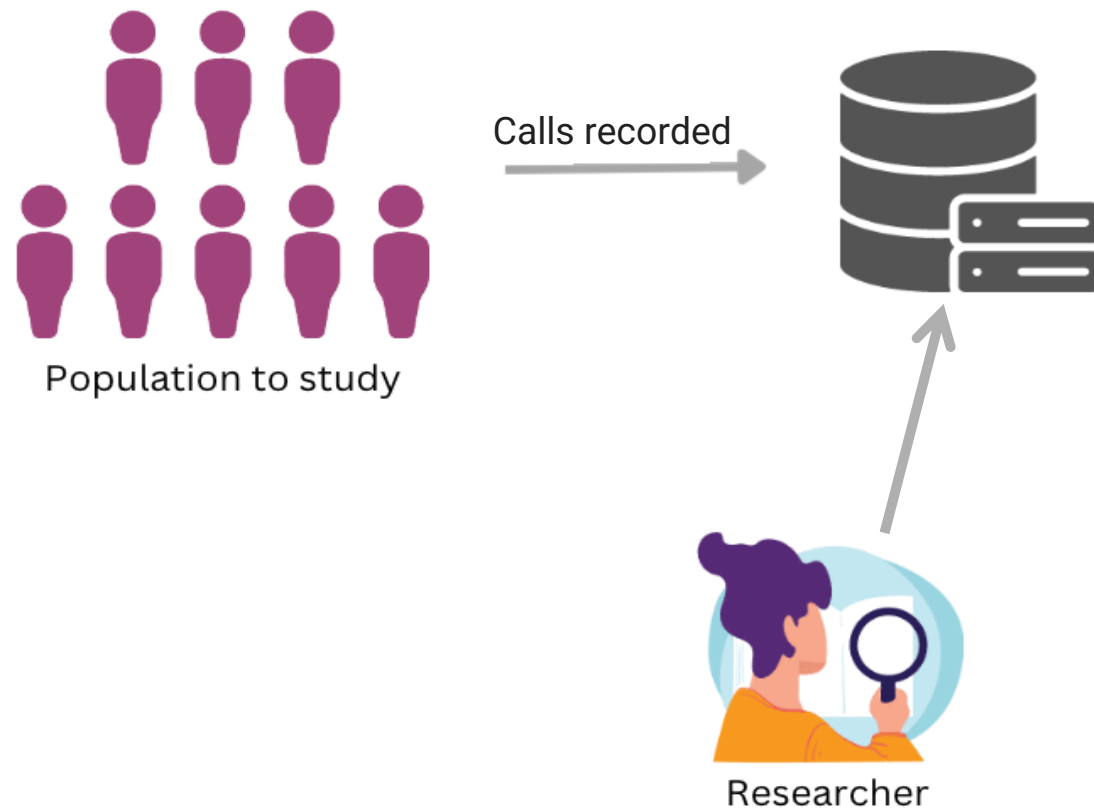


(Blumenstock et al., 2015)

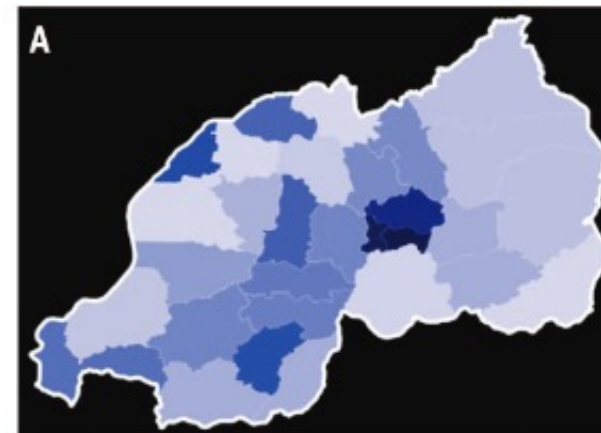
How do we understand human behavior/societies?

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

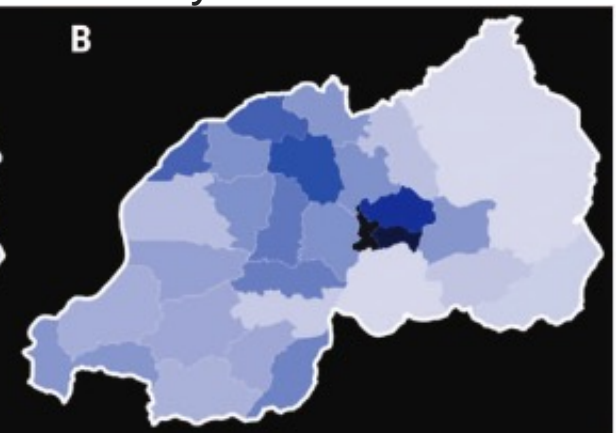
Using Digital Trace Data:



Predicted



Survey



(Blumenstock et al., 2015)

Advantages

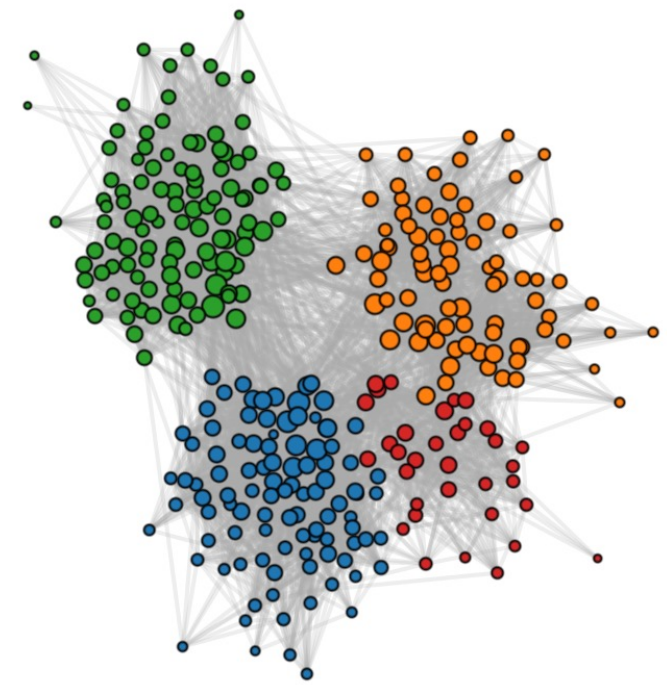
Unprecedented level of granularity: study small groups

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



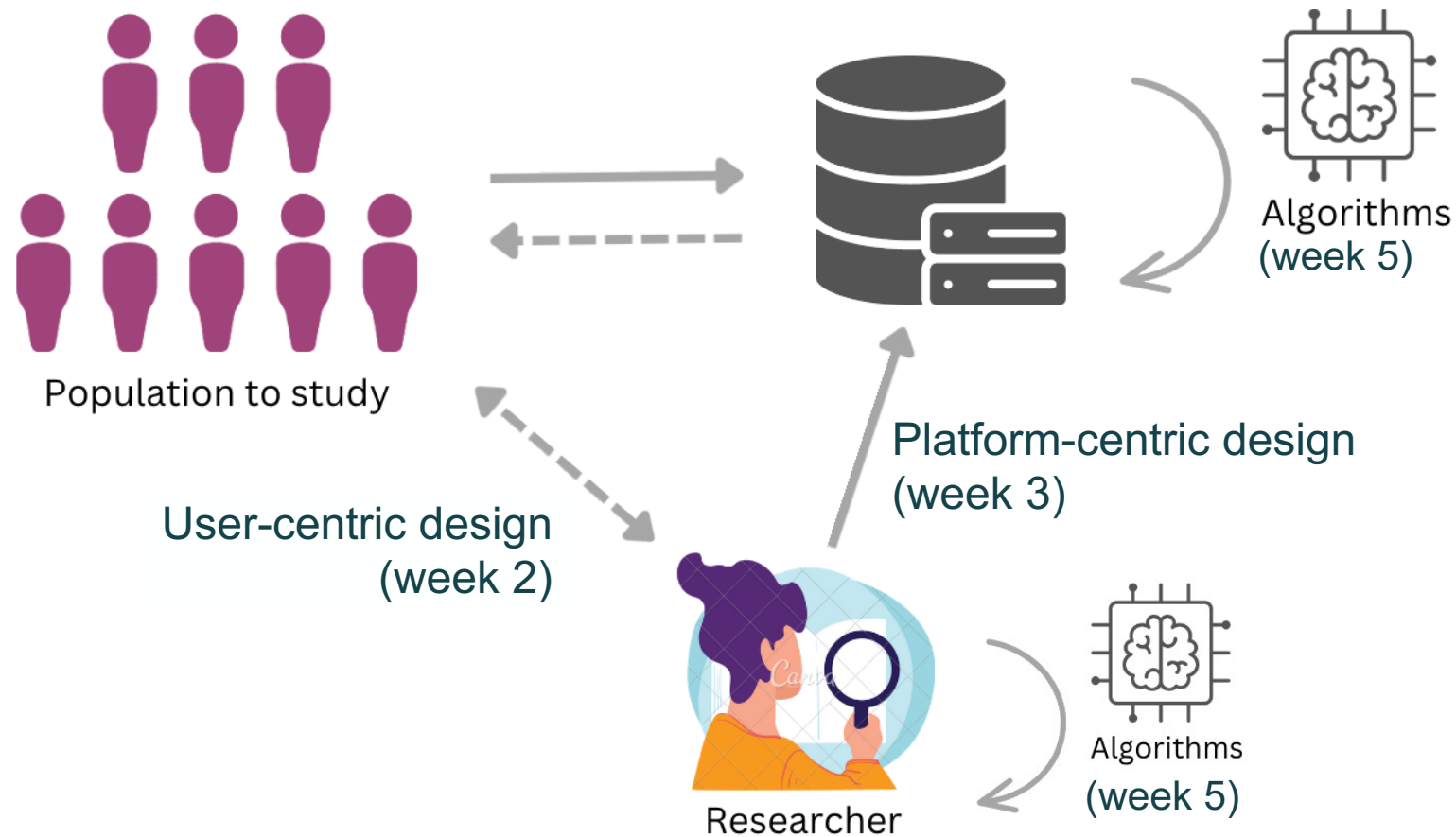
Disadvantages

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: Sensors, apps and data download packages

Platform-centric: APIs 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

Example MC question

What is the main purpose of an API key accessing web APIs?

- a) **To authenticate and authorize access to the API, often tracking usage and enforcing rate limit**
- b) To bypass rate limits and access unlimited data from the API
- c) To provide direct access to a database without authentication
- d) To automatically download data without restrictions

Example MC question

Why might a researcher choose a platform-centric design over a user-centric design?

- a) To collect data with minimal consideration for participant consent
- b) To obtain data from a large, diverse population without needing individual participation**
- c) To ensure the measurement errors are minimized
- d) To obtain highly personalized data supplemented with self-reported information

Student question

Can you give more examples on measurement and representation errors?

- 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*?



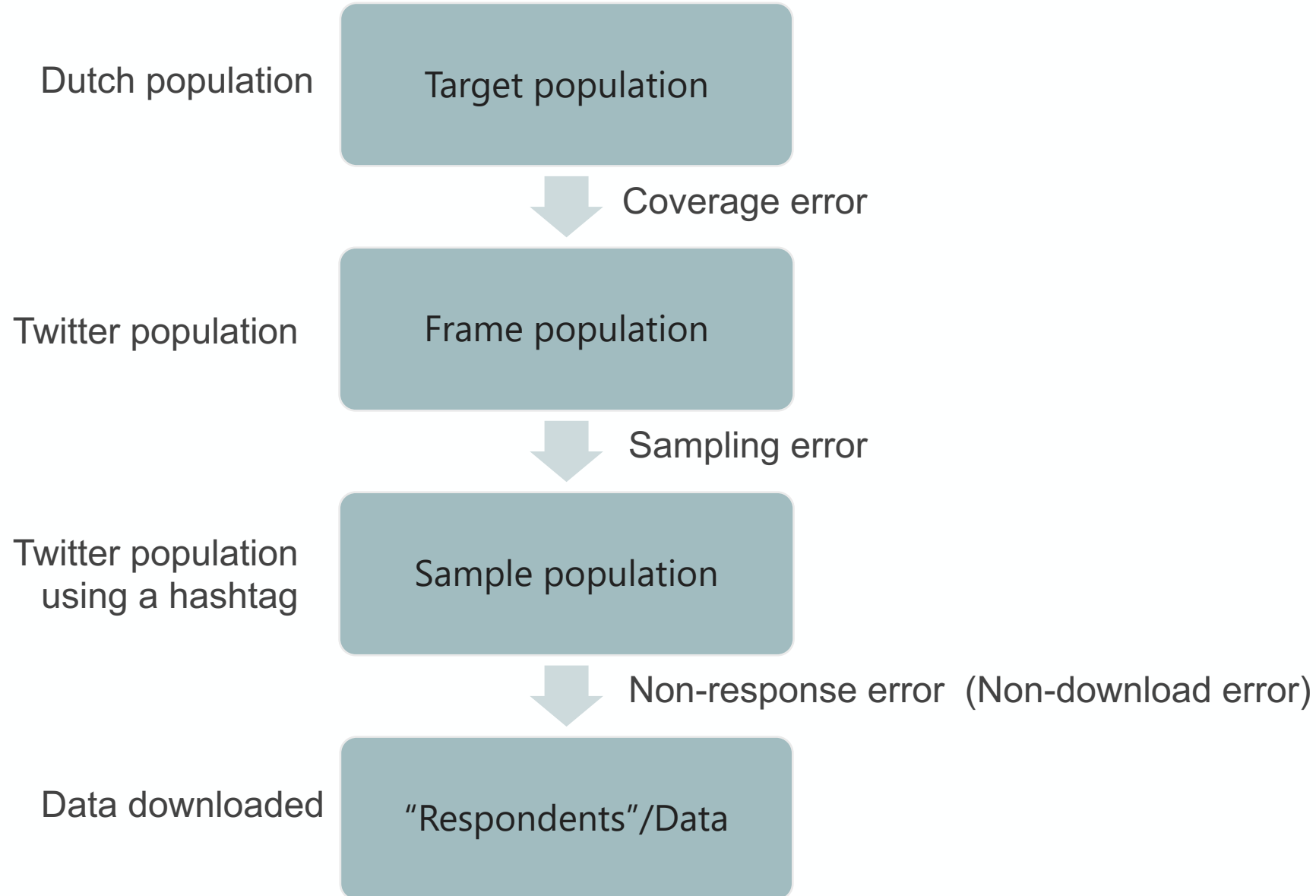
Number of likes from left-wing
and right-wing users



Affective polarization

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

Errors (week 4): representation, platform-centric



Errors 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 measurements correct? (maybe only the first 10 friends can be extracted)

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

Example MC question

In a study asking participants to install an app that tracks sleep patterns, which of the following is NOT considered a form of nonresponse error?

- a) Some participants never install the app after being invited
- b) Some participants stop using the app halfway through the study
- c) Some participants skip the daily survey questions in the app
- d) **People without smartphones are excluded from being invited to participate.**

Student question

About nonresponse error: What are:

- Missingness at random
- Missingness not at random

1. Do the people who do not donate their data differ from the people who do donate their data?

- In terms of what you try to measure in the data donations?
- For example: In the amount of time they spend on TikTok?
- Hint: you don't know because it is not observed
- Missing Not at Random → this is a problem

2. Do you have a variable observed for donators and non-donators that correlates with amount of time spend on TikTok?

- For example: you ask them to self-report the amount of time spent on TikTok
- This has error, but it correlates equally well for the donators and non-donators
- Missing at random → this is a problem, but we can fix it!

So much detail is not needed for the exam, but in block 3 there is a course about this!

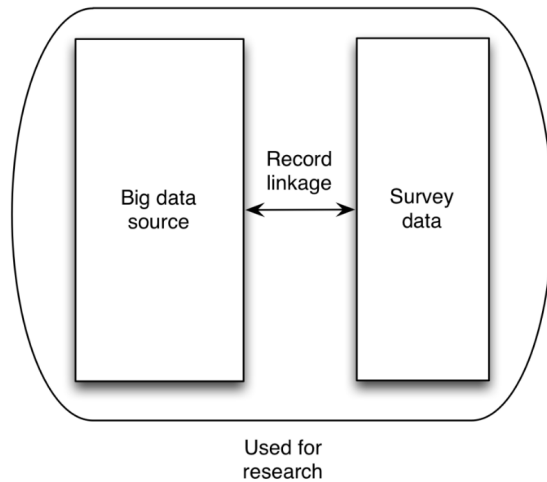
Designed big data (week 4)

Merging survey and DTD

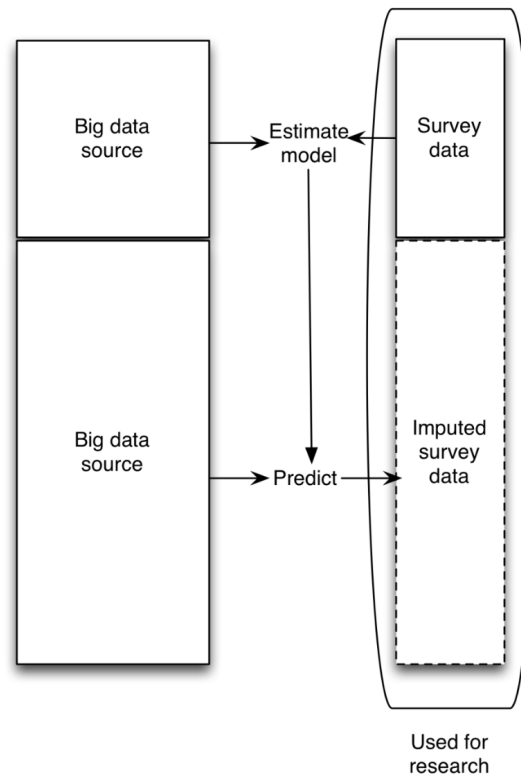
- Answering questions that cannot be answered with only one type of data (enriched asking)
- Creating predictions for a much larger population (amplified asking)

It can also help correct for errors of representation.

Enriched asking



Amplified asking



Example MC question

Which of the following scenarios best illustrates amplified asking?

- a) A sociologist merges social survey data with census records to explore demographic patterns in social attitudes
- b) A researcher combines school attendance data with student test scores to understand the relationship between attendance and academic performance.
- c) **A healthcare researcher links individual-level data from a survey on mental health with local crime statistics to estimate the mental health impact of neighborhood safety on a larger population.**
- d) An economist links employment and job satisfaction data to investigate the relationship between job stability and employee well-being.

The role of AI (week 5)

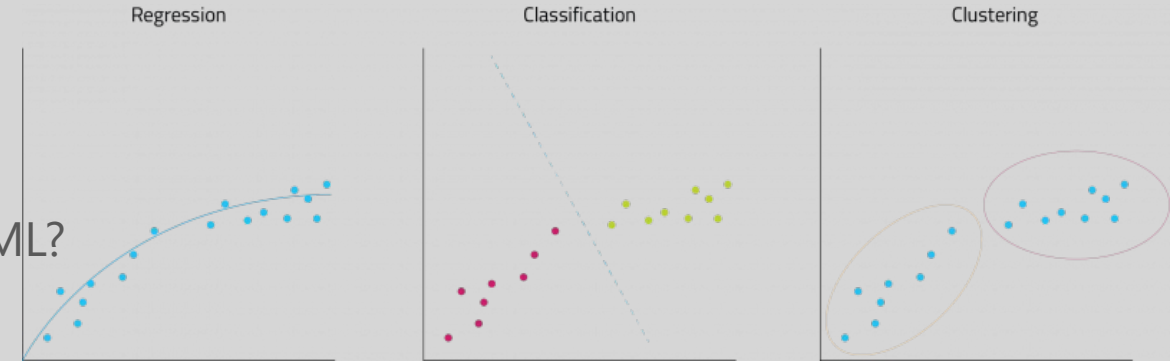
AI is used:

- 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?

Student question

Can you give more examples of supervised and unsupervised ML?



Supervised: We have inputs (features, independent variables) and an output (target, dependent variable)

Examples:

- Email spam detection: The model learns from labelled emails (“spam” or “not spam”) to make predictions for new emails coming in.
- House price prediction: The model learns from past data (“house size”, “location” and outcome “price”) to estimate the future house prices.

Unsupervised: We have inputs and (mostly) try to find groups

Examples:

- Customer segmentation: Grouping shoppers into clusters based on their behavior, without predefined labels.
- Topic modelling: Discovering themes in a large set of text documents without knowing the topics beforehand.

Examine power (week 5)

Examine power:

Who is involved in creating AI (and who is not)?

Whose goals/interests are prioritized (and whose are not)?

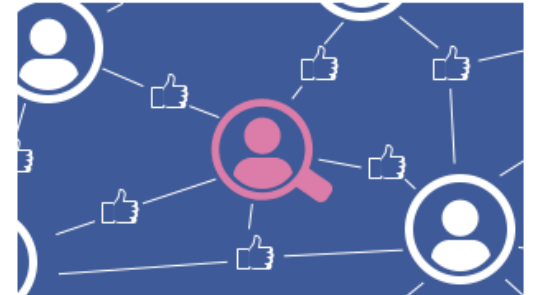
And who benefits (and who is either overlooked or actively harmed)?

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

Student question

What is the exact difference between the structural domain and the disciplinary domain in the matrix of domination? Do you have an example?

Table 1.1: The four domains of the matrix of domination ¹⁴	
Structural domain	Disciplinary domain
Organizes oppression: laws and policies.	Administers and manages oppression. Implements and enforces laws and policies.
Hegemonic domain	Interpersonal domain
Circulates oppressive ideas: culture and media.	Individual experiences of oppression.

Domain	Description	Example
Structural (the setup of a system)	Overarching system and institutions that organize access to power and resources	Credit laws and systems historically exclude women and racialized groups
Disciplinary (how the system enforces rules and norms daily)	Mechanisms that enforce norms, control and surveillance within those systems	Credit scoring algorithms penalize those same groups and limit opportunities

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	

Student question

Can you explain the FPR, FDR, FNR and FOR again?

	Predicted positive	Predicted negative
Actual positive	True positive (TP)	False negative (FN)
Actual negative	False positive (FP)	True negative (TN)

Basics of the confusion matrix:

TP: Correctly identified positive cases

FP: Incorrectly labelled as positive (false alarm)

FN: Incorrectly labelled as negative (a miss)

TN: Correctly identified negative case

Student question

Can you explain the **FPR**, **FDR**, **FNR** and **FOR** again?

	Predicted positive	Predicted negative
Actual positive	True positive (TP)	False negative (FN)
Actual negative	False positive (FP)	True negative (TN)

FPR: False positive rate:

Proportion of actual negatives that were incorrectly classified as positive: $FP / FP + TN$

- *Of all people who are innocent (negative), how many were wrongly flagged as "positive"?*
- *Risk: innocent people are being unfairly punished.*

FDR: False discovery rate:

Proportion of predicted positives that are actually false: $FP / TP + FP$

- *Of all people the model said were positive, how many were actually innocent (negative)?*
- *Risk: the positive predictions of the system cannot be trusted (many false alarms)*

Student question

Can you explain the **FPR**, **FDR**, **FNR** and **FOR** again?

	Predicted positive	Predicted negative
Actual positive	True positive (TP)	False negative (FN)
Actual negative	False positive (FP)	True negative (TN)

FNR: False negative rate:

Proportion of actual positives that were incorrectly classified as negative: $FN / TP + FN$

- *Of all people who are actually positive, how many were missed (labelled negative)?*
- *Risk: system overlooks real positives, missing people in need.*

FOR: False omission rate:

Proportion of predicted negatives that are actually positive: $FN / FN + TN$

- *Of all people the model predicted as negative, how many were actually positive?*
- *Risk: We tell too many people they are fine but they are actually at risk.*

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)

Application Bias: biases in how the model is applied (e.g. trusting an AI hiring men, but not to other genders)

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 on 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)

Application Bias: biases in how the model is applied (e.g. the researcher re-labels the social media posts of particular groups)

Student question

Can you explain the difference between feature and outcome bias

Feature Bias: Certain features have different meaning for different subpopulation

Outcome Bias: The outcome has different meaning for different subpopulation

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 on text from a right-wing forum labeled by UU students.

Feature bias: things intended as negative are classified as positive or vice versa.

Outcome bias: Classifications of the model reflect the views of the labellers and not of the population, bias in the outcome when trying to understand the public sentiment.

--> Difference is small but important, two different contexts!

Student question

Can you explain the difference between feature and outcome bias
... in the context of predictive policing

Feature Bias: Certain features have different meaning for different subpopulation

- *If police patrol some neighbourhoods more often, there will be more recorded arrests there, even if the people don't commit more crimes.*
- *The algorithms learns that those areas are "high crime", which reinforces over policing*

Outcome Bias: The outcome has different meaning for different subpopulation

- *If the model predicts where arrests happen instead of where crimes occur, its using a biased outcome, because arrests depend on who gets policed, not who commits crimes*
- *The model ends up predicting police activity, not real criminal activity.*

Example MC question

Which of the following best describes fairness in machine learning model predictions?

- a) **When the model's prediction error rates are similar across different demographic groups.**
- b) When the model consistently predicts higher probabilities for marginalized groups.
- c) When the model maximizes accuracy regardless of subgroup differences.
- d) When the model uses fewer features to minimize computation time.


Ethics (week 6)

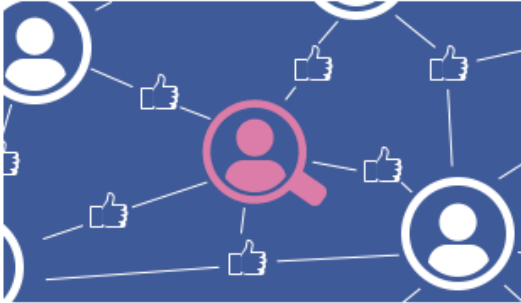
Four principles:


- 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



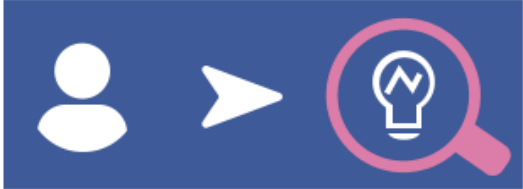
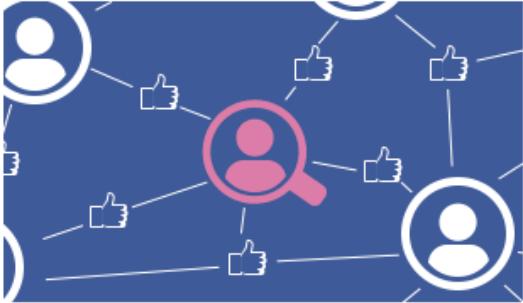


Ethics (week 6)

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*

Example MC question

Which ethical framework emphasizes the importance of fostering virtues like honesty and kindness in individuals rather than focusing strictly on duties or outcomes?

- a) **Virtue ethics**
- b) Deontology
- c) Consequentialism
- d) Utilitarianism

Example open question

In January 2023, two researchers, Sarah Mills and David Cho, conducted a study on public opinions regarding climate change by gathering data from the popular social media platform EcoTalk. EcoTalk is widely used for discussions on environmental issues and makes user comments and reaction data publicly visible. Mills and Cho accessed the platform's paid API, which allowed them to retrieve data, including usernames, post timestamps, user reactions (likes, shares), and comment text. They collected over 200,000 comments from users across various regions over two years.

After completing their study, Mills and Cho decided to publicly release the dataset to "support further research in public sentiment on climate change."

Questions:

- a) Compare briefly the advantages and disadvantages of using an API versus a data donation approach in this example. Identify 2-3 advantages of each method.
- b) Discuss the legal considerations Mills and Cho would need to address if they had chosen to scrape the data instead of using the API.
- c) List the four ethical principles and provide a reflection on how each principle applies to Mills and Cho's decision to release the EcoTalk dataset.

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).

A large white circle is centered on a dark gray background. The circle has a thin, light gray border. Inside the circle, the text "Other questions?" is written in a bold, dark gray font.

Other questions?

Evaluation

<https://caracal.uu.nl>

Good luck with the exam!