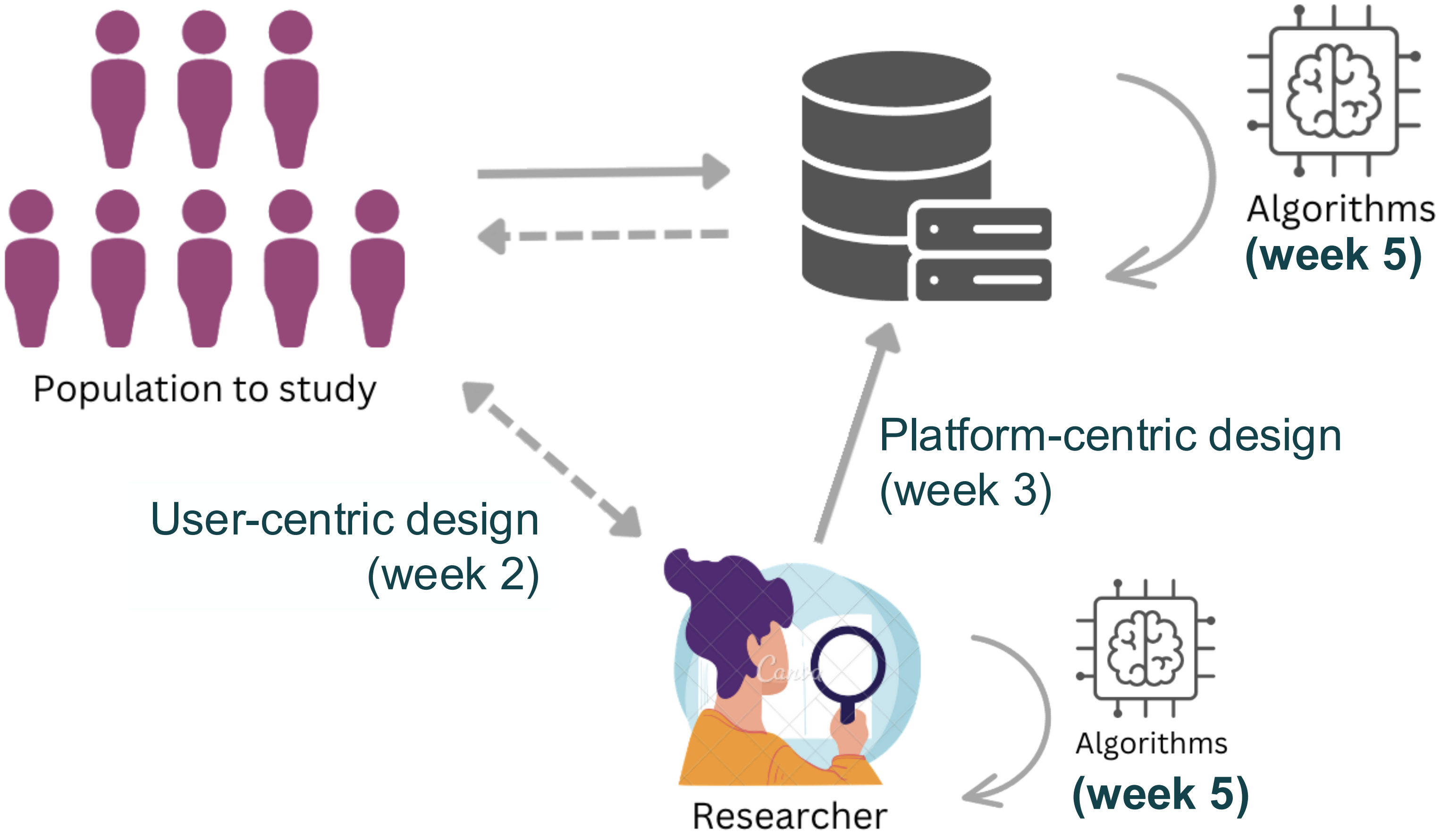
Introduction to Digital Trace Data: Quality, ethics, and analysis

Lecture 5: The role of AI in DTD

Javier Garcia-Bernardo
Assistant Professor
Department of Methodology and Statistics

Where are we?



Week 4: Errors in DTD
Week 6: Ethics and Legislation
Week 9: Beyond DTD and Q&A

Today's material

1. What are Algorithms/AI/Machine Learning?
2. Using ML to study societies
3. The impact of ML on Digital Trace Data
4. The impact of ML on societies
5. Dealing with bias in ML

TODAY

Lecture

1. Explain machine learning in your own words

2. Explain *why* machine learning models may be biased (sources of bias).

3. Understand the effects of ML on DTD and in society.

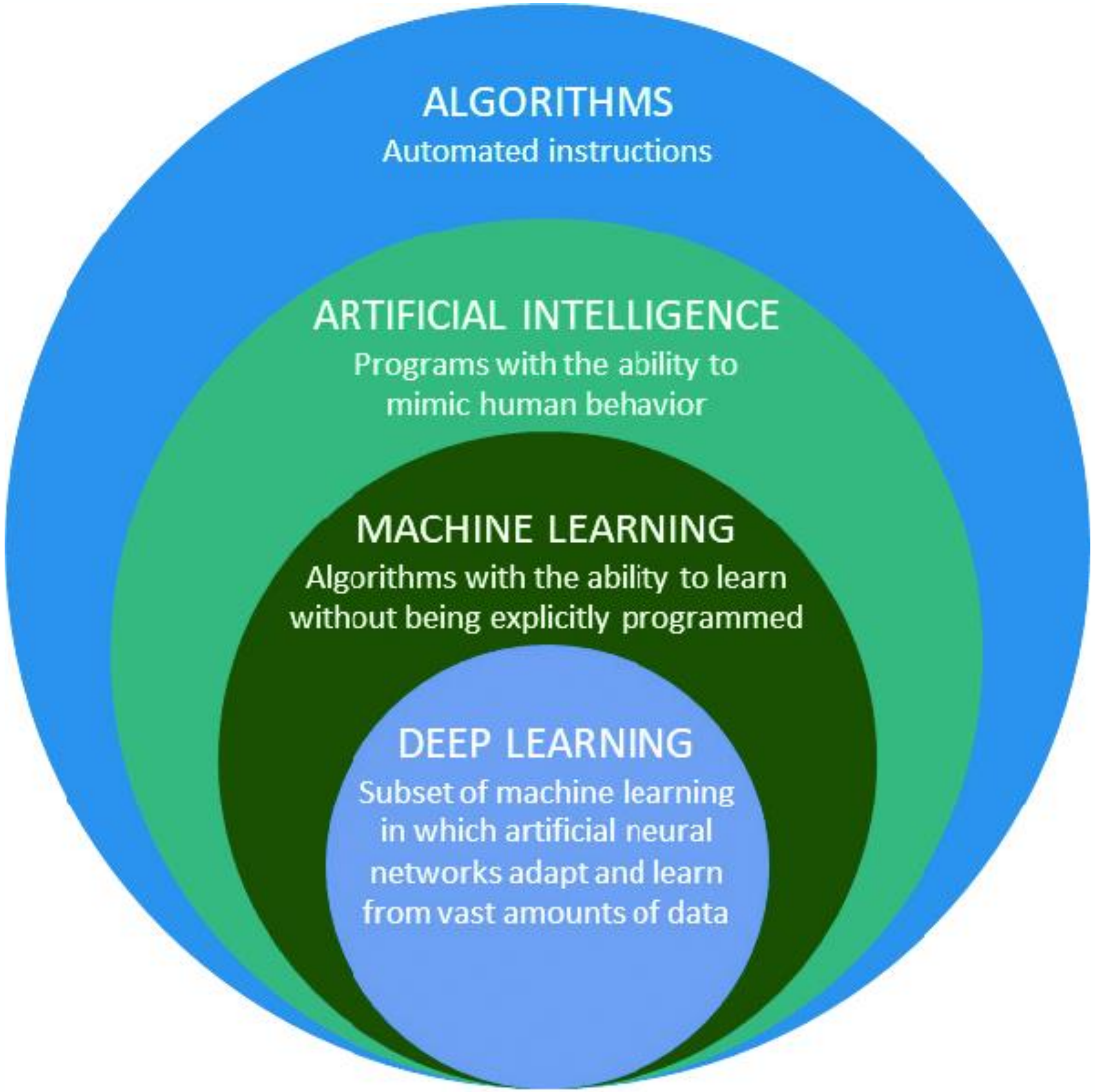
4. Assess bias in ML models

Lab

Apply a ML model to text data

Audit a ML model

1. What is machine learning?



(Author: Johannes Vrana, Vrana GmbH, Licenses: CC BY-ND 4.0)

Machine Learning

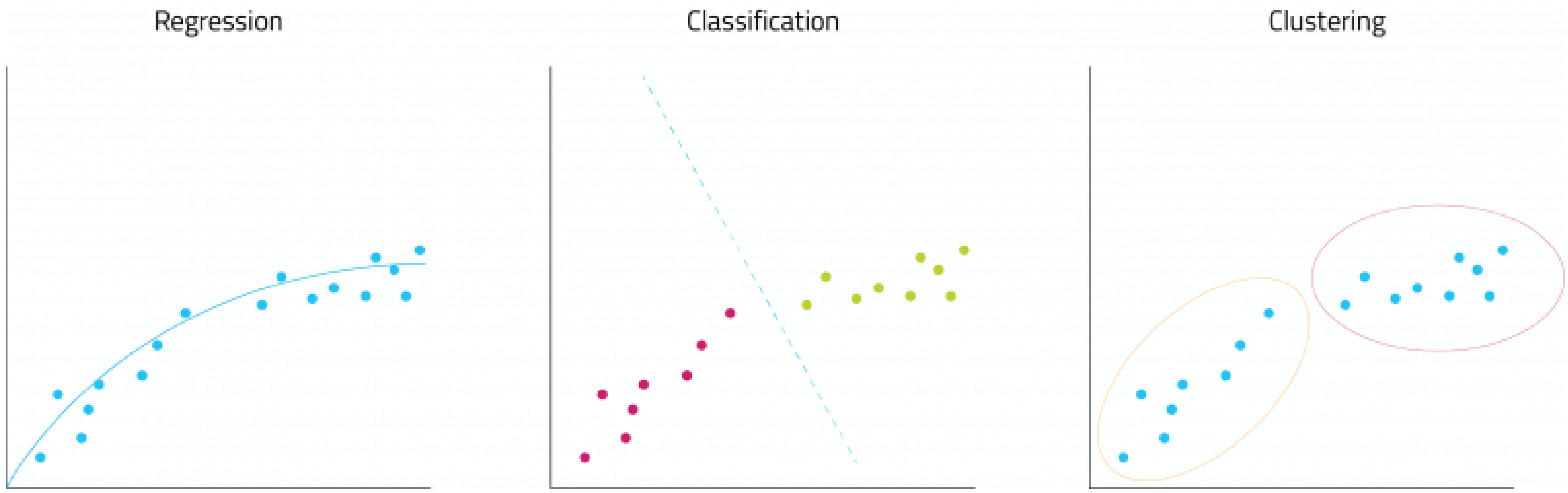
"A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at task *T*, as measured by *P*, improves with experience *E*." (Samuel/Mitchell, 1959/1997)

- Experience: Data (e.g. comments from TikTok)
- Task: Goal of the model (e.g. predict hate speech)
- Performance measure: Accuracy, R^2, etc

Is a linear regression a machine learning model?

Supervised vs unsupervised ML

Supervised ML: We have inputs (features, independent variables) and an output (target, dependent variable)
Unsupervised ML: We have inputs and (mostly) try to find groups



https://medium.com/@RayKleptzo/basic-machine-learning-concepts-regression-classification-and-clustering-66a216dac9cd

Examples

(A) You work at an advertisement company and want to customers into segments based on their purchasing behaviours, age, maximal educational degree attained, etc.

→ Unsupervised

(B) You work at a bank and want to develop a model that helps them predicting which loan applicants will default (not be able to pay the loan) based on their financial transactions.

→ Supervised (classification)

(C) You use news and social media analytics to predict changes in the stock market. You have access to historical stock and social media data and you want to predict the stock prices in the near future.

→ Supervised (regression)

Using ML to understand societies

- **Description**: The goal is to describe patterns or groupings in historical data.

- **Prediction**: The goal here is to predict outcomes. For example, predict missing data, predict risk of developing diseases, or label data.

- **Explanation**: The goal here is to understand *causal* relationships. For example, which gene is responsible of which disease.

In data analysis: Descriptive (unsupervised ML)

RESEARCH ARTICLE

Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter

Philipp Wicke[1]*, **Marianna M. Bolognesi**[2]

Topic #I: Communications and Reporting

Topic #II: Community and Social Compassion

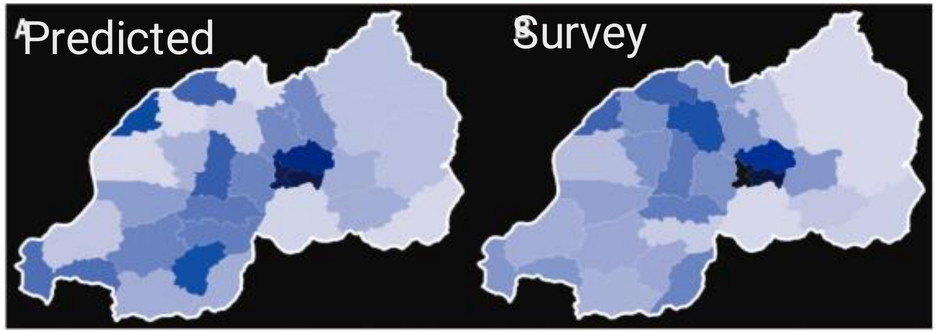
Topic #III: Politics

Topic #IV: Reacting to the epidemic

- **Question**: What is the framing of the COVID pandemic? Framing of WAR (fight, combat, battle), STORM (wave, storm, cloud), MONSTER (evil, horror, killer) or TSUNAMI (wave, tragedy, catastrophe).
- **Data**: Twitter around #Covid-19 (80 hashtags)

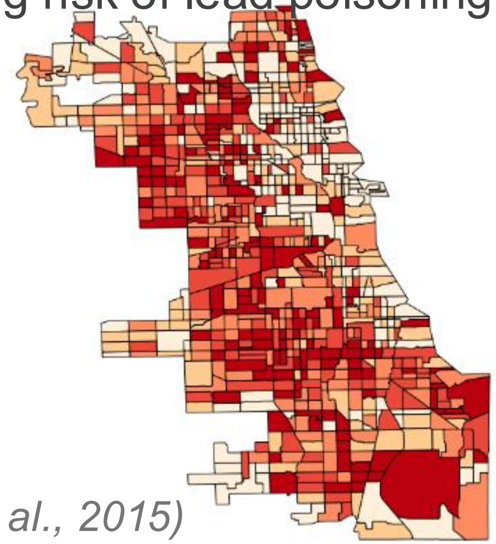
In data analysis: Prediction (supervised ML)

Predicting wealth/SES:

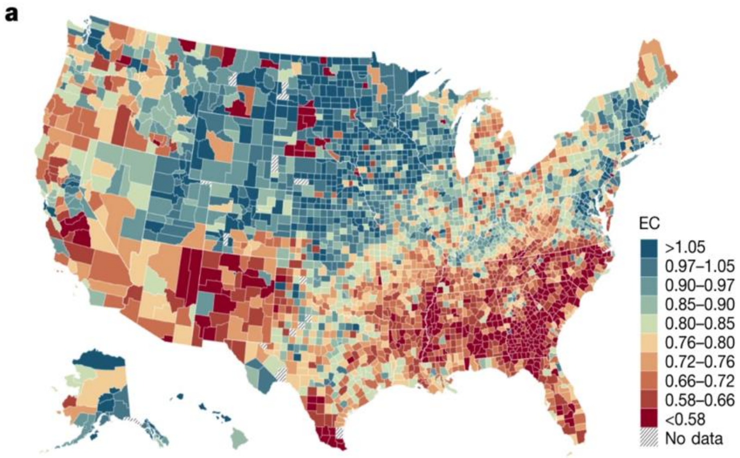


(Blumenstock et al., 2015)

Predicting risk of lead poisoning:

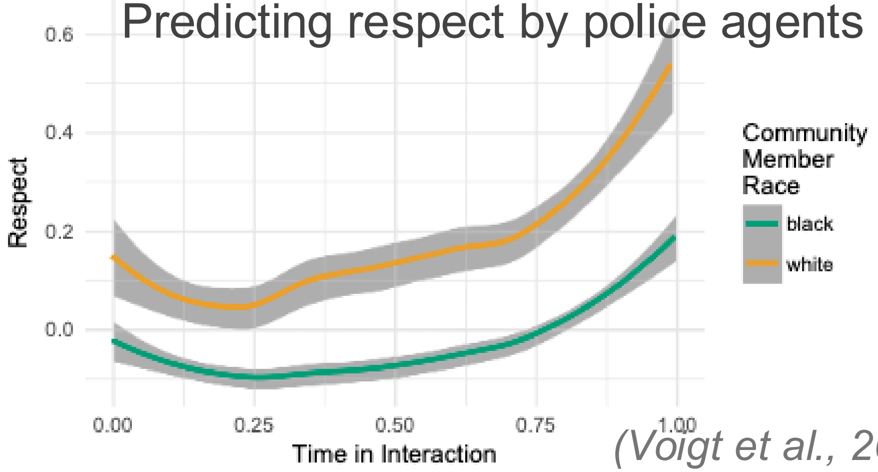


(Potash et al., 2015)



(Chetty et al, 2022)

Predicting respect by police agents



(Voigt et al., 2017)

2. Why are ML models often biased/unfair*?

*The model performance is different for different subgroups

Errors in algorithms

We need to consider the errors of every algorithm:

- How often they fail?
- For whom do they fail? (***bias/fairness***)

No model will be perfect, but we need to understand when and for whom they fail.

Fitting and using ML models (supervised ML)

1) Training data

X: Flights to Russia	Y: Criminal
3	✗
12	✓
1	✗
80	✓

2) Model

If X>10: criminal
If X<= 10: not criminal

3) New data

X: Flights to Russia
20
6
3

How often do they fail? The confusion matrix

	Predicted criminal	Predicted not criminal
Criminal	True positive	False negative
Not criminal	False positive	True negative

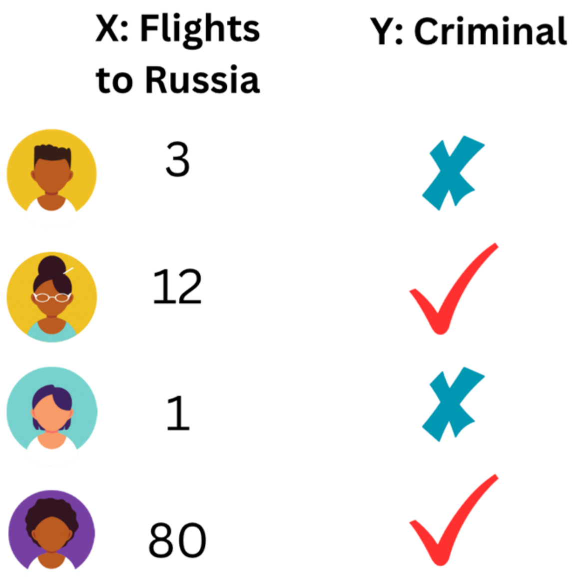
For whom do they fail?

Group A	Predicted criminal	Predicted not criminal
Criminal	10	10
Not criminal	1	100

Group B	Predicted criminal	Predicted not criminal
Criminal	10	1
Not criminal	10	100

Five main sources of bias in a ML model

Sample bias: the training data does not generalize to the prediction data.



Your sample may be only Dutch people, who don't go to Russia often unless they are criminals or have business. This could happen because:

- you are using data tracking only Dutch tourists.
- you are using a script that cannot handle Russian names, and they are dropped in the process
- you collected data during COVID times, or 10 years ago (drift)

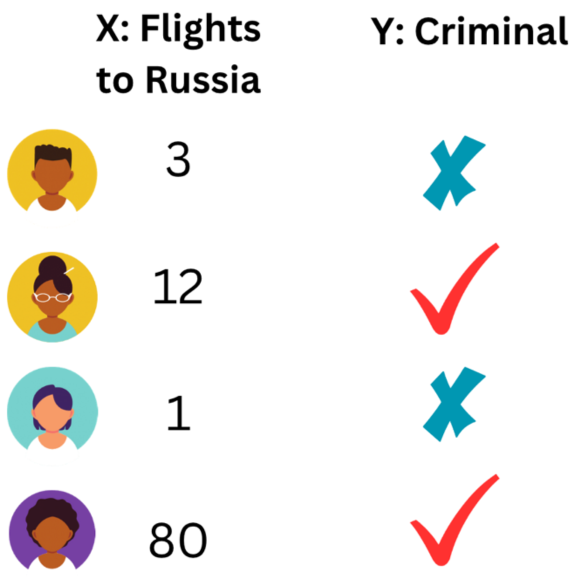
Five main sources of bias in a ML model

Outcome (label) bias: the label (criminal/not criminal) has different meaning from some subpopulation. For example if police mostly looks for criminals among foreigners, the label “non-criminal” for Dutch people mean something else (lack of policing, not lack of crime). *We can always expect label bias if the humans creating the labels are themselves biased*.

Features bias: the features (flights to Russia) have different meaning from some subpopulation. E.g., for Russians, flights may mean family visits.

ML pipeline bias: e.g., you are using a model or performance metric that focuses mostly on the majority class; there are errors in your code; you used a wrong model.

Application bias: the model and data is right, but it is applied in a bias way. E.g. a manager always trust the system for classifying foreigners as criminals but ignore it for Dutch.



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

What type of bias?

Dissecting racial bias in an algorithm used to manage the health of populations

ZIAD OBERMEYER, BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAINATHAN Authors Info & Affiliations

SCIENCE • 25 Oct 2019 • Vol 366, Issue 6464 • pp. 447-453 • DOI: 10.1126/science.aax2342

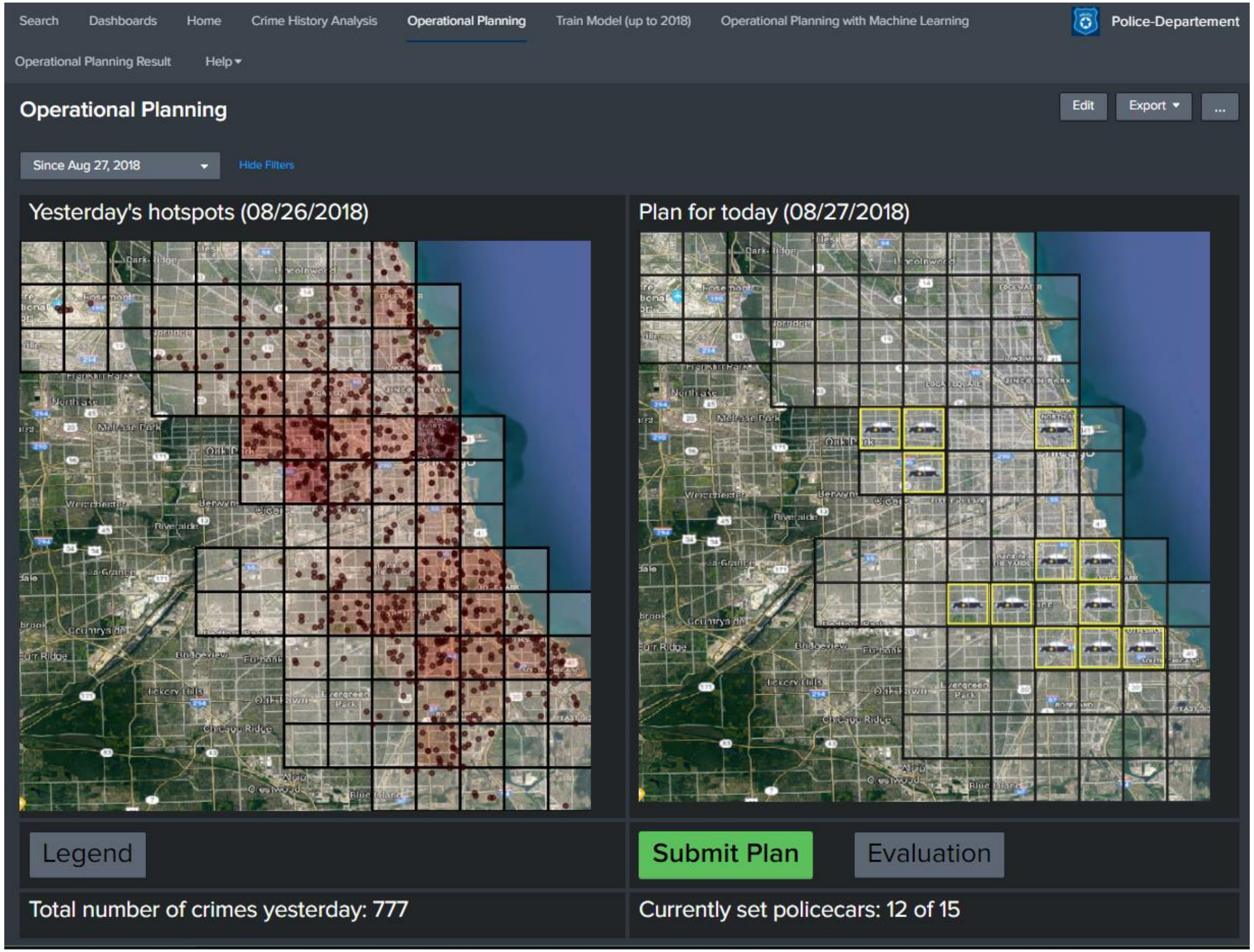
Racial bias in health algorithms

The U.S. health care system uses commercial algorithms to guide health decisions. Obermeyer *et al.* find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients (see the Perspective by Benjamin). The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half. Bias occurs because the algorithm uses health costs as a proxy for health needs. Less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients. Reformulating the algorithm so that it no longer uses costs as a proxy for needs eliminates the racial bias in predicting who needs extra care.

What type of bias?

Predictive policing

“Quickly connect the dots, identify hidden patterns and discover trends” (Splunk)



What type of bias?

Judge Rules $400 Million Algorithmic System Illegally Denied Thousands of People's Medicaid Benefits

Thousands of children and adults were automatically terminated from Medicaid and disability benefits programs by a computer system that was supposed to make applying for and receiving health coverage easier.

"The system often *doesn't load the appropriate data*, assigns beneficiaries to the wrong households, and makes incorrect eligibility determinations"

What type of bias?

https://gizmodo.com/judge-rules-400-million-algorithmic-system-illegally-denied-thousands-of-peoples-medicaid-benefits-2000492529

Humans are biased too

Case Study: Blind Auditions & Women in Orchestras (U.S.)
Goldin & Rouse, American Economic Review (2000); NBER Working Paper (1997)

In the top five U.S. orchestras, women were <5% of players in 1970 versus ~25% by the mid-1990s (now it is around 50%).

Why? Most major U.S. orchestras adopted 'blind' auditions (with a screen) in the 1970s–1980s;

Blind auditions increased the probability a woman advanced from prelim rounds by ~50%, and raised the chance she won the final round by severalfold.

"The biggest threat from artificial intelligence systems is not that they will become smarter than humans, but that they will hard-code sexism, racism, and other forms of discrimination into the digital infrastructure of our societies." — *Kate Crawford*

Exercise (in pairs)

You work at a Dutch bank, and you want to develop a model to help predict which loan applicants will default based on their financial transactions. You have labeled data from Utrecht customers from the last three years. You want to predict loan default for all new applicants.

In which parts of the process might bias be introduced? Think about

- The representativeness of the training data.
- Consider the quality of the outcome (default/non-default) in the training data for different subpopulations.
- Consider the quality of the features (financial transactions) in the training data for different subpopulations.
- Consider the machine learning pipeline.

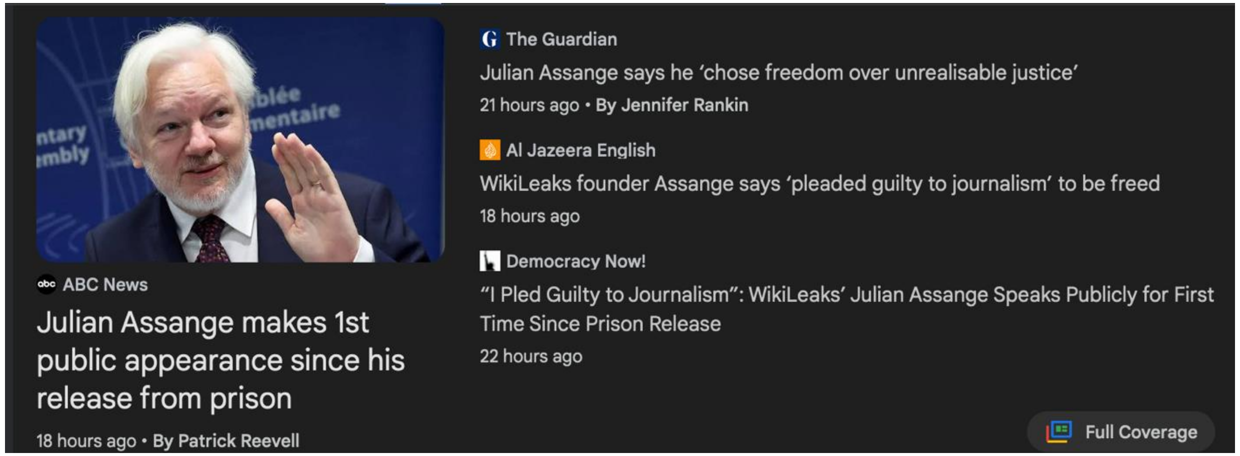
3a. How does machine learning affect DTD?

From the point of view of data quality and data analysis

ML is used in the collection, processing and interpretation of DTC

Main uses of ML in Digital Trade Data:

Recommendation algorithms

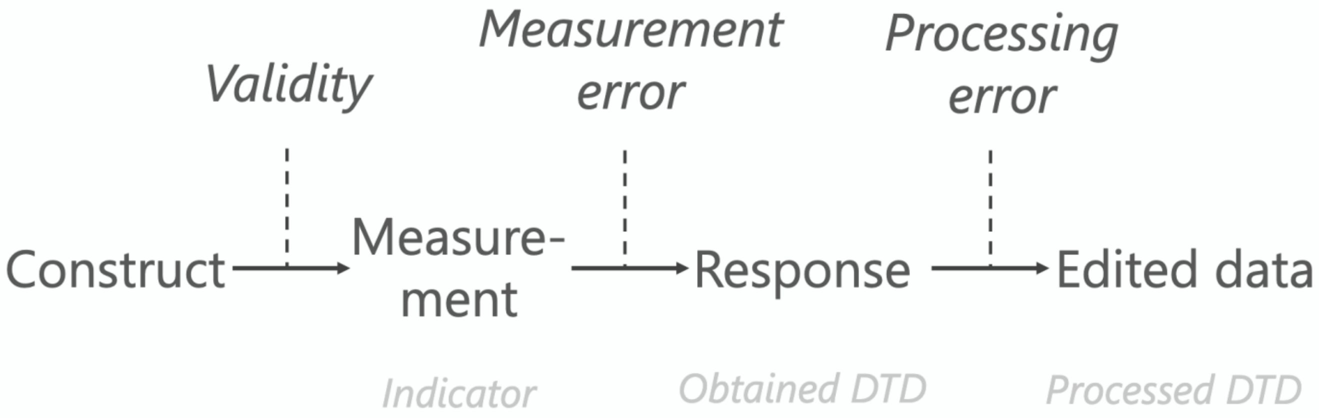


Data processing (interpretation and augmentation)


```
],
"genderInfo" : {
  "gender" : "male"
}
```

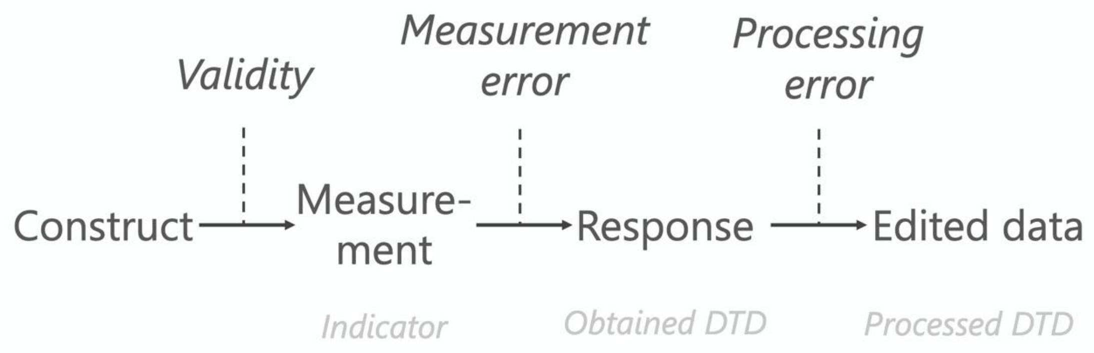
```
"inferredAgeInfo" : {
  "age" : [
    ">50"
  ],
  "birthDate" : ""
}
```

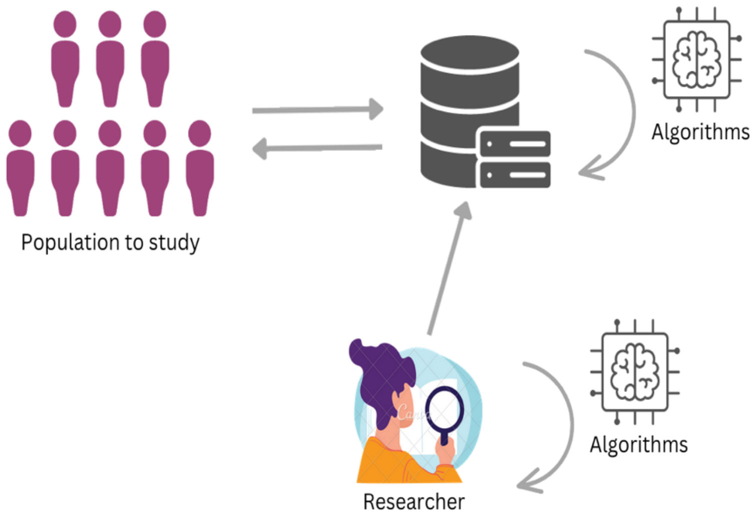
```
{
  "name" : "Rap",
  "isDisabled" : false
},
{
  "name" : "Retired life",
  "isDisabled" : false
},
{
  "name" : "Rom-com films",
  "isDisabled" : false
},
```

Errors introduced by ML

- Total error = measurement error + representation error
- ML impacts mostly measurement-side errors (from lecture 4):
    - On a conceptual level: **validity**
        - Data can be algorithmically confounded (recommendation algorithms)
    - **Measurement error:**
        - When the platform has augmented the data with ML (e.g. inferred gender)
    - **Processing error:**
        - When researchers use ML to process data

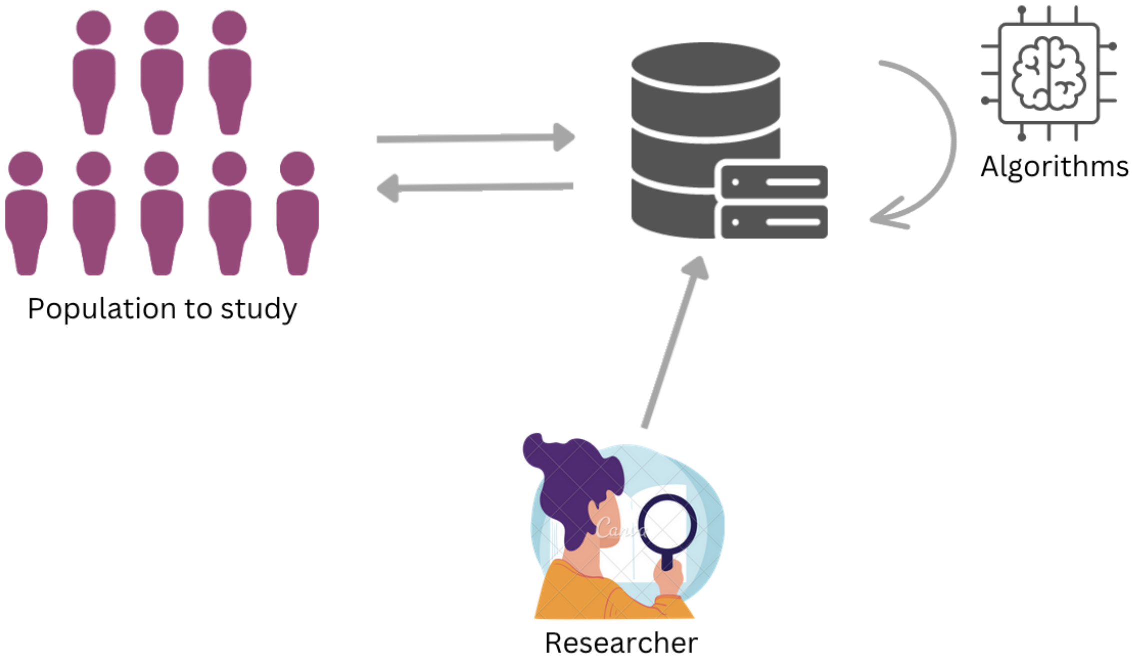




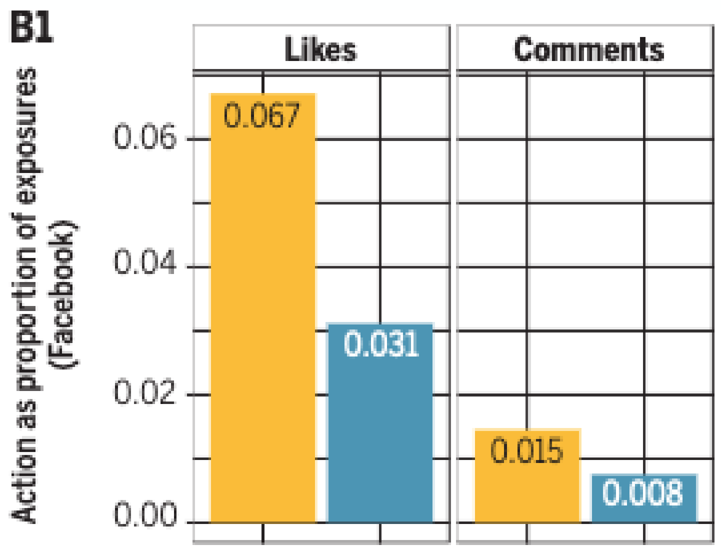
Problems with validity

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.

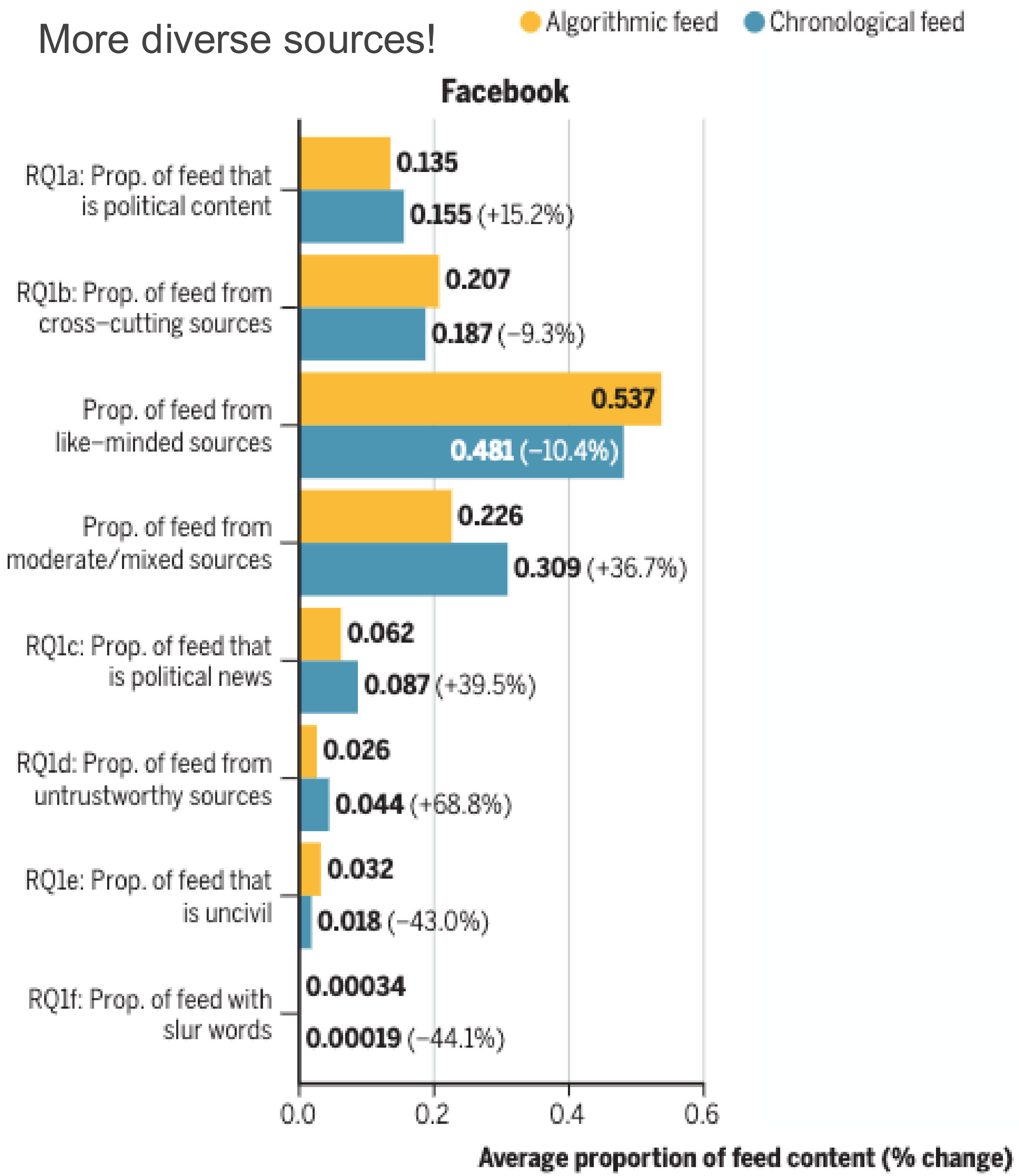


Problems with validity



People less engaged!

More diverse sources!



How do social media feed algorithms affect attitudes and behavior in an election campaign?
Guess et al., 2023, Science

In your group project

Think about how ML may be affecting:

- The text data you collected (validity)
- The labels you infer from the text (processing error)

3b. How does machine learning affect societies?

When machine learning models are being used to make decisions, they cannot be separated from the social and ethical context in which they are applied *(Kit T. Rodolfa, Pedro Saleiro, and Rayid Ghani, Big Data book)*

a) Through errors in algorithms impacting people
b) Through feedback loops
c) By reinforcing power structures

A) Errors in algorithms

We need to consider the errors of every algorithm:

- How often they fail?
- For whom do they fail? (***bias/fairness***)

Remember there are people behind the data:

- What are the costs of those failures?
- What are the long-term effecst? (feedback effects)

Algorithms make high-impact decisions

ML is used in many crucial areas for human wellbeing:

- Who to hire – CV screening
- Who to promote – performance reviews
- Who to insure – US life insurance premium based on social media
- Who to jail – predictive policing
- Who to kill – “we kill people based on metadata”, Gospel/Lavender systems in Gaza.

Those algorithms are often (1) “opaque,” (2) “beyond dispute or appeal,” and (3) disproportionally impact the underprivileged (Cathy O’Neal)

Chaty O’Neil
Mathematician (Harvard/MIT/Barnard College)
Worked four years in finance and advertisement

B) Algorithms create feedback effects

We often cannot separate ML from societies

Measuring algorithmically infused societies
Wagner et al, 2021, Nature

Facebook told the White House to focus on the 'facts' about vaccine misinformation. Internal documents show it wasn't sharing key data.

The tech giant meticulously tracked how misleading medical information spread — but didn't tell policymakers, even as they demanded it do so.

https://www.washingtonpost.com/technology/2021/10/28/facebook-covid-misinformation
https://www.wsj.com/articles/facebook-mark-zuckerberg-vaccinated-11631880296?mod=article_inline

Restricted access | Research article | First published online September 12, 2020

Upvoting extremism: Collective identity formation and the extreme right on Reddit

Tiana Gaudette, Ryan Scrivens, [...], and Richard Frank +1 View all authors and affiliations

Volume 23, Issue 12 | https://doi.org/10.1177/1461444820958123

Contents | Get access

Abstract

Since the advent of the Internet, right-wing extremists and those who subscribe to extreme right views have exploited online platforms to build a collective identity among the like-minded. Research in this area has largely focused on extremists' use of websites, forums, and mainstream social media sites, but overlooked in this research has been an exploration of the popular social news aggregation site Reddit. The current study explores the role of Reddit's unique voting algorithm in facilitating "othering" discourse and, by extension, collective identity formation among members of a notoriously hateful subreddit community, r/The_Donald. The results of the thematic analysis indicate that those who post extreme-right content on r/The_Donald use Reddit's voting algorithm as a tool to mobilize like-minded members by promoting extreme discourses against two prominent out-groups: Muslims and the Left. Overall, r/The_Donald's "sense of community" facilitates identity work among its members by creating an environment wherein extreme right views are continuously validated.

Simil

C) ML reinforces power structures

Meredith Whittaker

- Employed by Google for 13 years
- Research Professor at New York University
- Co-founder and faculty director of the AI Now Institute.
- President of Signal

"Private computational systems marketed as artificial intelligence (AI) are threading through our public life and institutions, concentrating industrial power, compounding marginalization, and quietly shaping access to resources and information" *(in The Steep Cost of Capture, 2021)*

"The commoditization of our data enables an asymmetric redistribution of power that is weighted toward the actors who have access and the capability to make sense of information" (Sarah Myers West, 2017)

Even if the algorithms are unbiased!

Concentration of control

Market share of cloud computing providers

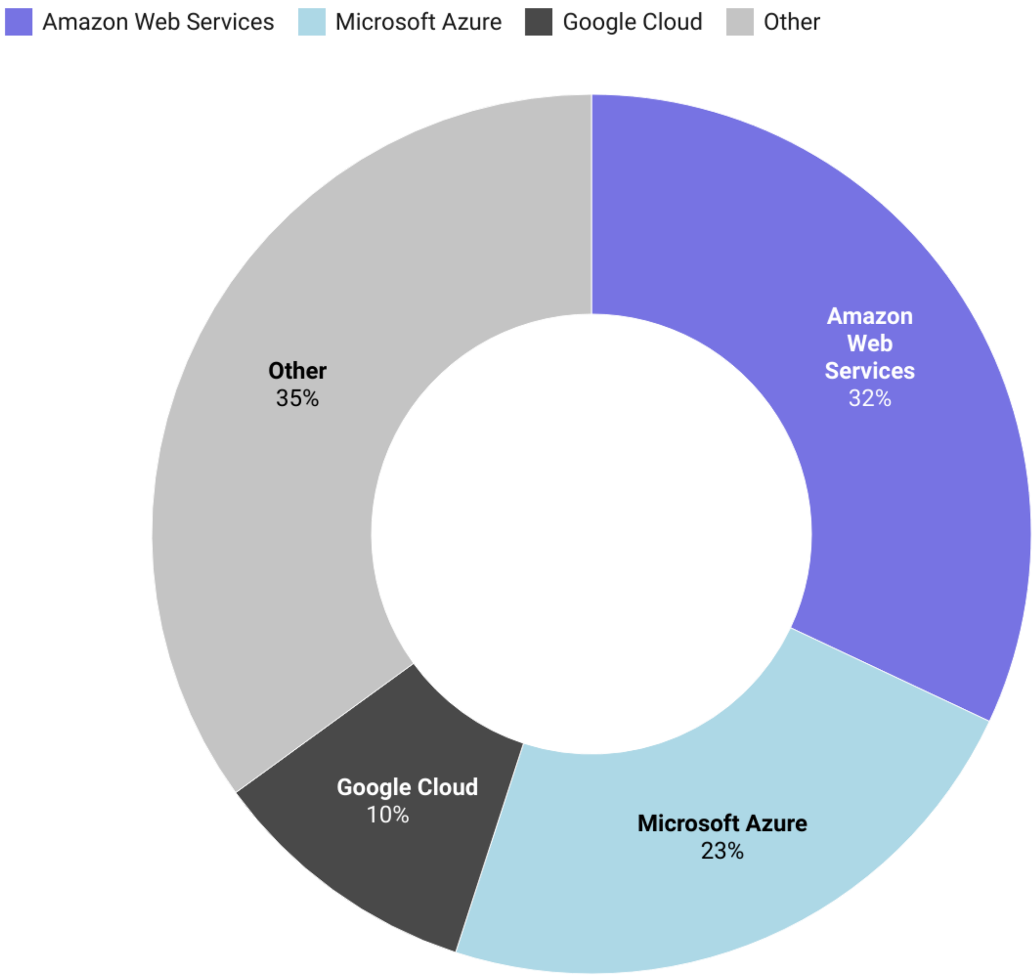


Chart: Exponential View • Source: CRN, AI Now Institute • Created with Datawrapper

Only hegemonic companies have the capital and power to thrive in the new era, reinforcing their power and dominance.

Revenue in 2024:

Amazon:	$638 bn.
Netherlands:	$440 bn. (407 bn. eur)
Apple:	$391 bn.
Alphabet:	$350 bn.
Microsoft:	$262 bn.
Meta:	$164 bn.

(updated from Babic et al., 2017)

- We understand social media platforms as ways to share and see content.
- Private companies see them as a way to make money, and they influence societal outcomes by controlling information flows and targeting ads and services.

Data feminism: data is never neutral

Data feminism = who has power, who doesn't, how power shapes data collection, interpretation, and use.

Who is doing the work of data science (and who is not)?

AI labs are dominated by men from elite institutions → limited perspectives and *privilege hazard*: those with privilege often don't see how data systems disadvantage others.

e.g. Face recognition not detecting black people

Whose goals are prioritized in data science (and whose are not)?

Why do we collect data on how to predict crime, instead of how to support those in need?

e.g., until recently, crash test dummies were designed in the size and shape of men, an oversight that meant that women had a 47 percent higher chance of car injury than men.

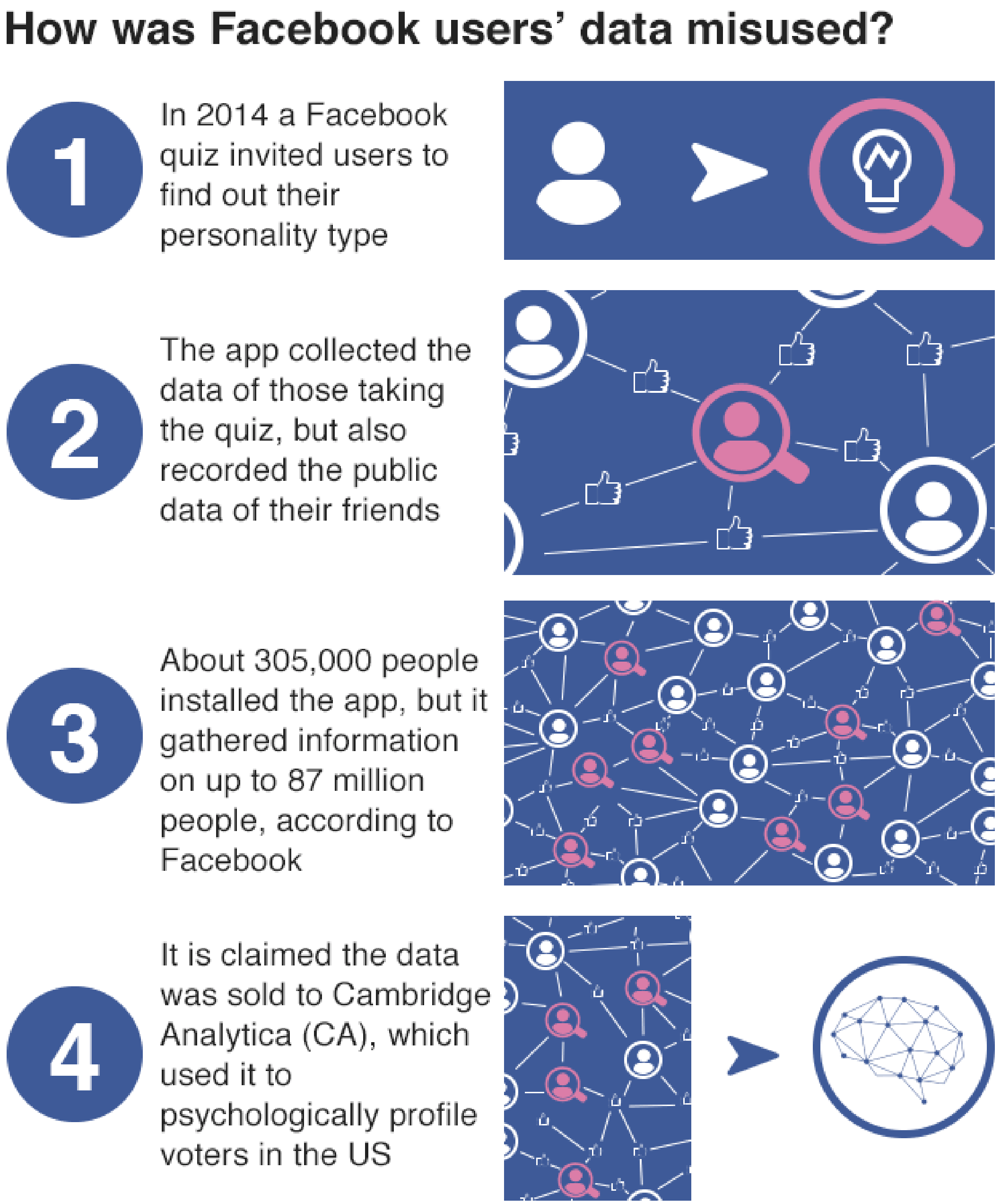
And who benefits from data science (and who is either overlooked or actively harmed)?
- Algorithms often surveil the poor more than the rich (Allegheny County child abuse risk scores).

- Target's pregnancy model exposed a teenager's private health information → profit over privacy.

- Facebook–Cambridge Analytica → corporate/political gain vs. public trust.

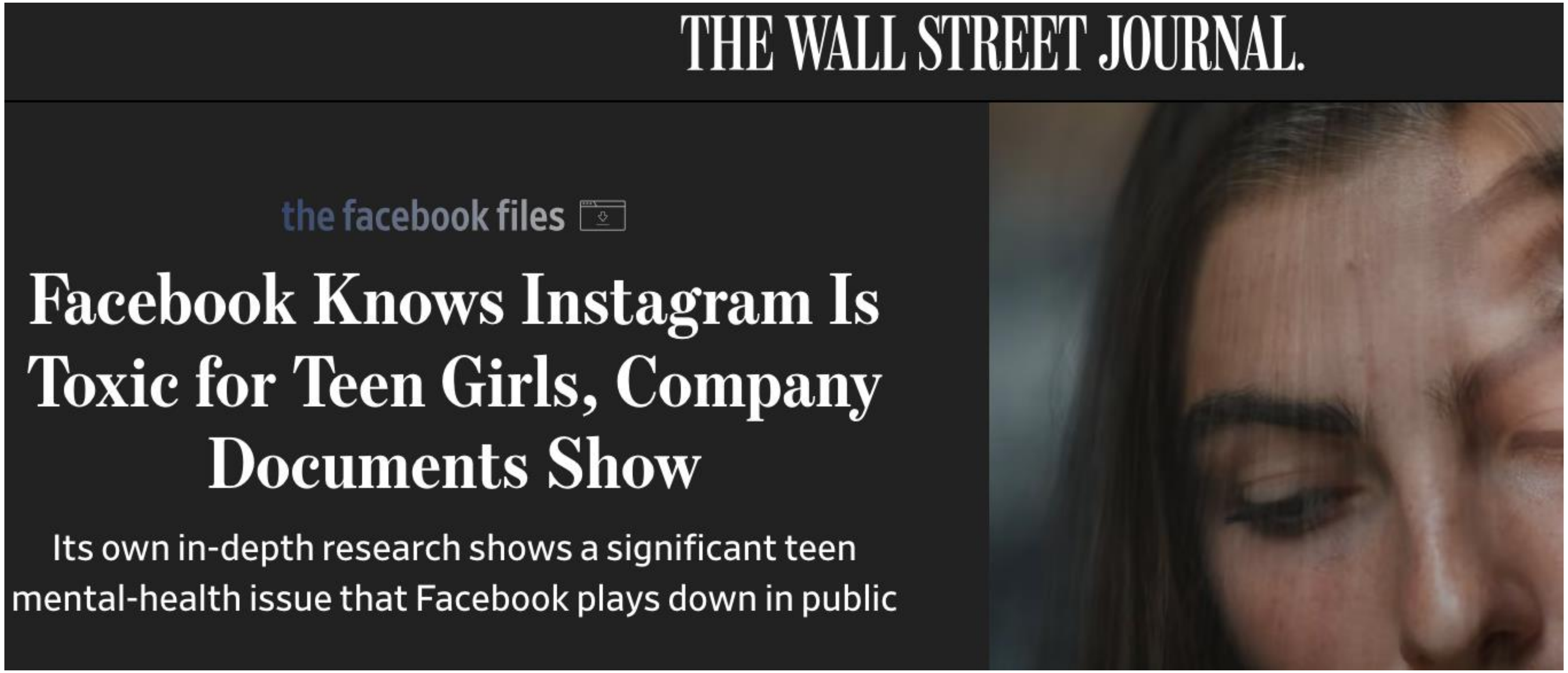
Whose goals are prioritized (and whose are not)?

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



Cambridge Analytica infographic

Harm to society is often accepted as part of the business model



For years they had focus groups, online surveys, diary studies - so this was not one chance finding.

- *A 2019 presentation slide said: "We make body-image issues worse for one in three teenage girls"*
- *Another slide said teenagers blamed Instagram for increased levels of anxiety and depression*
- *Some 13% of UK teenagers and 6% of US users surveyed traced a desire to kill themselves to Instagram*

Instagram response: "Based on our research and feedback from experts, we've developed features so people can protect themselves from bullying, we've given everyone the option to hide 'like' counts and we've continued to connect people who may be struggling with local support organisations."

https://www.bbc.com/news/technology-58570353

Performing Platform Governance: Facebook and the Stage Management of Data Relations

Karen Huang[1] · P. M. Krafft[2]

Received: 2 April 2021 / Accepted: 12 February 2024 / Published online: 4 April 2024

Abstract

Controversies surrounding social media platforms have provided opportunities for institutional reflexivity amongst users and regulators on how to understand and govern platforms. Amidst contestation, platform companies have continued to enact projects that draw upon existing modes of privatized governance. We investigate how social media companies have attempted to achieve closure by continuing to set the terms around platform governance. We investigate two projects implemented by Facebook (Meta)—authenticity regulation and privacy controls—in response to the Russian Interference and Cambridge Analytica controversies surrounding the 2016 U.S. Presidential Election. Drawing on Goffman's metaphor of stage management, we analyze the techniques deployed by Facebook to reinforce a division between what is visible and invisible to the user experience. These platform governance projects propose to act upon *front-stage data relations:* information that users can see from other users—whether that is content that users can see from "bad actors", or information that other users can see about oneself. At the same time, these projects relegate *back-stage data relations*—information flows between users constituted by recommendation and targeted advertising systems—to invisibility and inaction. As such, Facebook renders the user experience actionable for governance, while foreclosing governance of back-stage data relations central to the economic value of the platform. As social media companies continue to perform platform governance projects following controversies, our paper invites reflection on the politics of these projects. By destabilizing the boundaries drawn by platform companies, we

Exercise for the practical

Predictive policing is increasingly used (also by the Netherlands). Let’s imagine these systems are trained in historical and personal data. The use of personal data for predictive policing has recently been banned in the EU.

Where can biases enter those models? (sample/outcome/features/pipeline/application)
Who are the actors involved? Whose interests are prioritized (and who's not)?
Do you think the benefits outweight the biases?

How does ML affect societies? – Summary

Data and ML are never neutral
- They always carry **social, ethical, and political consequences**.

Power shift
- Growth of data & ML concentrates power in big tech companies.

Algorithms
- Enable functionality (e.g., search, recommendations, personalization)
- Are often opaque, can **marginalize groups**, and create **feedback loops**.

Data feminism
- Provides a critical lens: Who has power? Who is excluded? Who benefits, and who is harmed?

4. Dealing with bias in ML

What information do we need to understand bias?

Information on *protected attributes* (gender, sex, class, etc)

Information on *true labels* to create the confusion matrix (per attribute) → Often we need to manually label data, but be mindful of your own biases!

A definition of bias/fairness → There is no universally-accepted definition of what it means for a model to be fair. This is not an excuse for ignoring fairness!

Group A	Predicted criminal	Predicted not criminal
Criminal	10	10
Not criminal	1	100

Group B	Predicted criminal	Predicted not criminal
Criminal	10	1
Not criminal	10	100

Errors in algorithms

We need to consider the errors of every algorithm:

- How often they fail?
- For whom do they fail? (***bias/fariness***)

Remember there are people behind the data:

- **What are the costs of those failures?**
  - *Assistive intervention*: "Individuals may be harmed by being incorrectly included in the "low need" population that does not receive an intervention" (**harm = false negatives**, e.g., not supporting somebody in need)
  - *Punitive intervention*: "Individuals may be harmed by being incorrectly included in the "high risk" population that receives an intervention" (**harm = false positives,** e.g. jailing an innocent)

- What are the long-term effecst? (feedback effects)

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

Different definitions of bias, punitive example

- Among the general population (T), the probability of *being wrongly jailed* is independent of race

$$\frac{\text{False positive}}{\text{All}}$$

- Among the jailed population (those predicted criminals), the probability of being wrongly jailed is independent of race: Parity in **False Discovery Rate.**

→ Focuses on those affected by the intervention

$$\frac{\text{False positive}}{\text{Predicted criminal}}$$

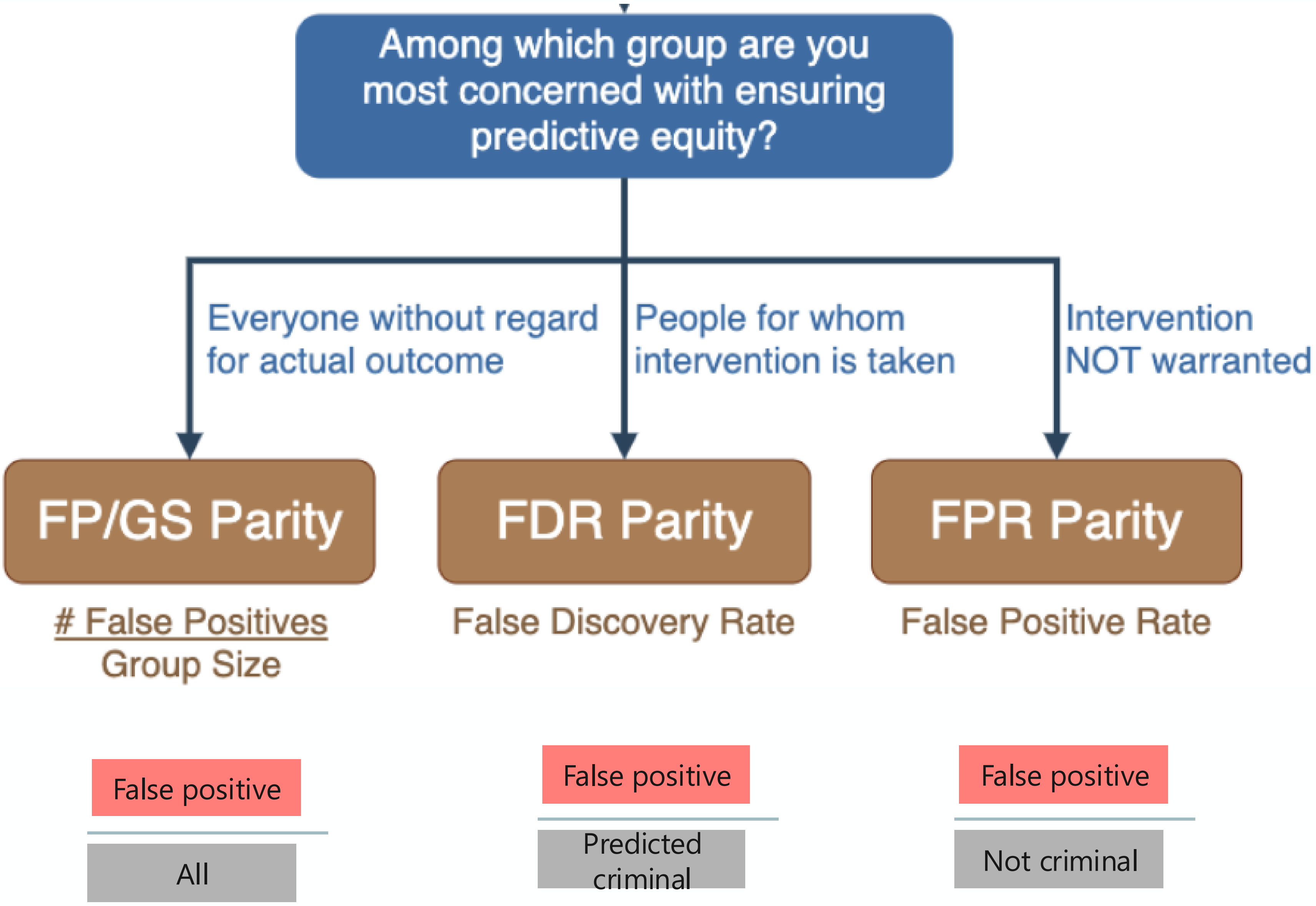
- Among innocents (the actual non-criminals), the probability of being wrongly jailed is independent of rage: Parity in **False Positive Rate** (Predictive Equality)

→ Focuses on those who should not be affected by the intervention

$$\frac{\text{False positive}}{\text{Not criminal}}$$

	Predicted criminal	Predicted not criminal
Criminal	True positive	False negative
Not criminal	False positive	True negative

No universal definition of fairness!


Among which group are you most concerned with ensuring predictive equity?
Everyone without regard for actual outcome
People for whom intervention is taken
Intervention NOT warranted
FP/GS Parity
FDR Parity
FPR Parity
False Positives
Group Size
False Discovery Rate
False Positive Rate
False positive
All
False positive
Predicted criminal
False positive
Not criminal

Models can be fair and unfair at the same time

	Predicted criminal	Predicted not criminal
Criminal	True positive	False negative
Not criminal	False positive	True negative

$$\frac{\text{False positive}}{\text{Not criminal}}$$

$$\frac{\text{False positive}}{\text{Predicted criminal}}$$

Table 11.1: COMPAS Fairness Metrics

Metric	Caucasian	African American
False Positive Rate (FPR)	23%	45%
False Discovery Rate (FDR)	41%	37%

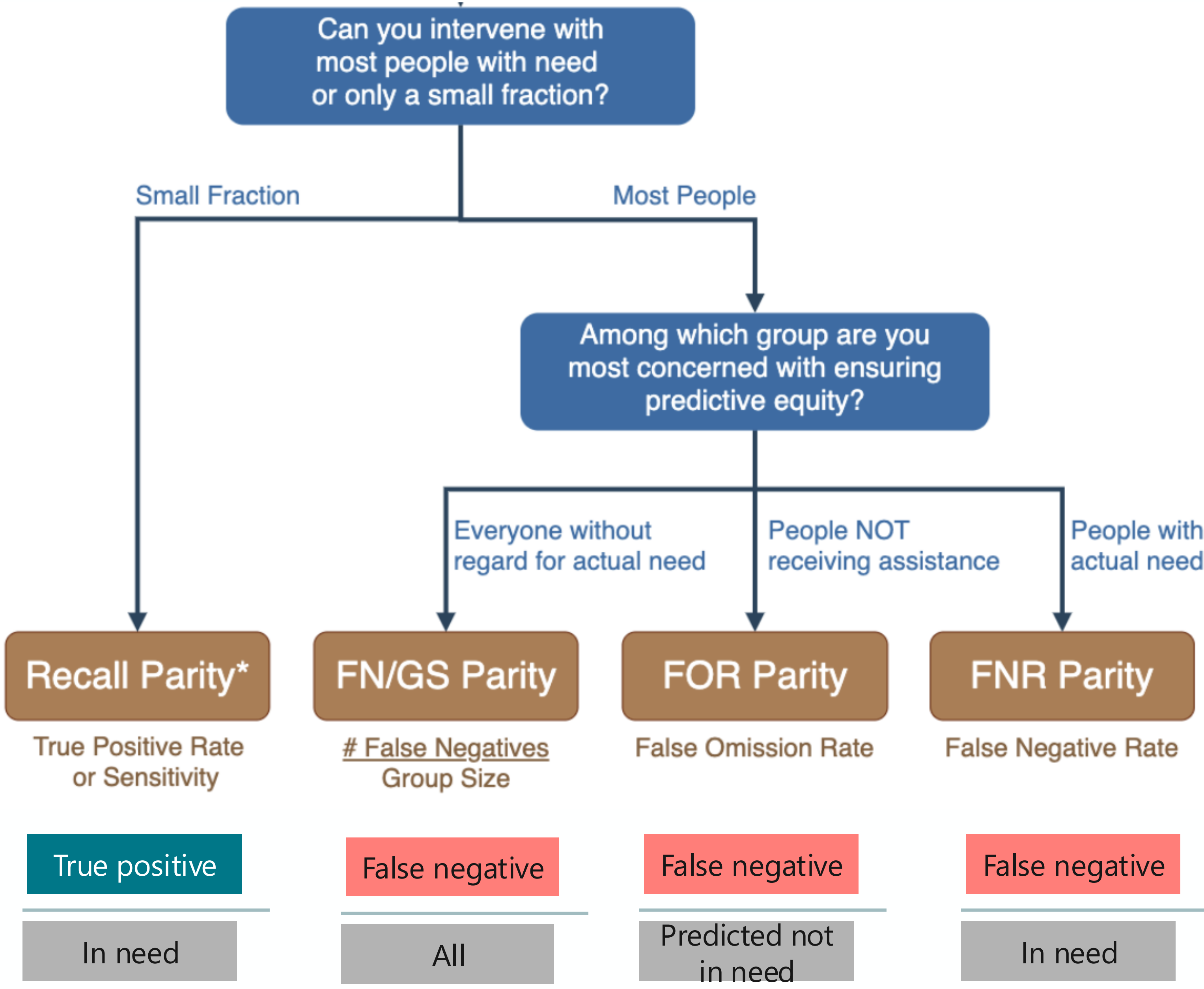
Correctional Offender Management Profiling for Alternative Sanctions (COMPAS): Evaluates the likelihood of an offender committing another crime in the future.

- *FPR*: Among black defendants who would did not end up with another arrest, 45% were labeled by the system as high risk, almost twice the rate for whites (23%).
- *FDR*: Among individuals labelled as high risk, a similar fraction of black and white defendants were arrested again.

It's generally impossible for a model to maximize both fairness criteria at the same time

Assessing bias, assistive example

	Predicted in need	Predicted not in need
In need	True positive	False negative
Not in need	False positive	True negative



Dealing with bias in ML

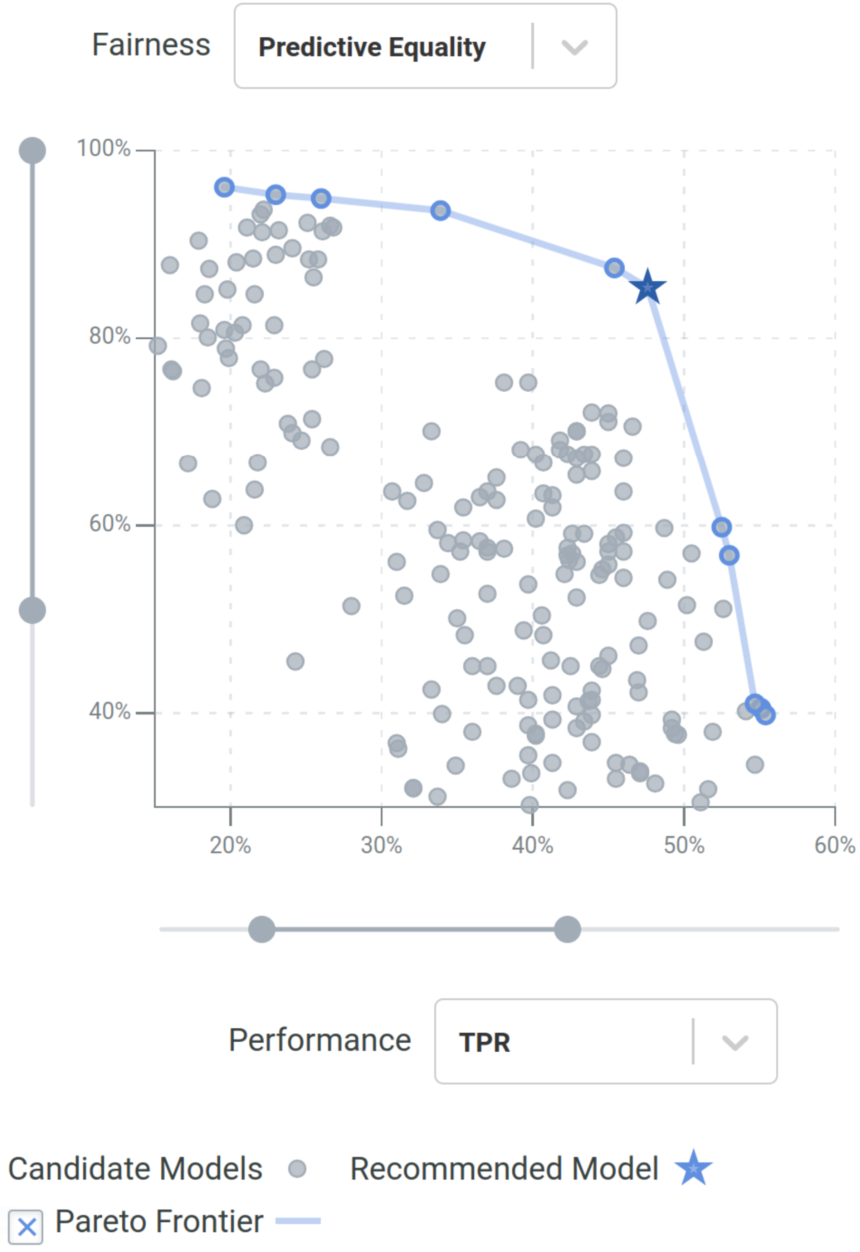
Audit the model to understand bias

Mitigate bias:
- Removing "protected attributes" (gender/race) generally does ***not*** mitigate bias. Other features are also correlated with those.
- Test different models and select one with strong performance across fairness and accuracy (Pareto optimality).
- Adjust thresholds to increase/decrease FP or FN. Example: Offer a subsidy to Group A if the model predicts a need with over 50% probability, and to Group B if the need is predicted with over 25% probability.

Consider Intersectionality: Optimizing for one attribute (e.g., gender) may introduce bias in another (e.g., class).

Regularly Test for Bias: Monitor for concept drift and ensure ongoing fairness.

Consider if Bias May Be Acceptable: For example, if the intervention is most useful to a specific subpopulation.



TODAY

Lecture

1. Explain machine learning in your own words

2. Explain *why* machine learning models may be biased (sources of bias).

3. Understand the effects of ML on DTD and in society.

4. Assess bias in ML models

Lab

Apply a ML model to text data

Audit a ML model

Block 2: https://textminingcourse.nl/

202400006

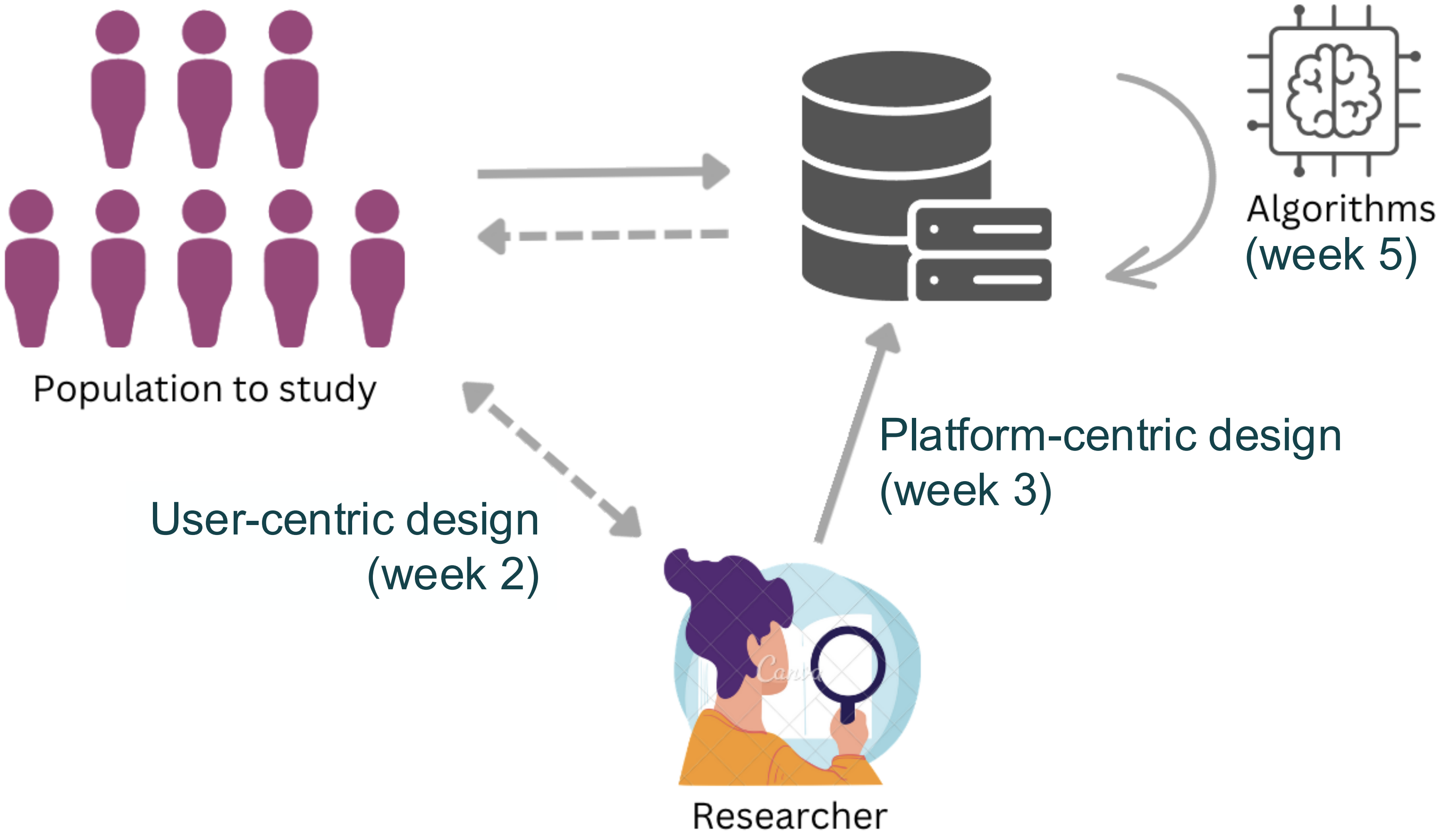
Materials of Text Mining, Transforming Text into Knowledge

Lectures

Week	Date	Topic
1	2025-02-03	Intro to text mining & regular expressions
2	2025-02-10	Text preprocessing
3	2025-02-17	Text classification
4	2025-02-24	Feature selection
5	2025-03-03	Clustering & topic modeling
6	2025-03-10	Word embedding
7	2025-03-17	Deep learning & LLMs
8	2025-03-24	Sentiment analysis
9	2025-03-31	Responsible text mining & applications

Labs

Summary of the course



Week 4: Errors in DTD
Week 6: Ethics and Legislation
Week 9: Recap and Q&A