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

Lecture 1: Introduction

Javier Garcia-Bernardo
Assistant Professor
Department of Methodology and Statistics

Instructors

Laura Boeschoten

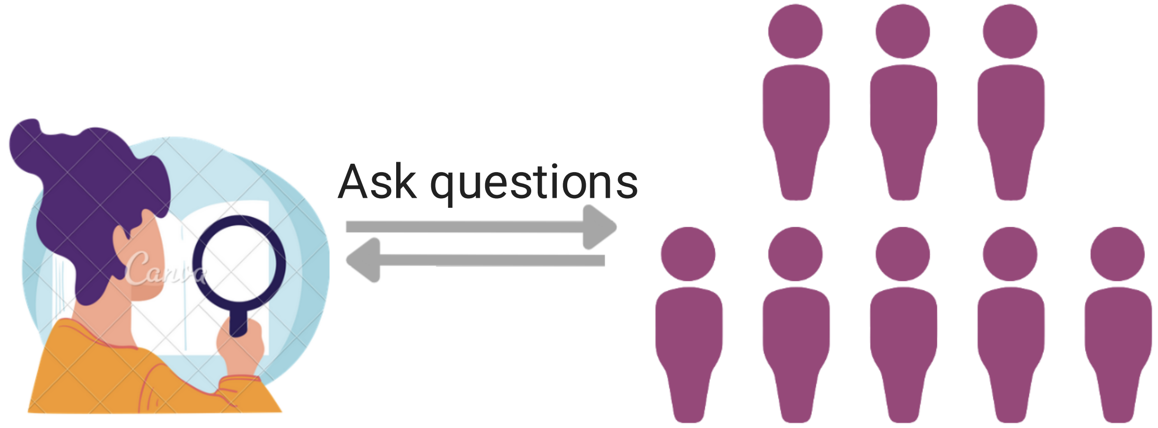
Javier Garcia-Bernardo

Thijs Carrière

How do we understand human behavior/societies?

e.g. determining poverty in Rwanda

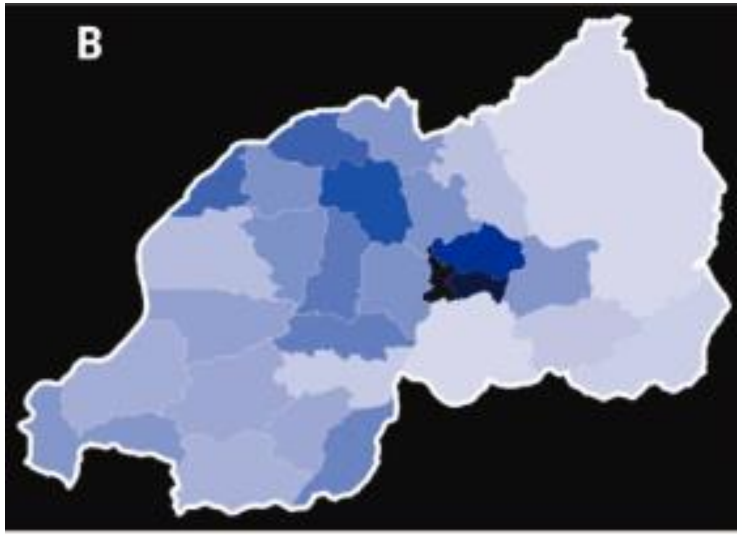
Our traditional approach:



Ask questions

Researcher

Population to study

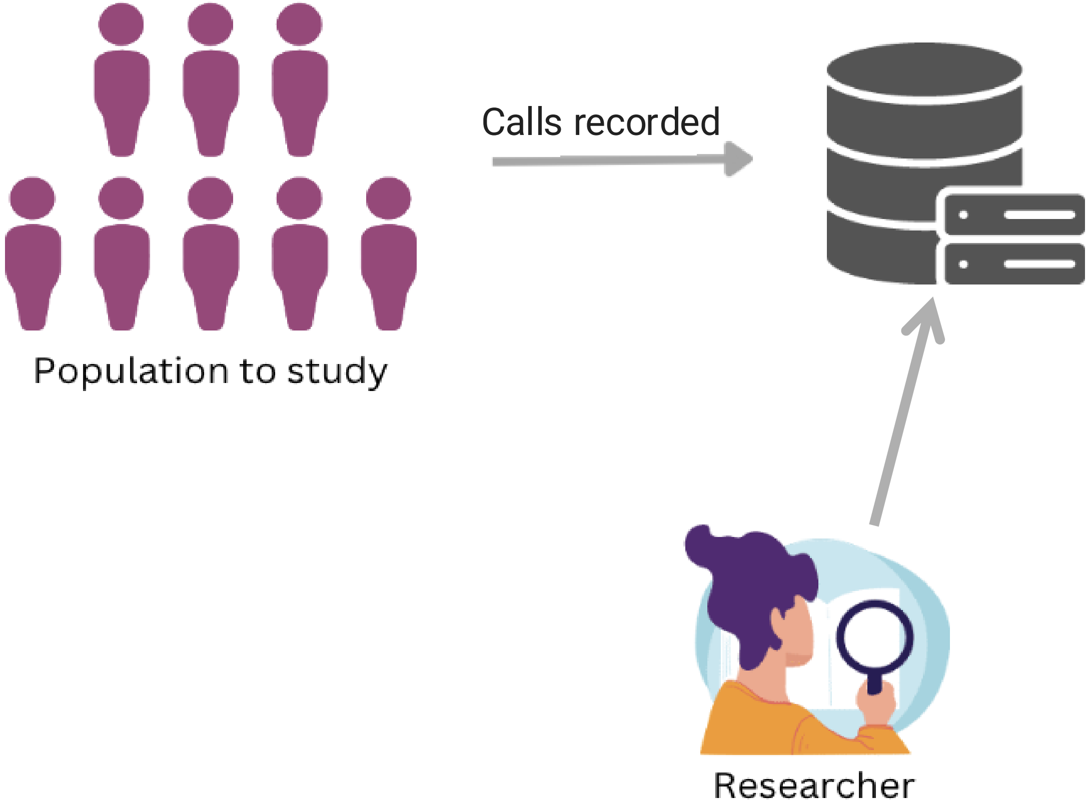


(Blumenstock et al., 2015)

How do we understand human behavior/societies?

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

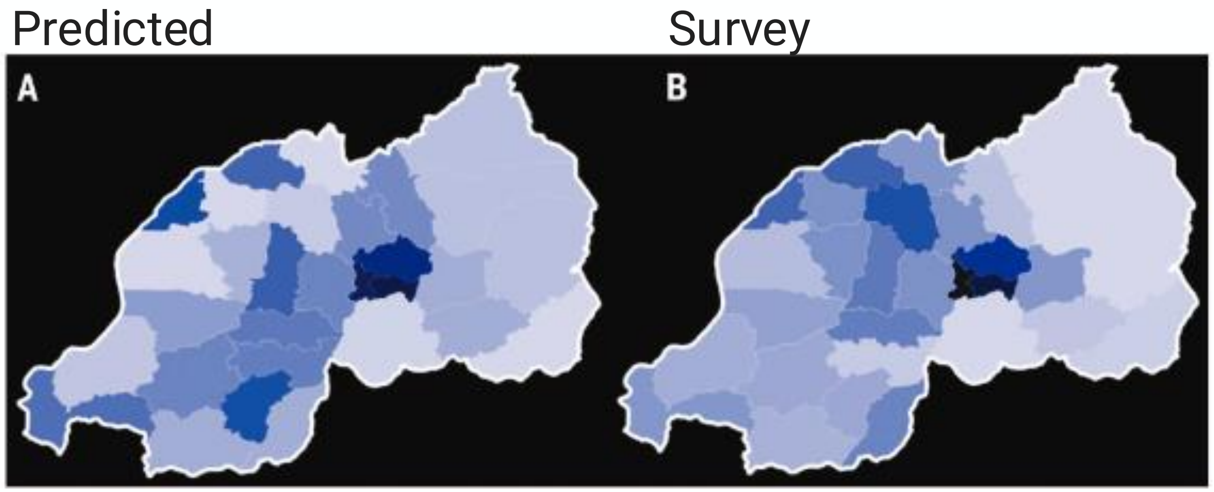
Using Digital Trace Data:



Calls recorded

Population to study

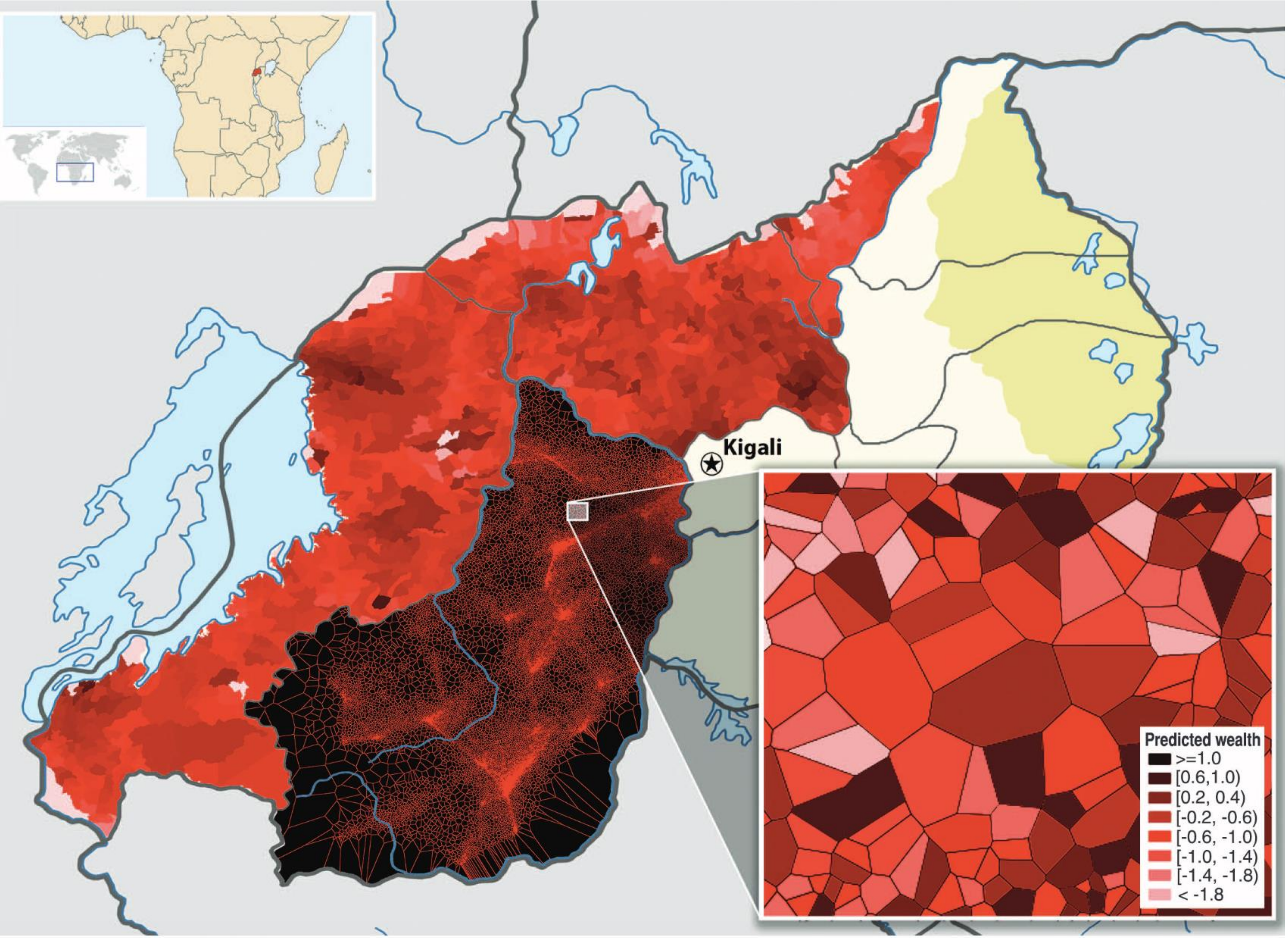
Researcher



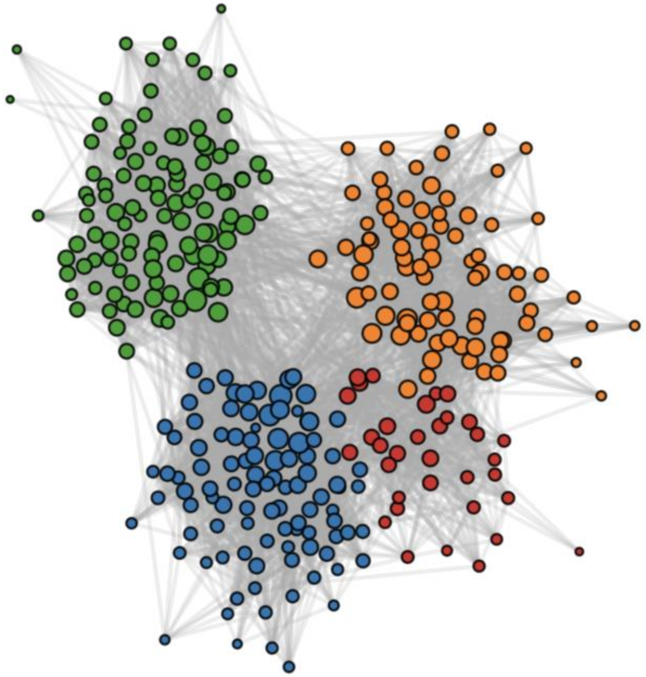
Predicted

Survey

(Blumenstock et al., 2015)


Kigali
Predicted wealth
>=1.0
[0.6,1.0)
[0.2, 0.4)
[-0.2, -0.6)
[-0.6, -1.0)
[-1.0, -1.4)
[-1.4, -1.8)
< -1.8

Great power



Unprecedented level of granularity: study small groups

Cheaper than surveys

Longitudinal data (dynamics!)

New research possible: e.g. social interactions (we have bad memory)

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

More examples (by Chris Bail): https://www.youtube.com/watch?v=uuSWQN7uYhk

Great responsibility

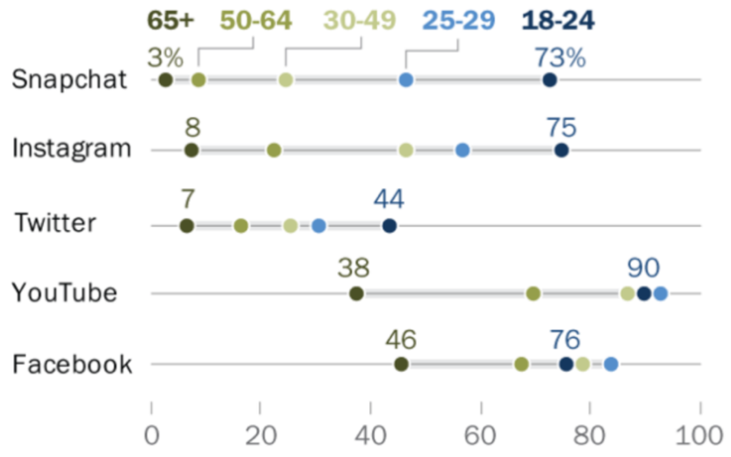
Can we keep the data safe?

Is the data representative for different groups?

Are we measuring what we want to measure? How can we validate our results?

Can our analyses harm people?

% of U.S. adults in each age group who say they ever use ...



Note: Respondents who did not give an answer are not shown.
Source: Survey conducted Jan. 8-Feb. 7, 2019.

Why is this important for you?

The age of surveillance capitalism

1. Data collection: Digital traces are thoroughly collected
2. Prediction products: Software uses that data to anticipate what a person likes and will do.
3. Behavioral markets: The software is used to target ads and products

Shoashana Zuboff,
Social psychologist at HBS

Understanding digital trace data will help you:

- Secure a job
- Be aware of the challenges associated with these data
- Prevent biases and reduce harm

Course set-up (digitaltracedata.github.io)

Wednesday:

- Lecture: 13:15 – 15:00
- Practical: 15:15 – 17:00

Group project:

- Oct 3th: Written report (30% of the grade)
- Oct 24th : Final presentation (30% of the grade) – deadline one day earlier

Feedback sessions: Fridays (Sep. 19th and 26th, Oct 10th and 17th)

Exam (Oct 31st) :

- 40% of the grade: mix of multiple choice and open questions

What would you expect/like to learn in this course?

app.wooclap.com/DTD25

Course overview

Week	Date	Content
1	**Sep 3 (We)**	**Lecture/lab: Introduction to digital trace data**
1	**Sep 3 (We)**	**Group project starts**
2	Sep 10 (We)	Lecture/lab: User-centric approaches to DTD
3	Sep 17 (We)	Lecture/lab: Platform-centric approaches to DTD
3	Sep 19 (Fr)	Group project feedback I
4	Sep 24 (We)	Lecture/lab: Errors in DTD collection
4	Sep 26 (Fr)	Group project feedback II
5	Oct 1 (We)	Lecture/lab: The role of AI in DTD
5	Oct 3 (Fr)	Deadline group project
6	Oct 8 (We)	Lecture/lab: Ethics
6	Oct 10 (Fr)	Group project feedback III
7	Oct 17 (Fr)	Group project feedback IV
8	Oct 24 (Fr)	Deadline: Group presentation
9	Oct 15th/29th (We)	Final recap and Q&A
9	Oct 31 (Fr)	Final exam
11	Nov 14 (Fr)	Exam inspection
12	Nov 28 (Fr)	Resit exam

What do we expect from you

Be on time to the lectures and labs.

Do the readings.

Work independently on the project and come prepared to the feedback moments.

Ask for support when needed.

Be kind (to yourself, to each other, and to us).

TODAY

Lecture

Explain what is Digital Trace Data (DTD)

Understand the main advantages and disadvantages of DTD

Distinguish user and platform-centric approaches to study DTD

Lab

Learn the difference between different types of data formats

Hands-on experience with unstructured data from Twitter

Explore a data analysis workflow

What is Digital Trace Data?

Digital Trace Data (DTD)

“Records of activity (trace data) undertaken through an online information system (thus, digital).” — Howison et al., 2011

Very diverse, but key characteristics:

- *Digital traces*: Interactions with technology (online information system).
- *Contains events* – i.e., interactions with the information system is recorded.
- *Ready-made* data: The data is a byproduct of people’s everyday actions, rather than produced for research. However, they are “designed” by someone for some other goal (often, profit-driven).

What are examples of DTD?

Social media posts

Web browsing history

GPS location data

Online purchase records

Email metadata

Mobile app usage logs

Mobile calls

Digital payment transactions

Cryptocurrency transactions

Fitness tracker data

Wi-Fi connection history

…

app.wooclap.com/DTD25

What do we need to study DTD?

1. **Understand the potential and the challenges associated with DTD**
    - The focus of this course

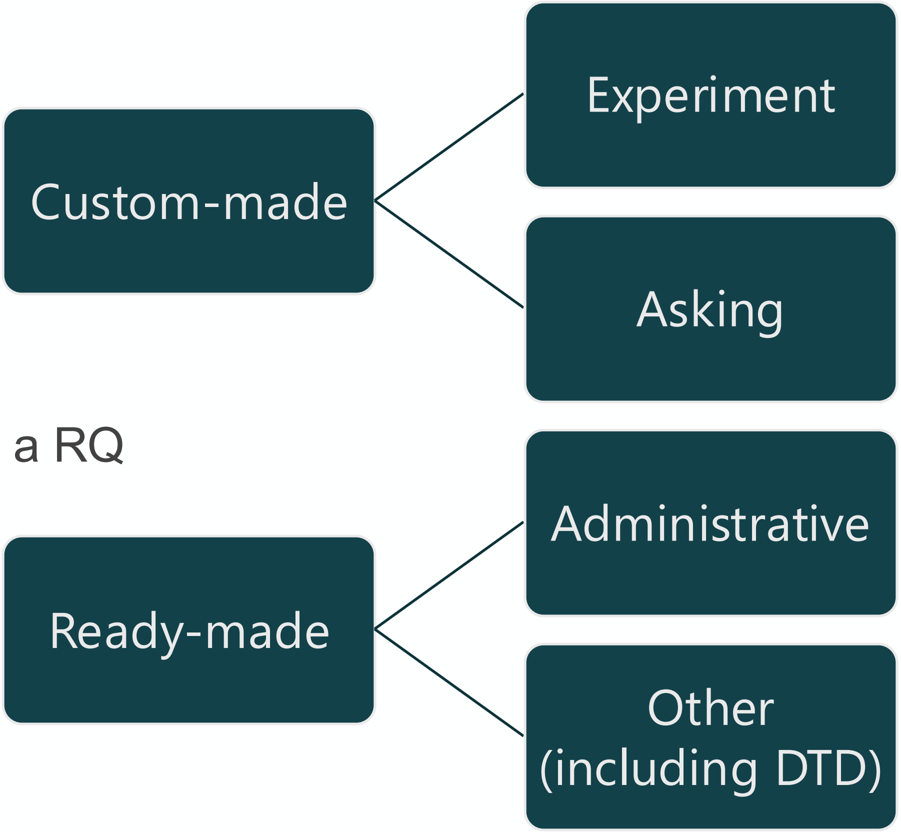
2. **Data Science skills**
    - Not the focus of this course
    - Strong programming skills help: the Applied Data Science minor

Characteristics of Digital Trace Data

Based on the Bit by Bit book (Salganik)

1. DTD is ready-made data

David (Michelangelo)
Data generated to answer a RQ



Answer RQ by changing the environment

Answer RQ by asking questions

Collected by governments to keep the country running.

Collected for a variety or reasons, mostly operational or profit-driven.

Bull’s Head (Pablo Picasso)
Data repurposed to answer a RQ

Readymade data is particularly affected by errors

Two main errors in DTD (more on this on week 4):

a. Measurement: i.e., does the data actually measure what you want to measure?

DTD allows us to study phenomena that are otherwise very difficult to examine:

- Concepts that are hard through other methods (e.g. social networks)
- Subpopulations that are difficult to track otherwise (e.g., conspiracies)

Example: suppose we are interested in social networks.

- Are phone calls a good way to measure social networks?
- How to test this? → Validation, does our data match (aggregated) estimates?

b. Representation: i.e., is it biased toward specific subpopulations?

(later)

Advise when repurposing data

Find as much as possible about how and for what purpose the data was created

Compare the characteristics of the data with the ideal data you would like to have

Exercise (in pairs)

Think about what type of DTD you could use to answer the following questions:

- How do social media influencers affect consumption of fake news?
- How did COVID-19 affect mobility patterns? (i.e., whether people travelled more/less to work, parks, do groceries, see friends)
- How do your friends and acquaintances affect job opportunities?
- Is exercise contagious? (if your friends exercise, do you start exercising?)

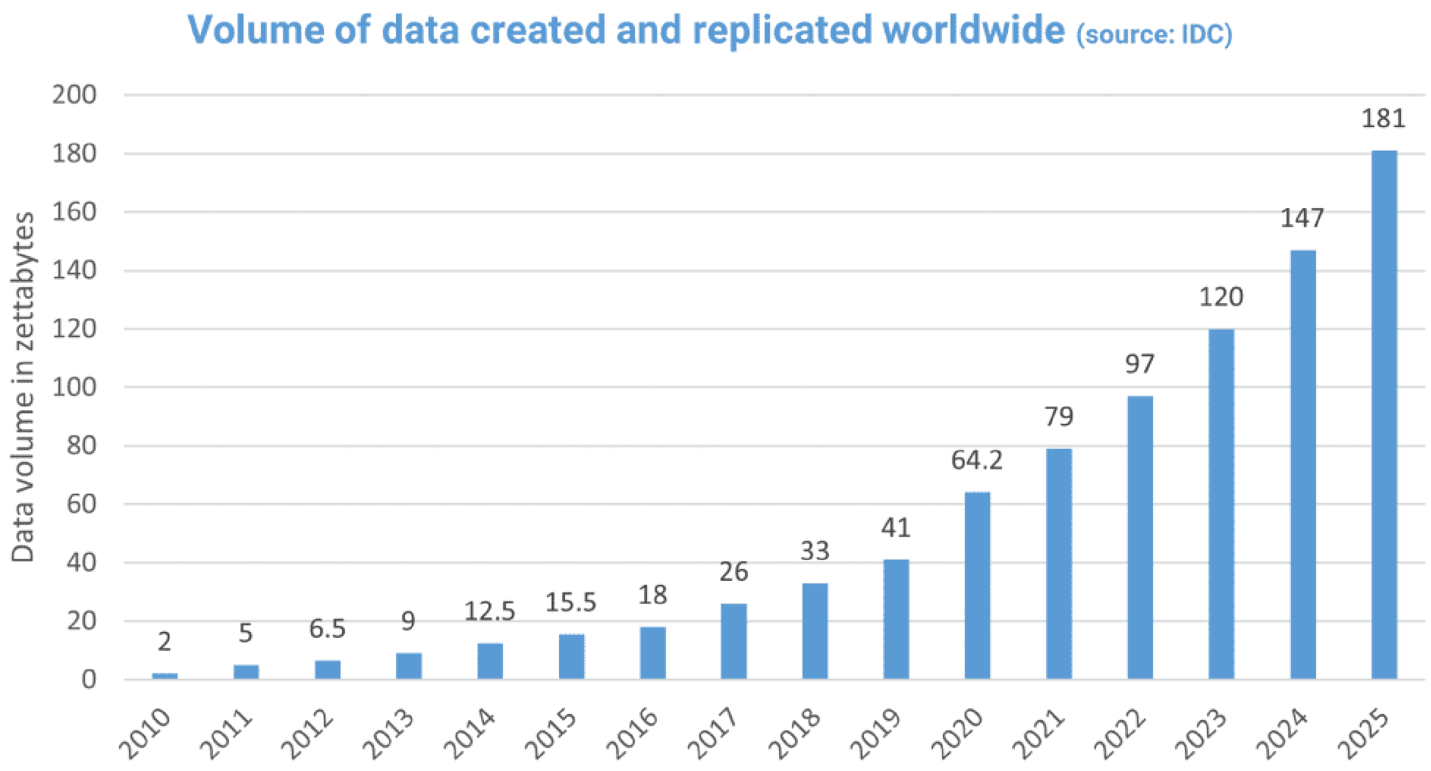
For one of the dataset, answer these questions:

- What was the original purpose of that data?
- How would you measure your dependent/independent variables (x and y variables)?
- Do you expect measurement error?

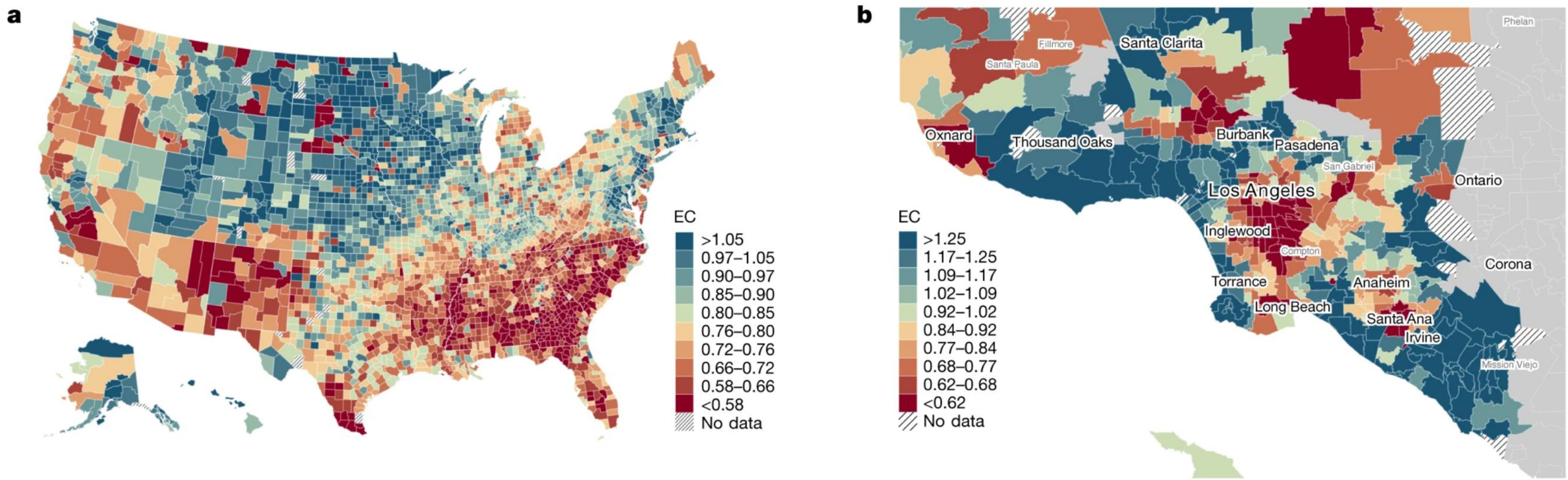
2. DTD is often large

2025: Equivalent of ~30 laptops full of data per person in the planet!

Each days, thousands of interactions are recorded

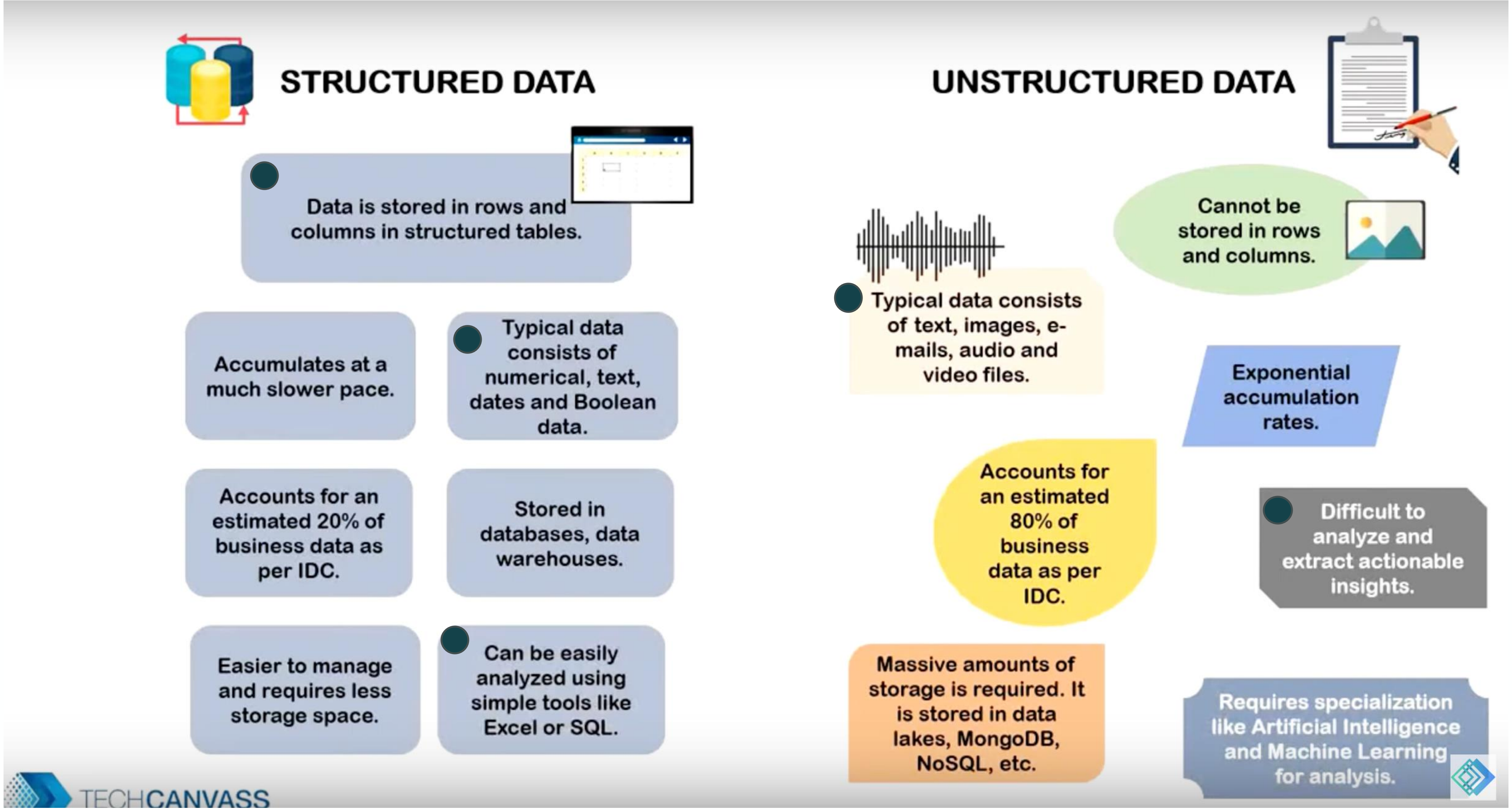


Big data allow us to study small subpopulations



Social capital I: measurement and associations with economic mobility
Chetty et al, 2022

Large/big data is often unstructured



STRUCTURED DATA

- Data is stored in rows and columns in structured tables.
- Accumulates at a much slower pace.
- Typical data consists of numerical, text, dates and Boolean data.
- Accounts for an estimated 20% of business data as per IDC.
- Stored in databases, data warehouses.
- Easier to manage and requires less storage space.
- Can be easily analyzed using simple tools like Excel or SQL.

UNSTRUCTURED DATA

- Cannot be stored in rows and columns.
- Typical data consists of text, images, e-mails, audio and video files.
- Exponential accumulation rates.
- Accounts for an estimated 80% of business data as per IDC.
- Difficult to analyze and extract actionable insights.
- Massive amounts of storage is required. It is stored in data lakes, MongoDB, NoSQL, etc.
- Requires specialization like Artificial Intelligence and Machine Learning for analysis.

Does having big data fix errors?

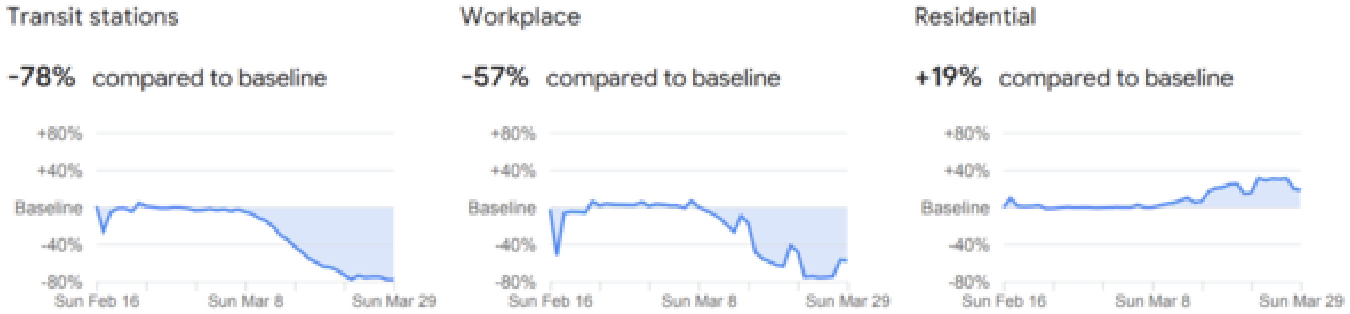
Imagine you have a database containing people that have been found to evade taxes, and that the prevalence of blue vs brown eyes in the population is 50:50.

- The data contains 10 people with blue eyes and 20 with brown eyes
- The data contains 10,000 people with blue eyes and 20,000 with brown eyes

Are people with brown eyes more likely to be involved in tax evasion?

3. DTD is always-on and non-reactive

Systems are constantly collecting data (they are *always-on*)

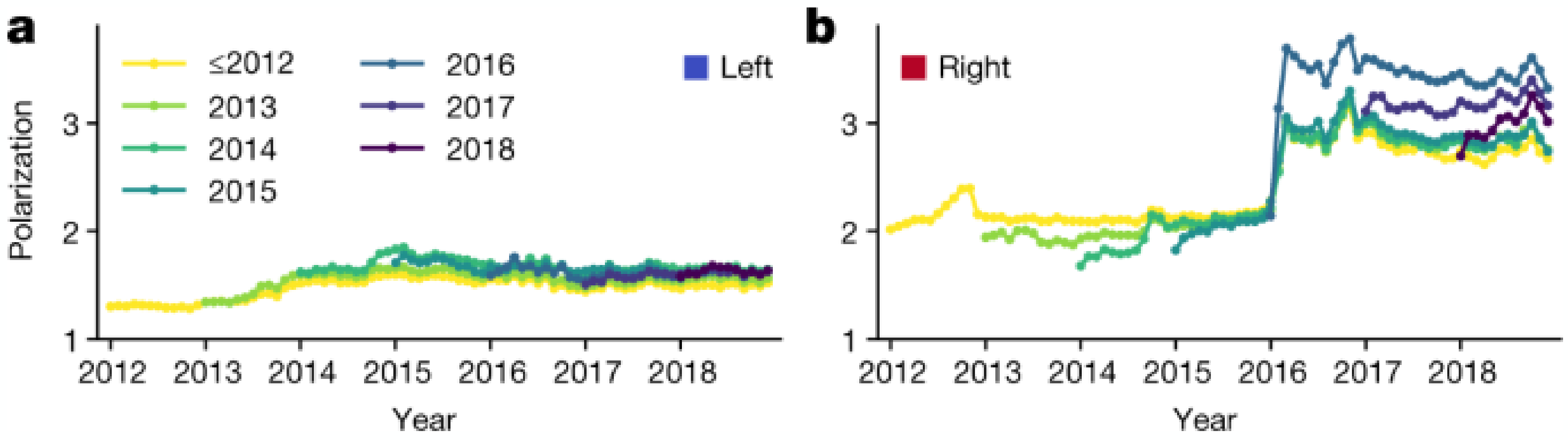


Google COVID-19 Community Mobility Reports

Advantages:

- Provide real-time estimates
- Allows to travel back in time (historical data)

Digital traces do not change when we study them (DTD is *non-reactive*).



Quantifying social organization and political polarization in online platforms
Waller and Anderson, 2021, Nature

4. Incomplete and non-representative

Incomplete: The *privacy paradox* (Golder & Macy, 2014): Data are at once too revealing in terms of privacy protection, yet also not revealing enough in terms *of providing the demographic background information* needed by social scientists, and in *measuring what the researchers want to measure*.

Two main errors in DTD (more on this on week 4):

a. Measurement: i.e., does it measure what you want it to measure?

b. Representation: i.e., is it biased towards specific subpopulations?

You are studying political polarization studying political tweets. You find large polarization. Why could this be the case?

How to deal with this issue?

→ Combining DTD with (aggregated) sources to understand/correct the biases

→ Studying within-person or within-group phenomena (e.g. do individuals polarize? vs. does society polarize?). This works if those phenomena are expected to be ~universal

5. DTD data is drifting

i.e., the measurement and representation can change over time.

Example: you find out that mobile calls are a great measurement of wealth (Blumenstock et al., 2015)

Measurement drift:

But as mobile internet becomes cheaper, people increasingly rely on Internet calls (e.g. via WhatsApp). Our measurement of wealth may have changed!

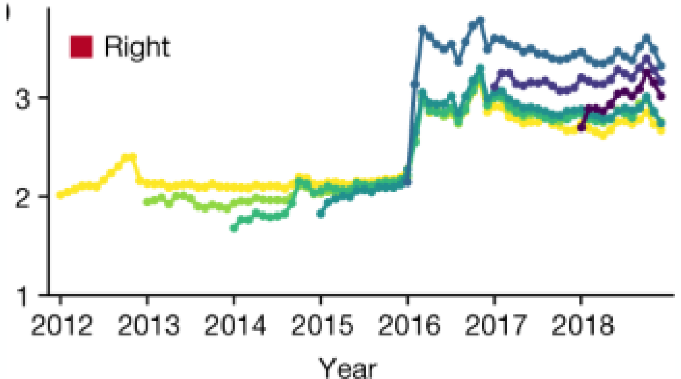
Representation drift:

And our data may not be representative anymore, as younger people may not call

Another example: polarization changed in Reddit because the new user inflow

How to deal with this issue?

→ Keep validating the results



Waller and Anderson, 2021, Nature

15 min break

6. Algorithmically confounded

We may be interested in understanding how being part of a tight-knit community (social closure) affects wellbeing.

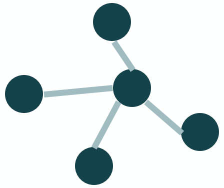
You cannot study this easily with surveys, so you decide to use Facebook data.

You decide to study social closure using the "clustering coefficient". This coefficient is the probability that two of your friends are also friends themselves.

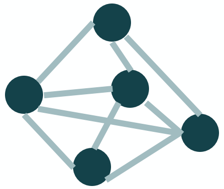
You find that the clustering coefficient is 14%, which is five times greater than expected (Ugander et al, 2011)

What may have happened?

What type of error is this? (measurement/representation)



Low clustering

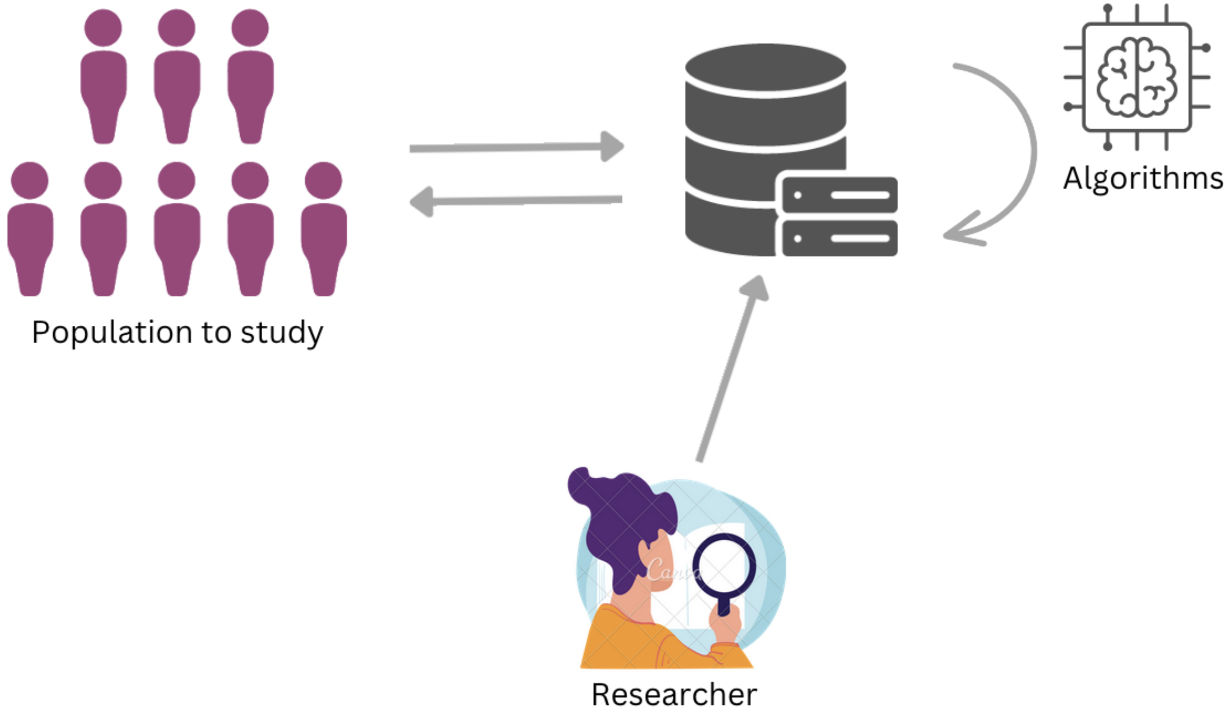


High clustering

Algorithms can cause measurement error

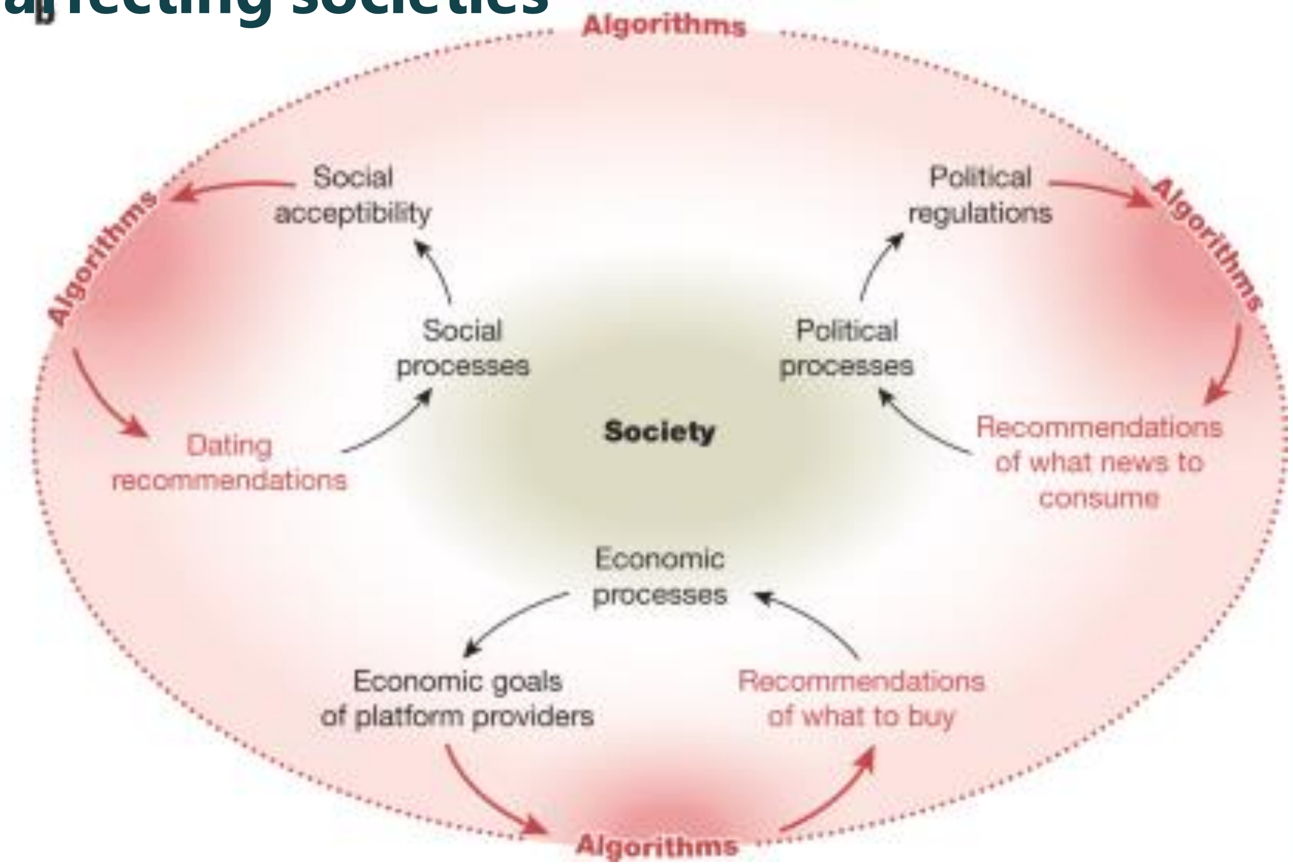
Facebook uses the “clustering coefficient” to recommend friends: e.g., if you have two friends, Sanne and Joep, that are not Facebook friends, Facebook will suggest Sanne and Joep to add each other as friends.

Your measurement of social closure (clustering coefficient) is measuring *both* social closure and the effect of the algorithm.



Algorithms create feedback effects, affecting societies

Sanne and Joep may become friends in real life.
Algorithms affect the world!



Measuring algorithmically infused societies
Wagner et al, 2021, Nature

Also:

- Who to hire – CV screening
- Who to promote – performance reviews
- Who to jail. – predictive policing
- Who to kill – “we kill people based on metadata” (US), “the Gospel” (Israel)

But what are the algorithms trying to optimize? Who profits?
e.g. Facebook knew its products damage teenagers’mental health, foment ethnic violence are allow misinformation to spread. But accepted those consequences as part of its business model (Frances Haugen scandal, 2021).

7. Dirty

DTD comes in a wide variety of formats

It contains many artifacts:

- Algorithmic effects
- Bots
- Organized groups (e.g. hackers)

8. Sensitive

What do your digital traces reflects about you?

- Age, gender, income, political and sexual preferences, beliefs, tastes, addictions, traumas, location

Two risks:

- **Individual privacy:** Protecting personal data
  - DTD reveal intimate details about a person's preferences/behavior/location
  - Main risk: Personal identity leaked
- **Group privacy:** Protecting the interests of a collective
  - DTD reveal sensitive information about groups
  - Example: mobile call data may show that refugees tend to call each other and foreign countries
  - Main risk: Unfair treatment of individuals for (allegedly) being part of a group – e.g., identification as a refugee

(more on this on weeks 5—6)

9. Inaccessible

DTD data is crucial to understand 21st century societies, especially the role of algorithms.

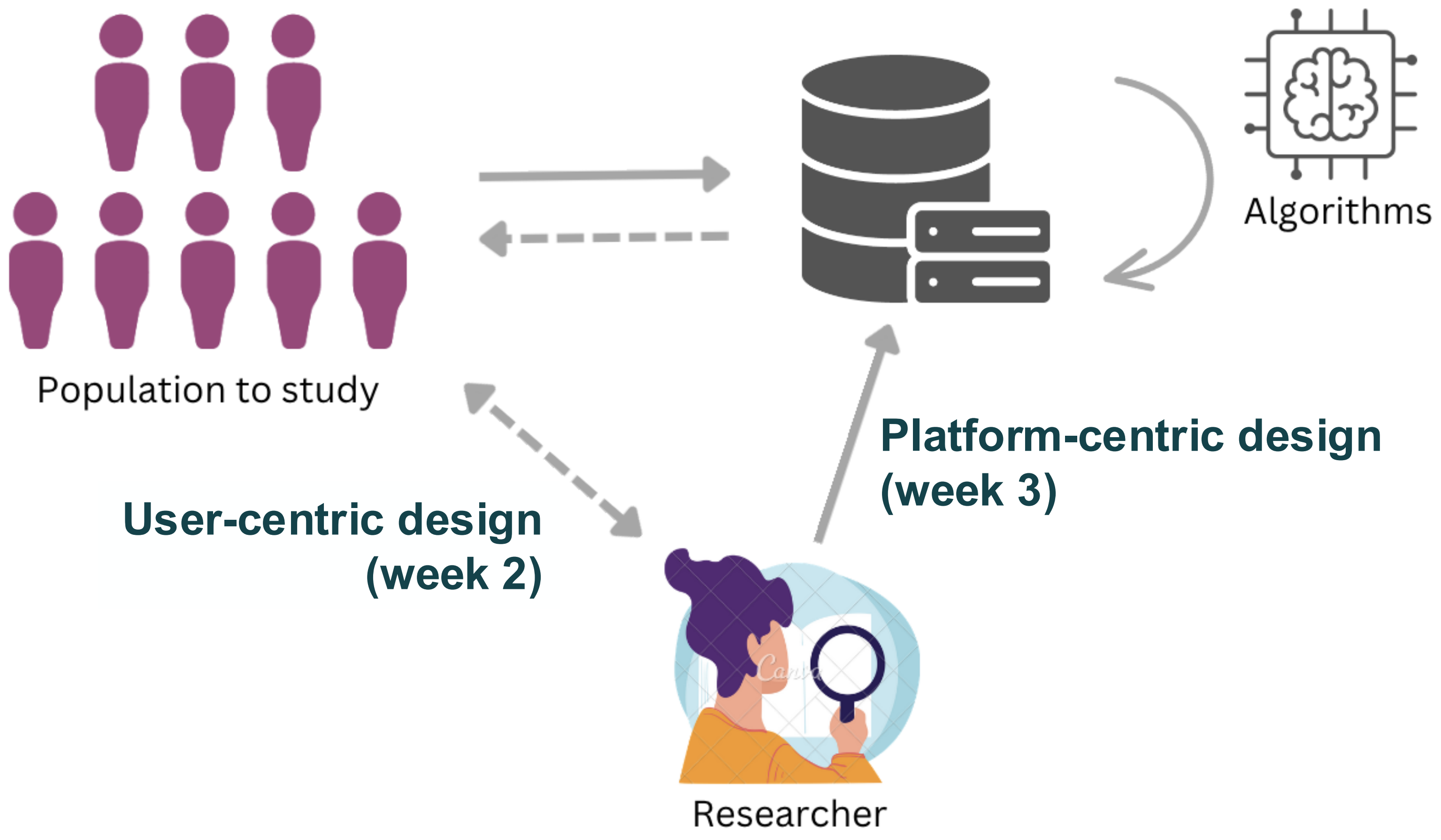
DTD is held (mostly) by companies and governments.

Sharing that data is difficult: legal, ethical and business barriers

How to access the data:

- **User-centric approaches:** Rely on the users to collect the data
- **Platform-centric approaches:** Use the information that platforms provide

Digital Trace Data Collection



Population to study

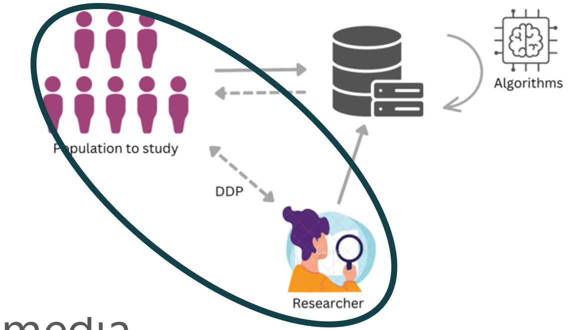
Algorithms

User-centric design (week 2)

Platform-centric design (week 3)

Researcher

User-centric design

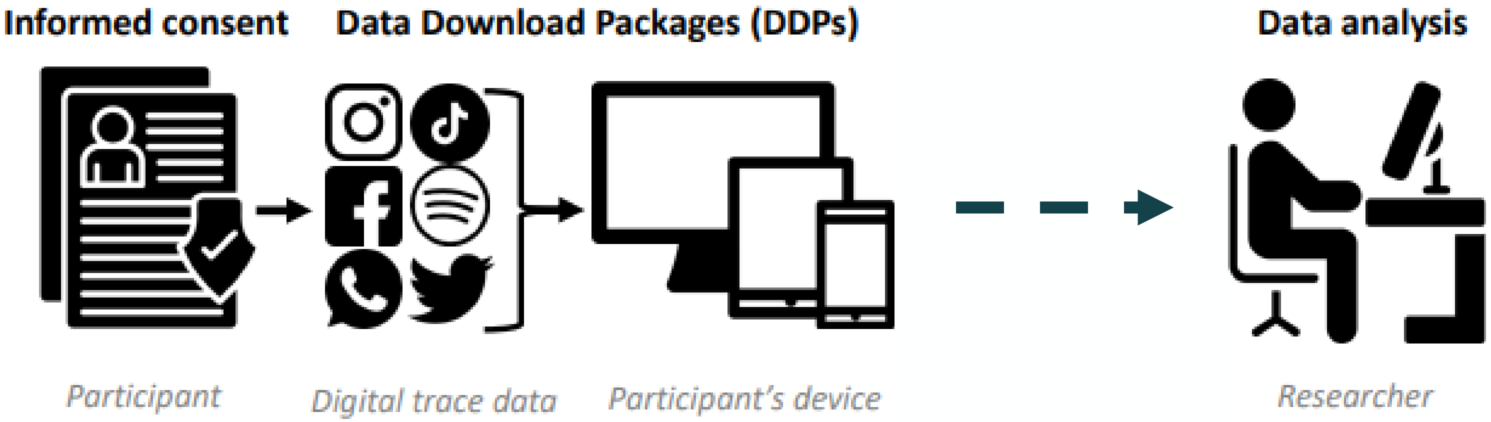


Tracking with wearables, apps and sensors

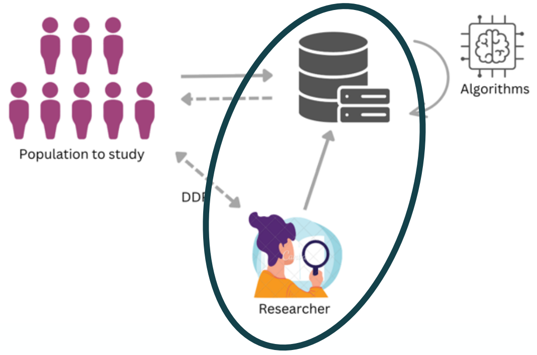
- E.g. data from browser plugins, collecting data on how people interact with social media

Data donation:

- Ask people to donate their DTD. It takes advantage of the *right of access* by the data subject and *right to data portability* (General Data Protection Regulation (GDPR)).
- Researchers receive the data the platform has collected on people in (semi)structured, commonly used, and machine-readable format (“Data Download Package“; DDP).



Boeschoten et al. (2023)



Platform-centric design

Collaboration with the organization holding the data

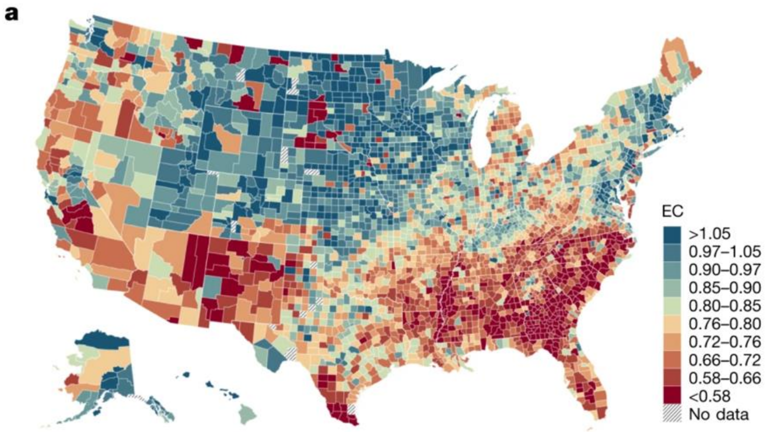
- The organization provides the data

APIs (Application Programming Interfaces)

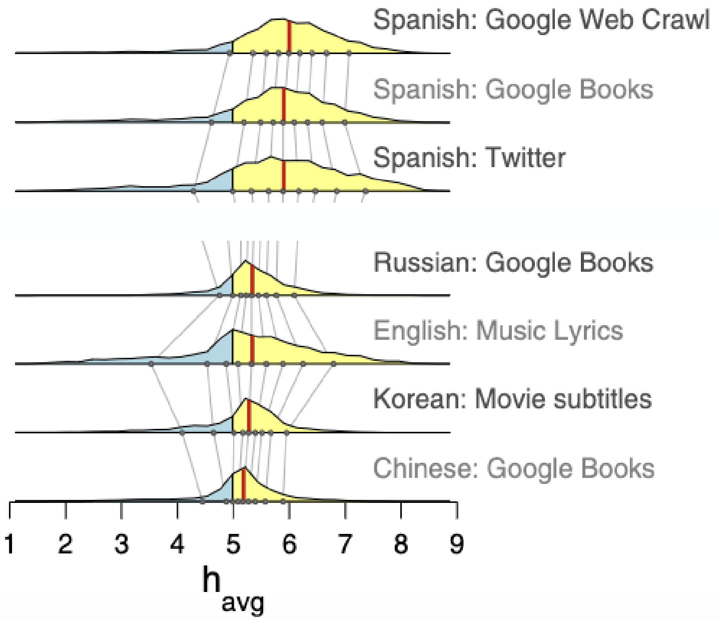
- The organization provide direct access to (parts) of the data.

Web-Scraping

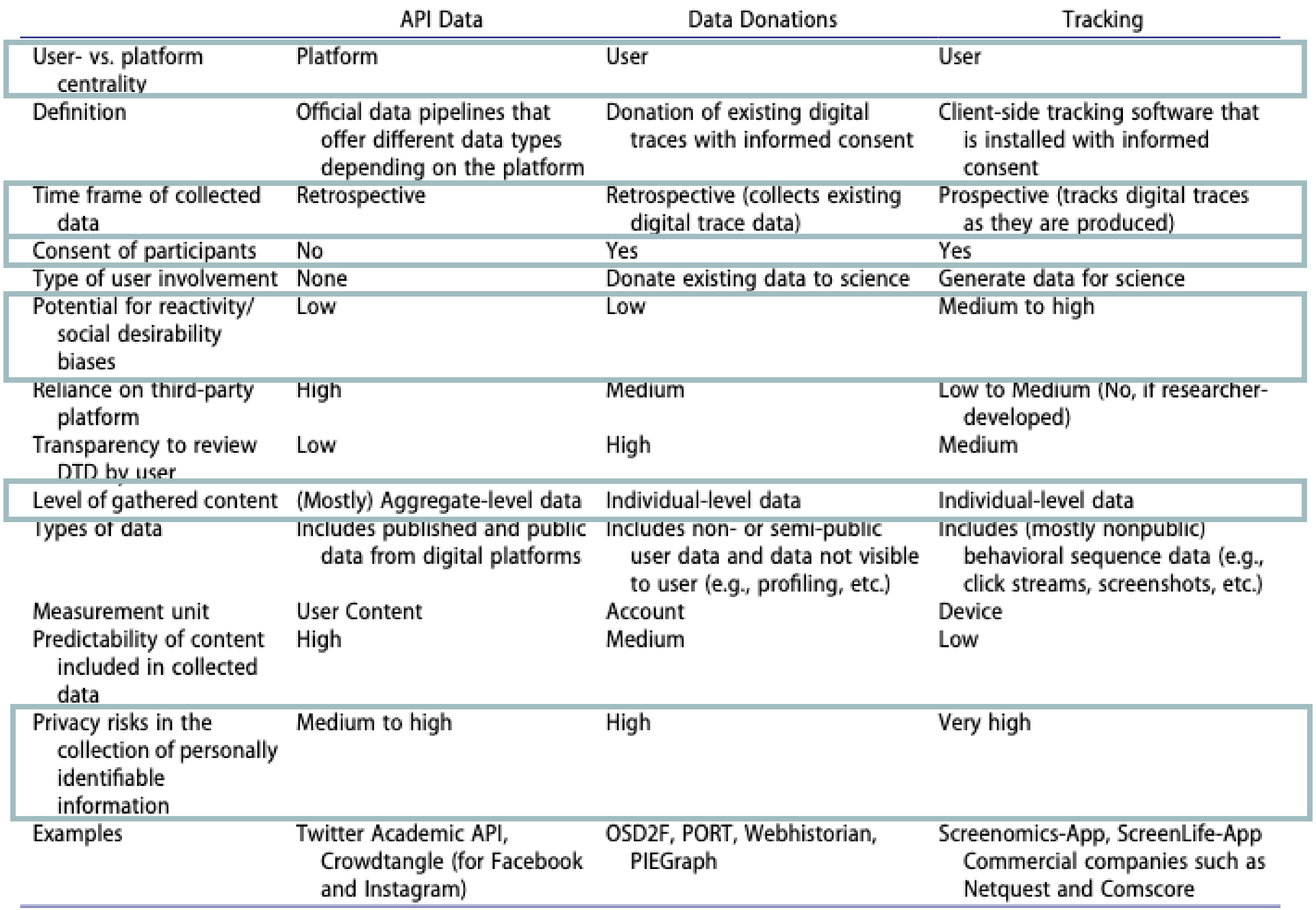
- Downloads the information shown in website



Chetty et al, 2022, Nature



Dodds et al., 2015, PNAS



	API Data	Data Donations	Tracking
User- vs. platform centrality	Platform	User	User
Definition	Official data pipelines that offer different data types depending on the platform	Donation of existing digital traces with informed consent	Client-side tracking software that is installed with informed consent
Time frame of collected data	Retrospective	Retrospective (collects existing digital trace data)	Prospective (tracks digital traces as they are produced)
Consent of participants	No	Yes	Yes
Type of user involvement	None	Donate existing data to science	Generate data for science
Potential for reactivity/ social desirability biases	Low	Low	Medium to high
Reliance on third-party platform	High	Medium	Low to Medium (No, if researcher-developed)
Transparency to review DTD by user	Low	High	Medium
Level of gathered content	(Mostly) Aggregate-level data	Individual-level data	Individual-level data
Types of data	Includes published and public data from digital platforms	Includes non- or semi-public user data and data not visible to user (e.g., profiling, etc.)	Includes (mostly nonpublic) behavioral sequence data (e.g., click streams, screenshots, etc.)
Measurement unit	User Content	Account	Device
Predictability of content included in collected data	High	Medium	Low
Privacy risks in the collection of personally identifiable information	Medium to high	High	Very high
Examples	Twitter Academic API, Crowdtangle (for Facebook and Instagram)	OSD2F, PORT, Webhistorian, PIEGraph	Screenomics-App, ScreenLife-App Commercial companies such as Netquest and Comscore

Ohme et al, 2024, Communication Methods and Measures

Exercise (in pairs)

You want to study how people consume news (i.e., given a series of news articles, which one do they choose to read).

How would you study this using:

- A user-centric approach?
- A platform-centric approach?

Think also about:

- Can the analysis be (easily) replicated by other researchers?
- Is your approach easy?
- Is your approach compliant with regulations?

Summary

Main take away message

Advantages of DTD

*

Disadvantages of DTD

*

app.wooclap.com/DTD25

TODAY

Lecture

Explain what is Digital Trace Data (DTD)

Understand the main advantages and disadvantages of DTD

Distinguish user and platform-centric approaches to study DTD

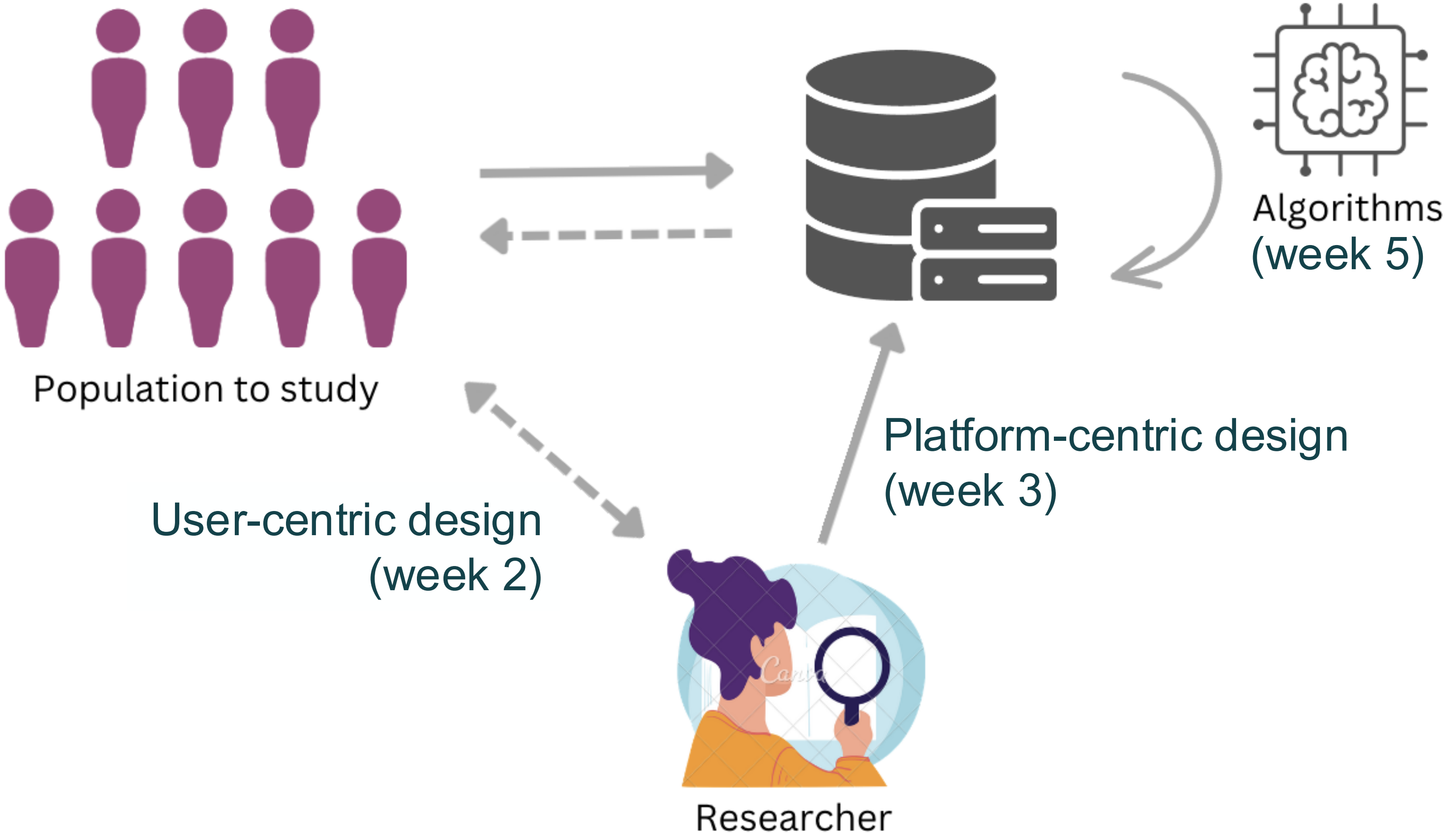
Lab

Learn the difference between different types of data formats

Hands-on experience with unstructured data from Twitter

Explore a data analysis workflow

Summary of the course



Week 4: Errors in DTD
Week 6: Ethics and Legislation
Week 7: Beyond DTD and Q&A

See you after lunch!

RUPPERT - 011