Künstliche Intelligenz in der US-Strafjustiz:

Fairness und Bias in Vorhersagemodellen

A Social Data Science Perspective
Frauke Kreuter LMU Munich & UMD

KI Nutzung bei Entscheidungen der Justiz

Ziele:

- Steigerung der Effizienz,
- Verringerung der Subjektivität,
- ohne Verlust der Effektivität.

Letztlich:

Verringerung der Insassenzahl.

Library of Congress Prints and Photographs Division Washington, D.C. 20540 USA (intermediary roll film) thc 5a38062 https://hdl.loc.gov/loc.pnp/thc.5a38062

Law Enforcement Use of Artificial Intelligence

December 15, 2023

The use of artificial intelligence (AI) has expanded in a variety of arenas, including by law enforcement. AI has been broadly conceptualized as computerized systems operating in ways often thought to require human intelligence. It is defined in the *U.S. Code* (15 U.S.C. §9401(3)) as

> a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to-
>
> (A) perceive real and virtual environments;
>
> (B) abstract such perceptions into models through analysis in an automated manner; and
>
> (C) use model inference to formulate options for information or action.

AI involves a host of technologies and applications. In the law enforcement realm, researchers note that while the use of AI is not yet widespread, existing tools may be enhanced with AI to expand law enforcement *capabilities* and increase their *efficiency*. Examples include the following:

- Automated license plate readers can be leveraged to employ machine, or computer, vision for capabilities such as automating the issuance of red-light violation tickets.
- Security cameras outfitted with certain AI-embedded hardware can be used for real-time

equivant

Our Solutions | Who We Support | About Us | Resources | Blog | Contact Us

OUR PRODUCTS

NORTHPOINTE SUITE RISK NEEDS ASSESSMENTS

Successfully managing justice-involved people requires an individualized lens to identify risk and needs just for that person. Designed to support objective decision-making based on known data elements, a completed Risk and Need Assessment supports community-based supervision management (probation, parole), jail and institution intake for initial treatment planning, and problem-solving courts designed to divert people from jail to long-term healing. Design your assessment using our nationally validated COMPAS Assessment options, or employ other available assessments based on your needs.

KEY FEATURES

- 3 Risk Need Assessment Tools
- Alternative Screenings [Sex Offense, Criminal Thinking]
- Integrated Case Plans
- Ad Hoc Reporting

WHO IT SERVES

- Probation Officers
- Parole Officers
- Jail and Institution Staff
- Problem-Solving Courts
- Pretrial Supervision Agents

NORTHPOINTE SUITE CASE MANAGER

Northpointe Suite Case Manager is a person-based supervision case management solution that links person data together with individual assessment outcomes, treatment plans and long-term progress reports while offering critical functionality for Pre-Sentence Investigation

NORTHPOINTE SUITE PROBLEM-SOLVING COURTS

This module manages all court participant processing and case/court activities. Created specifically for problem-solving courts with Risk Need Assessments and Phase Management built right in, this solution is applicable for all types of

NORTHPOINTE SUITE CUSTODY MANAGEMENT

The Northpointe Suite Custody Management solution is not your typical 'book and release' jail management system. This best-in-class software incorporates three key pillars of success for any custody oriented agency: intake, facility and

COMPASS Lifecycle

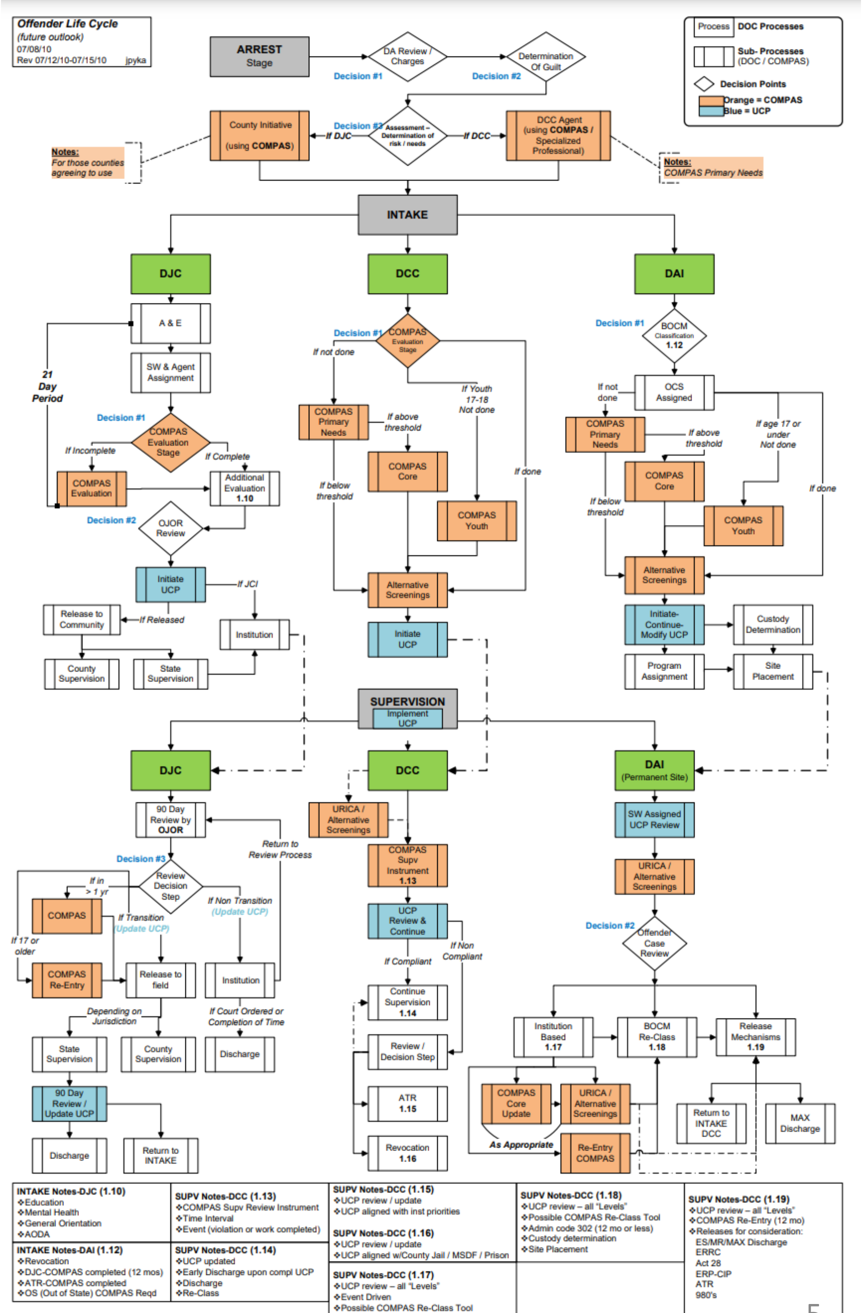
Correctional Offender Management Profiling for Alternative Sanctions

"**Evidence Based Practices** clearly state that having a sound assessment that accurately identifies an offender's risk to reoffend is the cornerstone of effective supervision.

…Services are targeted for moderate to **higher-risk offenders** who are likely to reoffend if appropriate interventions are not available.

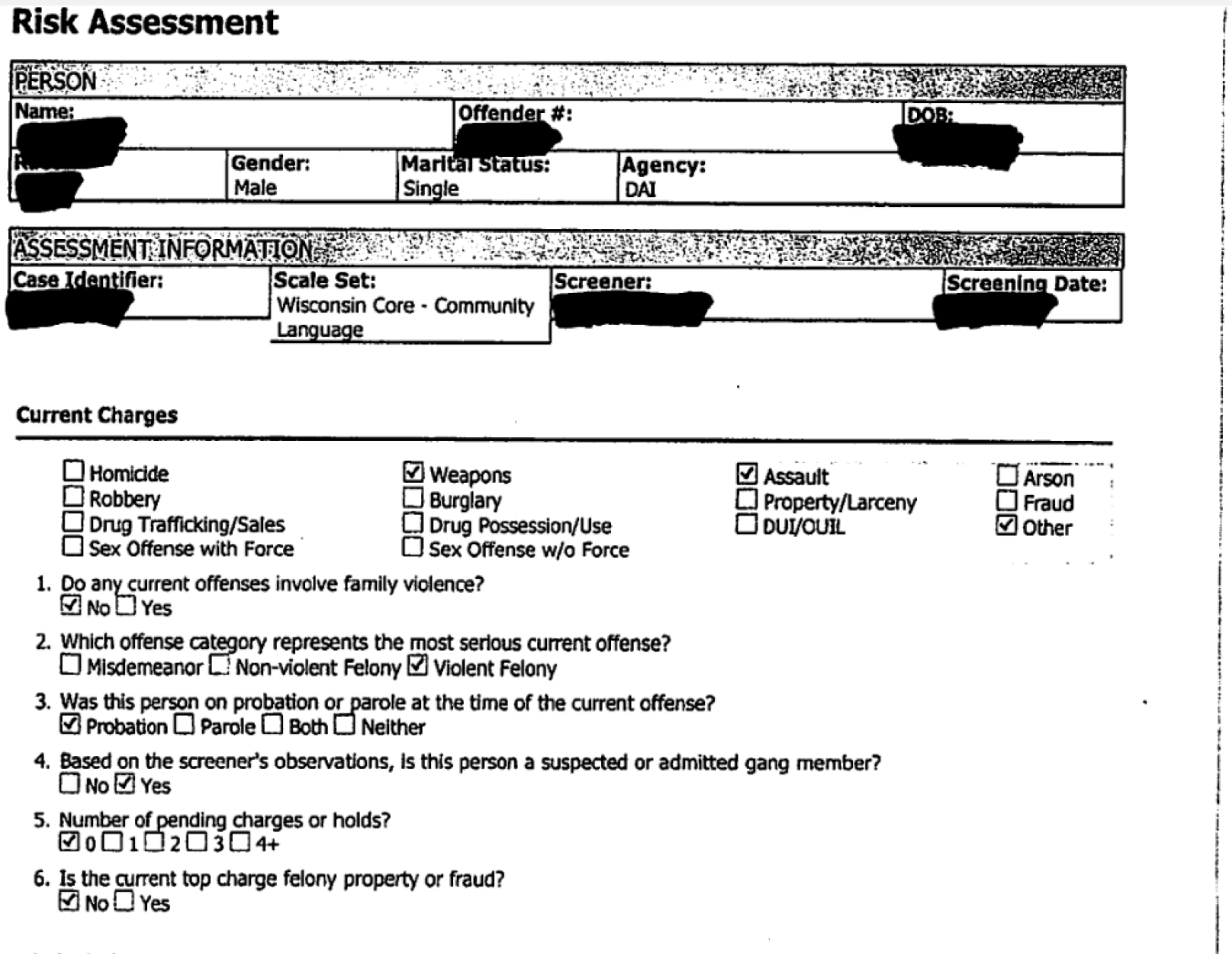
Research shows **low risk offenders** are less likely to commit new crimes and should be given minimal services and be excluded from intensive treatment and supervision"

https://doc.wi.gov/Pages/AboutDOC/COMPAS.aspx



COMPASS

- Risikoabschätzung (erneute Straffälligkeit)
- Basiert auf 137 Eigenschaften
- Kommerzielles Produkt von Northpointe
  Algorithmus nicht bekannt
- Genutzt zur Einschätzung von Strafmaß, Art des Strafvollzugs, Bewährung etc.



Risk Assessment

PERSON

Name: [redacted]	Offender #: [redacted]		DOB: [redacted]
[redacted]	Gender: Male	Marital Status: Single	Agency: DAI

ASSESSMENT INFORMATION

Case Identifier: [redacted]	Scale Set: Wisconsin Core - Community Language	Screener: [redacted]	Screening Date: [redacted]

Current Charges

☐ Homicide ☑ Weapons ☑ Assault ☐ Arson
☐ Robbery ☐ Burglary ☐ Property/Larceny ☐ Fraud
☐ Drug Trafficking/Sales ☐ Drug Possession/Use ☐ DUI/OUIL ☑ Other
☐ Sex Offense with Force ☐ Sex Offense w/o Force

1. Do any current offenses involve family violence?
   ☑ No ☐ Yes
2. Which offense category represents the most serious current offense?
   ☐ Misdemeanor ☐ Non-violent Felony ☑ Violent Felony
3. Was this person on probation or parole at the time of the current offense?
   ☑ Probation ☐ Parole ☐ Both ☐ Neither
4. Based on the screener's observations, is this person a suspected or admitted gang member?
   ☐ No ☑ Yes
5. Number of pending charges or holds?
   ☑ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4+
6. Is the current top charge felony property or fraud?
   ☑ No ☐ Yes

Entscheidungen im Arbeitsmarktkontext

Job-Center

- Ziel: Intervention zur Verhinderung von Langzeitarbeitslosigkeit
- Limitierte Ressourcen (Hilfsprogramme sind teuer), hoher Personaleinsatz
- Unnötig für alle mit niedrigem Risiko
- **Priorisierung und gezielte Intervention**

Job-Börsen

- Automatisierte Textanalyse
- Automatisierte Vorschläge
- Faire Erstauswahl
  Hangartner et al. 2021. Nature, Vol 589

Photo by Souvik Banerjee on Unsplash

Profiling Arbeitssuchender

Loxha and Morgandi, 2014; Desiere et al. 2019

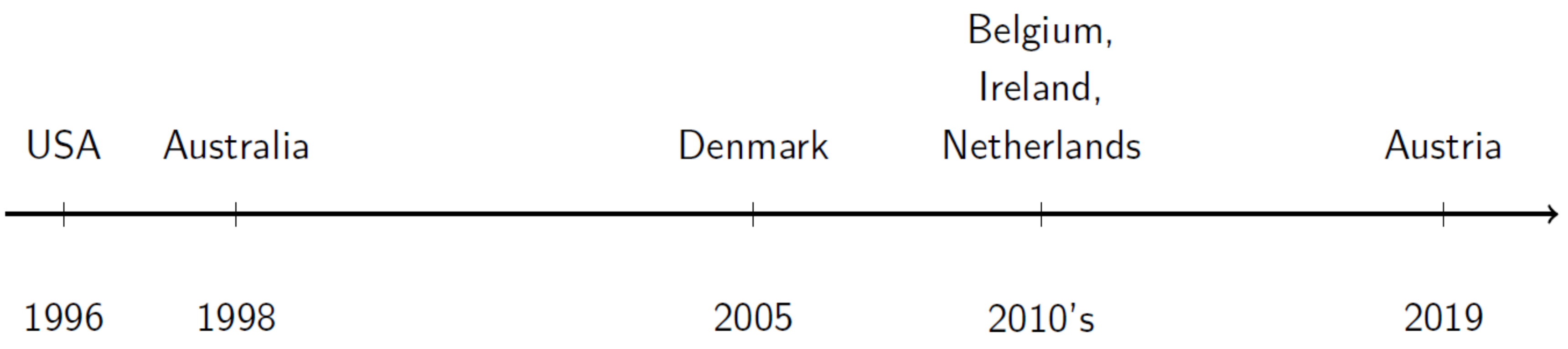
Menschliche Einschätzung: Angestellte entscheiden wer, wann und wie Unterstützung erhält
Problem: ineffizient, subjektiv, …

Regelbasierte Einschätzung: Programmqualifizierung erfolgt nach festgesetzten Regeln

Problem: ineffektiv, …

Algorithmus-basierte Entscheidungen: Modelle bestimmen Programm und/oder Programmqualifizierung
Hoffnung: effizient, effektiv, objektiv

Figure: Stylized timeline of **algorithmic profiling** across OECD



Vorhersagen im Gesundheitskontext

HEALTHY CHICAGO
CHICAGO DEPARTMENT OF PUBLIC HEALTH

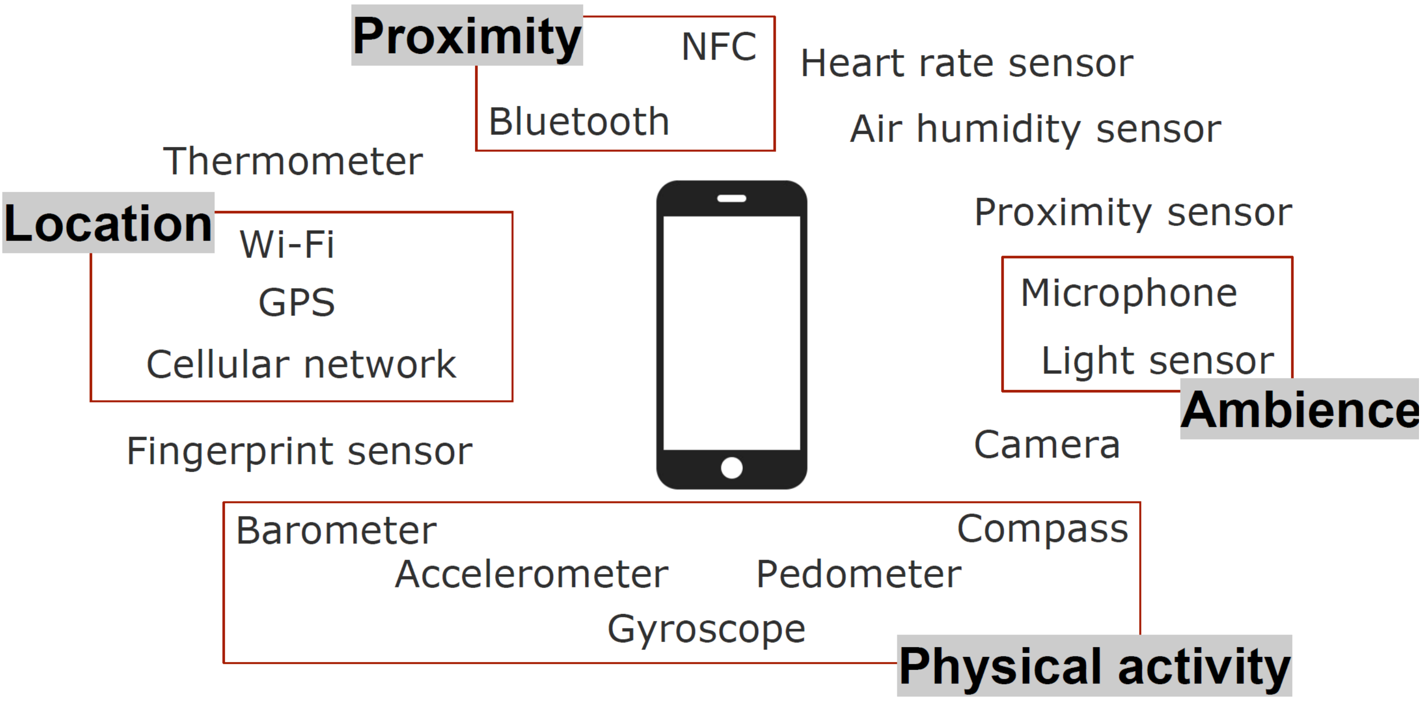
- In den US verbleiben weniger als die Hälfte der HIV-Positiven in Versorgungsprogrammen.
- KI hilft ambulanten vorherzusagen, wer aus Vorsorge herausfällt.
- Ressourcen für Interventionen werden entsprechend des Ausfallrisikowertes geplant

Ramachandran, A., Kumar, A., Koenig, H. et al. Predictive Analytics for Retention in Care in an Urban HIV Clinic. Sci Rep 10, 6421 (2020). https://doi.org/10.1038/s41598-020-62729-x

- Kooperation: Johnson County, Kansas
  Carnegie Mellon University
- Ziel: Spirale unbehandelter psychiatrischer Erkrankungen und Verhaftungen zu durchbrechen.
- KI-gestützte Vorhersage der Wahrscheinlichkeit ins Gefängnis zu kommen
- Personen mit hohem Risikowert werden je nach vorhandenen Ressourcen priorisiert behandelt

Rodolfa, K.T., Lamba, H. & Ghani, R. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. Nat Mach Intell 3, 896–904 (2021). https://doi.org/10.1038/s42256-021-00396-x

Vorhersagen im COVID-19 Kontext



KI-gestützte Vorhersagen für COVID Erkrankungen aus tragbare Sensoren und persönlichen Angaben

Quer, Radin, Gadaleta, 2021, Wearable sensor data and self-reported symptoms for COVID-19 detection. Nature Medicine

Keusch, Conrad, 2021, Using Smartphones to Capture and Combine Self-Reports and Passively Measured Behavior in Social Research, *Journal of Survey Statistics and Methodology*

Geschützte Attribute

Table: **Protected attributes** in the US as defined by the Fair Housing Act (FHA) and Equal Credit Opportunity Act (ECOA)

Attribute	FHA	ECOA
Race	✓	✓
Color	✓	✓
National origin	✓	✓
Religion	✓	✓
Sex	✓	✓
Familial status	✓	
Disability	✓	
Exercised rights under CCPA		✓
Marital status		✓
Recipient of public assistance		✓
Age		✓

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Fair und Unfair?

Table: Confusion matrix

		Prediction 0	Prediction 1	
Reference	0	TN	FP	N’
	1	FN	TP	P’
		N	P	

Table: COMPAS Fairness Metrics (Rodolfa et al. 2020)

Metric	Caucasian	African American
False Positive Rate (FPR)	23%	45%
False Negative Rate (FNR)	48%	28%
False Discovery Rate (FDR)	41%	37%

- ProPublica focused on FPR and FNR
- Northpointe’s response[2] put forward FDR
- Is the model “*...simultaneously fair and unfair*”?

Incompatibility Between Fairness Metrics

$$FPR = \frac{p}{1-p}\left(\frac{FDR}{1-FDR}\right)(1-FNR)$$

<u>False Positive Rate</u>
Among all actual 0’s, fraction predicted to be 1

<u>Prevalence</u>
Fraction of actual 1’s in population

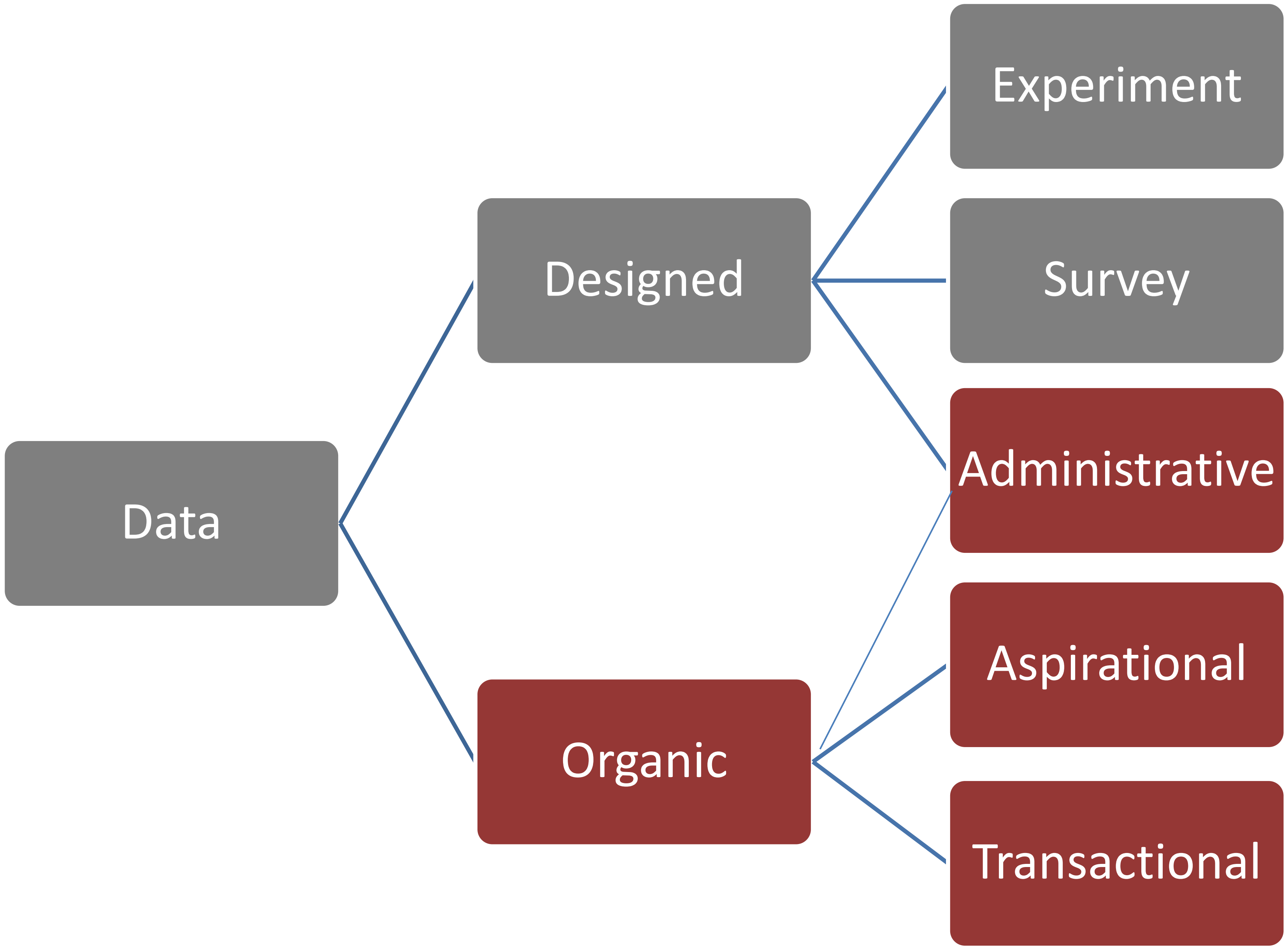
<u>False Discovery Rate</u>
Among all predicted 1’s, fraction that are actual 0’s =(1 – precision)

<u>False Negative Rate</u>
Among all actual 1’s, fraction predicted to be 0

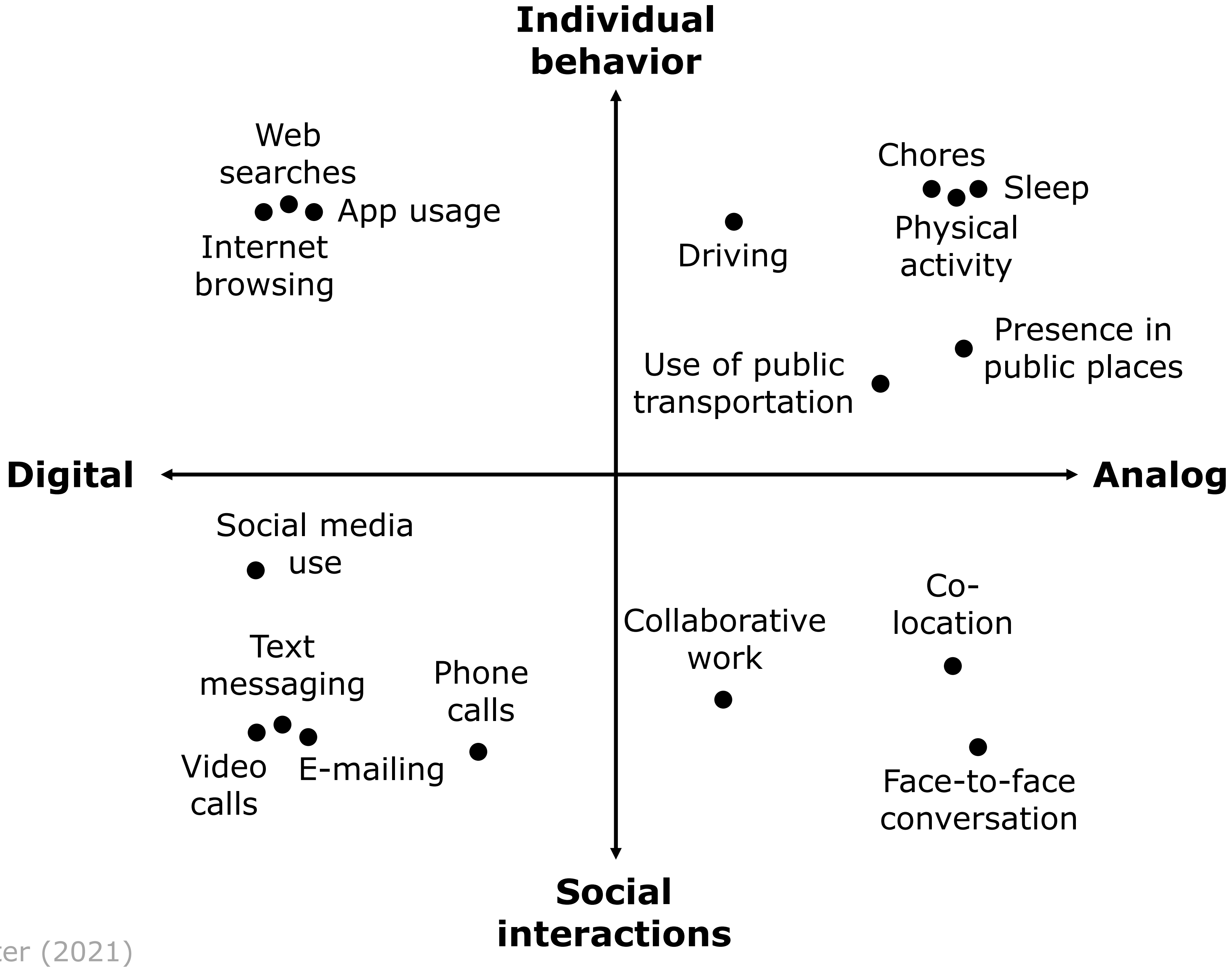
Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153-163.

- 2 https://www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/

Datengenerierende Prozesse

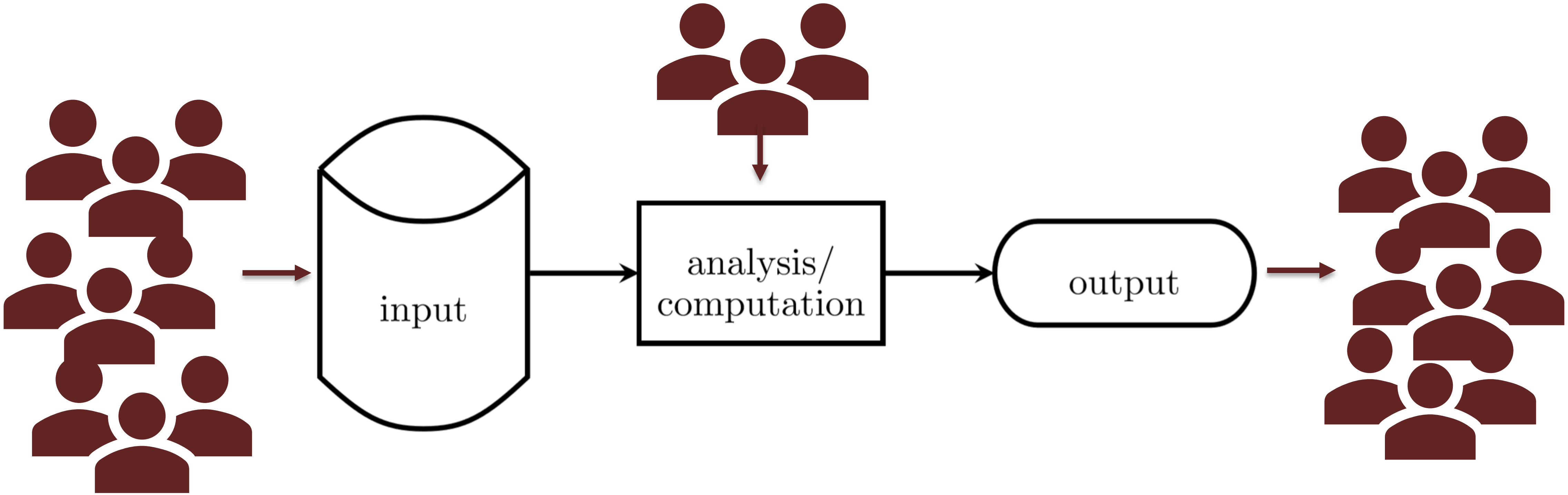


Source: Roberto Rigobon, [Discussion on Applications and Issues with Using Commercial Data in Research](), BEA Expert Meeting on Exploiting Commercial Data for Official Economic Statistics November 19, 2015

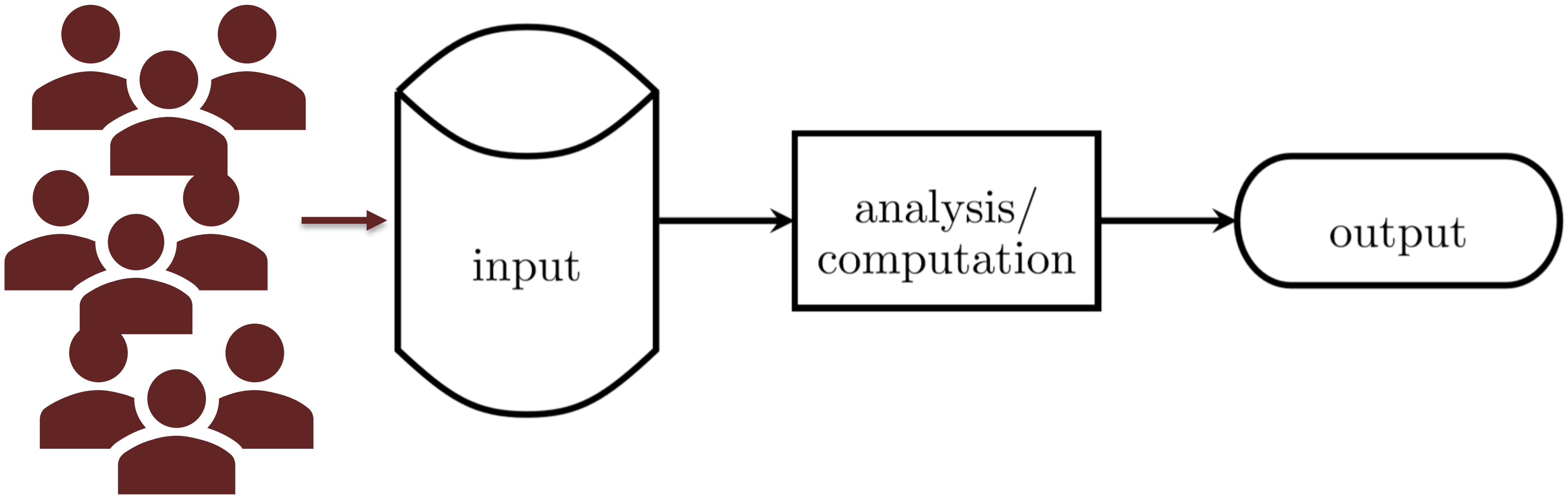


Keusch & Kreuter (2021)

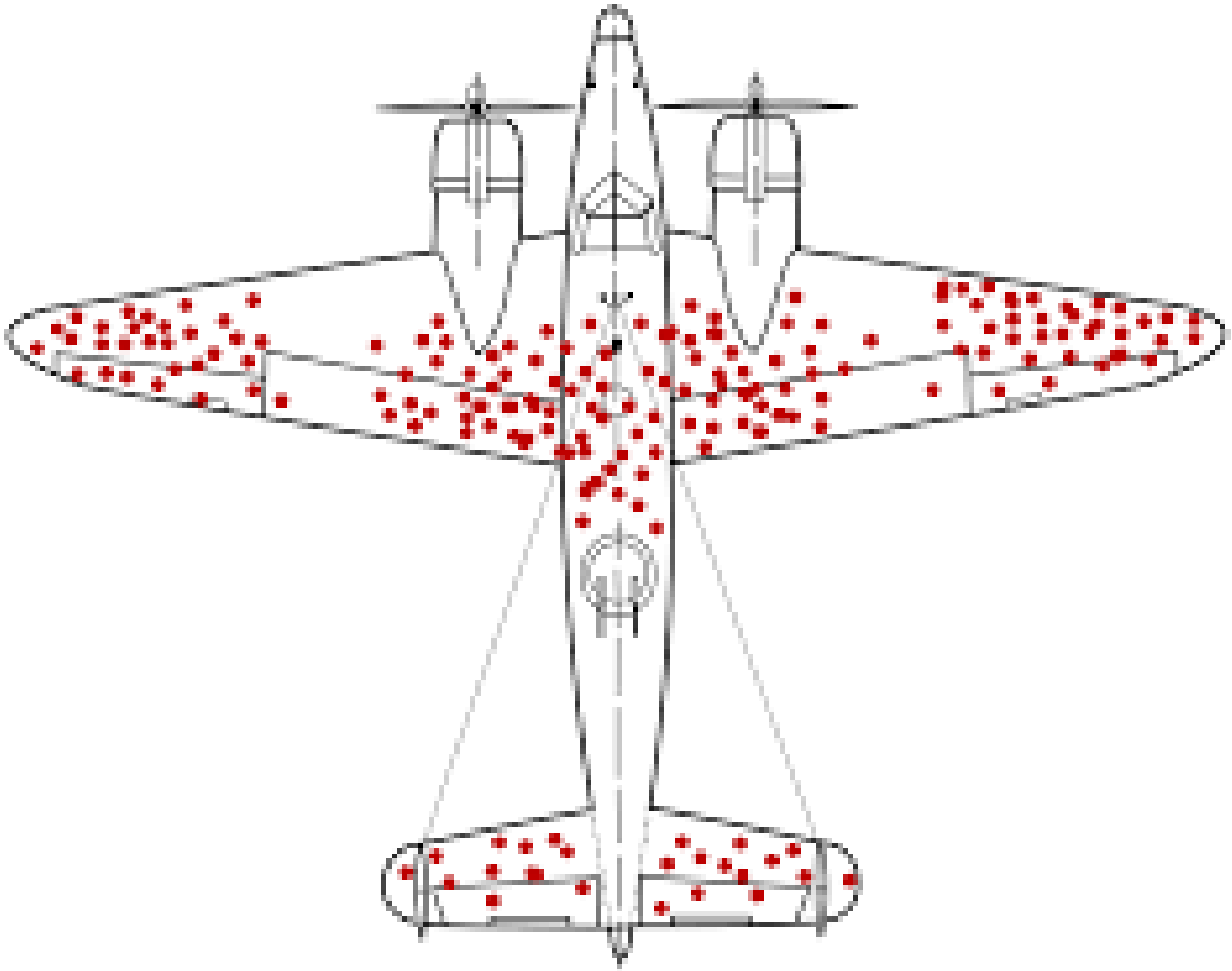
Entscheidungen



Herausforderung: Verfügbarkeit von Daten


input
analysis/
computation
output

Boston Street Bumps



Wald, Abraham. (1943). *A Method of Estimating Plane Vulnerability Based on Damage of Survivors*.

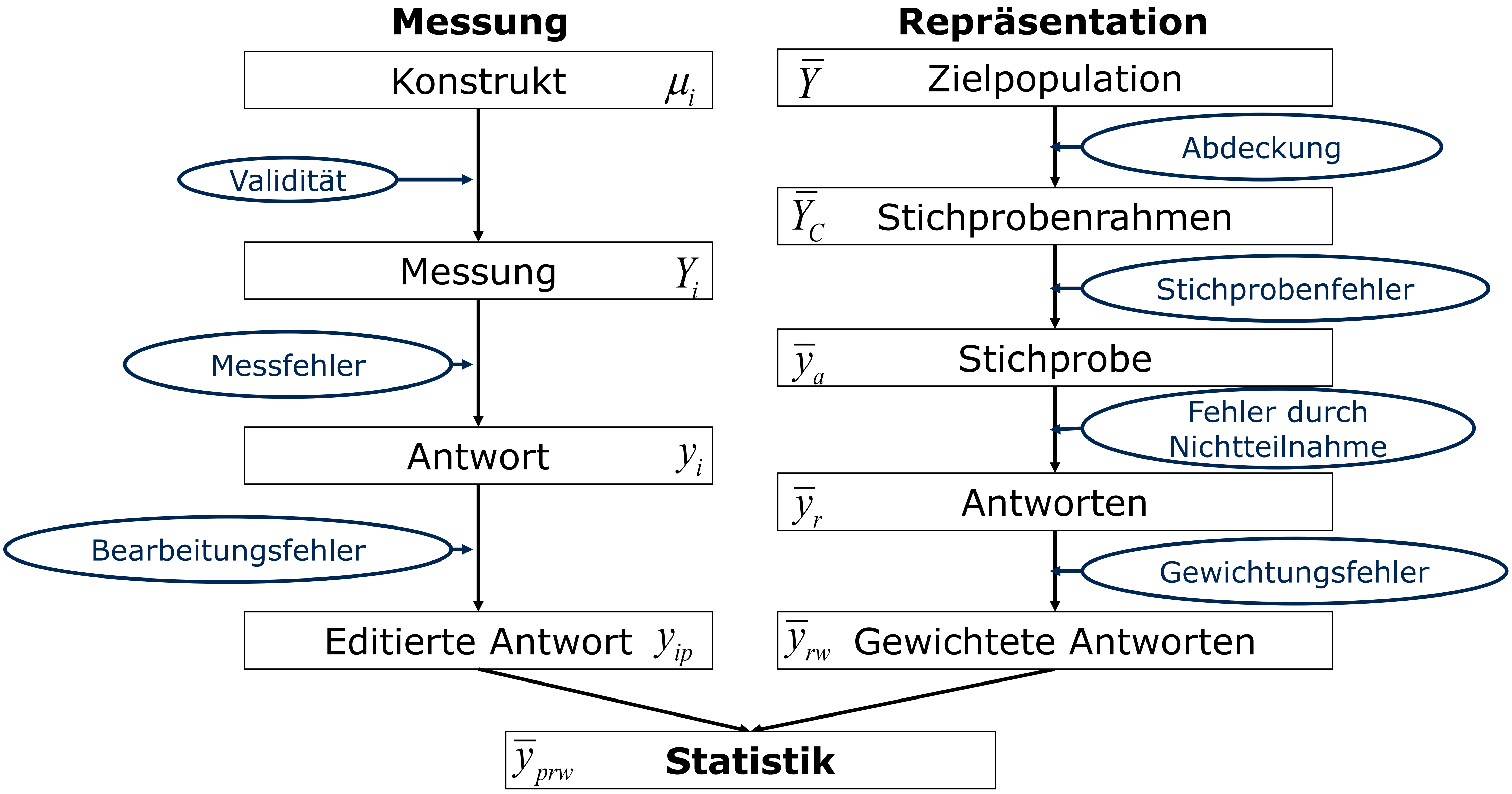
Statistical Research Group, Columbia University.

accessed 8.5.2022
https://apps.dtic.mil/docs/citations/ADA091073

Illustration of hypothetical damage pattern on a WW2 bomber, based on report above; picture concept by Cameron Moll (2005, claimed on Twitter and credited by Mother Jones), new version by McGeddon based on a Lockheed PV-1 Ventura drawing (2016)


SURVEY
EXCELLENT
GOOD
AVERAGE
POOR


SURVEY
EXCELLENT
GOOD
AVERAGE
POOR



Groves et al. (2009): Survey Methodology. 2nd Edition, Wiley

Fairness Challenge

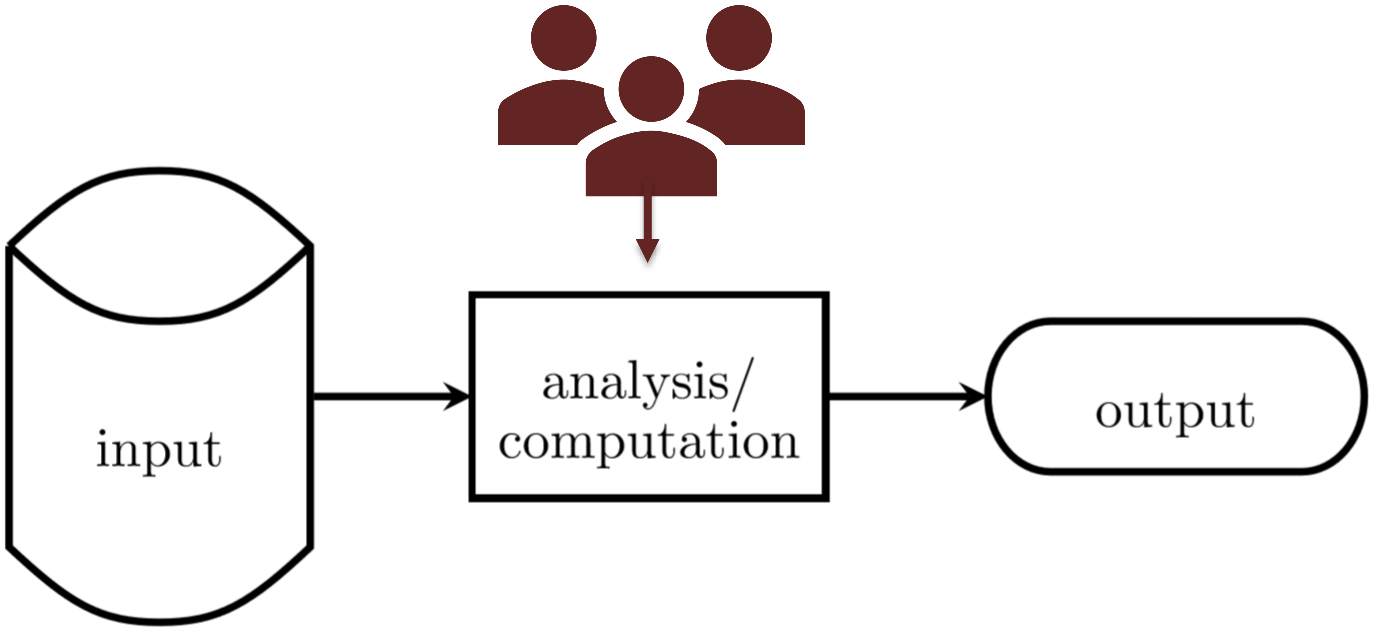
Imperfect training data → many subpopulations

- *Misrepresentation of subpopulations affects accuracy and can have considerable real-world consequences (Buolamwini 2019)*

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%

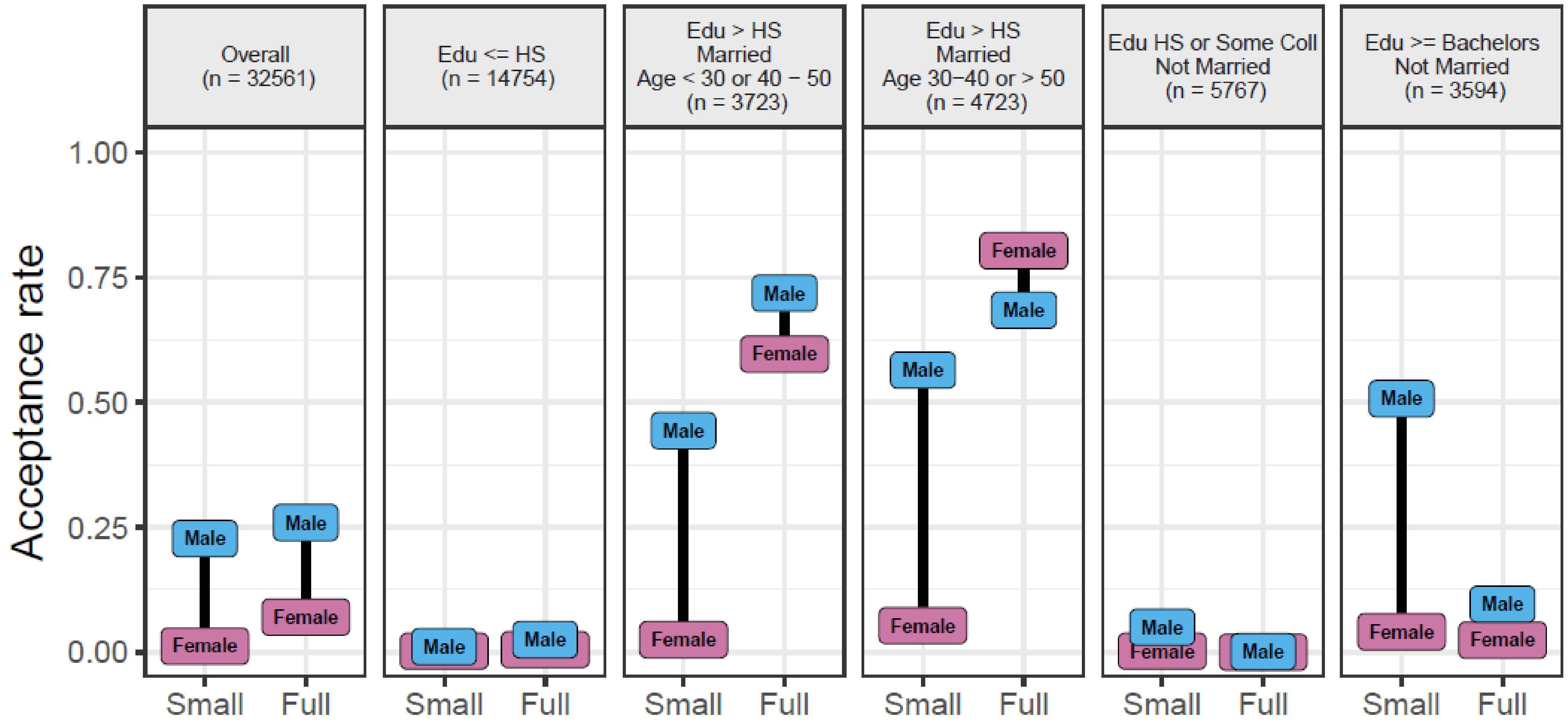
Herausforderung: Entscheidungen in der Analyse

Entscheidungen

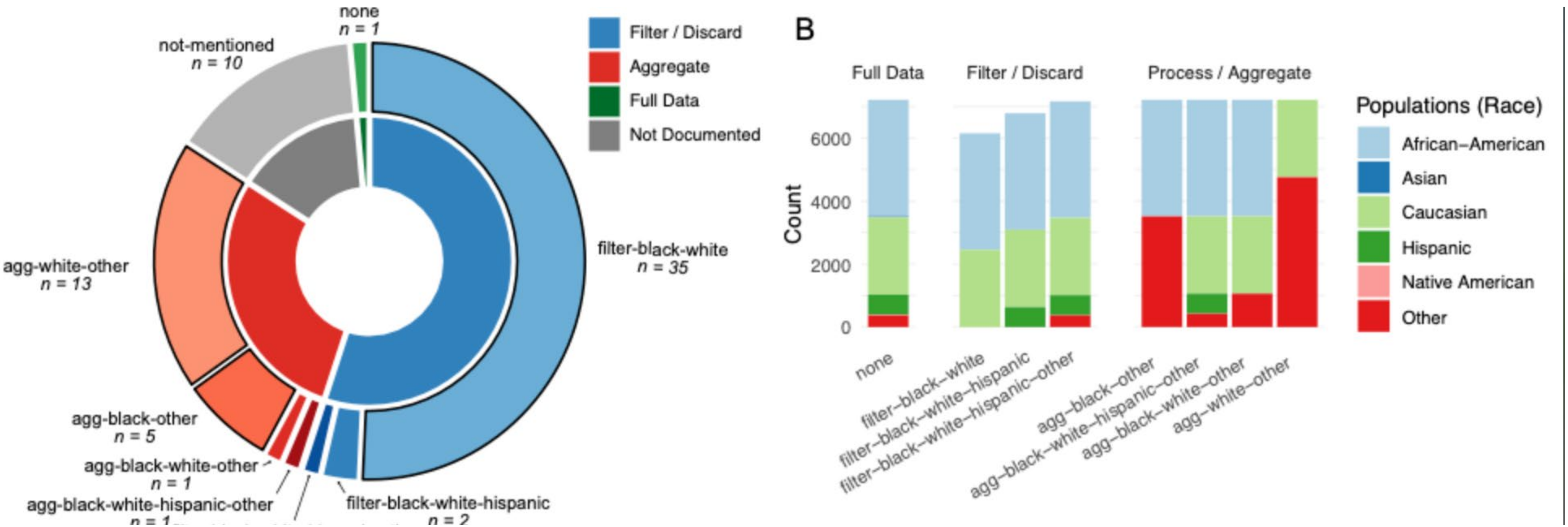


Beispiel: Kreditvergabe

Modelle mit (Full) und ohne (Small) Familienstand und afro-amerikanischer Herkunft



Chouldechova & G'Seel (2017). Fairer and more accurate but for whom? FATML Workshop, 8 2017, Halifax, NS, CA
arXiv:1707.00046 CC-by NC 4.0



none
n = 1
not-mentioned
n = 10
filter-black-white
n = 35
agg-white-other
n = 13
agg-black-other
n = 5
agg-black-white-other
n = 1
agg-black-white-hispanic-other
filter-black-white-hispanic
Filter / Discard
Aggregate
Full Data
Not Documented
B
Full Data
Filter / Discard
Process / Aggregate
Count
0
2000
4000
6000
none
filter−black−white
filter−black−white−hispanic
filter−black−white−hispanic−other
agg−black−other
agg−black−white−hispanic−other
agg−black−white−other
agg−white−other
Populations (Race)
African−American
Asian
Caucasian
Hispanic
Native American
Other

Fairness Definitionen

- Equality of Treatment
  “An algorithm is fair as long as any protected attributes are not explicitly used in the decision-making process”
- Equality of Outcomes
  “People in both protected and unprotected groups should have equal probability of being assigned to a positive outcome”
- Equality of Errors
  “Given a model, evaluation data and groups defined by protected attributes, how can we “define” fairness (more adequately)?
- …… [and many more]

Fairness zwischen Gruppen – welche Metrik?

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$ F_1 score = $\frac{1}{\frac{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}{2}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	

Ghani, 2021, Value of Science Conference

Fairness Metriken sind inkompatibel

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$	F_1 score = $\frac{1}{\frac{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}{2}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR–) $= \frac{FNR}{TNR}$		

$$FPR = \frac{p}{1-p}\left(\frac{FDR}{1-FDR}\right)(1-FNR)$$

False Positive Rate
Among all actual 0's, fraction predicted to be 1

Prevalence
Fraction of actual 1's in population

False Discovery Rate
Among all predicted 1's, fraction that are actual 0's =(1 – precision)

False Negative Rate
Among all actual 1's, fraction predicted to be 0

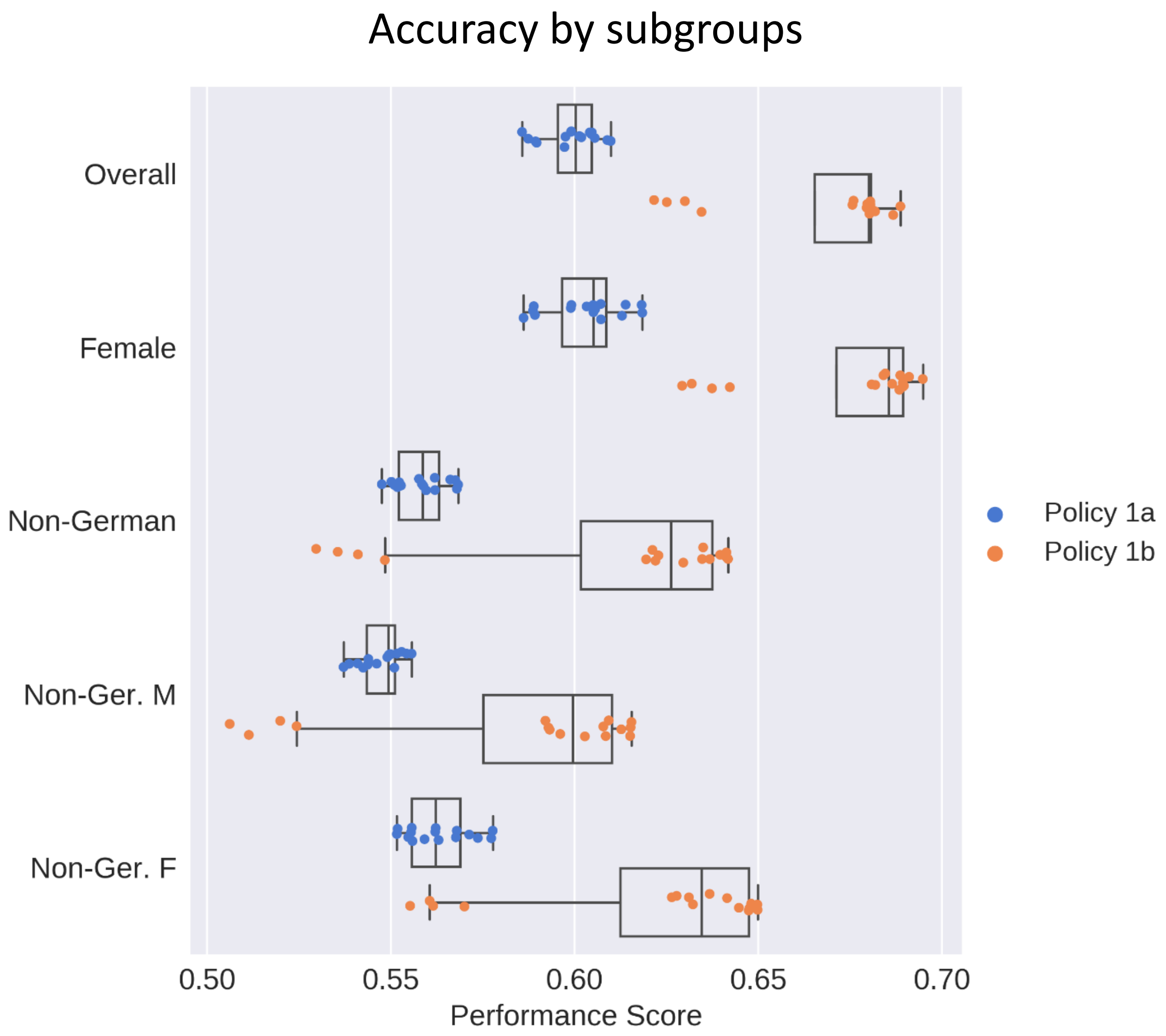
"Of the lessons that can be taken … perhaps the most important for policy is that

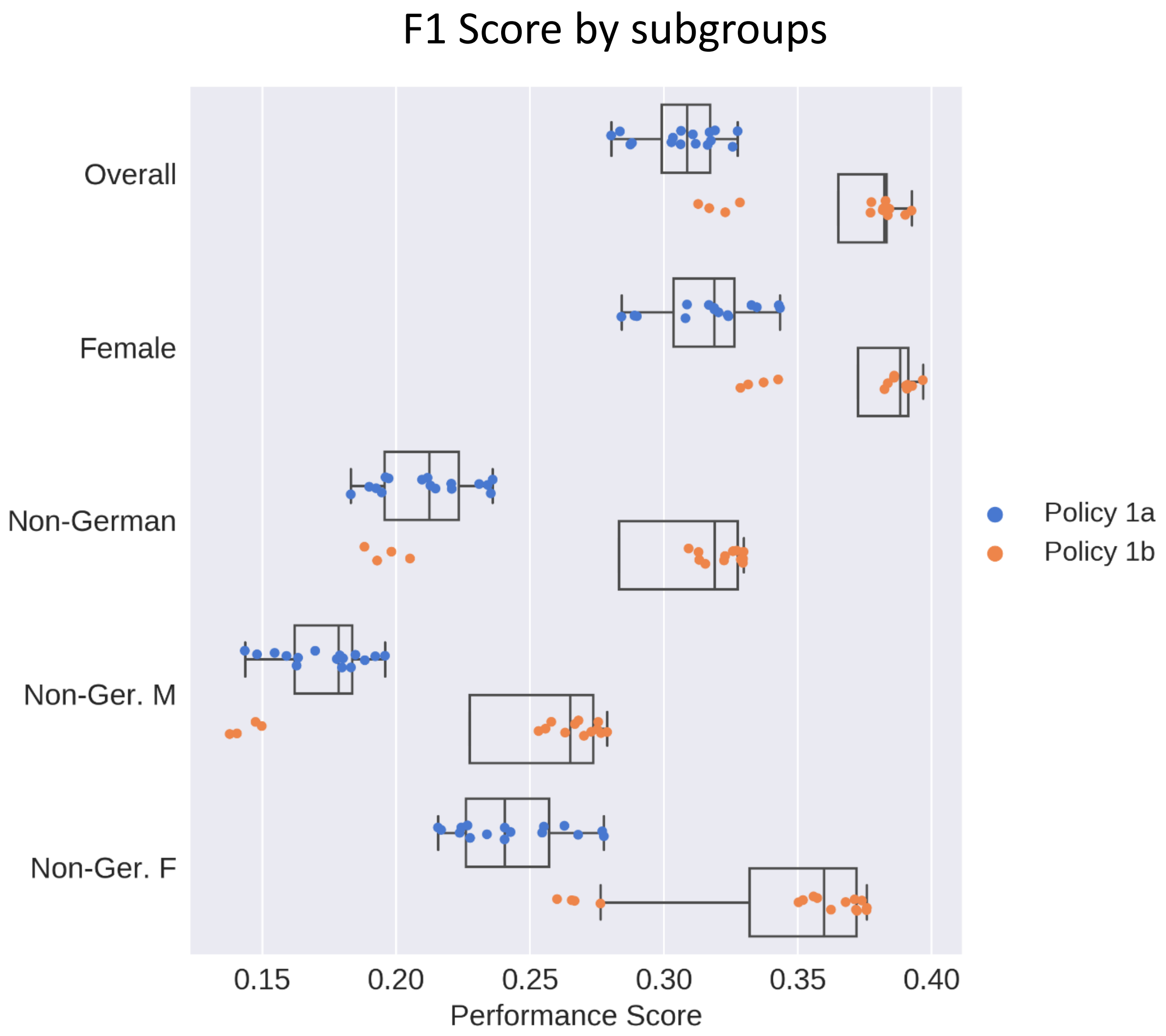
- when there is a lack of separation - and different base rates across group categories,
- a key tradeoff will be between the false positive and false negative rates on one hand and conditional use accuracy equality on the other"

Berk et al. 2017 arXiv:1703.09207: 24

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153-163.

Profiling of Jobseekers

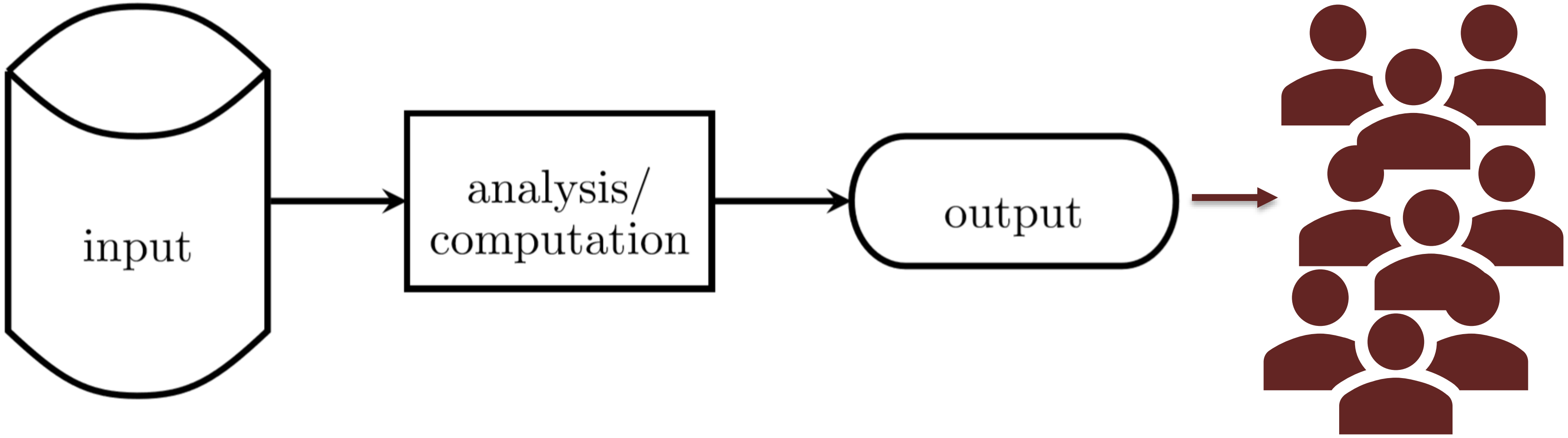




Accuracy = correctly classified observations (positive and negative)

F1 score = balancing precision and recall on the positive class

Entscheidungen



June 12, 2018; July 15, 2019

Google Image Search: “University Professor”

July 19, 2019

Google Image Search: “University Professor”

Eine Behörde hat ein Computerprogramm zur Entscheidung über die vorzeitige Entlassung von Strafgefangenen entwickelt.

Dieses Programm verwendet Daten über den Lebenslauf der Person sowie im Internet verfügbare Informationen über die Person. Das Programm vergleicht diese Informationen mit denen von anderen Personen, die bereits frühzeitig entlassen wurden.

Das Programm schlägt einem Beamten vor, ob die Person vorzeitig aus der Haft entlassen werden sollte. Die endgültige Entscheidung wird vom Richter getroffen.

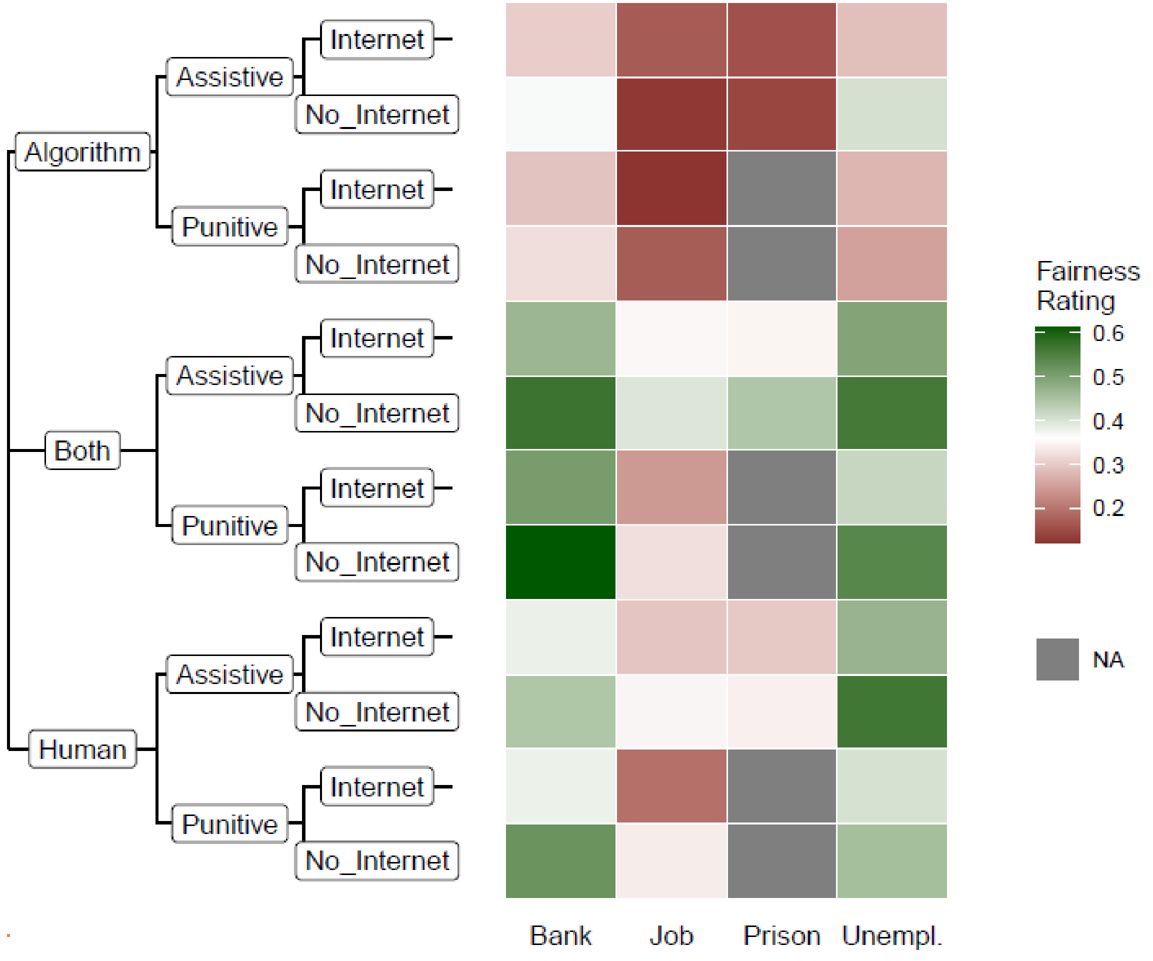
Context matters

- Probability-based longitudinal online survey
- July 2021; n = 4,108 respondents

Vignette Dimensions

1. The **context** in which an ADM system is applied
2. Type of **action** the decision effects
3. Type of **data** used to inform decision
4. The degree of **human involvement** in decision-making

Average fairness ratings of respondents, depending on the experimentally varied vignette dimensions: decision−maker, type of decision, type of data used, and context. Displayed as relative frequencies of respondents that rated a scenario as 'Fair' (i.e., either 'Somewhat fair' or 'Very fair').



Kern, C., Gerdon, F., Bach, R. L., Keusch, F., and Kreuter, F. (2022). Humans vs. machines: Who is perceived to decide fairer? Experimental evidence on attitudes towards automated decision-making. *Patterns*.

Datenschutz und Datennutzung

*The data you **already provided** to us whould be **much more (gain frame) / much less (loss frame)** valuable if you would allow us to link them with .... Do you agree?*

Web	Back	Total
% agree: gain	62.4	520
% agree: loss	75.4	489
Total	498	1009

Phone	Front	Back	Total n
% agree	90.8	78.7	598

Web	Front	Back	Total
% agree	82.6	62.4	520

*The data you are about to provide (front) / already provided (back) to us would be **much more** valuable if you would allow us to link them with .... Do you agree?*

Sakshaug et al. 2019, Public Opinion Quarterly

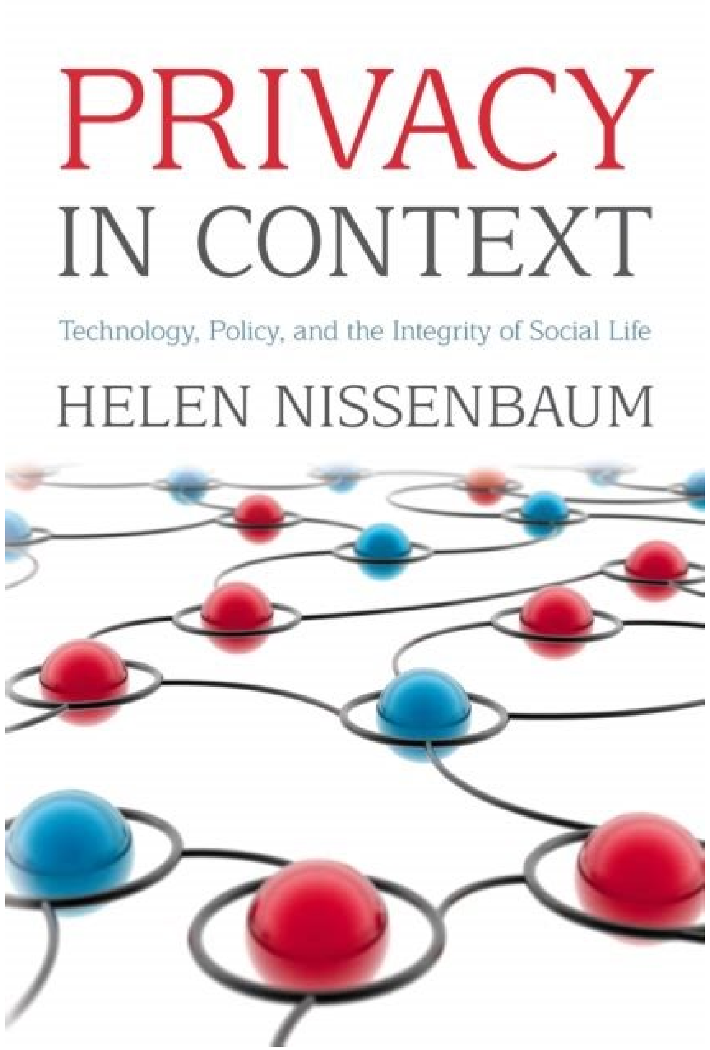
Recipient:
- Public authority
- Company

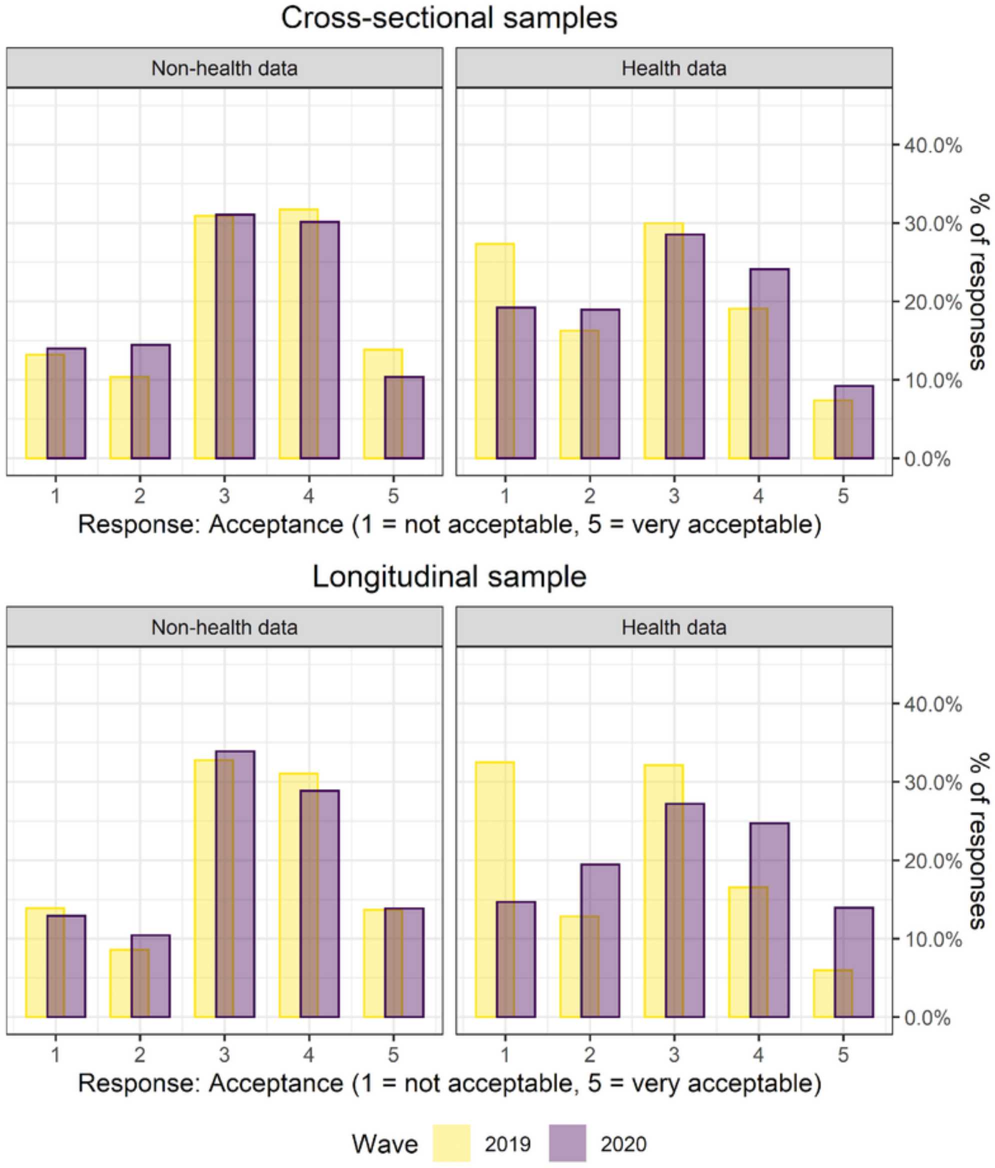
Data type:
- Health: Sensors on a smartphone collect data on the health condition
- Location: Smartphones collect location data during car rides
- Energy: Intelligent power meters collect data on the energy consumption

Sensors installed on a smartphone collect data on the health condition of the holders (e.g., heart rate). With consent of the holder, these data are transmitted to a public authority. The public authority uses these data to detect the spread of infectious diseases in the population early and to develop solutions to their containment. The data are safe, anonymous, and protected from misuse.

Depending on data type:	**Private purpose**	**Public purpose**
Health	… personal recommendations on health behavior	…to detect outbreaks of infectious diseases early and to develop solutions to their containment.
Location	… personal recommendations on driving behavior and routes	…to develop improvements of the local infrastructure
Energy	… personal recommendations on the optimization and reduction of the own energy consumption	… to develop a more efficient energy distribution system

Figure 1. Example vignette as well as dimensions and levels of the other vignettes.
The vignettes varied along the indicated data type, recipient, and data use.

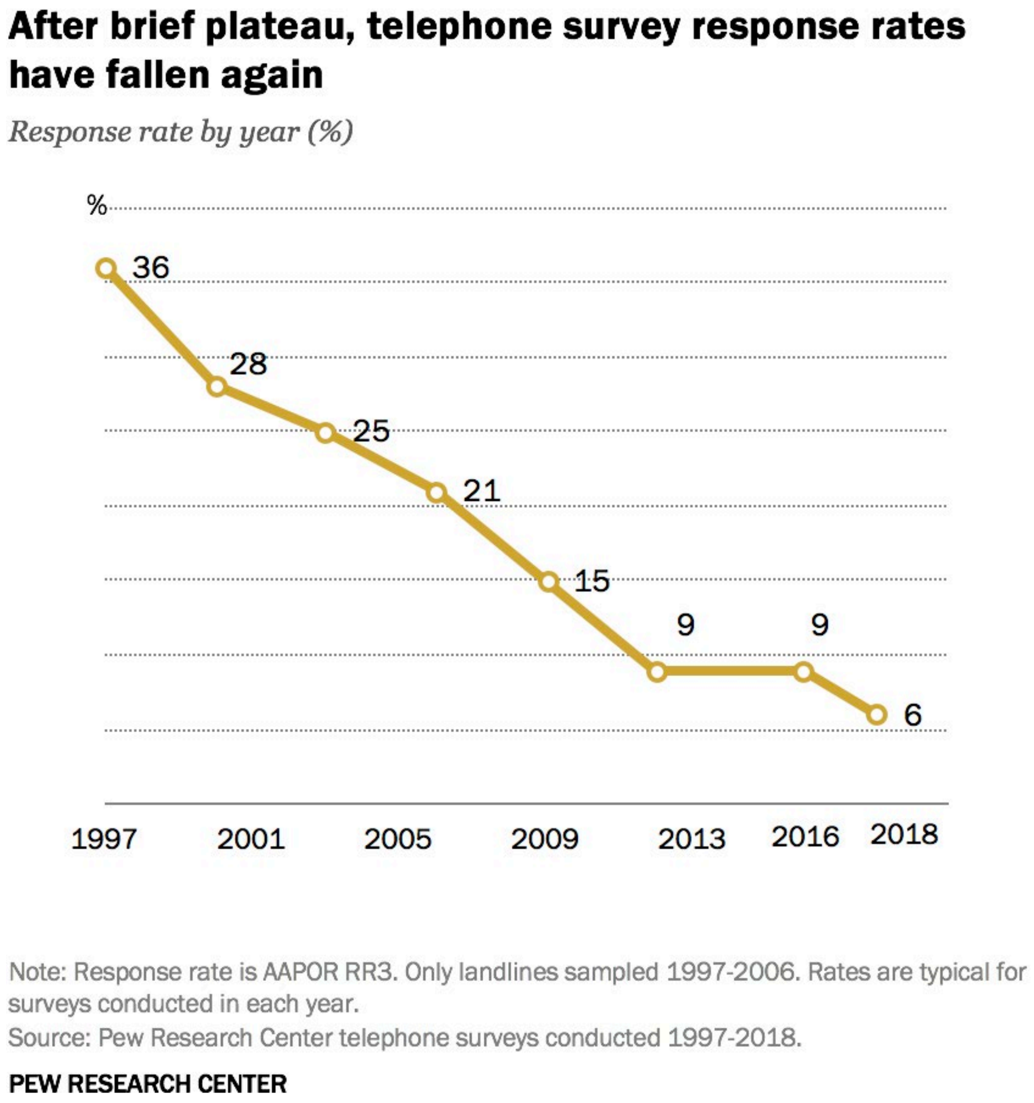


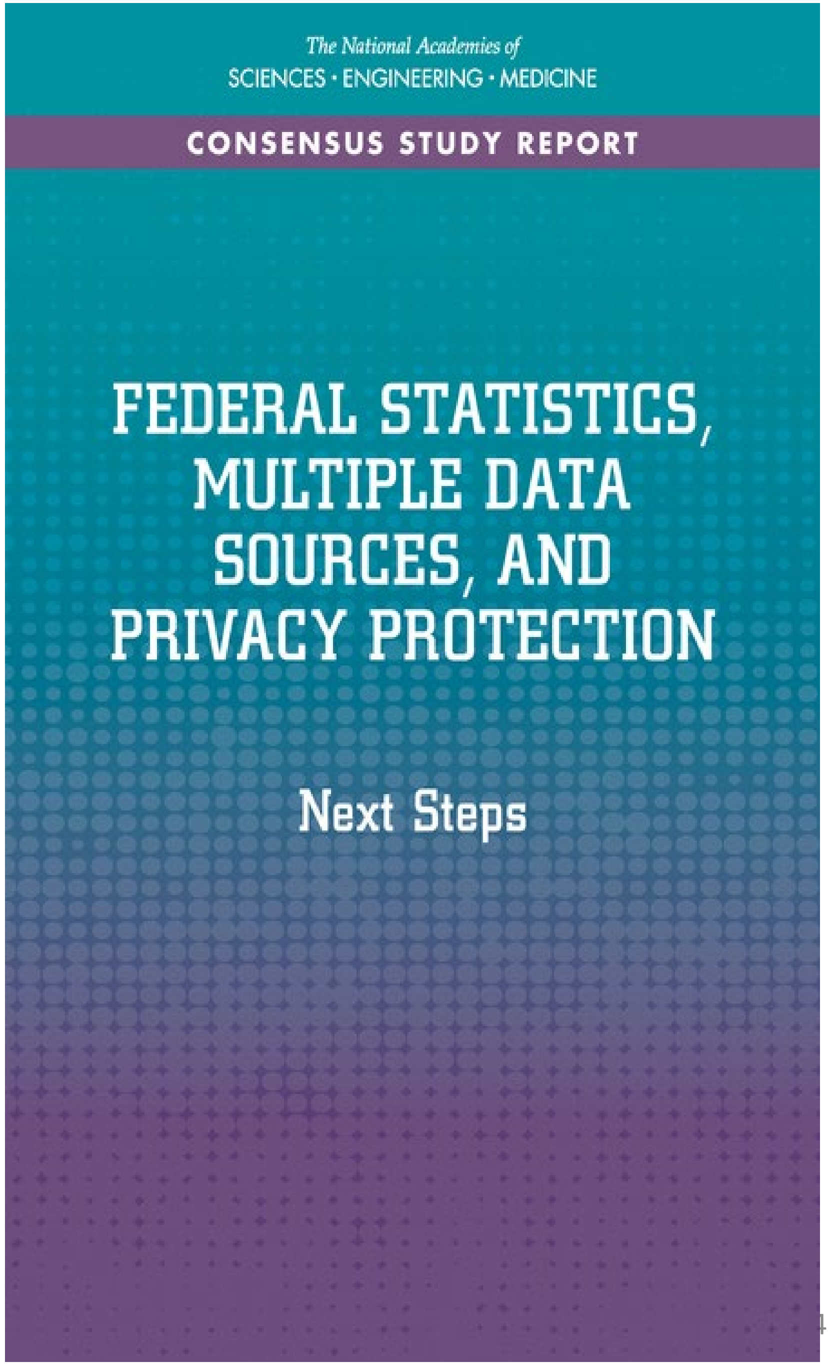


Gerdon, Nissenbaum, Bach, Kreuter & Zins. 2021. Harvard Data Science Review

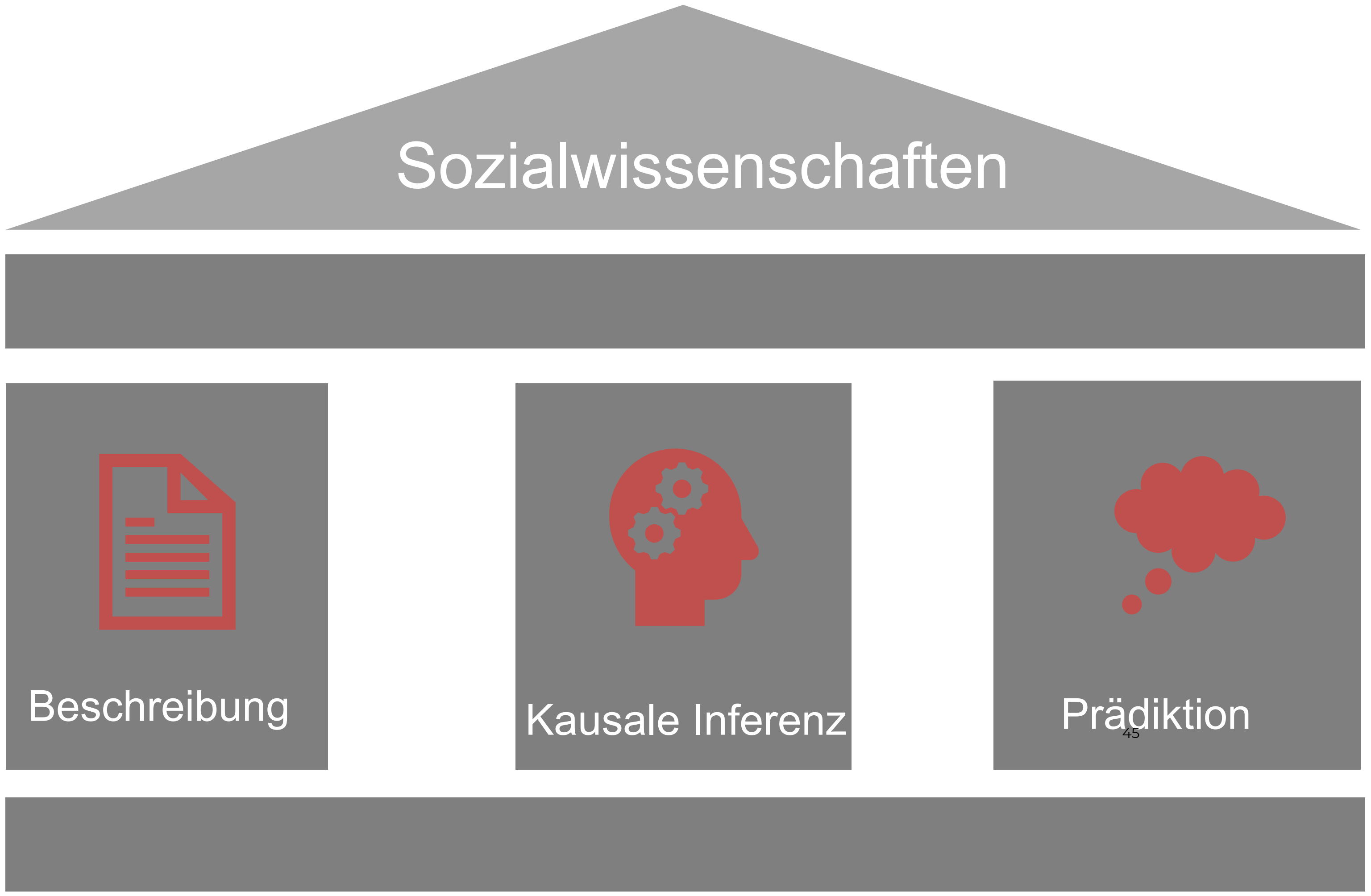
Entwicklung neuer Verfahren

1. Nutzung multipler Quellen





2. Stärkere Betrachtung der Inferenzziele

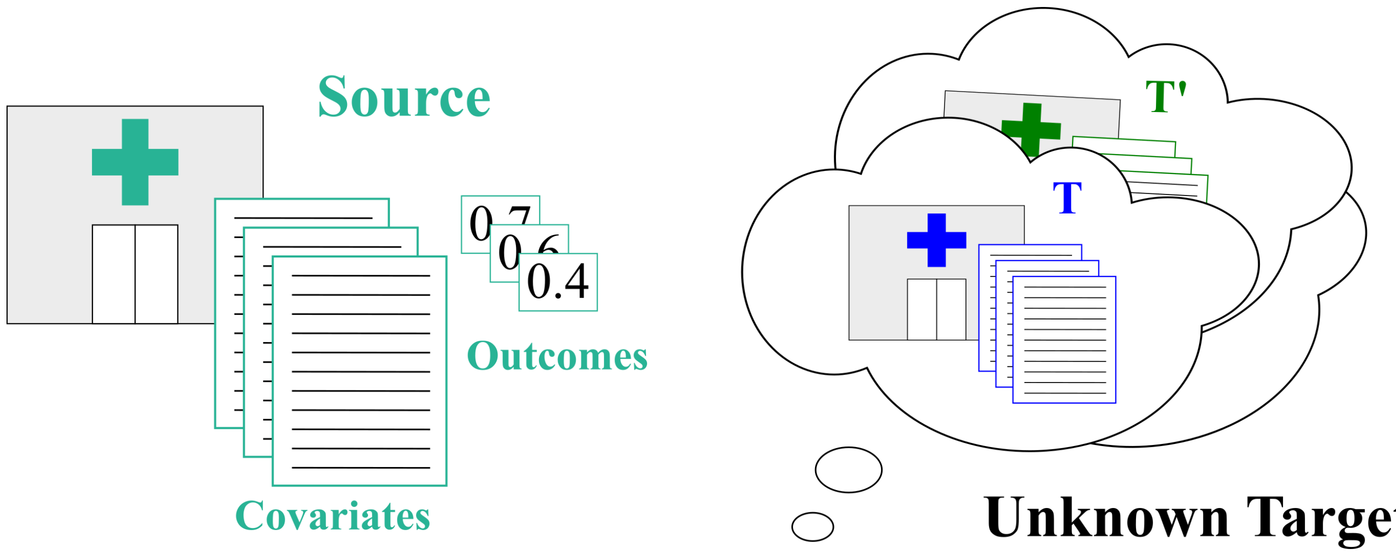


Kohler et al. (2019), Annual Review of Statistics and its Application, 6, 149-172

3. Robuste Inferenz über Populationen hinweg

Eine Datenquelle→ nutzbar für viele Zielpopulationen!

- *s: Beispiel Studie in Berlin Mitte*
- *t: Übertragung der Ergebnisse auf das ländliche Bayern*

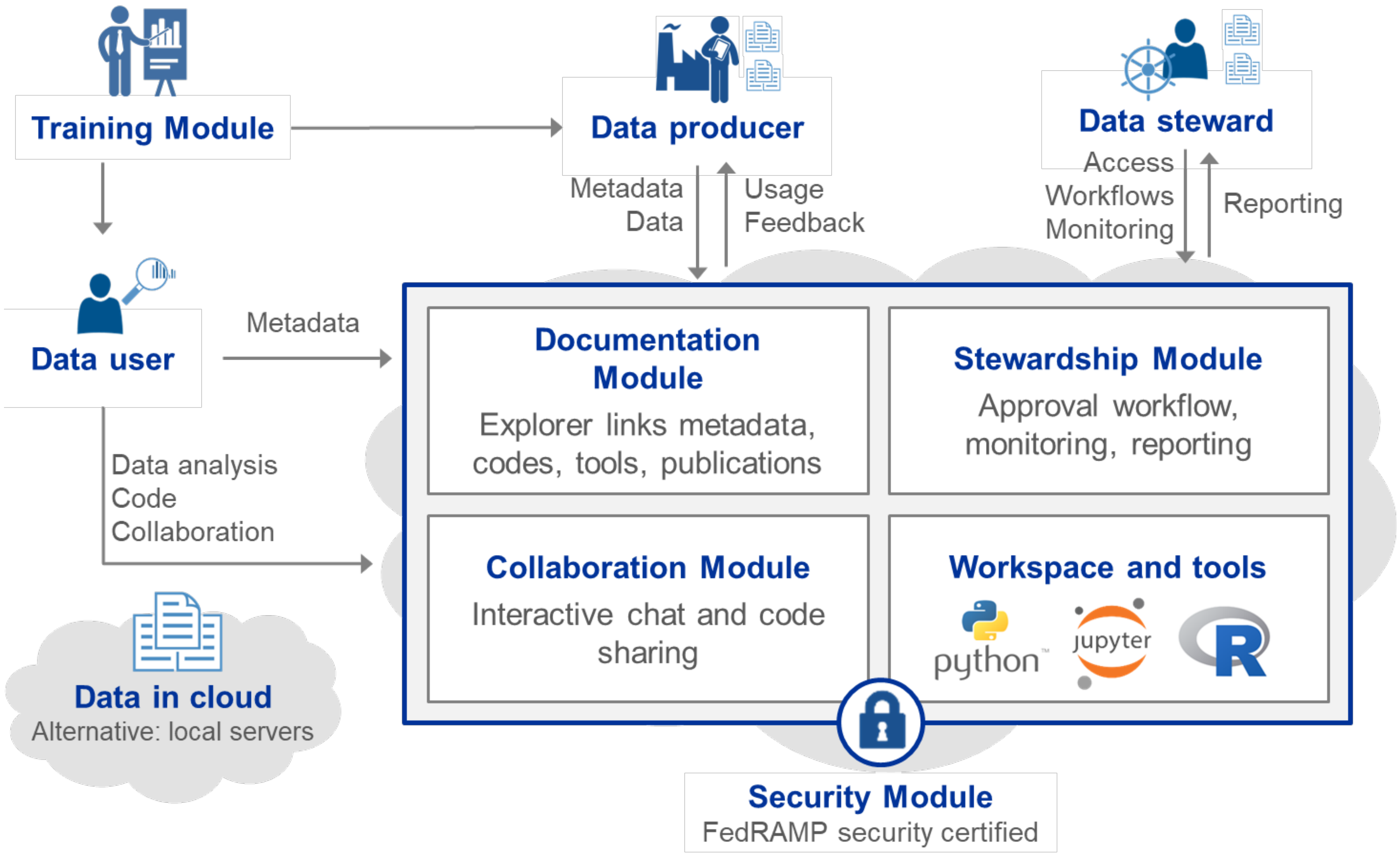


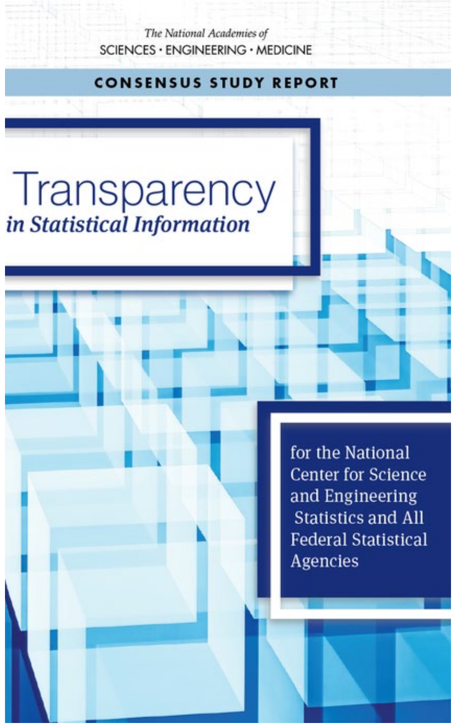
Kim, M. P., Kern, C., Goldwasser, S., Kreuter, F. and Reingold, O. (2022). Universal Adaptability: Target-Independent Inference that Competes with Propensity Scoring. Proceedings of the National Academy of Sciences of the United States of America (PNAS) 119(4). doi:10.1073/pnas.2108097119

(Forschungs)-dateninfrastrukturen


COLERIDGE INITIATIVE
Home
Training
ADRF
Projects & Research
Collaborators
Events
Democratizing our Data: A Challenge to Invest in Data and Evidence-based Policy
Learn More
WHO WE ARE

Context matters for AI model training and sharing





Kreuter, F., Ghani, R., & Lane, J. (2019). Change Through Data: A Data Analytics Training Program for Government Employees. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.ed353ae3

https://coleridgeinitiative.org/

Angewandte Datenanalyse für die öffentliche Verwaltung

ADA Bayern

In der öffentlichen Verwaltung liegen **vielfältige Daten** vor. Daneben gibt es eine Vielzahl **öffentlich nutzbarer externer Datenquellen**. Zusätzlich ist **offene und frei verfügbare Software** (zum Beispiel R)—vom Daten sammeln, für statistische Analysen, Machine Learning Modelle, Anwendungen von

https://ada-oeffentliche-verwaltung.de/

1. Entscheidungen über Fairness sind unabhängig von KI

2. Gute Daten sind essentiell für gute KI

3. Contextual integrity hilfe als Framework für Datenschutz

4. Letztlich muss das Vertrauen (in Fairness) gestärkt werden.

Wichtig sind dabei geteilte Datenräume und Transparenz:

Safe projects	Safe people	Safe settings	Safe data	Safe exports

Frauke Kreuter
frauke.kreuter@lmu.de · THANK YOU