

Tag 1: Der Daten-Lifecycle und: Woher kommen Daten?

Session 3: Datengenerierung und was dabei schief gehen kann

Simon Munzert
Hertie School

1. Woher kommen die Daten?
2. Was bei der Datengenerierung schiefgehen kann
3. Das Problem mit der Repräsentativität
4. Problematische Messung

Woher kommen die Daten?

AI-Produkte und ihre Datenbasis

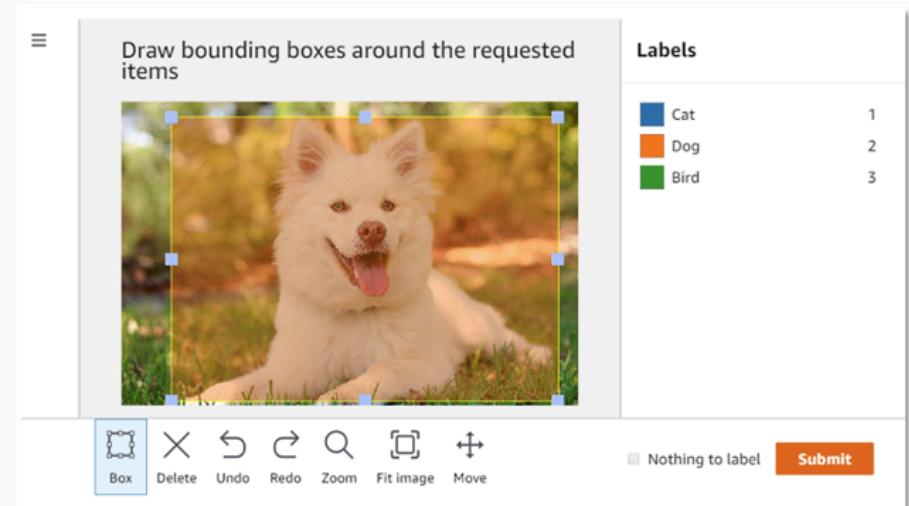
AI-Tool/Anwendung	Funktion	(vermutete) Datenquellen
ChatGPT (OpenAI)	Sprachmodell für Textgenerierung, Dialoge, Programmierung	???
Stable Diffusion	Open-Source-Bildgenerator	???
Alexa (Amazon)	Sprachbasierter Assistent für Smart Home	???
Perspective API	Automatisierte Erkennung und Blockierung unerwünschter Inhalte	???
Patternizr	Vorhersage von Straftaten, Einsatzplanung der Polizei	???

AI-Produkte und ihre Datenbasis

AI-Tool/Anwendung	Funktion	(vermutete) Datenquellen
ChatGPT (OpenAI)	Sprachmodell für Textgenerierung, Dialoge, Programmierung	Web-Texte, Bücher, Wikipedia, Foren, Code-Repositories (z. B. GitHub)
Stable Diffusion	Open-Source-Bildgenerator	Bilder aus dem LAION-5B-Datensatz (Internetbilder mit Textbeschreibungen)
Alexa (Amazon)	Sprachbasierter Assistent für Smart Home	Sprachaufnahmen, Amazon-Konten, Smart-Home-Geräte-Daten
Perspective API	Automatisierte Erkennung und Blockierung unerwünschter Inhalte	Online-Kommentare (z. B. aus Foren, News-Seiten), moderierte Textkorpora
Patternizr	Vorhersage von Straftaten, Einsatzplanung der Polizei	Polizeiliche Falldatenbanken, Tatortdaten, Kriminalstatistiken

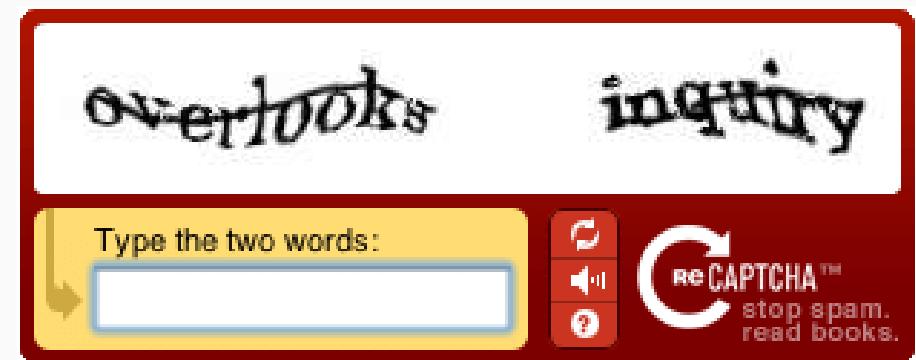
Menschlicher Input für KI-Modelle

- **Menschliche Expertise** ist oft unverzichtbar, um Daten zu generieren, zu annotieren und zu validieren.
- **Menschliche Expertise** ist auch unverzichtbar, um die Qualität von Daten zu beurteilen und zu verbessern.



Beispiele

- **Crowdsourcing:** z. B. Amazon Mechanical Turk
- **Content Moderation:** z. B. Facebook, Twitter
- **CAPTCHA** ("Completely Automated Public Turing test to tell Computers and Humans Apart"): z. B. Google reCAPTCHA



Clickworking-Plattformen

- Plattformen designt für Mikrojobs, z. B. Texterstellung, Datenerfassung, Kategorisierung
- Beispiele: Amazon Mechanical Turk, Clickworker, Upwork, Prolific
- **Clickworker**: ca. 6 Mio. registrierte Clickworker weltweit
- **Amazon Mechanical Turk**: 100Tsde. registrierte Worker, nur ein Bruchteil aktiv



Ethische Fragen

- **Fairness**: Bezahlung, Arbeitsbedingungen
- **Ausbeutung benachteiligter Gruppen**: z. B. Clickworker in Entwicklungsländern
- **Datenschutz**: Schutz sensibler Daten
- **Transparenz**: Offenlegung von Auftraggebern, Zweck der Datenerhebung

Tagging Instructions (Click to expand)

Highlight the **name** in the description

An issue was discovered in the base64d function in the SMTP listener in Exim before 4.90.1 . By sending a handcrafted message , a buffer overflow may happen . This can be used to execute code remotely .

Undo Reset

N(e)ame
V(e)rsion
P(r)otocol

Product name
 There is no name

Product version
 There is no version

Protocol
 There is no protocol

Submit

Übung

Wie sieht das Leben eines Clickworkers aus?

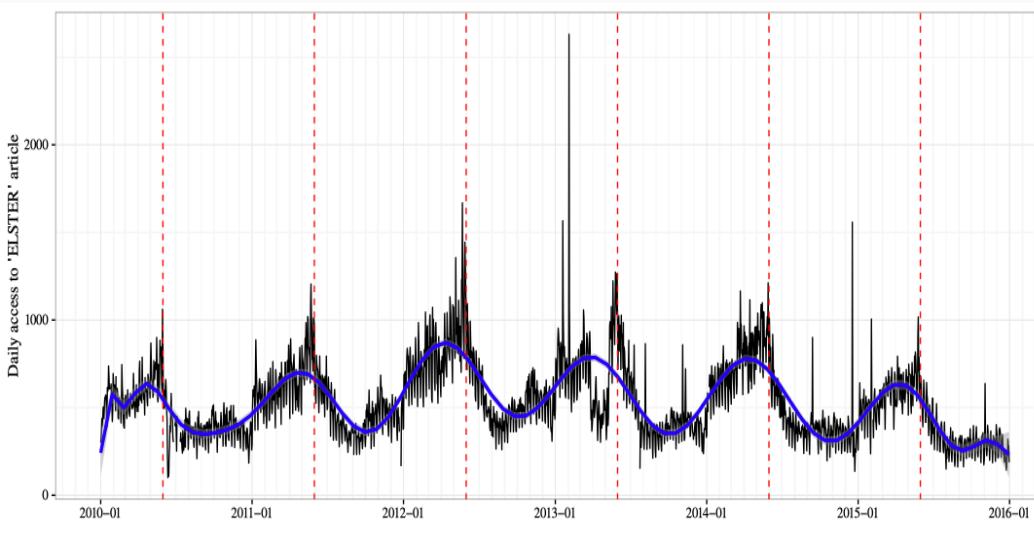
Zum Ausprobieren:

Amazon Mechanical Turk (Sandbox; benötigt Amazon-Account)

Prozessproduzierte vs. aktiv generierte Daten

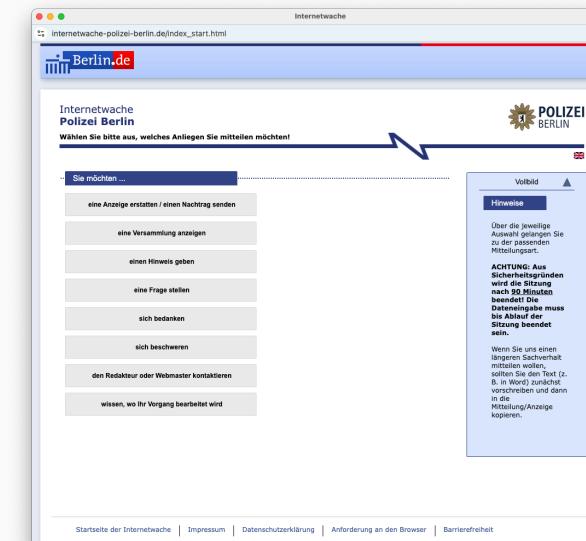
Prozessproduzierte Daten

- Daten, die als Nebenprodukt von Prozessen generiert werden
- Beispiele: Transaktionsdaten (z. B. E-Commerce), Log-Daten (z. B. Website-Clicks), Emails, Verwaltungsakten
- Ein Großteil der Daten, die in KI-Anwendungen verwendet werden, sind prozessproduzierte Daten



Aktiv generierte Daten

- Daten, zu einem bestimmten Zweck erhoben
- Beispiele: Umfragen, Experimente, Beobachtungen (z. B. manuelle Zählungen), Formulare
- Aktiv generierte Daten sind oft teurer und aufwendiger zu erheben, aber passgenauer für bestimmte Fragestellungen



Prozessproduzierte vs. aktiv generierte Daten

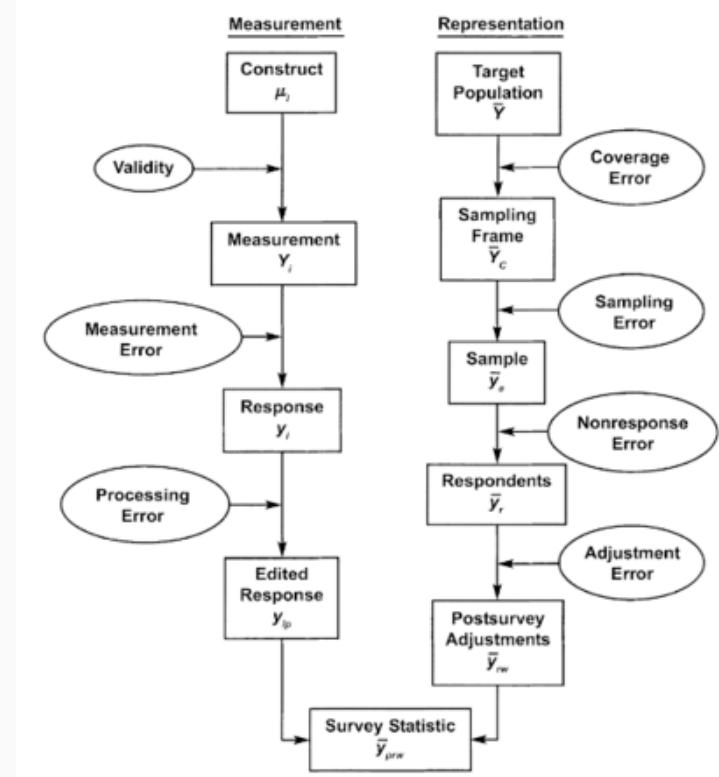
Datenart	Vorteile	Nachteile
Prozessproduzierte Daten	<ul style="list-style-type: none">• Kontinuierlich und in großem Umfang verfügbar• Realistische, nicht-beeinflusste Daten	<ul style="list-style-type: none">• Daten oft unstrukturiert oder unvollständig• Keine Kontrolle über Erhebungsmethode
Aktiv erhobene Daten	<ul style="list-style-type: none">• Präzise Daten, speziell für den Anwendungsfall• Gute Dokumentation der Erhebungsmethoden	<ul style="list-style-type: none">• Aufwand und Kosten der Datenerhebung hoch• Risiko von Verzerrungen durch Antwortverhalten

Was bei der Datengenerierung und Verarbeitung schiefgehen kann

Zwei grundsätzliche Fehlerquellen bei der Datensammlung

- **Messfehler:** was man misst, ist nicht das, was man messen will
- Fehler der **Repräsentation:** Die Gruppe, die Sie beobachten, ist nicht verallgemeinerbar auf die interessierende Population

Total survey error framework



Quelle Groves et al. 2009, Survey Methodology

Überrepräsentation und falsche Angaben in Wahlumfragen

- Umfragestatistiken überschätzen die Wahlbeteiligung oft erheblich.
- Zwei unterschiedliche Phänomene sind für diese Diskrepanz verantwortlich:
 1. Überrepräsentation der tatsächlichen Wähler
 2. Falsche Angaben zur Wahlbeteiligung durch Nichtwähler unter den Umfrageteilnehmern.
- Studien zur Validierung der Wahlbeteiligung helfen, das Problem auf individueller Ebene zu identifizieren.
- Eine Verzerrung der Wahlbeteiligung kann sich auch auf Analysen nachgelagerter Variablen (z.B. Wahlverhalten) auswirken.

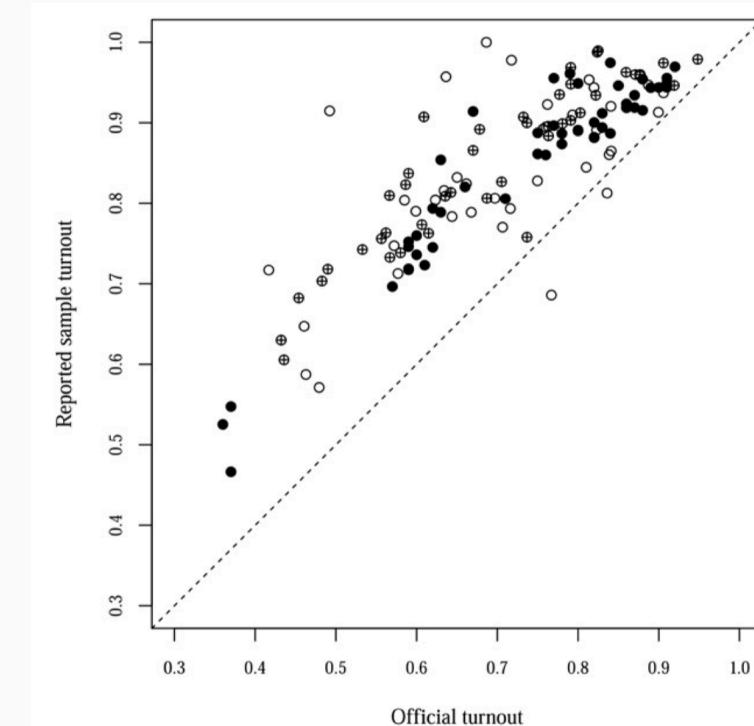


Fig. 1. Reported sample turnout rates from 130 postelection surveys versus official turnout. Data are taken from Modules 1–3 of the *Comparative Study of Electoral Systems* (CSES), and from a collection of election surveys for which vote validation studies (VVS) are available. For more detailed information, see the [Appendix](#) to this paper.

Quelle [Selb and Munzert 2013, Electoral Studies](#)

Beispiele für Datengenerierungs-Fails

Kategorie	Probleme	Beispiele
Unvollständige Daten	<ul style="list-style-type: none">• Fehlende Datenpunkte durch technische Ausfälle• Daten werden nur für bestimmte Zeiträume erhoben	<ul style="list-style-type: none">• Sensoren fallen zeitweise aus, sodass Verkehrsdaten für einige Tage fehlen• Wichtige Umfragedaten werden nicht erhoben, weil bestimmte Fragen übersprungen wurden
Selection Bias (Verzerrte Stichproben)	<ul style="list-style-type: none">• Stichprobe nicht repräsentativ für die Zielpopulation• Ungeeignete Auswahlkriterien für Datenquellen	<ul style="list-style-type: none">• Online-Umfragen schließen ältere Menschen ohne Internetzugang aus• Ein KI-Modell wird nur mit Daten von Großstädten trainiert, was ländliche Regionen vernachlässigt
Fehlerhafte Messung	<ul style="list-style-type: none">• Messgeräte liefern ungenaue Daten• Menschliche Fehler bei manueller Datenerfassung	<ul style="list-style-type: none">• GPS-Tracking liefert falsche Standortdaten in Gebieten mit schlechtem Empfang• Manuelle Zählungen von Besuchern eines Events führen zu Doppelzählungen

Beispiele für Datengenerierungs-Fails

Kategorie	Probleme	Beispiele
Probleme beim Datenimport	<ul style="list-style-type: none">• Inkompatible Formate erschweren den Import• Fehler bei der Zuordnung von Datenfeldern während des Imports	<ul style="list-style-type: none">• Daten aus verschiedenen Quellen sind unterschiedlich formatiert (z. B. unterschiedliche Datumsformate)• Digitalisierung von Papierakten erfolgt qualitativ unzureichend über OCR (auch Messproblem)
Ethische und rechtliche Probleme	<ul style="list-style-type: none">• Daten werden ohne Einwilligung erhoben• Sensible Daten werden ohne angemessene Schutzmaßnahmen erfasst	<ul style="list-style-type: none">• Datensammlung in einer Smart City ohne Zustimmung der Bürger• Gesundheitsdaten werden ohne ausreichenden Schutz gesammelt, was zu Datenschutzverletzungen führen kann

Das Problem mit der Repräsentativität

Zweifelhafte „Repräsentativität“



Donald J. Trump ✅

@realDonaldTrump · 1d



Quelle citechdaily.com

Aus welchen Medien beziehen Sie Ihre Information über das politische Geschehen? (1/2)

Glaubwürdigkeit der Medien | Mai 2025 | Angaben in Prozent | Veränderungen zu November 2023 | * kein Vergleichswert



Quelle: infratest dimap



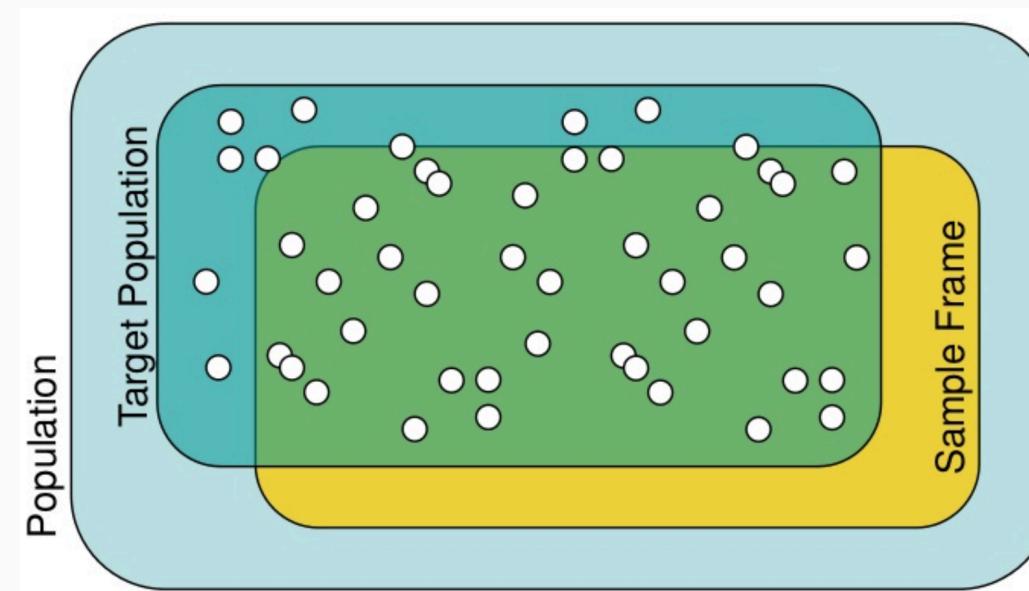
Quelle [tagesschau.de/infratest dimap](https://tagesschau.de/infratest-dimap)

Eine substantielle Definition von Repräsentativität

Eine Stichprobe (oder Daten im Allgemeinen) ist „repräsentativ“, wenn die aus der Stichprobe gezogenen Schlüsse verallgemeinert werden können auf die Grundgesamtheit von Interesse.

Eine formalere Definition

Eine Stichprobe ist repräsentativ, wenn sie so gezogen wird, dass sie statistisch nicht von der interessierenden Grundgesamtheit unterscheidbar ist.



Warum „Repräsentativität“ ein problematischer Begriff ist

1. Ob eine Stichprobe repräsentativ ist, hängt von Ihrem Interesse ab.
2. Man kann eine Stichprobe nicht a priori als „repräsentativ“ bezeichnen.
3. Die Beurteilung der Repräsentativität einer Stichprobe erfordert starke Annahmen über Ihr Wissen über die Grundgesamtheit und Ihre Messungen der Merkmale, die „repräsentativ“ sein sollten.

Ein Beispiel

- Sie führen eine Umfrage zur Wahlabsicht durch.
- Wen möchten Sie repräsentieren? Die Wahlbevölkerung oder die Gesamtbevölkerung?
- Wie erreichen Sie eine repräsentative Stichprobe? Zufallsauswahl? Gewichtung?



Was bedeutet das für Sie?

- Nehmen Sie die angegebene „Repräsentativität“ nicht für bare Münze.
- Der Stichprobenumfang allein garantiert keine Repräsentativität (siehe Big Data!).
- Lassen Sie sich nicht von „großen Datenmengen“ täuschen (nicht per se repräsentativ).
- Lassen Sie sich nicht von „Zufallsstichproben“ täuschen (nicht per se repräsentativ).
- Schlechte Stichproben sind nicht auf Erhebungen beschränkt (denken Sie z.B. an Daten aus sozialen Medien, die Auswahl von Fällen für eine medizinische Studie oder die Auswahl von Ländern für eine politische Studie).

Achten Sie stattdessen auf:

1. **Transparenz** über das Auswahlverfahren.
2. **Validierung** der Stichprobe anhand externer Benchmarks.
3. **Ihren gesunden Menschenverstand:** Ist die Datengrundlage systematisch verzerrt, d.h. auf relevanten Kriterien unterschiedlich von der Grundgesamtheit?

Übung: Repräsentativität

Diskussion in 2er-Gruppen, jeweils 2-3 Minuten

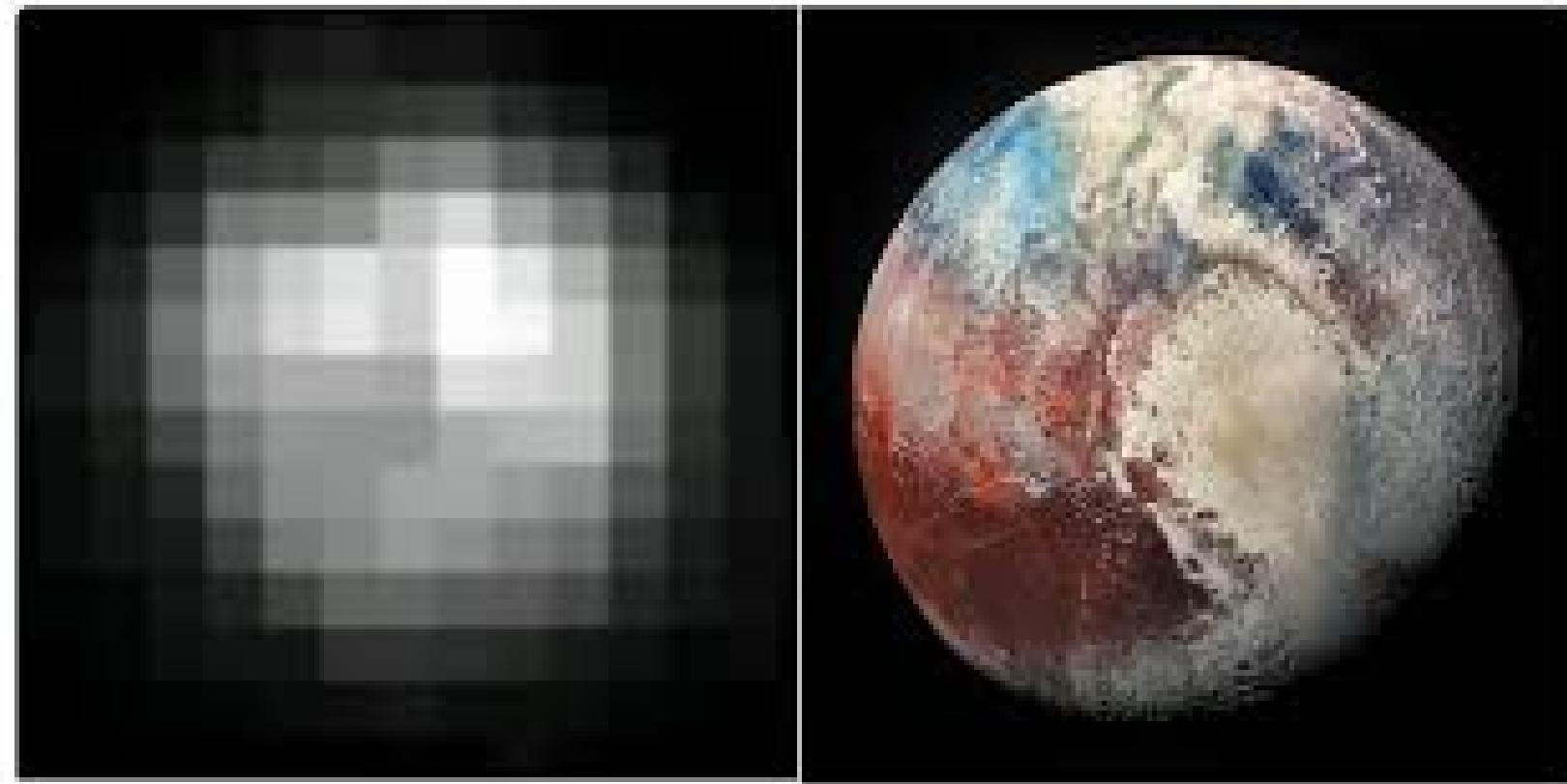
Finden Sie sich als Paar zusammen und erklären sie in 1-2 Minuten jeweils eine der Fragen.

Partner 1: Was bedeutet "repräsentative Daten" oder "repräsentative Stichprobe"?

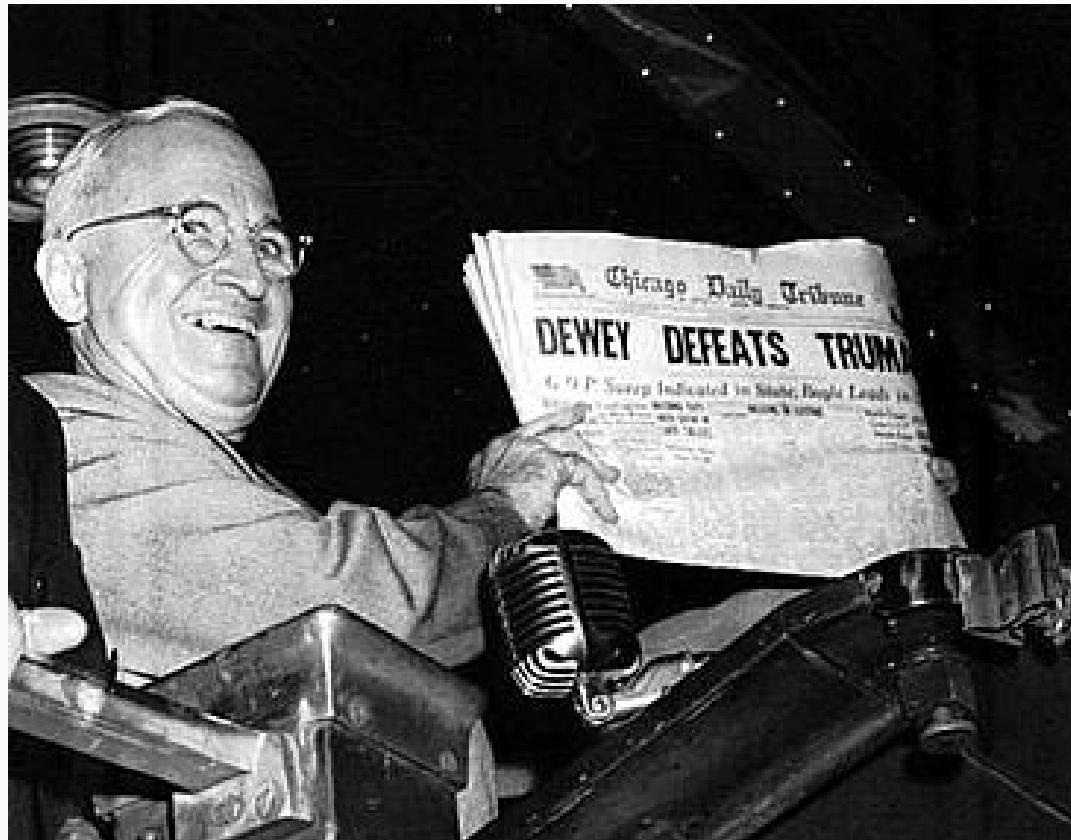
Partner 2: Warum ist "repräsentative Stichprobe" meistens Quatsch?

Problematische Messung

1994 (Hubble) vs. 2018 (New Horizons)



US-Präsidentswahlen 1948 (Dewey) vs. 2016 (Clinton)



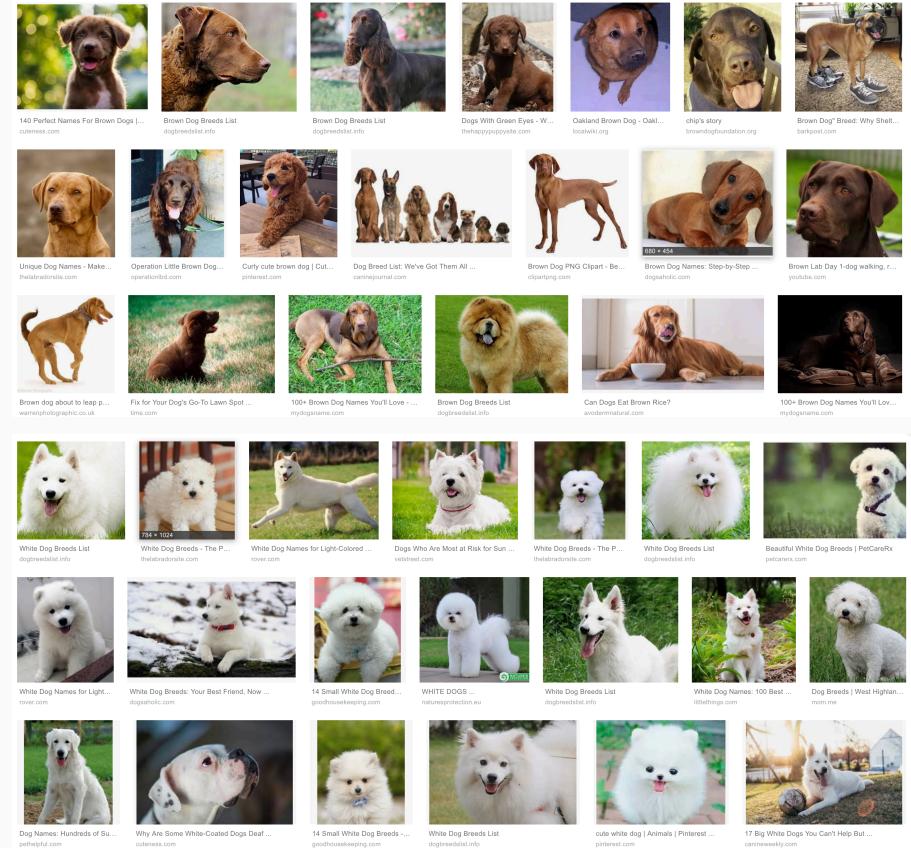
Messung ist ein zentraler Bestandteil von AI-Anwendungen

- AI-Modelle lernen von Daten, die gemessen werden.
- Gleichzeitig können AI-Modelle auch messen (z. B. Gesichtserkennung, Spracherkennung).

💡 Messung ist gleichzeitig Grundlage (Input) für als auch Output von AI-Anwendungen.

Messfehler in AI-Anwendungen

- Messfehler können systematisch sein (z. B. durch soziale Erwünschtheit).
- Messfehler können zufällig sein (z. B. durch technische Probleme).
- Messfehler können auch durch die AI-Anwendung selbst entstehen (z. B. durch schlechte Datenqualität).



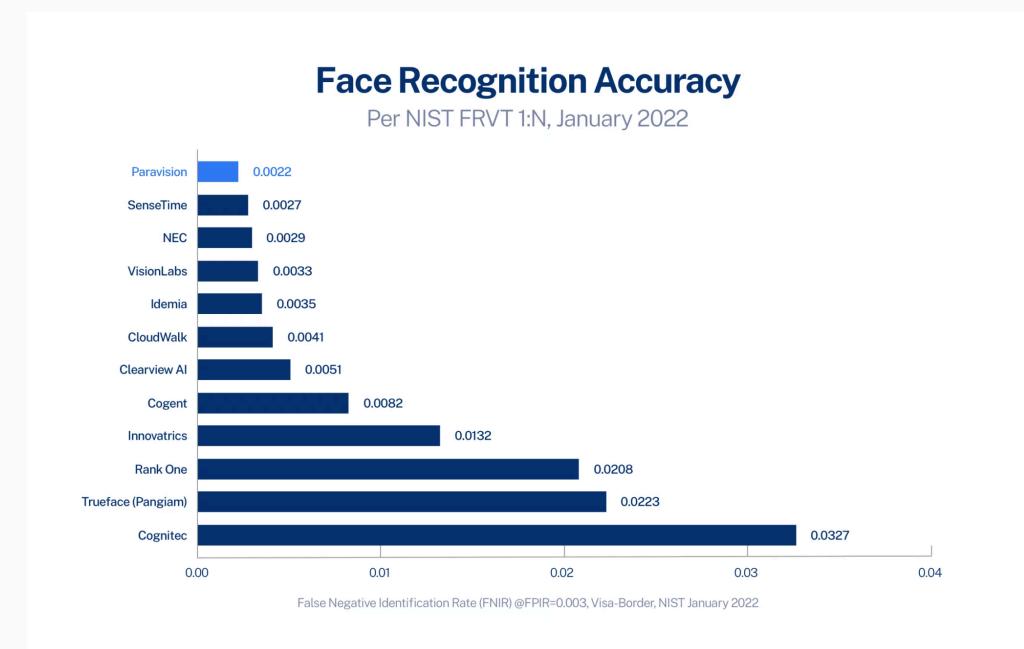
Gut funktionierende Messung: Gesichtserkennung

Messung von Gesichtsmerkmalen

- Gesichtsmerkmale sind oft gut messbar.
- Gesichtsmerkmale sind oft stabil über die Zeit.
- Massive Trainingsdaten für Gesichtserkennung verfügbar (Social Media, Fotoportale, etc.)

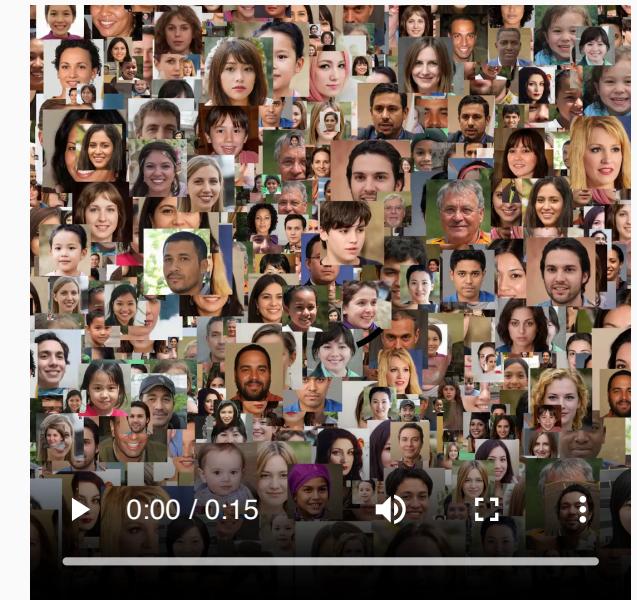
Beispiel: Paravision AI

"In this Visa / Border test, Paravision ranks #1 globally for accuracy with an error rate that is 18% lower than the second-ranked vendor, delivering a False Negative Identification Rate of 0.22% at a False Positive Identification Rate of 0.3% across a dataset of 1,600,000 images." **Paravision AI**



- Tool für Gesichtserkennung beliebiger Personen anhand von Trainingsdaten aus Social Media, Zeitungen, Online-Plattformen, PayPal Transaktionen, etc.
- **Behauptung:** 99,85% Accuracy bei Stichprobe von 12 Millionen Fahndungsfotos ([s. Homepage](#))
- **Belege:** Gutes Testergebnis beim US Nationalen Institut für Standards und Technologie
- **Anfechtbarkeit:** Tool ist weder offen zugänglich, noch können Nutzer/innen ihre Fotos entfernen.
- **Training der Modelle:** Web Scraping von Online-Plattformen, Webseiten und Medien ohne Einverständnis der Nutzer/innen
- Hintergrund: NYT-Recherche: Staatliche Behörden benutzen Clearview.ai auch während Gerichtsverfahren weiter ([Hill, 2020](#))

Quelle [Kashmir Hill, 2024](#)



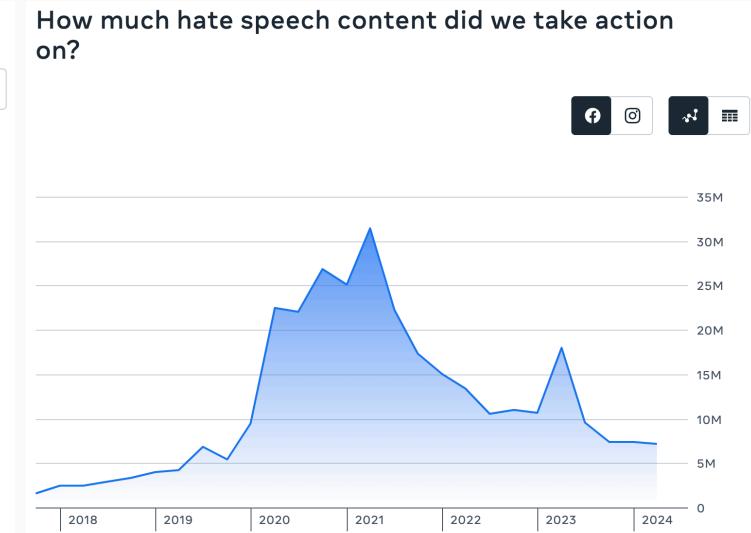
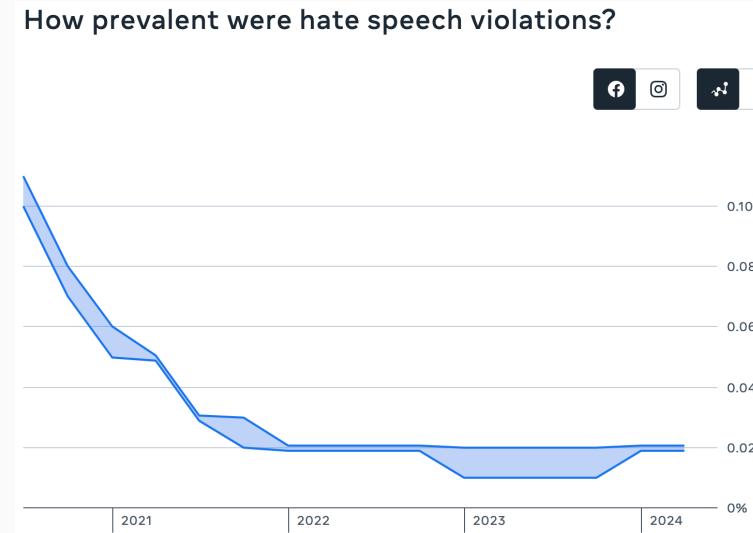
Quelle [New York Times](#)

Nicht so gut funktionierende Messung: Hate Speech

Das Problem

- **Hate Speech** kann sich in verschiedenen Formen äußern.
- **Hate Speech** ist schwer zu messen, weil es oft subtil und kontextabhängig ist.
- Es herrscht fundamentale Uneinigkeit darüber, was genau **Hate Speech** ist und wie es gemessen werden sollte.

Beispiel: Facebook



Quelle [Facebook](#)

Menschliche Hate-Speech-Ratings

Message	Hatefulness				Distribution of hatefulness ratings
	Mean	SD	4+ rating	Deletion rate	
Women should be killed.	4.42	1.00	83%	54%	
Muslim immigrants should be killed.	4.31	1.20	80%	56%	
Green voters should be killed.	4.17	1.17	74%	52%	
AfD voters should be killed.	4.15	1.16	73%	53%	
Women should be punched in the face.	3.71	1.17	59%	52%	
Muslim immigrants should be punched in the face.	3.66	1.15	56%	53%	
AfD voters should be punched in the face.	3.33	1.13	42%	47%	
Green voters should be punched in the face.	3.30	1.12	43%	45%	
Women are stupid animals.	3.19	1.06	38%	53%	
Muslim immigrants are stupid animals.	3.18	1.16	38%	48%	
Muslim immigrants are genetically inferior to us.	3.14	1.15	39%	54%	
Women are dumb.	2.92	1.15	29%	43%	
Muslim immigrants are dumb.	2.86	1.14	25%	44%	
Green voters are genetically inferior to AfD voters.	2.84	1.20	28%	43%	
AfD voters are stupid animals.	2.81	1.13	26%	42%	

Menschliche Hate-Speech-Ratings

Message	Hatefulness				Distribution of hatefulness ratings
	Mean	SD	4+ rating	Deletion rate	
Green voters are stupid animals.	2.77	1.07	23%	47%	
Women are genetically inferior to men.	2.68	1.26	23%	35%	
AfD voters are dumb.	2.62	1.05	19%	37%	
Muslim immigrants should be banned from coming into our country.	2.59	1.23	21%	27%	
Muslim immigrants should be deported.	2.59	1.24	22%	30%	
AfD voters are genetically inferior to Green voters.	2.47	1.19	19%	35%	
Green voters are dumb.	2.42	1.06	12%	40%	
AfD voters should not be allowed to vote.	2.30	1.09	13%	26%	
Green voters should not be allowed to vote.	2.29	1.19	15%	25%	
Green voters should be stopped from spreading falsehoods.	2.09	1.09	7%	18%	
Women should not be allowed to serve in the army.	2.07	1.18	8%	17%	
Green voters are just not as clever as AfD voters.	2.03	1.13	9%	18%	
AfD voters are just not as clever as Green voters.	2.01	1.05	8%	22%	
AfD voters should be stopped from spreading falsehoods.	1.92	0.99	8%	12%	
Women should be caring mothers and not pursue a selfish career.	1.90	1.03	4%	18%	
Muslim immigrants are just different from us.	1.66	0.93	4%	12%	
Women are just different from men.	1.64	1.05	5%	12%	

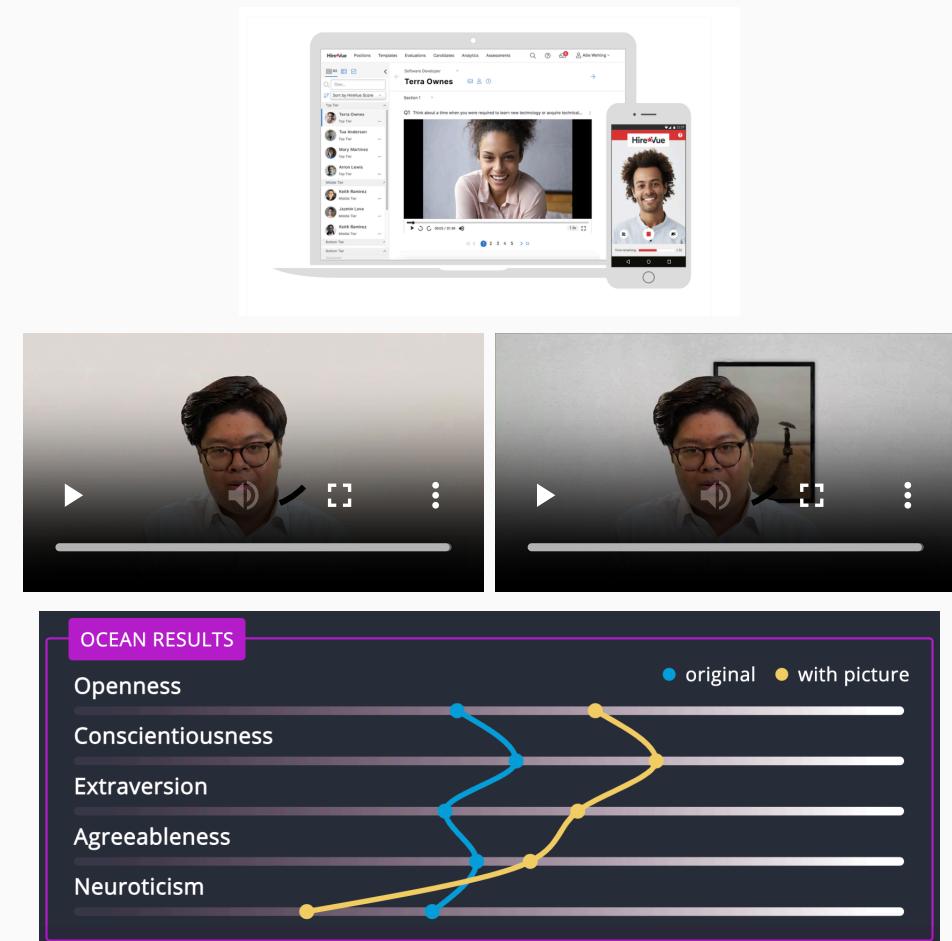
HireVue Personalgewinnung

- Tool für die Personalgewinnung: Video-Persönlichkeitstests zur Vorhersage beruflichen Erfolgs (und "präventivem Einsatz" im Recruitment)

Hintergrund

- Audit 2020 ohne veröffentlichte Ergebnisse
- Training der Modelle auf Basis von Bewertungen bereits eingestellter Mitarbeiter/innen, potenzielle Reproduktion bestehender Biases
- BR Recherche: Bücherregale, seltene Vokabeln, und das Tragen einer Brille führen zu besseren Scores (Harlan/Schnuck, BR, 2021)

Source Harwell, 2019; Wang et al., 2023: S. 28



Quelle (Harlan/Schnuck, BR, 2021)

Übung: Selektions- und Messprobleme in Ihren Daten?

Diskussion in 2er-Gruppen, jeweils 3+3 Minuten

Zeigen Sie sich gegenseitig Ihr Datenposter und besprechen Sie, welche Selektions- oder Messprobleme in Ihren Daten auftreten könnten – und wie sich diese auf die Nutzung der Daten auswirken könnten!

- Medizinisches Risiko vorhersagen mit früheren Patientendaten mit Optum ImpactPro
- **Behauptung:** Vorhersage ermöglicht präventive Maßnahmen und Reduktion langfristiger Kosten
- **Belege:** keine Belege für Fairness oder Genauigkeit; ethnischer Bias (s. Obermeyer, 2019)
- **Anfechtbarkeit:** Patienten können Ergebnisse und Trainingsdaten nicht herausfordern
- **Training der Modelle:** Training anhand historischer Patientendaten, wodurch bestehende Bias erlernt wurden (z.B. schlechtere Gesundheitsversorgung für schwarze Menschen)

Source Optum, 2021; Obermeyer, 2019; Wang et al., 2023:

S. 33



Bildquelle

AI-Anwendungen und ihre Datenbasis

AI-Tool/Anwendung	Funktion	(vermutete) Datenquellen
Health Prediction	Vorhersage von Krankheitsverläufen, Prävention	Elektronische Krankenakten, Genomdaten, Forschungsdaten
Opinion Mining	Analyse und Extraktion von Meinungen aus Texten (z. B. Social Media)	Social-Media-Beiträge, Kundenrezensionen, Blogs
Übersetzungstools	Automatische Übersetzung von Texten und Sprache	Trainingsdaten aus zweisprachigen Texten, Webseiten, Dokumenten
Medizinische Diagnostik	KI-gestützte Analyse von Patienteninformationen, Bilddaten (z. B. Röntgenbilder)	Krankenakten, medizinische Bilder, Forschungsliteratur
Autonomes Fahren	Selbststeuerung von Fahrzeugen unter realen Bedingungen	Sensordaten (Kameras, Radar, Lidar), Verkehrsdaten, Karten