Was kann „Künstliche Intelligenz“ (KI) - was nicht?

09.00 - 10.30 Uhr

Legal Tech - Workshop
am 07.05.2019 in Königs Wusterhausen

Tobias Krafft

TU Kaiserslautern

Trusted AI GmbH
@NetworkTobi

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

…zwei verurteilte Kriminelle….

Das sind Brisha und Vernon,….

Wer wird es wieder tun?

ACLU (American Civil Liberties Union) fordert:

2011

eine genaue Datenanalyse um das Risiko zu kalkulieren, ob Straftäter tatsächlich rückfällig und zu einer Gefahr für die Gesellschaft werden

2019

keine genaue Datenanalyse um das Risiko zu kalkulieren, ob Straftäter tatsächlich rückfällig und zu einer Gefahr für die Gesellschaft werden

Chettiar, I. M., & Gupta, V. (2011). Smart Reform is Possible: States Reducing Incarceration Rates and Costs While Protecting Communities. *Available at SSRN 1934415*.

https://civilrights.org/2018/07/30/more-than-100-civil-rights-digital-justice-and-community-based-organizations-raise-concerns-about-pretrial-risk-assessment/

Menschen – so irrational!

- Richter müssen vorzeitige Haftentlassungsanträge begutachten.
- Studie: je weiter von der letzten Pause weg, desto weniger risikoreiche Entscheidungen[1].
- Eine Vielzahl solcher Studien scheint zu beweisen:
  - Menschen sind irrational und vorurteilsbeladen.

1 Danziger, S.; Levav, J. & Avnaim-Pesso, L.: “Extraneous factors in judicial decisions”, Proceedings of the National Academy of the Sciences, 2011 , 108 , 6889-6892

Das kleine ABC der Informatik

Können

Algorithmen,

Big Data und

Computerintelligenz

Menschen besser bewerten und richten als Menschen?

A wie Algorithmus

Ein Algorithmus ist ein Problemlöser

Mathematisches Problem

Der PUTPUT der uns sagt, wie Input mit Output zusammenhängt.

INPUT

OUTPUT

Beispiel für ein Problem: Navigation

Navigation

Gegeben das Karten-material und weitere Daten, berechne die kürzeste Route zwischen Start und Ziel

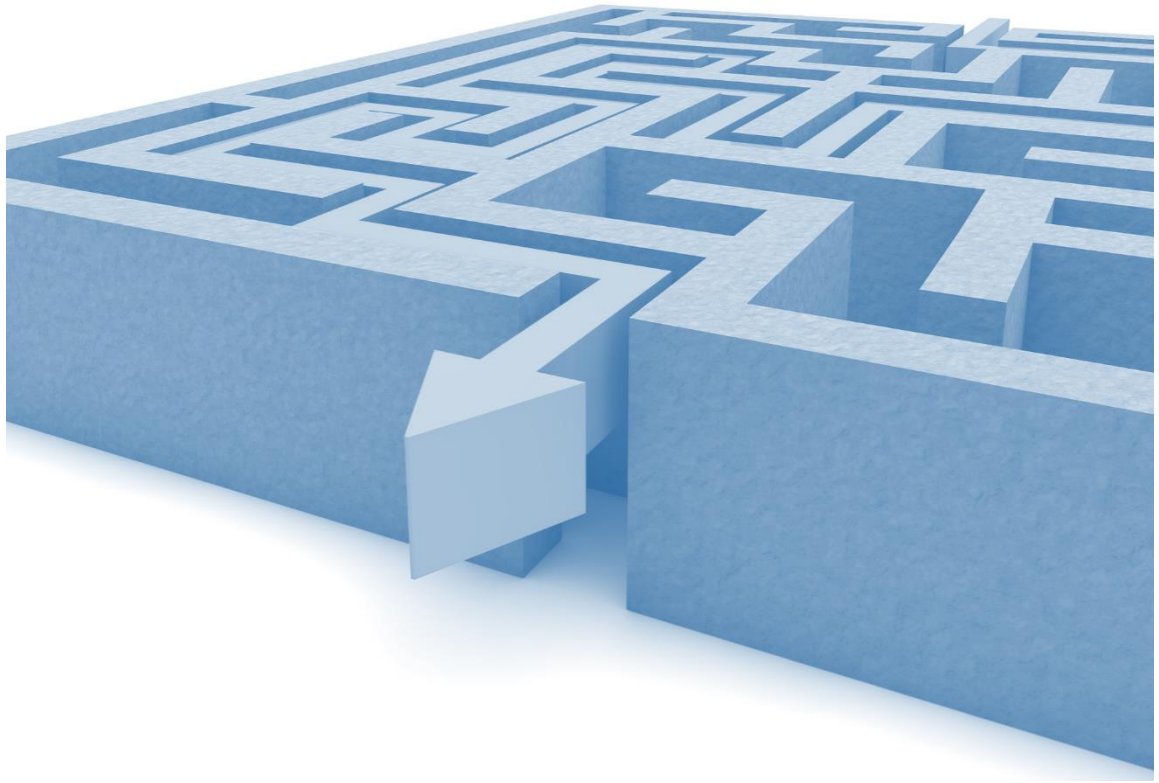
Das **Problem** sagt nicht, wie man die Lösung **findet**.

Input: Straßen, Länge, Staus, …
Start und Ziel

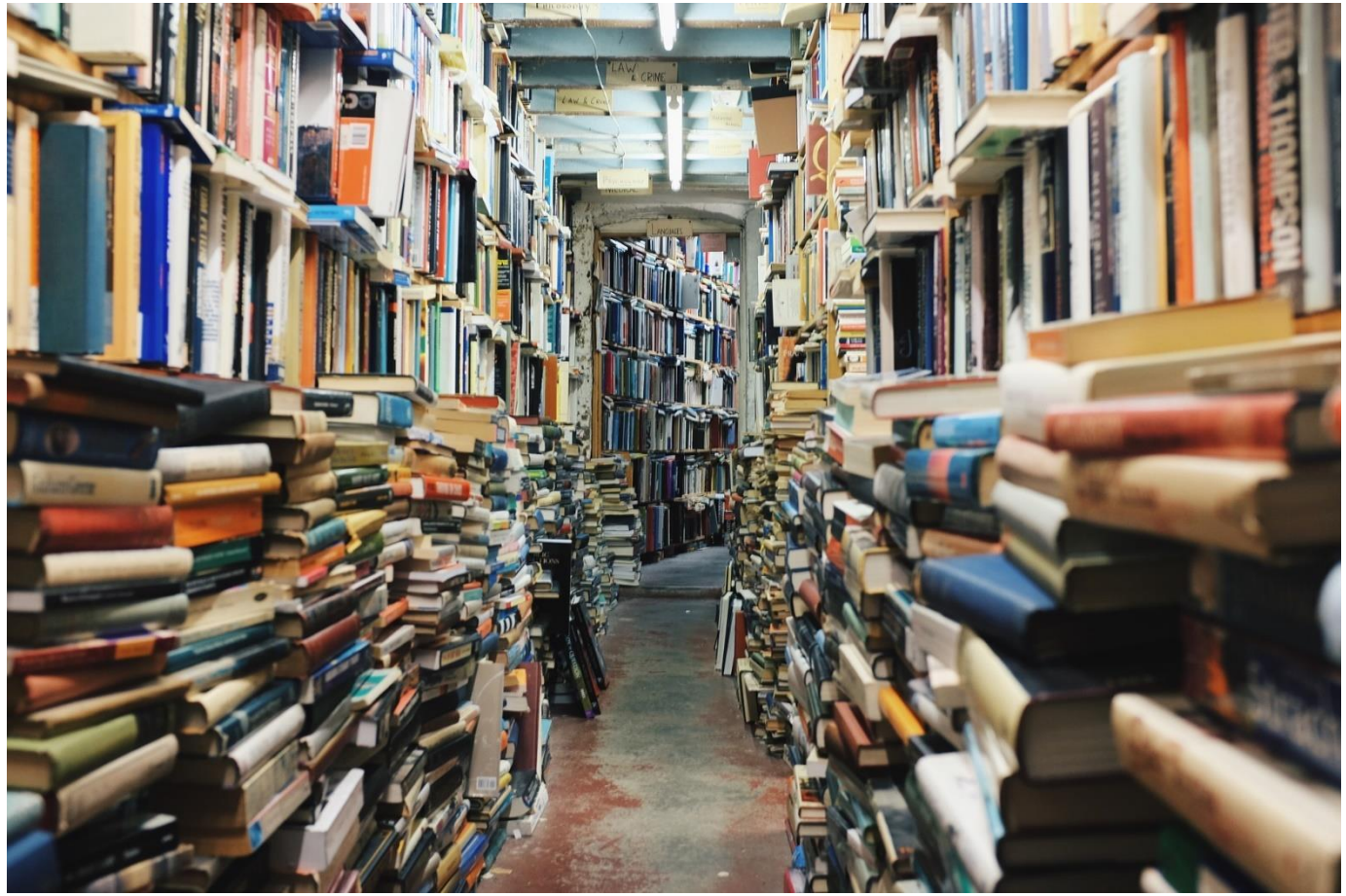
Output: optimale Route

Ein Algorithmus ist…

…eine für jede **erfahrene Programmiererin ausreichend detaillierte Lösungsvorschrift**, so dass bei **korrekter Implementierung** der Computer **für jede korrekte Inputmenge den korrekten Output** berechnet – in endlicher Zeit.



Beispiel: Sortieren



Sortieren 1: „Sortieren durch Einfügen“

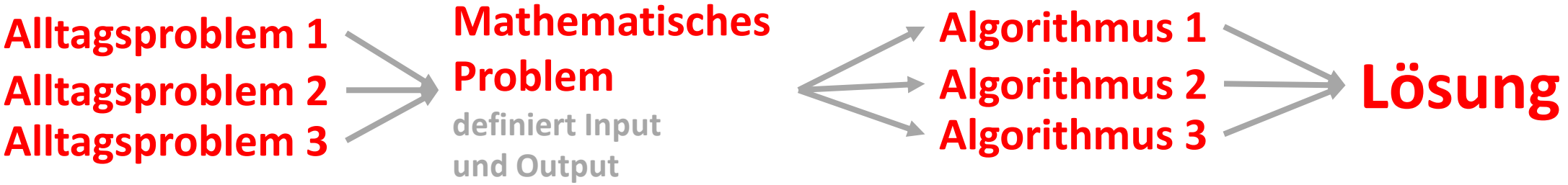
- Fange mit einem Buch an, stelle es ins Regal.
- Solange es noch Bücher gibt,
  - nimm das nächste,
  - geh am Regal entlang und sortiere es an der passenden Stelle ein.

- Alle Bücher, die schon im Regal stehen, sind in der richtigen, relativen Reihenfolge.
- Daher: wenn alle im Regal stehen, sind sie vollständig sortiert.

Sortieren 2: Aufsteigendes Sortieren

- Stelle alle Bücher irgendwie ins Regal.
- Gehe das Regal entlang – wenn dabei zwei Bücher in der falschen Reihenfolge nebeneinander stehen, vertausche sie. Tue dies bis zum Ende des Regals und gehe wieder zum Anfang.
- Laufe solange immer wieder am Regal entlang, bis im letzten Durchgang kein Tausch mehr nötig war.
- Wenn kein Tausch mehr nötig war, sind alle Bücher sortiert.

Problem-Algorithmus-Lösung



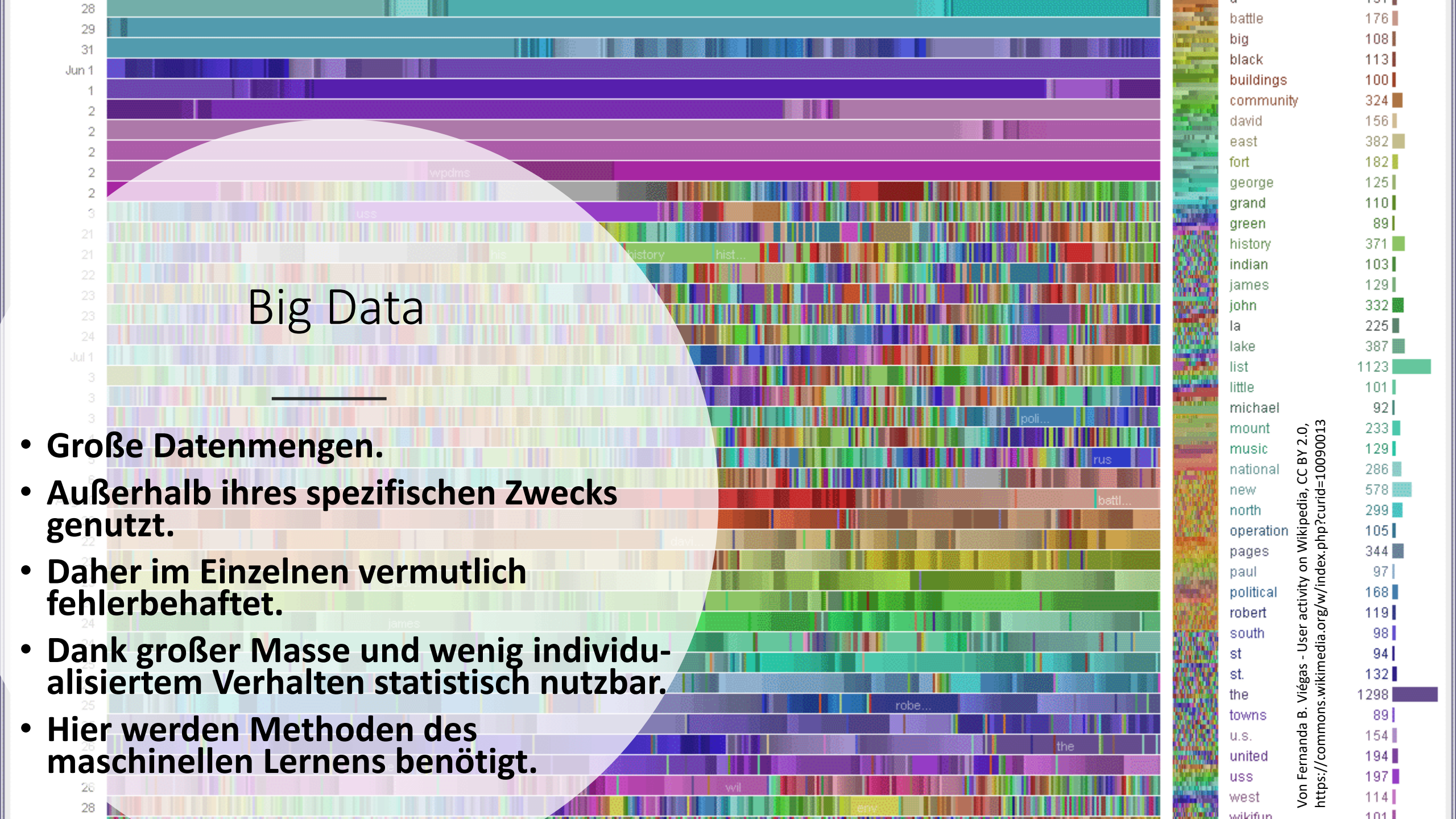
- Ein mathematisches Problem kann also meist durch mehrere Algorithmen gelöst werden.
- Jeder Algorithmus löst nur genau ein mathematisches Problem.
- Im Sinne von „Alltagsproblemen“ löst derselbe Algorithmus sehr viele verschiedene Probleme:
  - Sortieren von Personen nach Anzahl ihrer Follower auf Twitter;
  - Anzeige von Nachrichten, sortiert nach Publikationsdatum;
  - Suchmaschineneinträge sortieren nach Bewertung durch Suchmaschinenalgorithmus;

Und worüber reden dann gerade alle?

Maschinelles Lernen aus Big Data

B wie Big Data

Daten als Grundlage



- **Große Datenmengen.**
- **Außerhalb ihres spezifischen Zwecks genutzt.**
- **Daher im Einzelnen vermutlich fehlerbehaftet.**
- **Dank großer Masse und wenig individu-alisiertem Verhalten statistisch nutzbar.**
- **Hier werden Methoden des maschinellen Lernens benötigt.**

C wie Computerintelligenz

Was heißt Lernen?

Einfach:

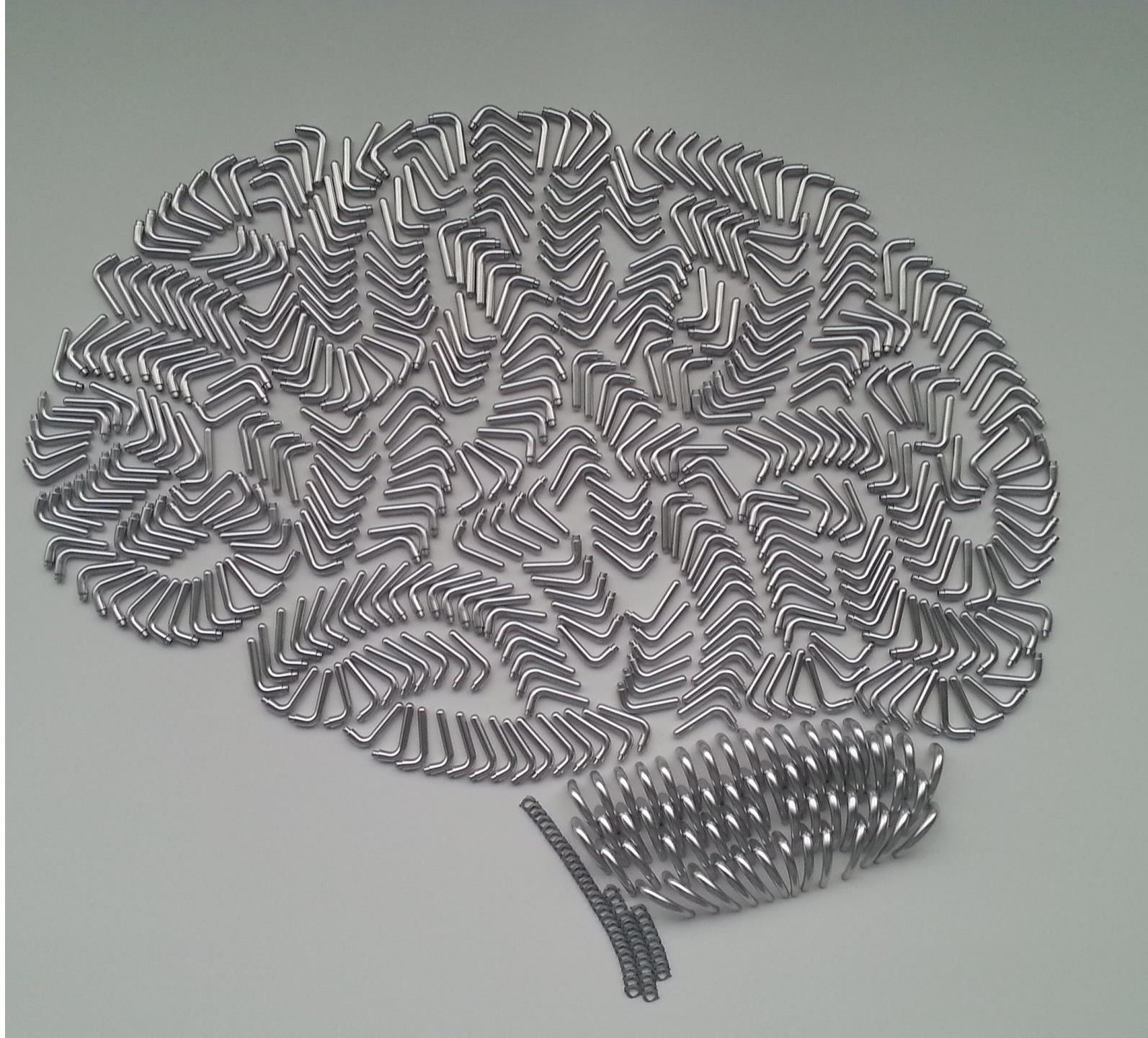
In derselben Situation ein vorher gezeigtes Verhalten wiederholen.

Generalisiert:

In derselben Art von Situation das richtige Verhalten aus einer Reihe von Möglichkeiten auswählen.

Sebastian lernt…

- Durch **Rückkopplung**: unerwartet heiß, unerwartet kalt
- Durch **Speicherung in einer Struktur:** in Neuronen und deren Verknüpfung.
- Durch viele **Datenpunkte.**
- Durch **Generalisierung des Gelernten**.

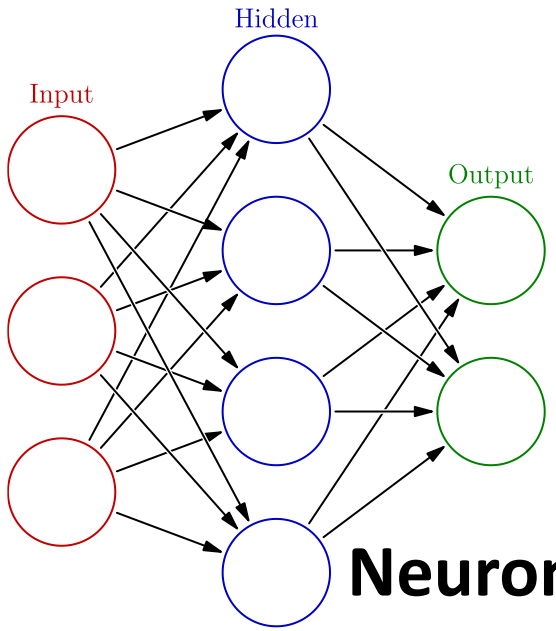


Computer lernen

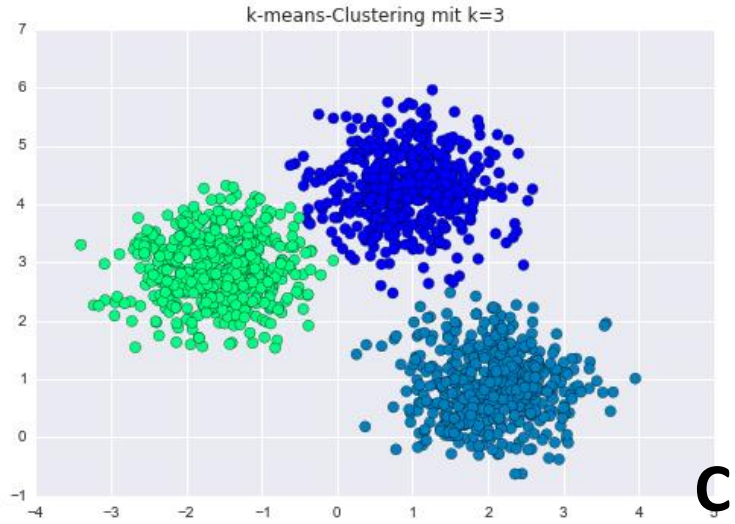
Damit ein Computer lernen kann, benötigt er ebenfalls eine **Struktur,** um Gelerntes abzuspeichern.

Optimal auch **Rückkopplung**.

Er lernt **generelle Regeln**.



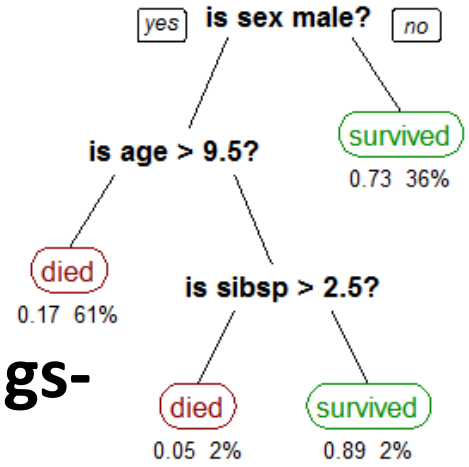
Neuronales Netz



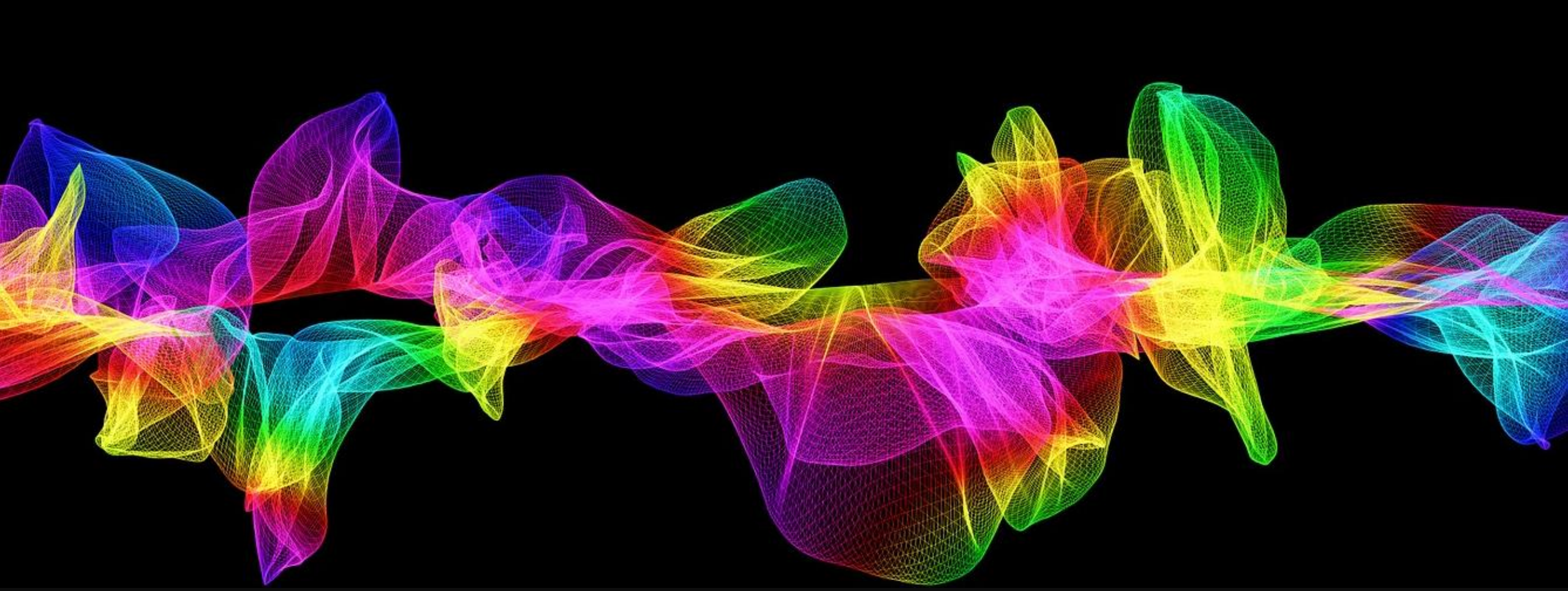
Clustering

$$w_1 * \#Vh - w_2 * \#dayiVh + w_3 * I[g = male] \quad B] * 1.0 + \cdots$$

Formel

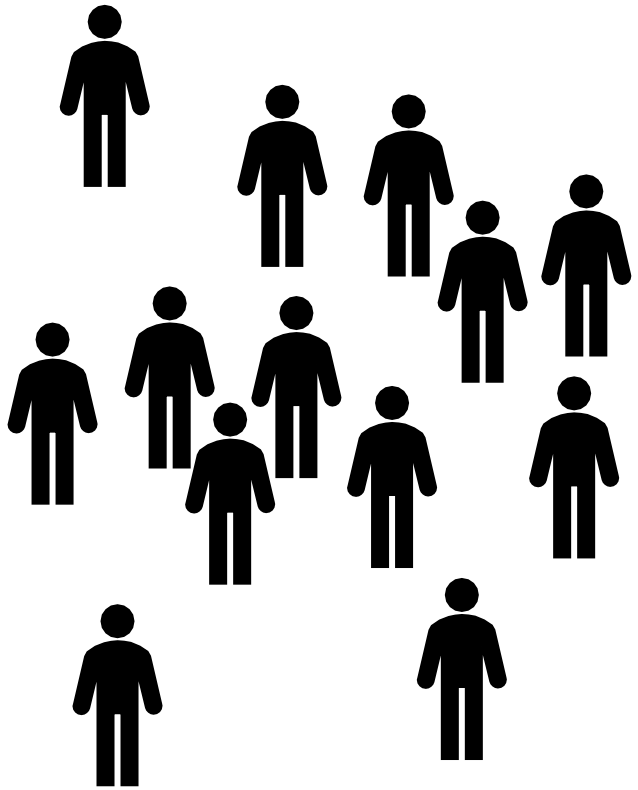


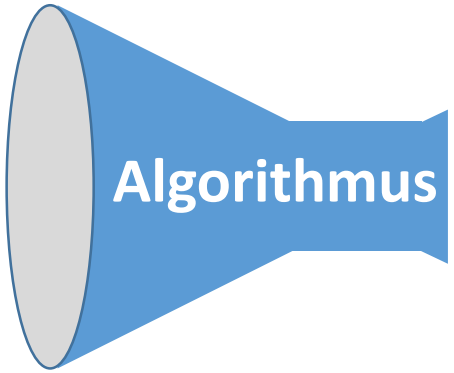
Entscheidungs-bäume

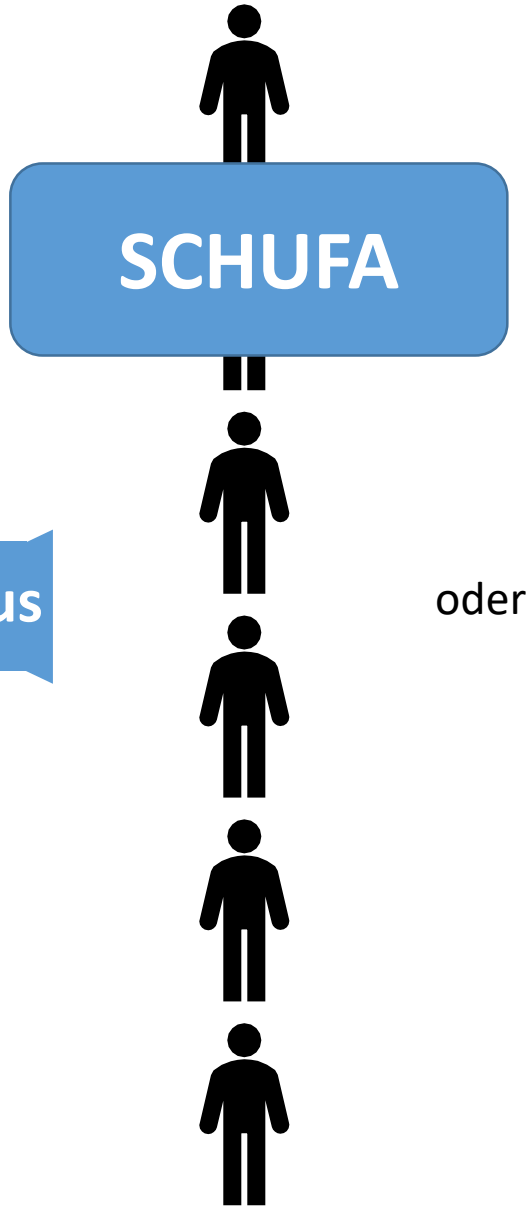


“Lernen” mit Korrelationen

Algorithmische Entscheidungssysteme







Scoring-Verfahren

oder

Klassifikation

Lernen mit Formeln

Rückfälligkeitsvorhersage für (schon verurteilte) Kriminelle

Datengrundlagen

- Data Mining Methoden nutzen, z.B.:
  - Alter der ersten Verhaftung
  - Alter des Delinquenten (der Delinquentin!)
  - Finanzielle Lage
  - Kriminelle Verwandte
  - Geschlecht
  - Art und Anzahl der Vorstrafen
  - Zeitpunkt der letzten kriminellen Akte
  - Extra-Fragebogen
  - Aber bspw. nicht die (in den USA eindeutig zugeordnete) ethnische Zugehörigkeit.

- Wichtig: Beim Trainingsset ist bekannt, ob die Person rückfällig geworden ist oder nicht.

Regressionsansatz

$$
\begin{aligned}
&3 * \text{ bisherige Verhaftungen} \\
&- 2 * \text{ Anzahl Tage seit letzter Verhaftung} \\
&+ 3 * (\text{Wenn Mann, dann 1, sonst 0}) \\
&+ 2,5 * (\text{ Wenn Raubüberfall, dann 1, sonst 0}) + ...
\end{aligned}
$$

$$
\begin{aligned}
&w_1 * \text{ bisherige Verhaftungen} \\
&- w_2 * \text{ Anzahl Tage seit letzter Verhaftung} \\
&+ w_3 * (\text{Wenn Mann, dann 1, sonst 0}) \\
&+ w_4 * (\text{Wenn Raubüberfall, dann 1, sonst 0}) + ...
\end{aligned}
$$

Der Computer bestimmt die Gewichte und bekommt ein Feedback (Rückkopplung), inwieweit die damit resultierende Bewertung tatsächlich mit dem (beobachteten) Verhalten übereinstimmt.

Computer sagt: Hochrisikokategorie 8

Computer sagt: Niedrigrisikokategorie 3

Wer wird es wieder tun?

Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner: Machine Bias, ProPublica, 23.5.2016
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Wer hat es wieder getan?

Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner: Machine Bias, ProPublica, 23.5.2016
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Wie kommt die Ethik in den Rechner?

ALGORITHM ACCOUNTABILITY LAB

Maschinelles Lernen

Software, die aus Daten der Vergangenheit Entscheidungsregeln ableitet für zukünftige Daten.

Die Software trifft dann mit den gelernten Regeln Entscheidungen über neue Situationen.

Wann muss das auf technischer Ebene kontrolliert und reguliert werden?

Wie „lernt“ das System von Daten?

DIY:
Sie sind heute meine
„Support Vector Machine“

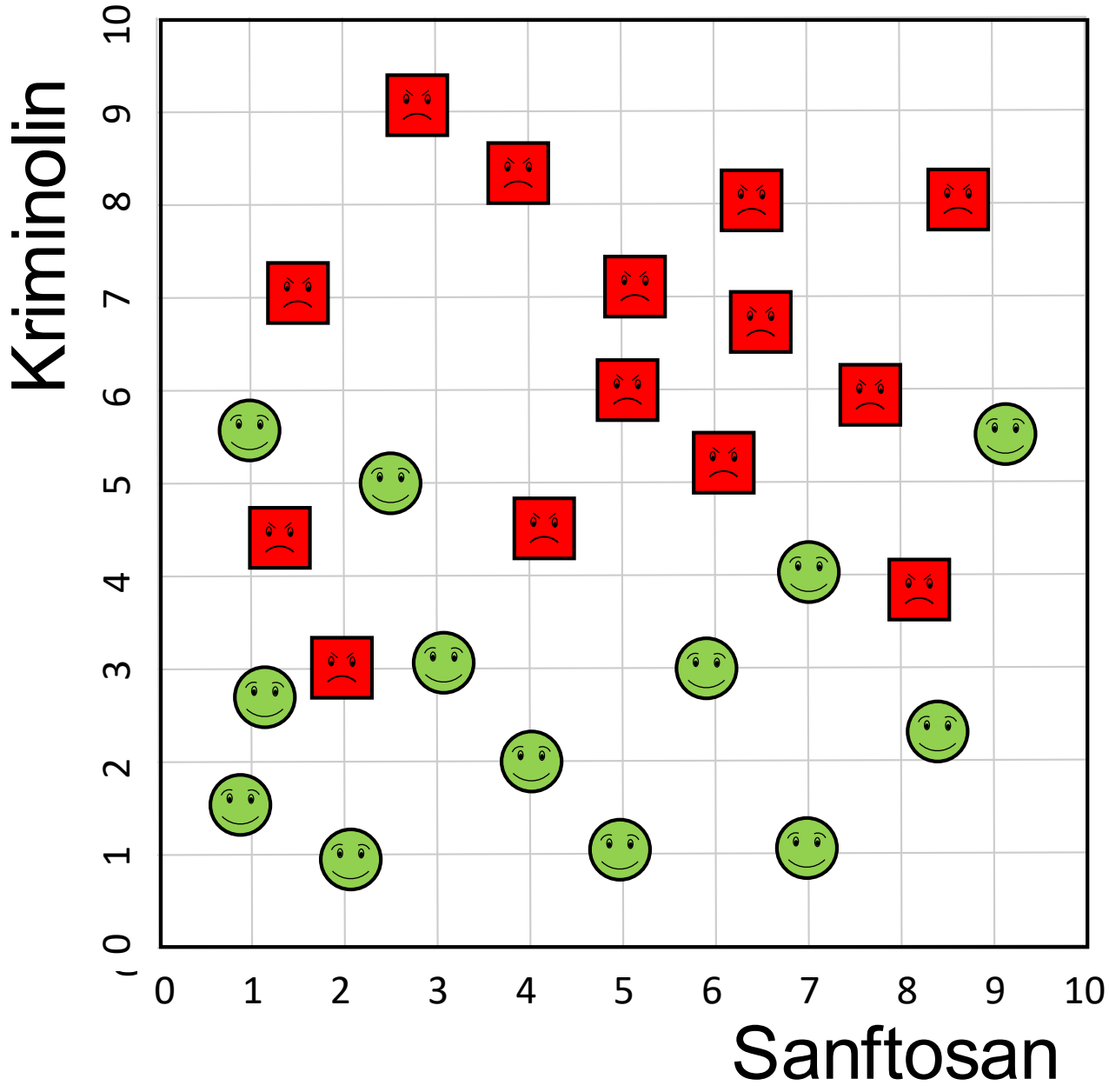
Bösartige Kriminelle

Unschuldige Bürger

Legen Sie den Holzspieß so zwischen die Smileys, dass die roten möglichst gut von den grünen getrennt sind. Kleben Sie ihn fest.

Gratulation: Sie haben eine Support Vector Machine trainiert!

Der Holzspieß dient nun als Entscheidungsregel, ob eine Person als kriminell gilt oder unschuldig zu sein scheint.

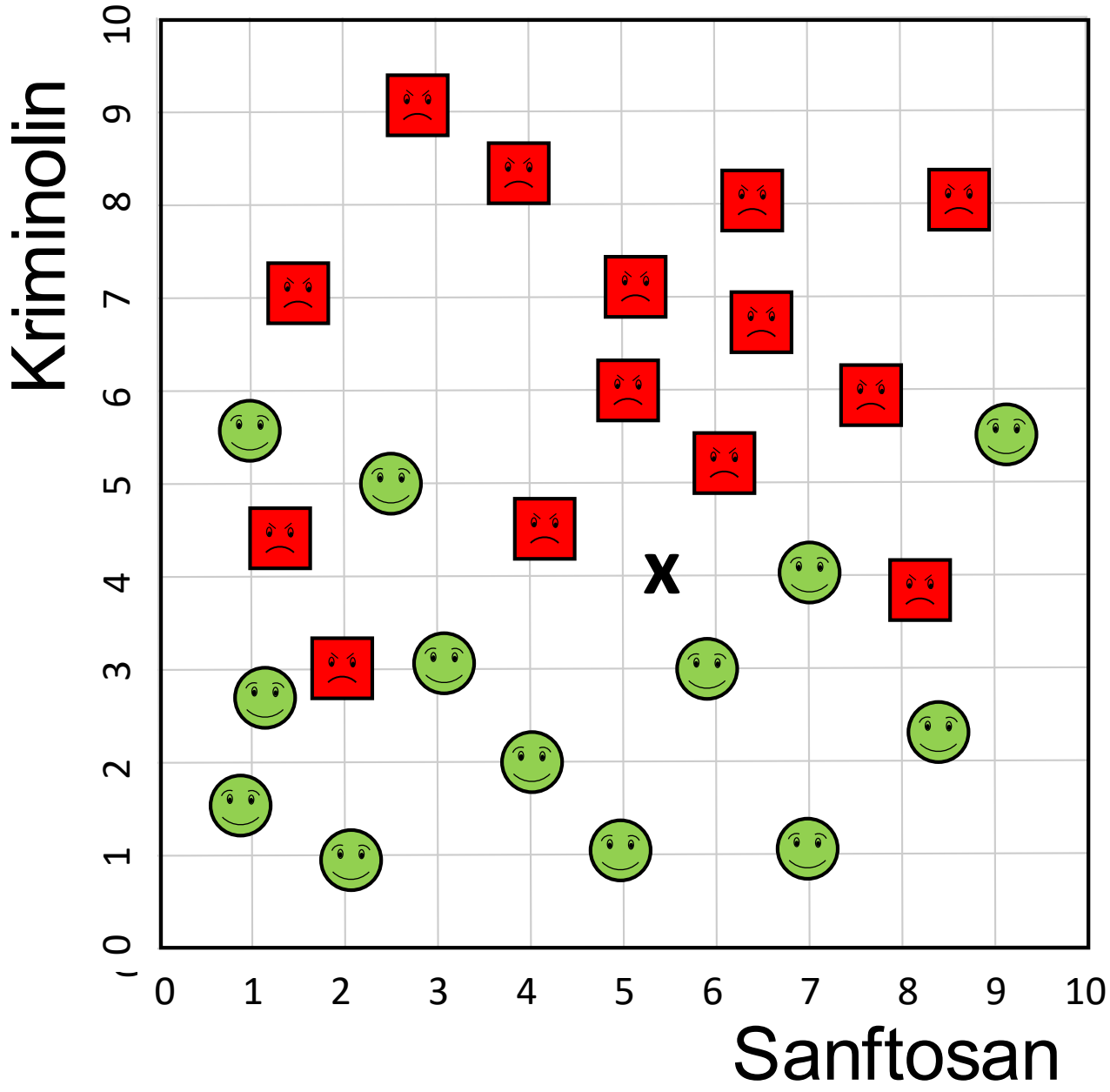


Bösartige Kriminelle

Unschuldige Bürger

Bewerten Sie Frau Müller:

5.5 Sanftosan
4.0 Kriminolin

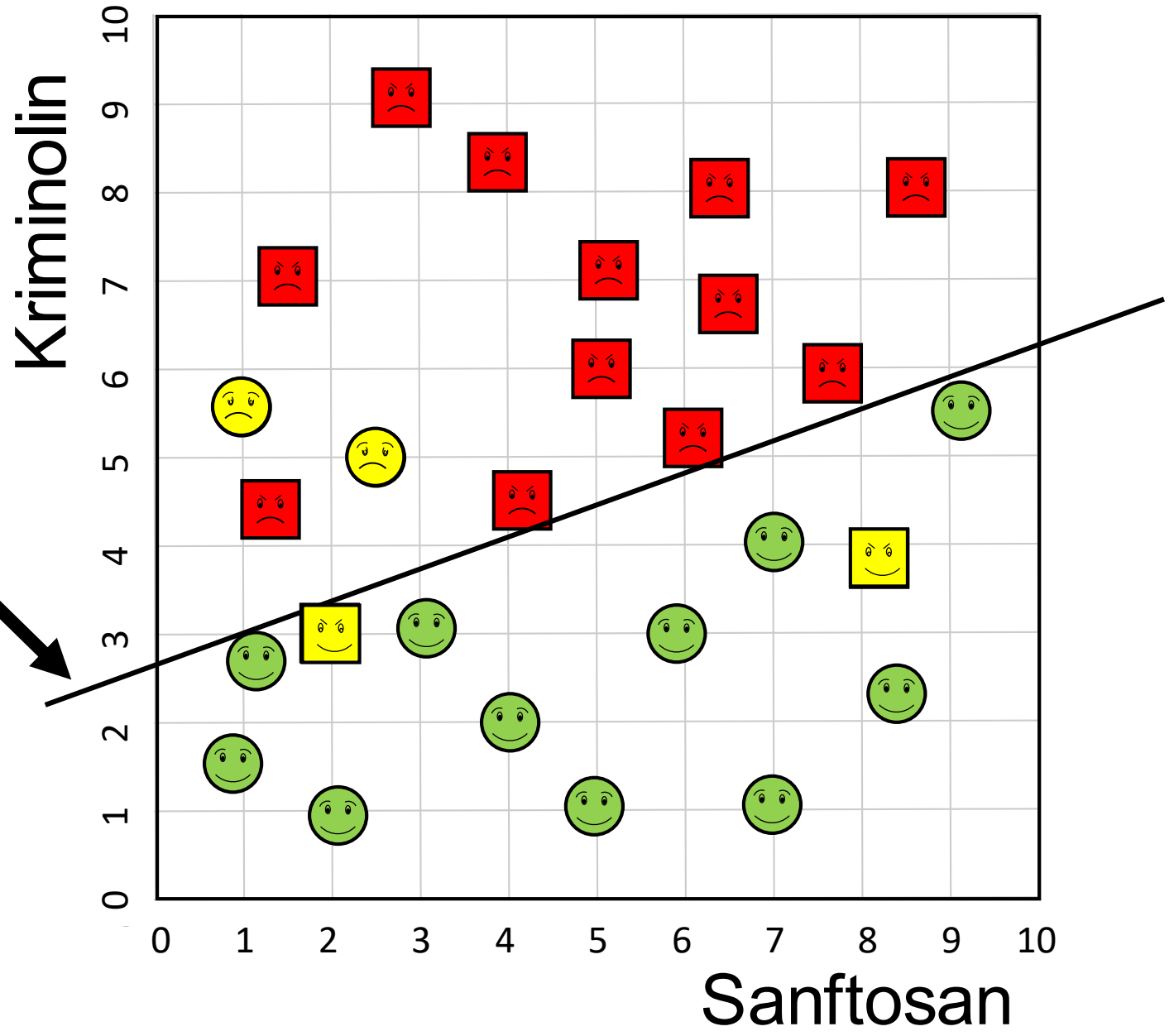


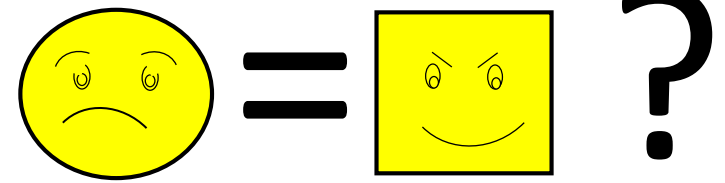
Eine der möglichen Trennlinien

Alle möglichen Trennlinien erzeugen Fehler:

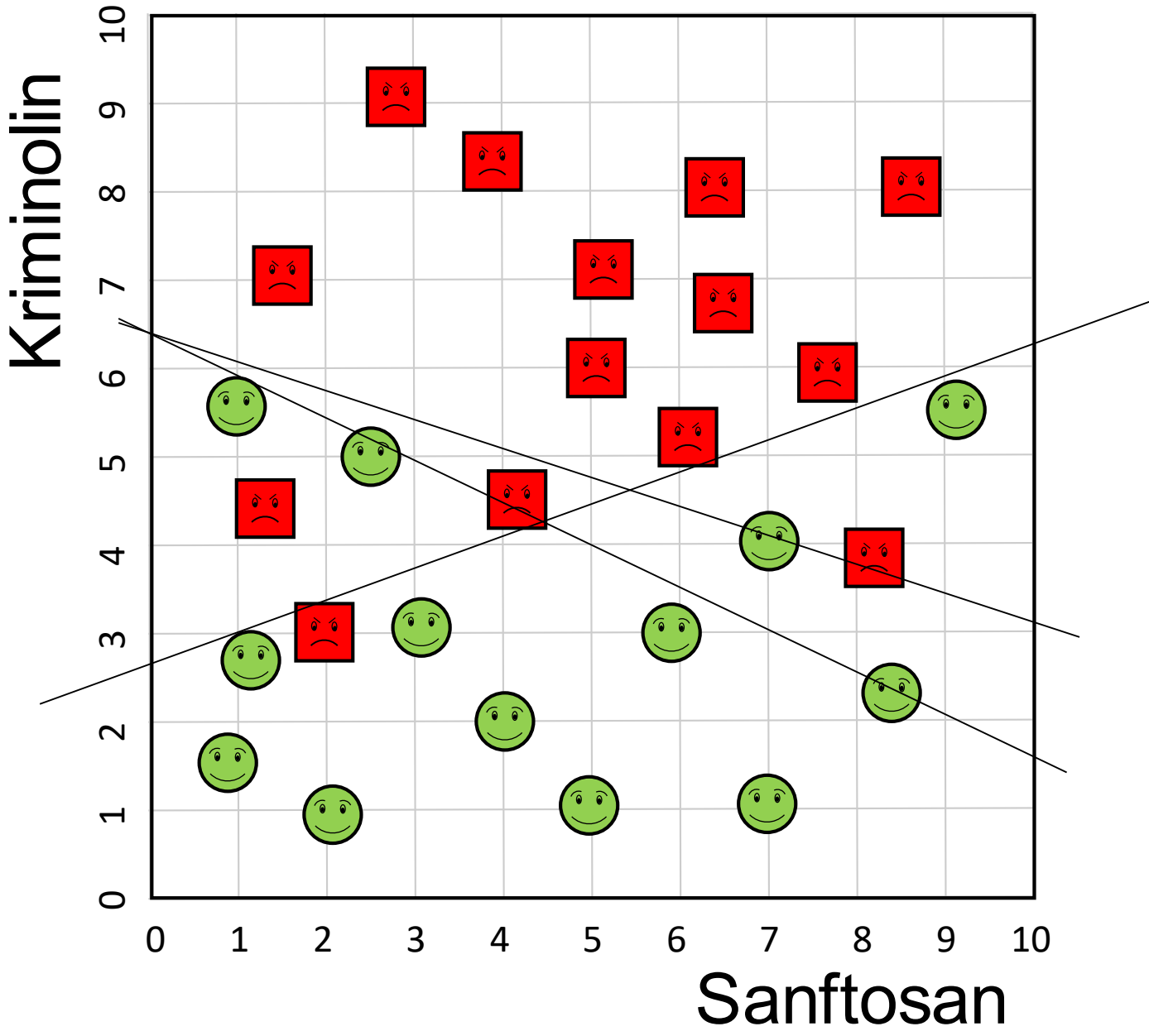
Bösartige Kriminelle, die unentdeckt bleiben

Unschuldige Bürger, die für kriminell gehalten werden





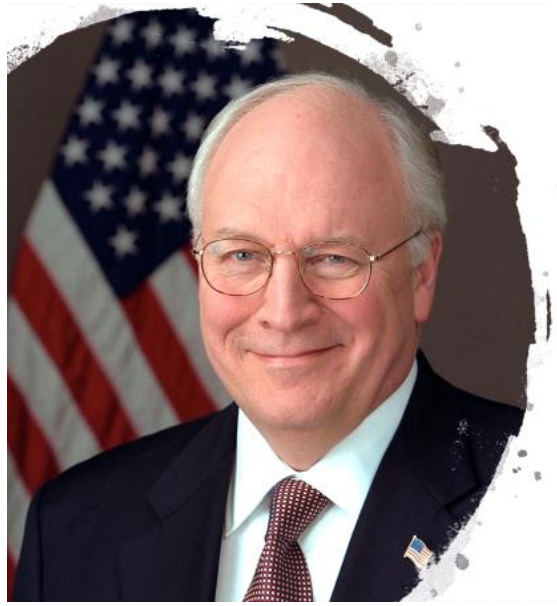
Wenn beide Fehler als gleich schlimm gelten, gibt es mehrere optimale Trennlinien mit möglichst wenigen Fehlern.




Sind beide Arten von Fehler gleich zu bewerten?

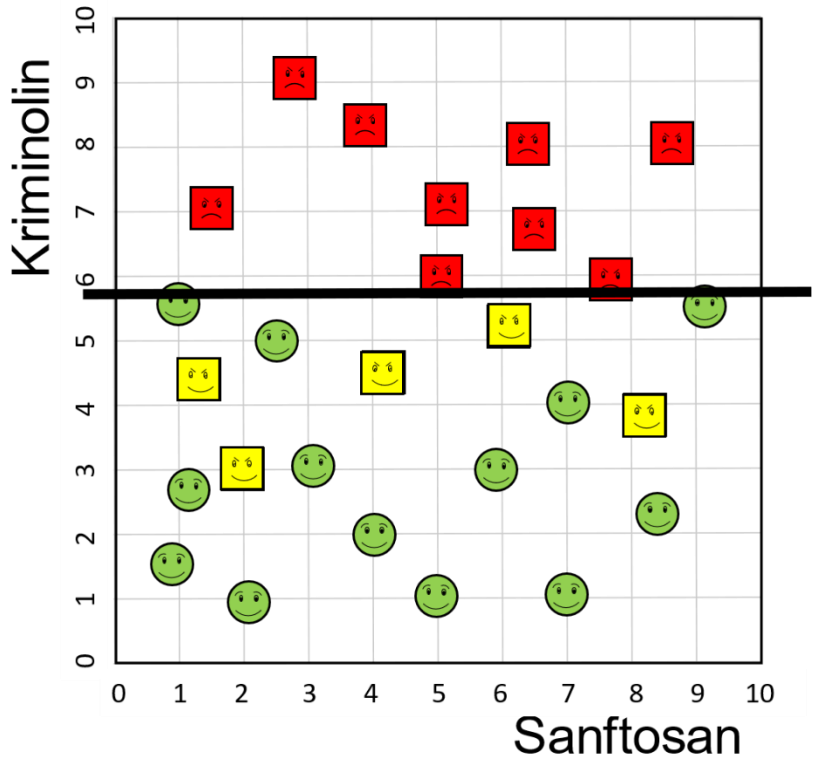
„It is better that ten guilty persons escape than that **one** innocent suffer.“

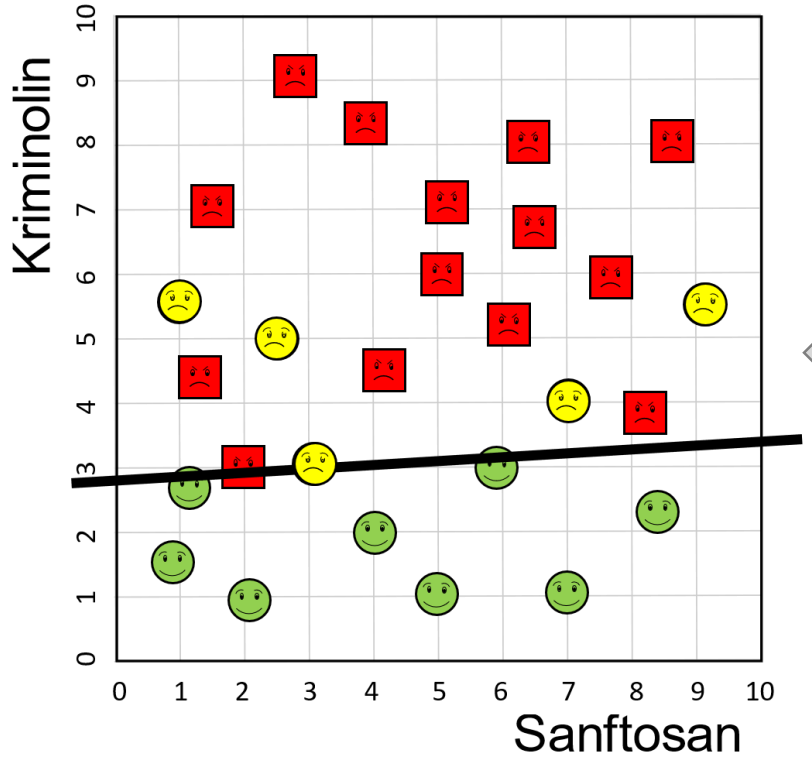
William Blackstone,
Rechtsphilosoph, 1760



"I am more concerned with bad guys who got out and released than I am with a few that, in fact, were innocent."

Dick Cheney, ehemaliger Vizepräsident der USA,





1. Beobachtung

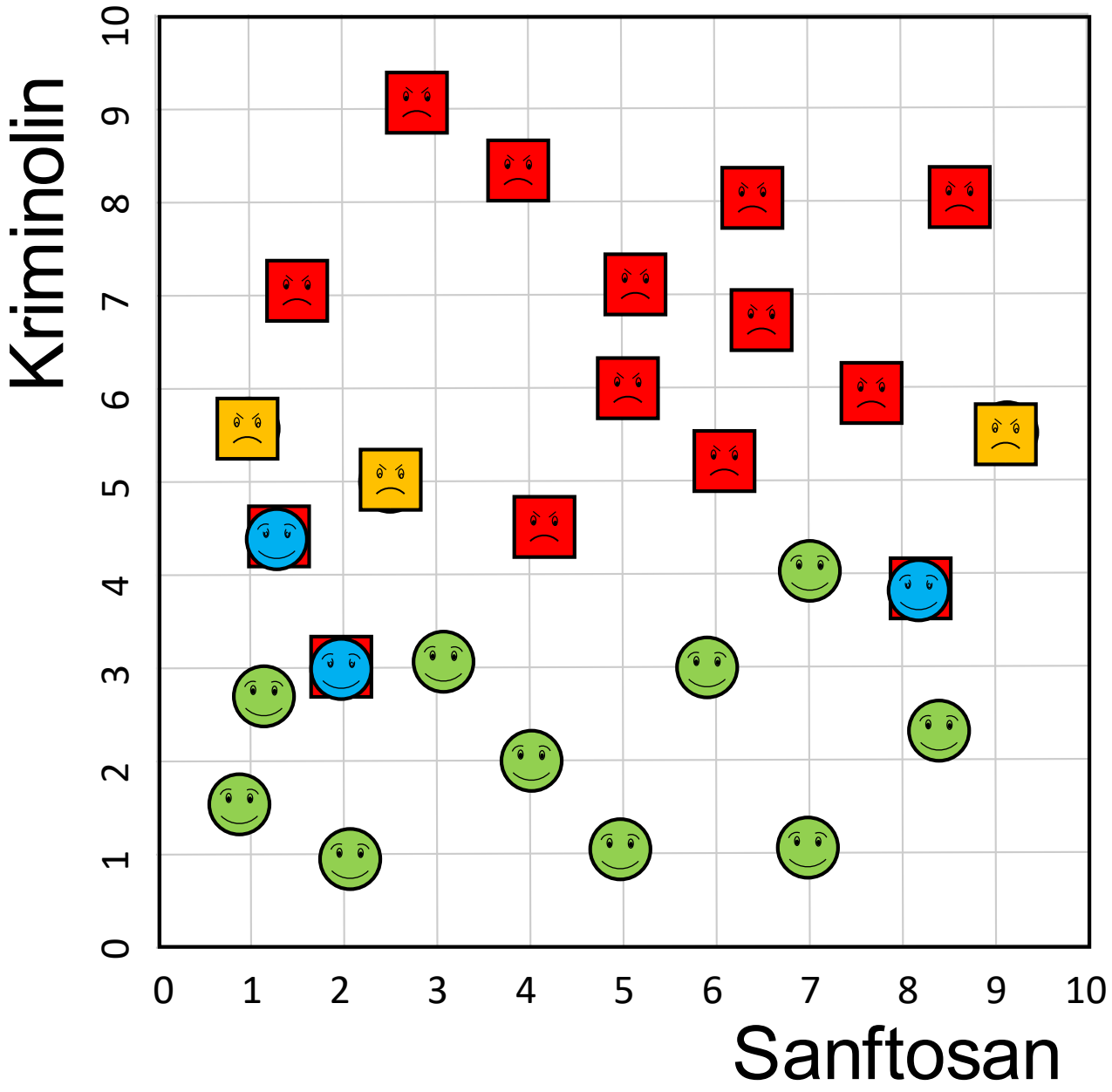
Was durch eine künstliche Intelligenz
optimiert werden soll,
ist eine gesellschaftliche Entscheidung!

Datenqualität

Noch nicht entdeckte Steuerbetrüger

Unschuldig im Gefängnis

Falsche Datenpunkt-zuordnungen haben Einfluss auf das Training der Support Vector Machine und damit auf die nachfolgenden Entscheidungen.



2. Beobachtung

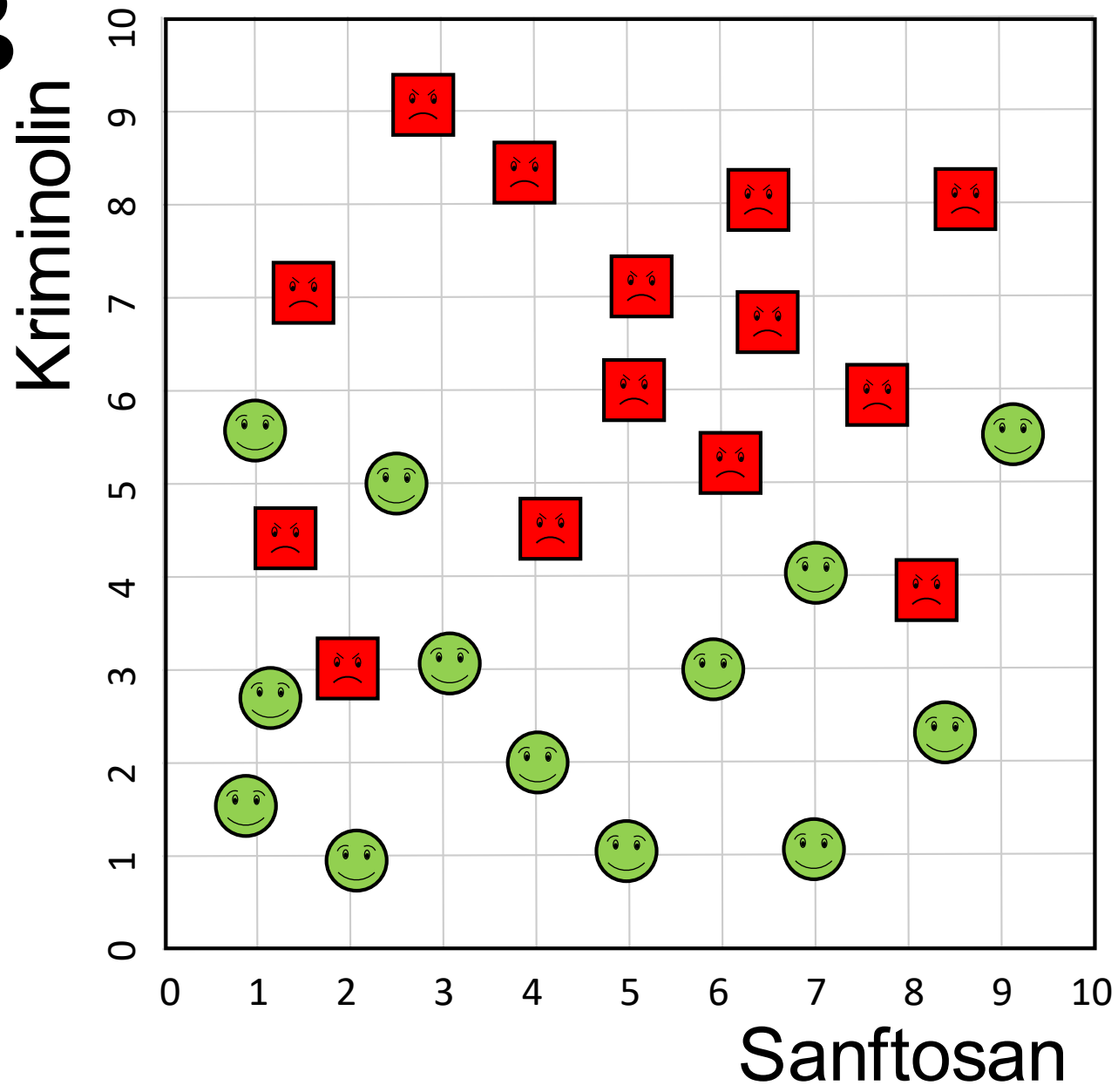
Wie gut die Maschine lernt, ist direkt abhängig von der Qualität der Daten.

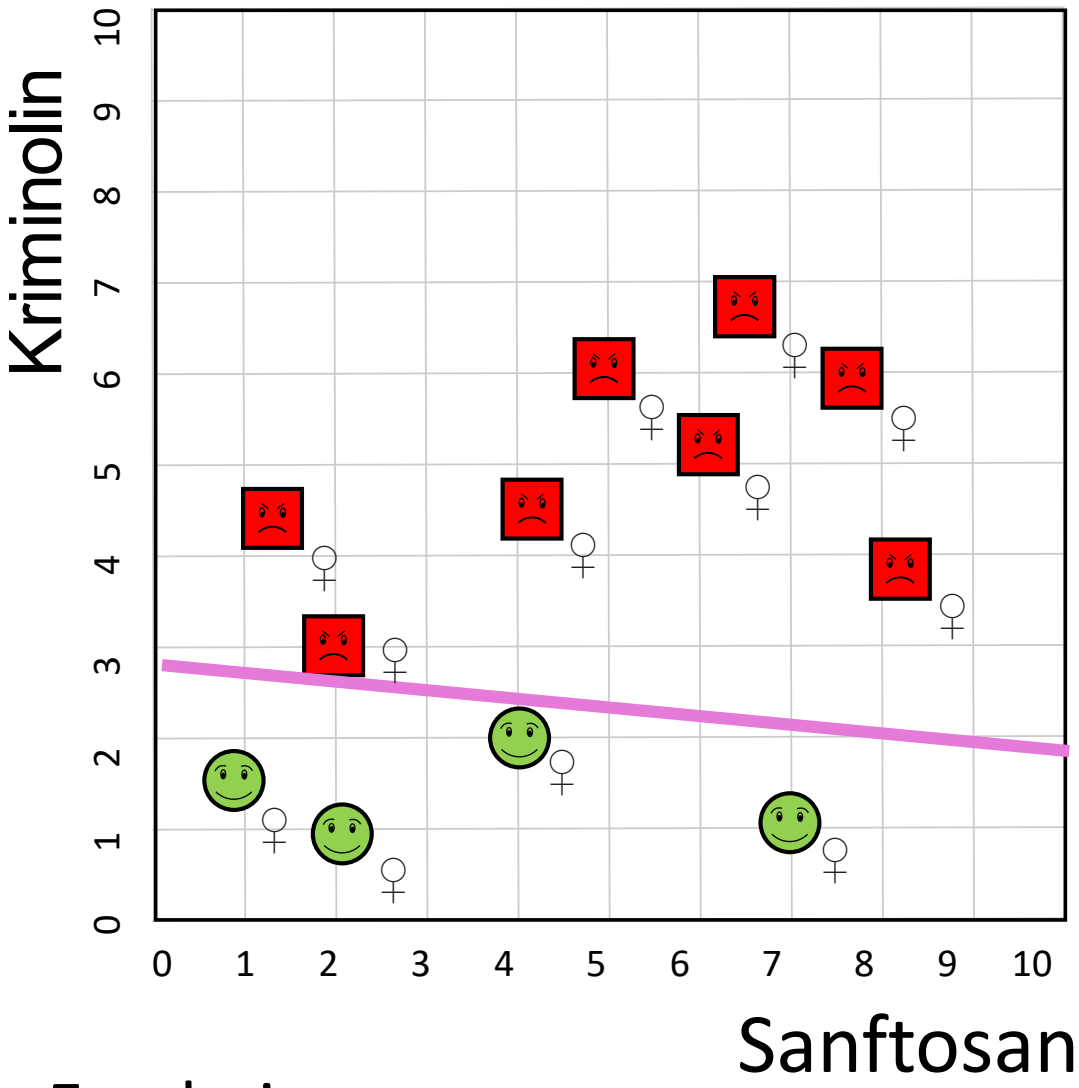
Diskriminierung

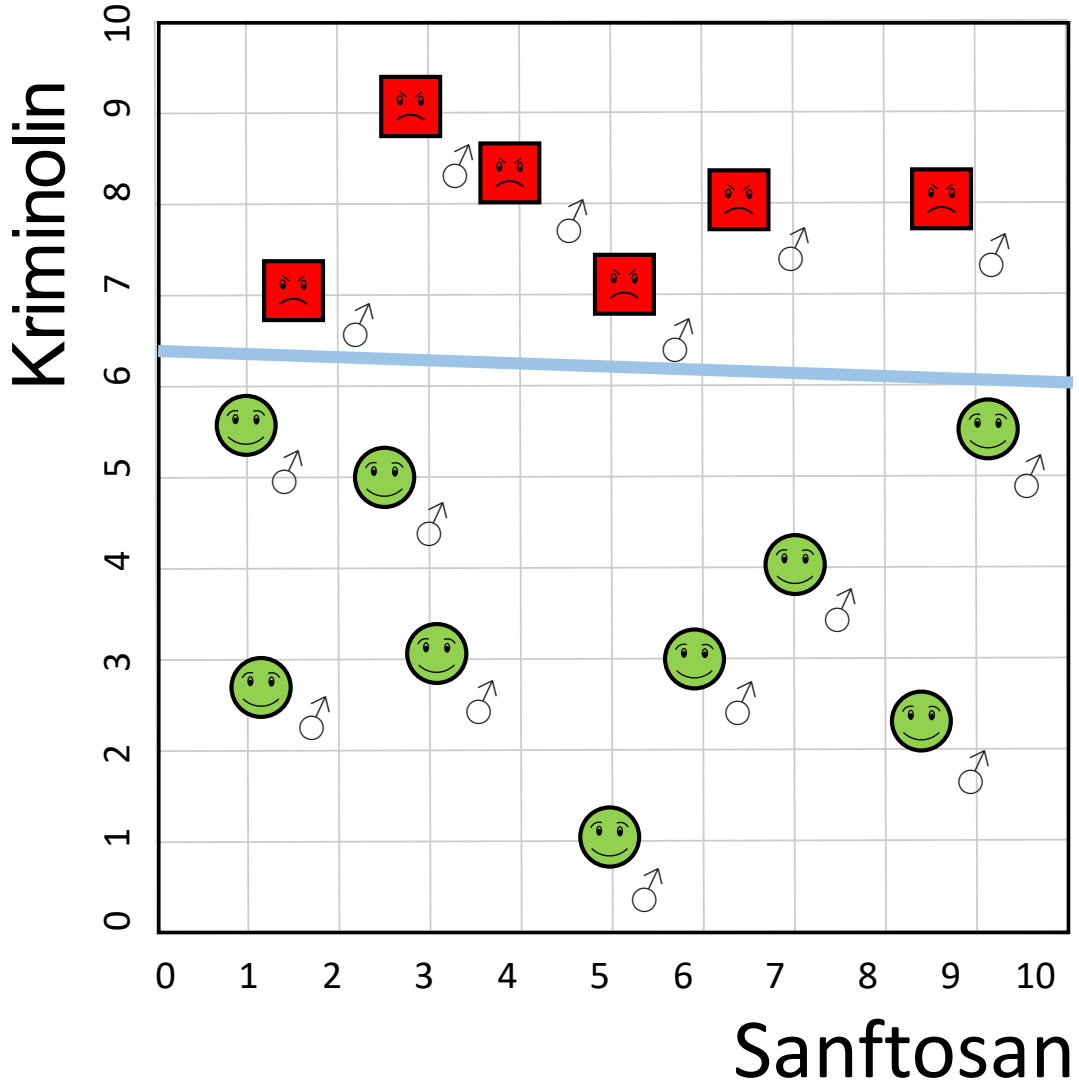
Kann ein Computer diskriminieren, wenn maschinelles Lernen verwendet wird?

Ja!

Auf der nächsten Seite trennen wir die Datenpunkte auf der rechten Seite auf in männliche und weibliche Personen und trainieren für jede Teilmenge jeweils eine Support Vector Machine.





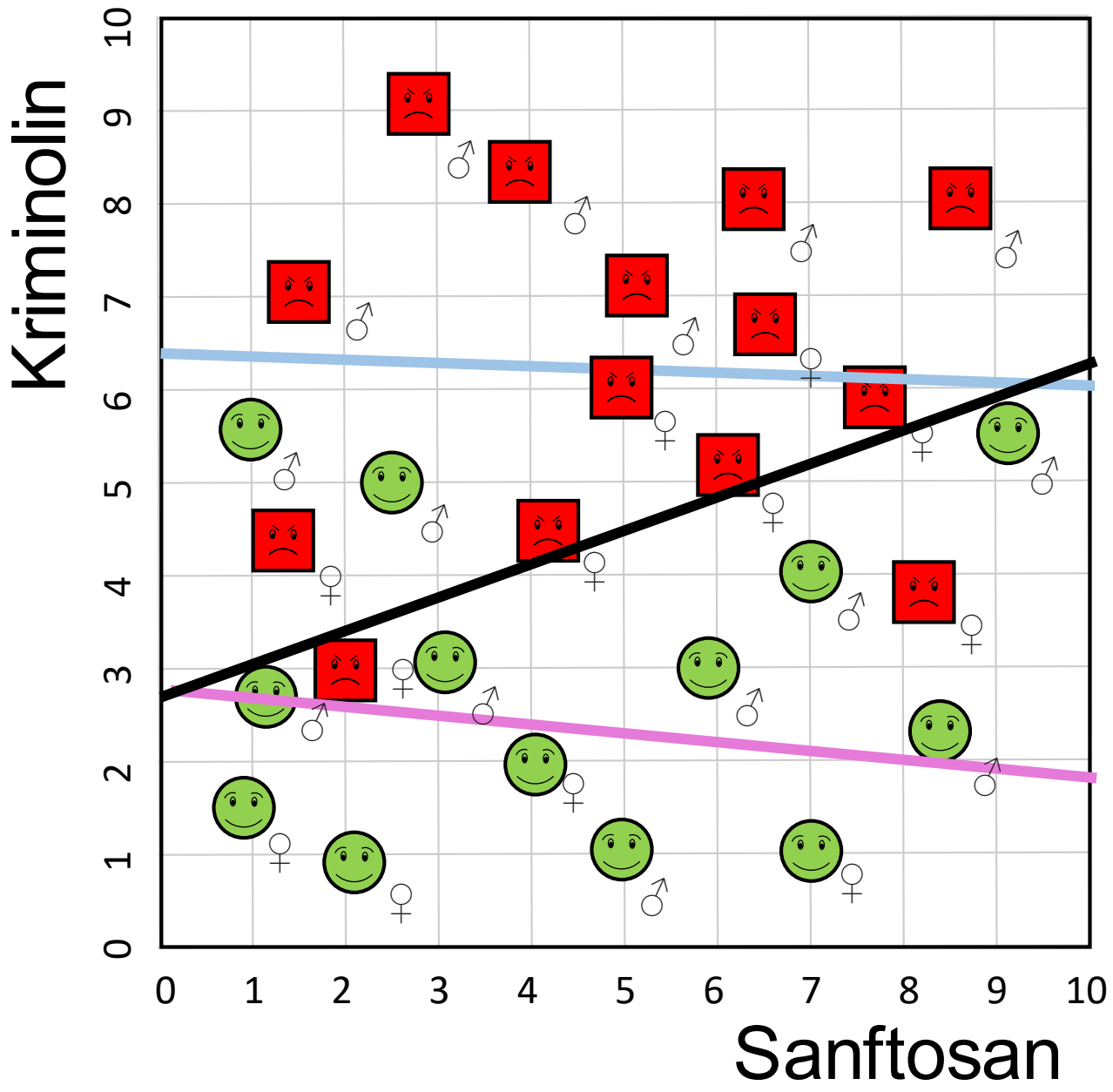


Ergebnis:

In diesem fiktiven Beispiel wird für jede Teilgruppe eine optimale Entscheidungsregel ohne Fehler gefunden.

Legt man dagegen beide Gruppen zusammen, diskriminiert die trainierte Support Vector Machine **Männer:**

Zwei weibliche Kriminelle gelten als unschuldig, zwei unschuldige Bürger als kriminell.



3. Beobachtung

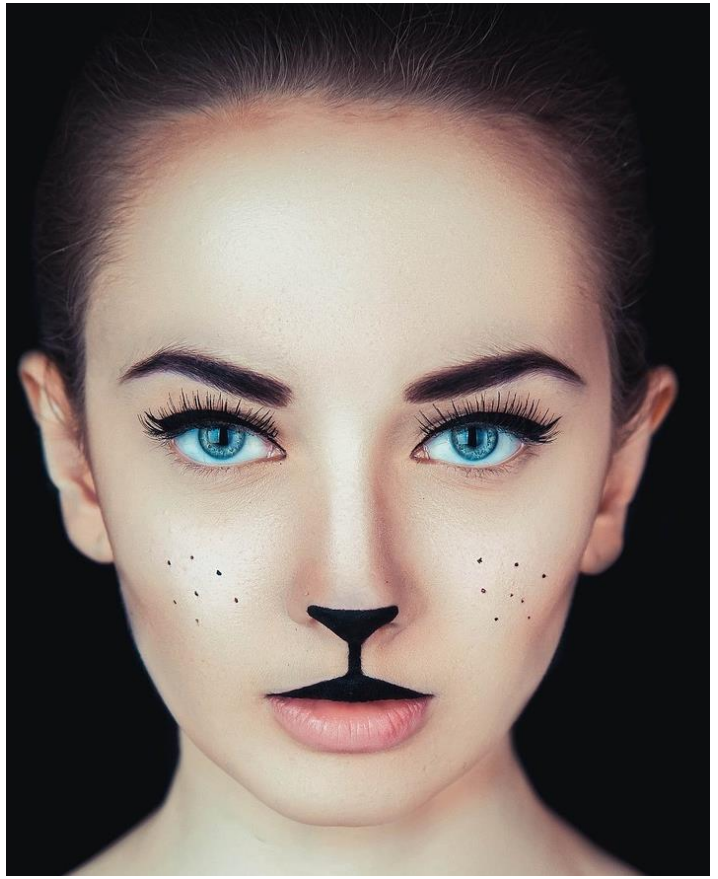
Eine geschützte Information kann wichtig sein,
um bessere Entscheidungen zu treffen.
Diskriminierung wird nicht per se dadurch
vermieden, dass die Information vorenthalten wird.

Statistische Vorhersagen über Menschen

Statistische Prognosen beim Wetter

Zu 40% ein Krimineller….

- Wenn dieser Mensch eine Katze wäre und 7 Leben hätte, würde er in 3 davon wieder rückfällig werden…
- Nein!
- **Algorithmische Sippenhaftung**
  - Von 100 Personen, die „genau so sind wie dieser Mensch“, werden 40 wieder rückfällig;
  - Wir folgen einem *algorithmisch legitimierten Vorurteil*.



https://pixabay.com/de/m%C3%A4dchen-katze-k%C3%A4tzchen-mode-1048728/

Können Algorithmen diskriminieren?

Diskriminierung bei Bewerbungen

- Lebensläufe mit „deutschen“ Namen bekommen 14% mehr Vorstellungsangebote als solche mit „türkischen“ Namen[1].
- US-amerik. Studie: Frauen mit Kopftuch erhalten weniger Jobangebote als solche ohne[2].

1 Kaas, L. & Manger, C.: “Ethnic Discrimination in Germany's Labour Market: A Field Experiment”, German Economic Review, 2011 , 13 , 1-20

2 Ghumman, S. & Ryan, A. M.: “Not welcome here: Discrimination towards women who wear the Muslim headscarf , human relations, 2013 , 66(5) , 671-698

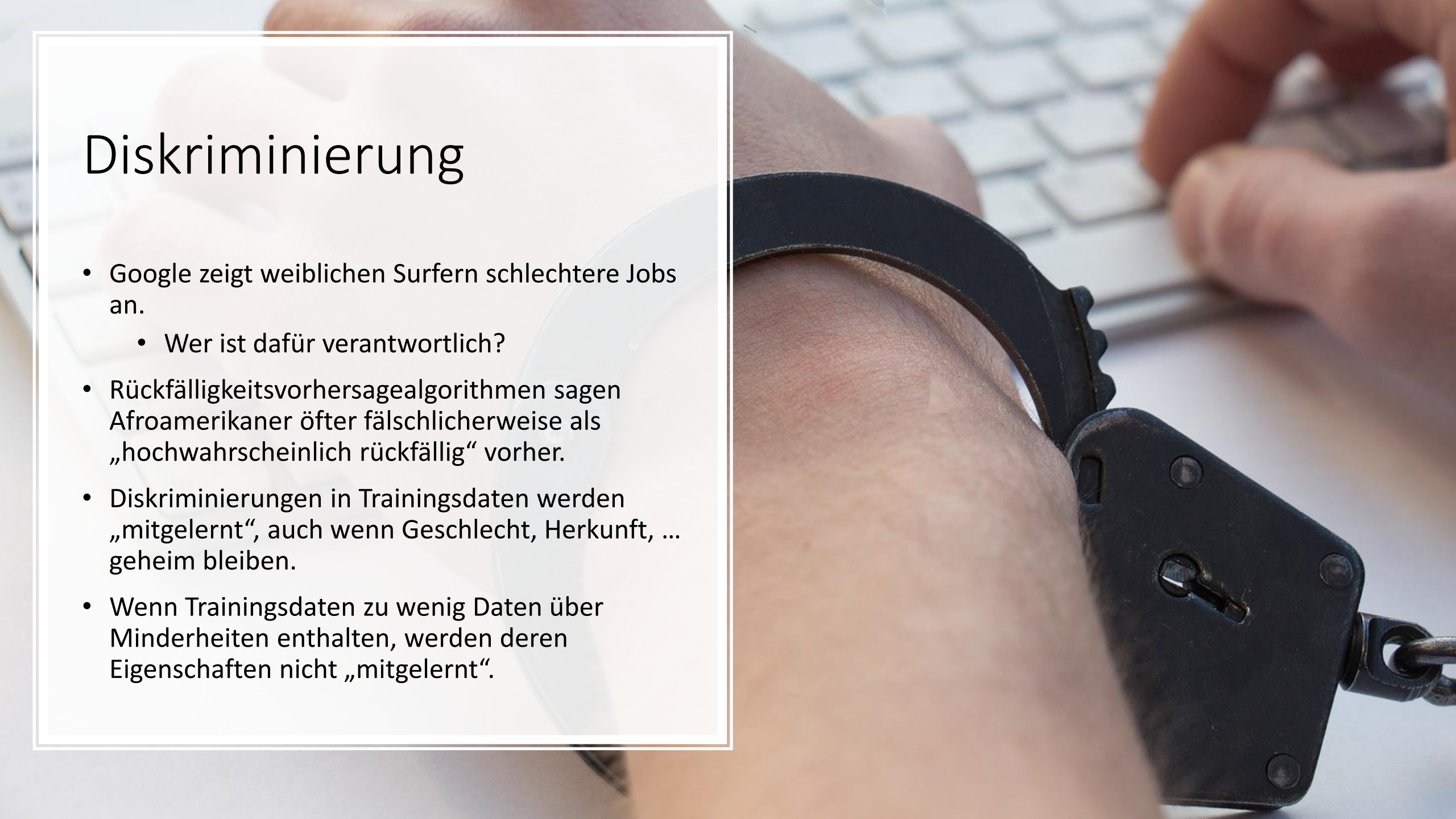
Magie der Algorithmen?

Wenn man auf Google nach „CEO“ sucht…

Und das, wenn ich auf Pixabay nach „Chef“ suche…

Diskriminierung

- Google zeigt weiblichen Surfern schlechtere Jobs an.
  - Wer ist dafür verantwortlich?
- Rückfälligkeitsvorhersagealgorithmen sagen Afroamerikaner öfter fälschlicherweise als „hochwahrscheinlich rückfällig“ vorher.
- Diskriminierungen in Trainingsdaten werden „mitgelernt“, auch wenn Geschlecht, Herkunft, … geheim bleiben.
- Wenn Trainingsdaten zu wenig Daten über Minderheiten enthalten, werden deren Eigenschaften nicht „mitgelernt“.



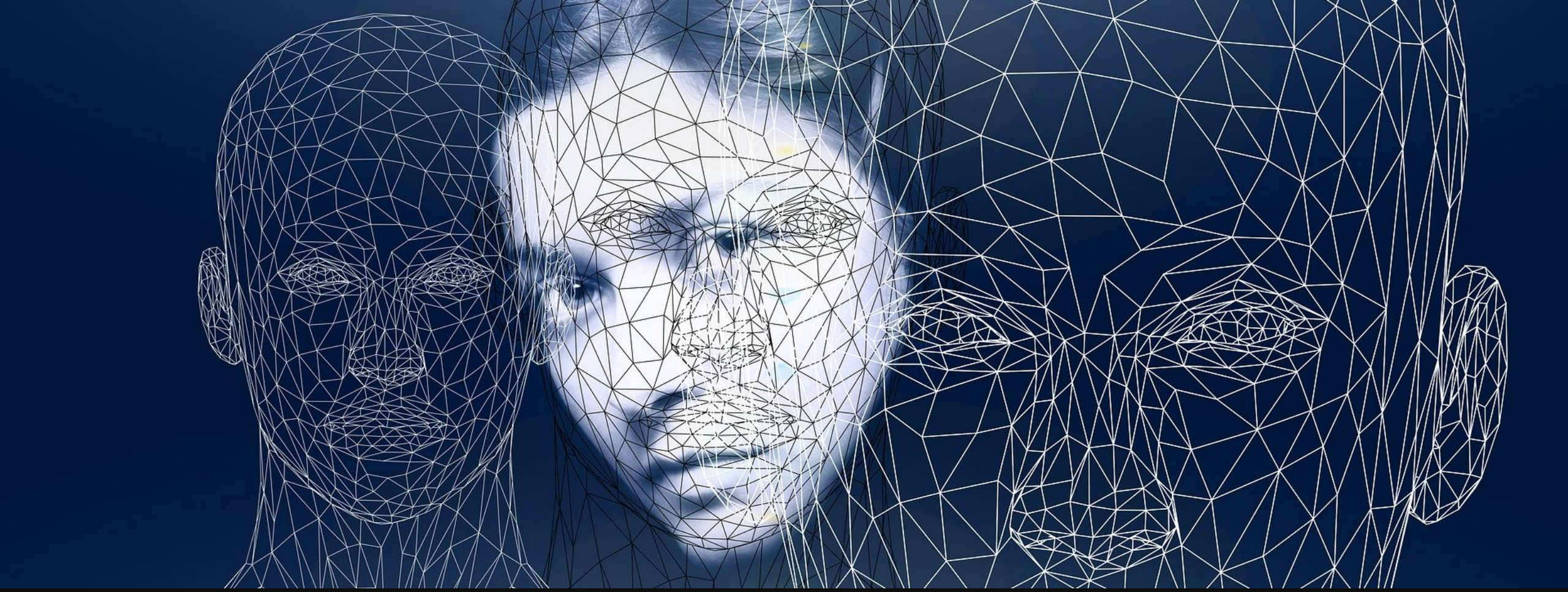
Regel

Algorithmen der künstlichen Intelligenz werden da eingesetzt, wo es **keine einfachen Regeln** gibt.

Sie suchen **Muster** in hoch-verrauschten Datensätzen.

Die Muster sind daher grundsätzlich **statistischer Natur.**

Versuchen fast immer, eine **kleine Gruppe** von Menschen zu identifizieren (Problem der **Unbalanciertheit**)



Sozio-informatische Gesamtbetrachtung

Probleme der Einbettung der ADM in den sozialen Prozess

- **Aufmerksamkeitsökonomie** von Entscheiderinnen und Entscheidern.
- **„Best practice“** erfordert Nutzung der Software.
- Eine Nichtbeachtung der Empfehlung und gleichzeitige Fehleinschätzung wirkt oft schwerer als eine Beachtung der (falschen) Empfehlung. **Delegierung von Verantwortung!**
- Manchmal kann ein(e) falsch-negativ Beurteilte(r) **die Vorhersage prinzipiell nicht entkräften!**
  - Z.B. abgelehnte Bewerberin, eingesperrte Kriminelle

Algorithmen in einer demokratischen Gesellschaft

Einschätzung

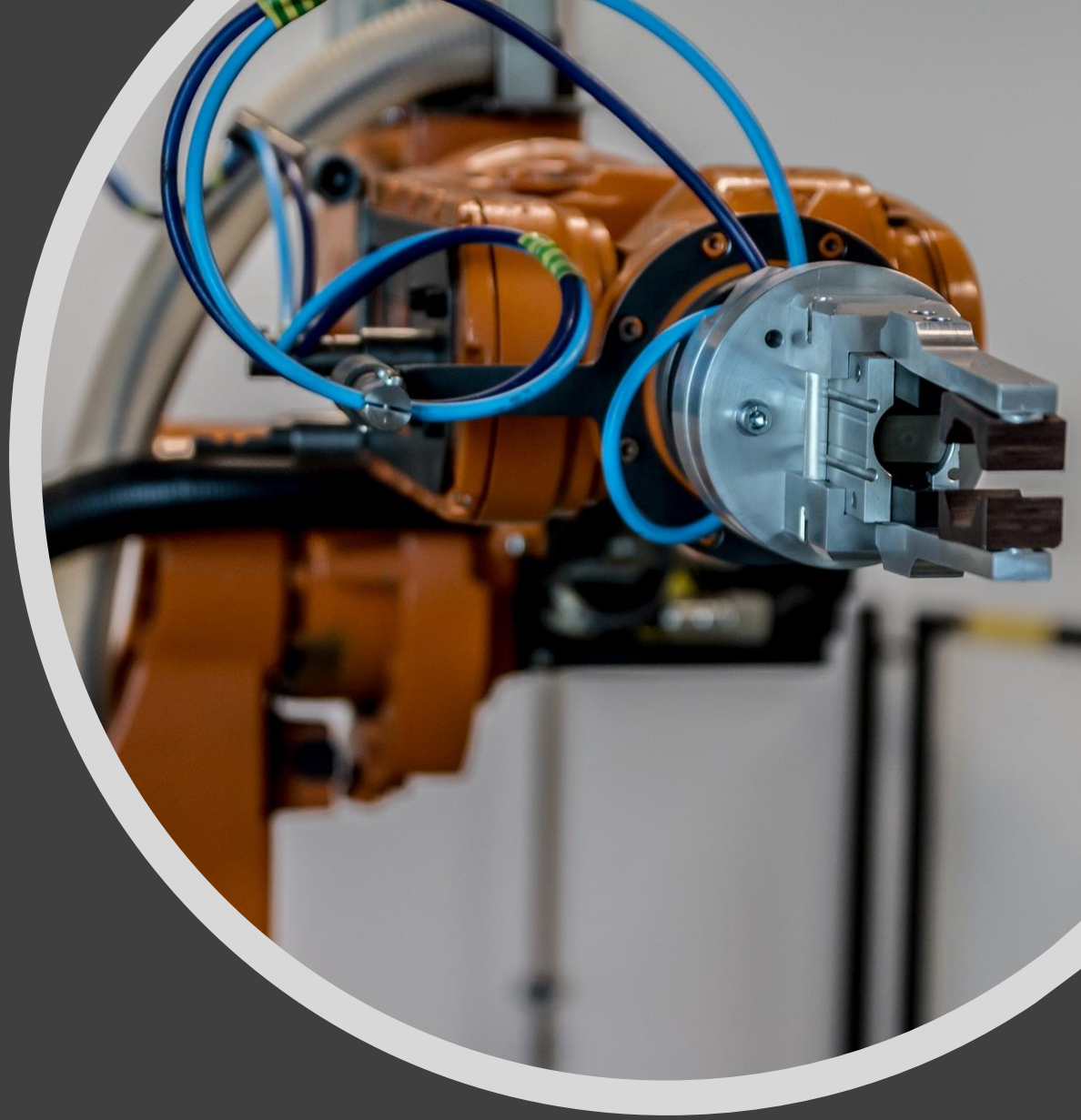
- Algorithmen **könnten** dabei helfen, bessere Entscheidungen zu treffen.
  - Sie sind zuverlässig.
  - Können Entscheidungs-wege transparenter machen.
  - Könnten Diskriminierung vermeiden.
- Allerdings sind sie heute oft noch nicht gut genug.

Probleme von algorithmischen Entscheidungssystemen (ADM Systemen) im Risk Assessment

1. **Wer entscheidet, wann ein ADM System „gut“ ist?**
2. **ADM Systeme ergeben nur Wahrscheinlichkeiten, keine Wahrheiten.**
3. **ADM Systeme können diskriminieren.**
4. **ADM Systeme können soziale Prozesse verändern.**

@NetworkTobi
Tobias Krafft
TU Kaiserslautern

Kontrolle von algorithmischen Entscheidungssystemen



Maschinelles Lernen muss um so stärker kontrolliert und reguliert werden, je höher das durch die Software mögliche individuelle und gesamtgesellschaftliche Schadenspotenzial ist.

Im Allgemeinen sind Entscheidungen über Objekte, z.B. im Produktionsprozess, nicht kritisch und bedürfen keiner Kontrolle und Regulierung auf Ebene der Software.

Einordnung auf Risikomatrix

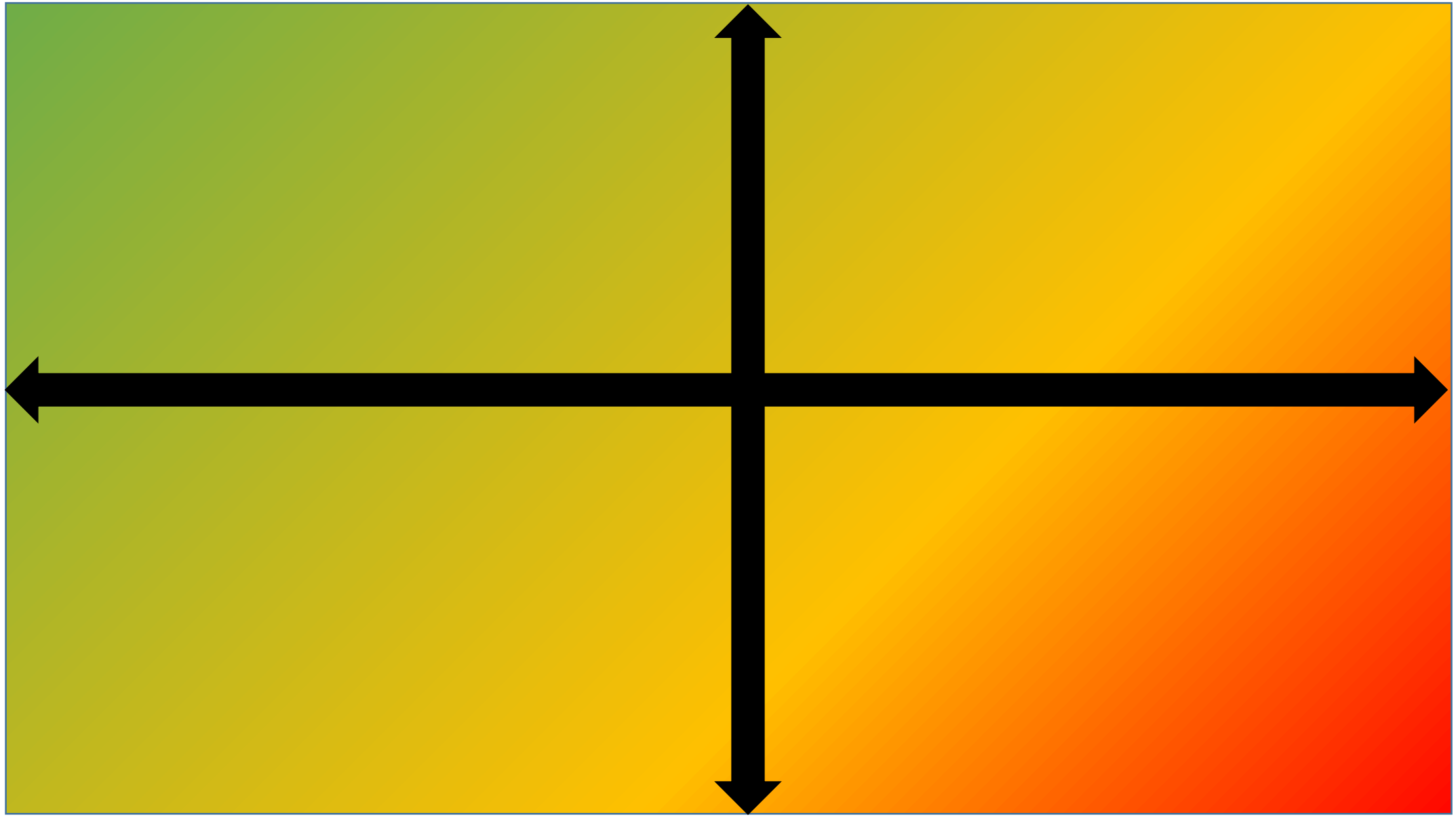
1. Schadenstiefe

$$\sum \begin{array}{c} \text{Schaden für Individuum(Fehlurteil)} \\ +\text{Schaden für Gesellschaft(Fehlurteil)} \end{array}$$

2. Anbietervielzahl, Wechselmöglichkeiten, Möglichkeiten der Anfechtbarkeit, Revisionen durch Menschen, etc.

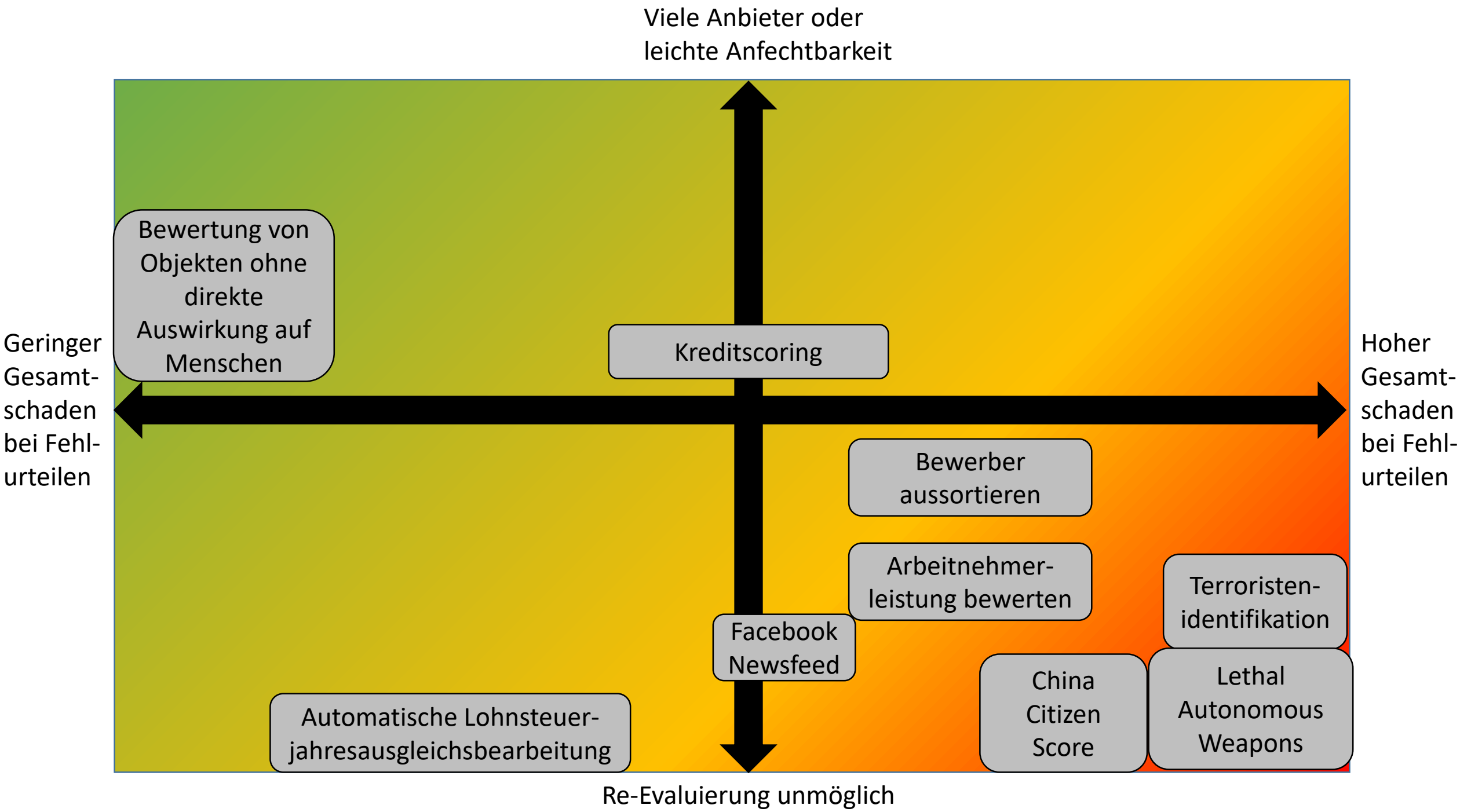
Viele Anbieter oder
leichte Anfechtbarkeit

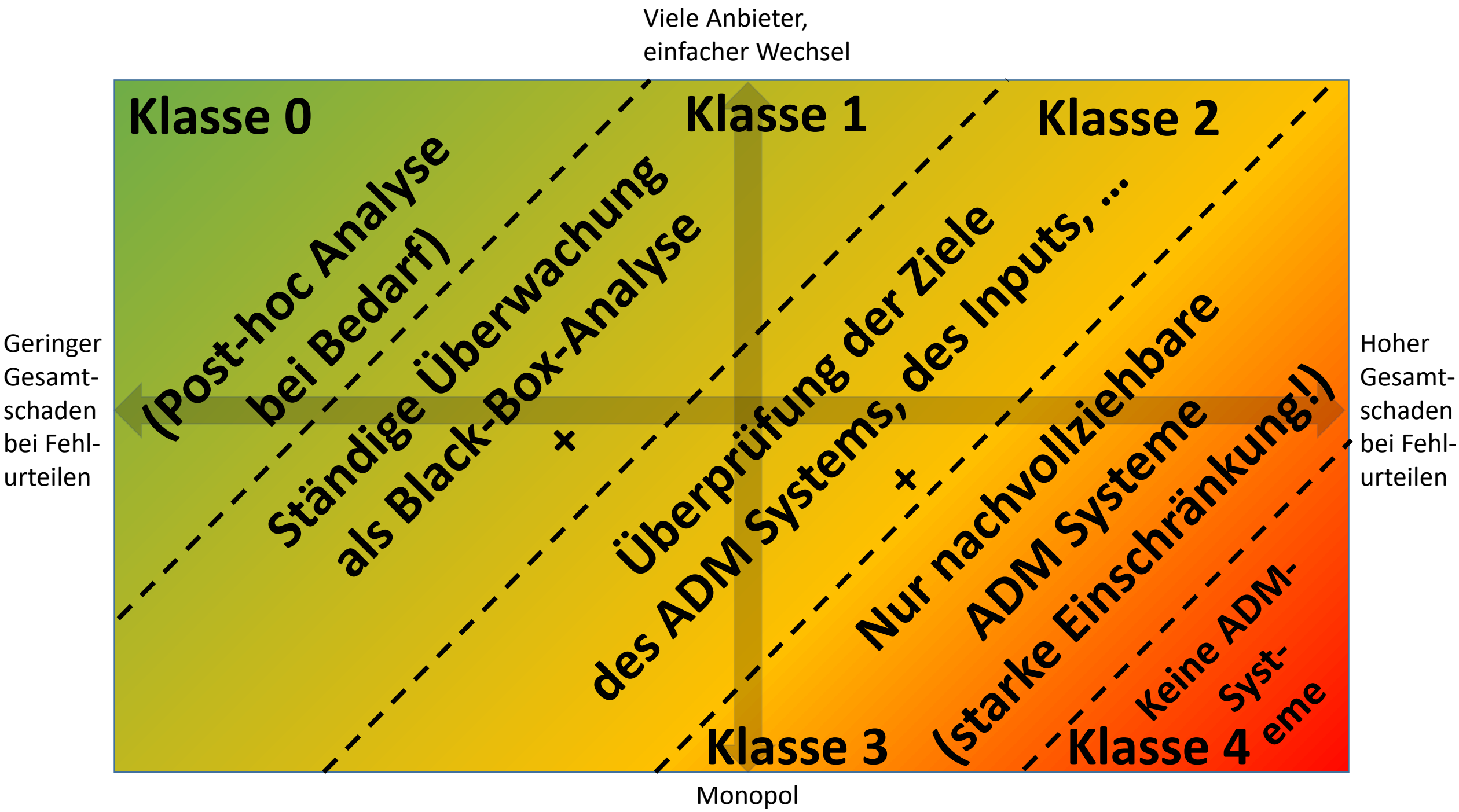
Geringer Gesamt-schaden bei Fehl-urteilen



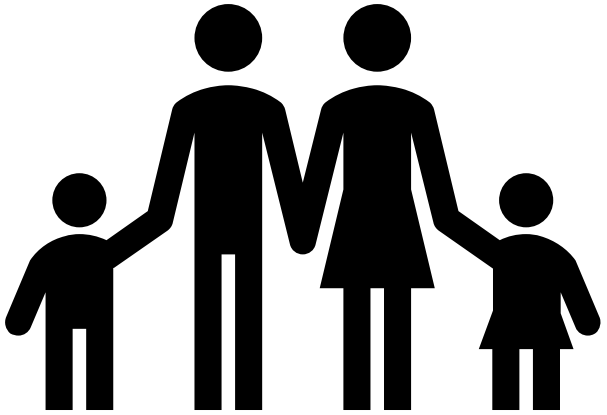
Hoher Gesamt-schaden bei Fehl-urteilen

Re-Evaluierung unmöglich

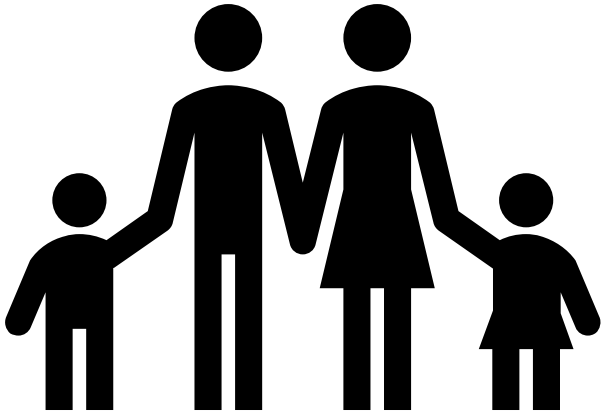

Viele Anbieter oder
leichte Anfechtbarkeit
Geringer Gesamt-schaden bei Fehl-urteilen
Hoher Gesamt-schaden bei Fehl-urteilen
Re-Evaluierung unmöglich
Bewertung von Objekten ohne direkte Auswirkung auf Menschen
Kreditscoring
Bewerber aussortieren
Arbeitnehmer-leistung bewerten
Terroristen-identifikation
Facebook Newsfeed
China Citizen Score
Lethal Autonomous Weapons
Automatische Lohnsteuer-jahresausgleichsbearbeitung


Viele Anbieter,
einfacher Wechsel
Klasse 0
Klasse 1
Klasse 2
(Post-hoc Analyse bei Bedarf)
Ständige Überwachung als Black-Box-Analyse
+
Überprüfung der Ziele des ADM Systems, des Inputs, …
+
Nur nachvollziehbare ADM Systeme (starke Einschränkung!)
Keine ADM-Syst-eme
Klasse 3
Klasse 4
Geringer Gesamt-schaden bei Fehl-urteilen
Hoher Gesamt-schaden bei Fehl-urteilen
Monopol

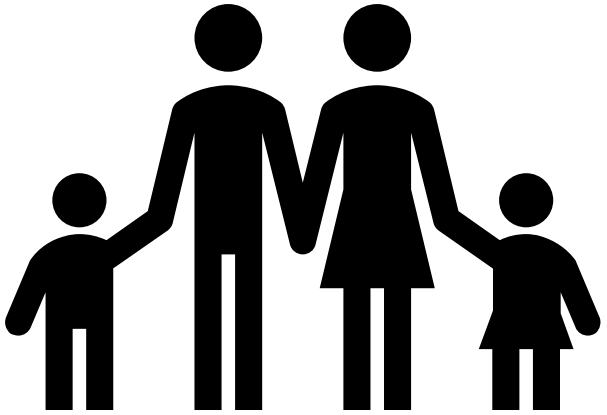
Der Fall YouTube

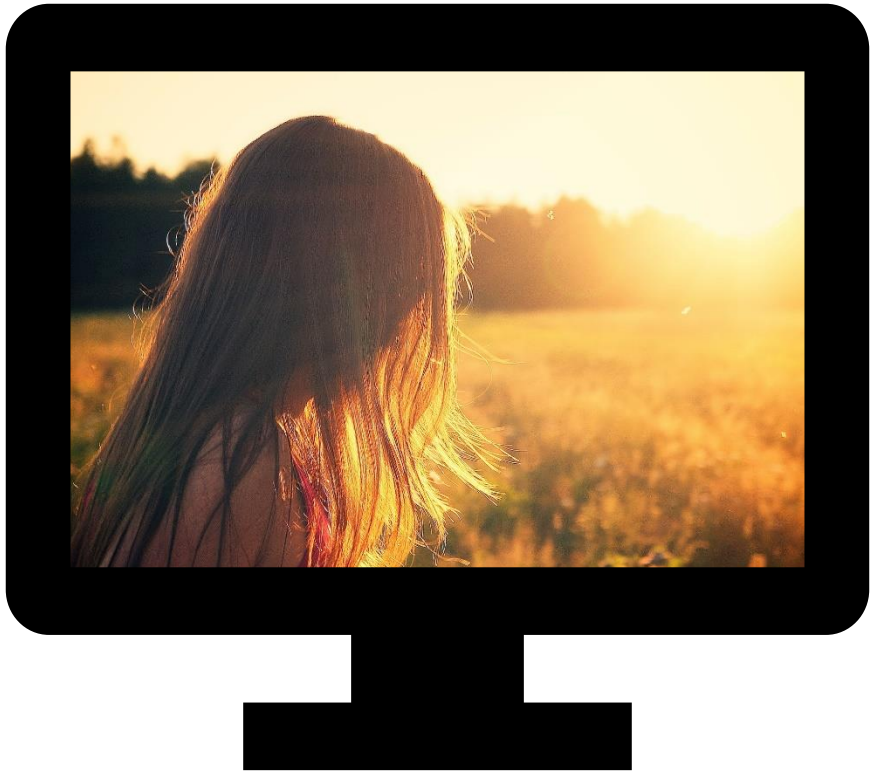


Der Fall YouTube



Der Fall YouTube

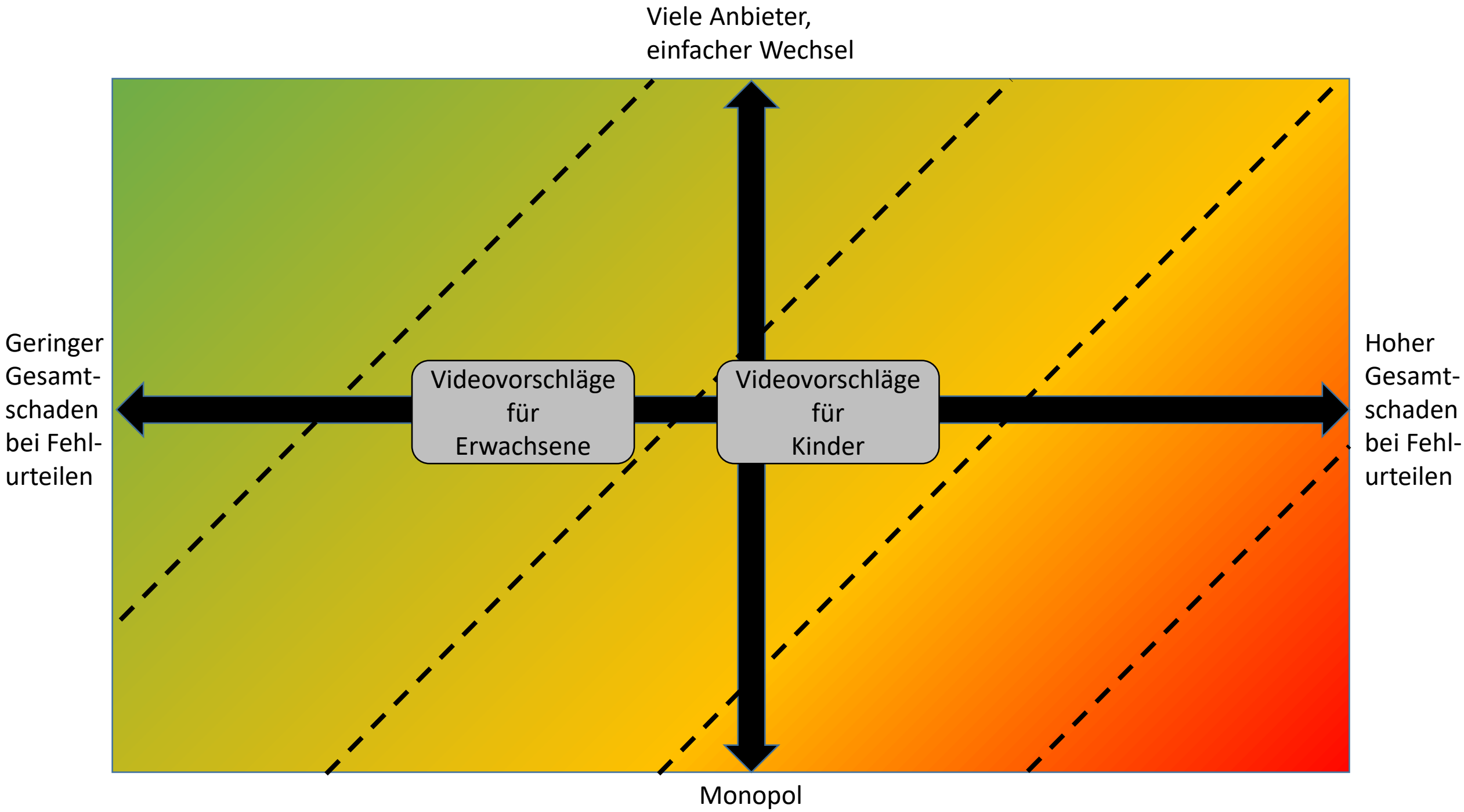




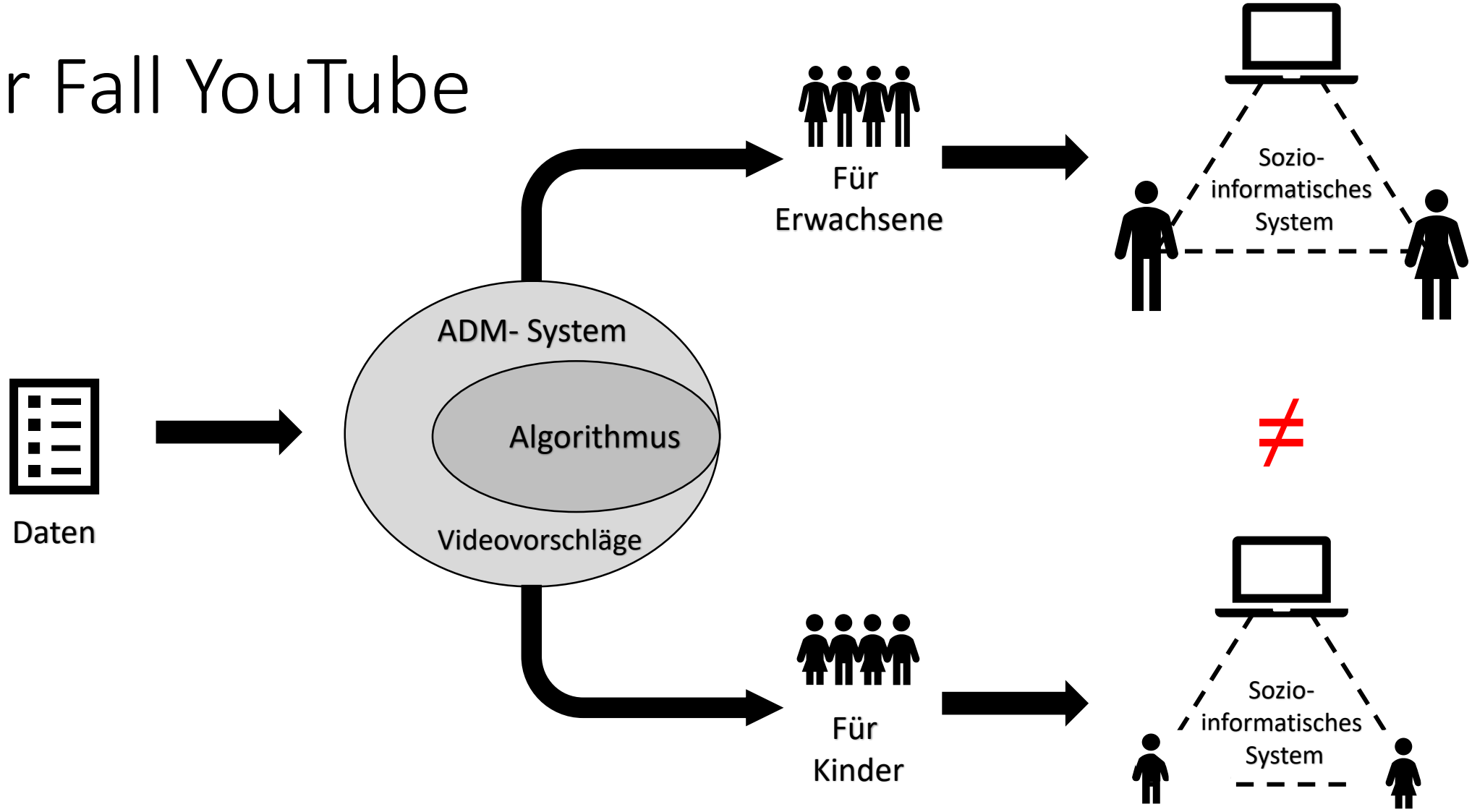
Der Fall YouTube

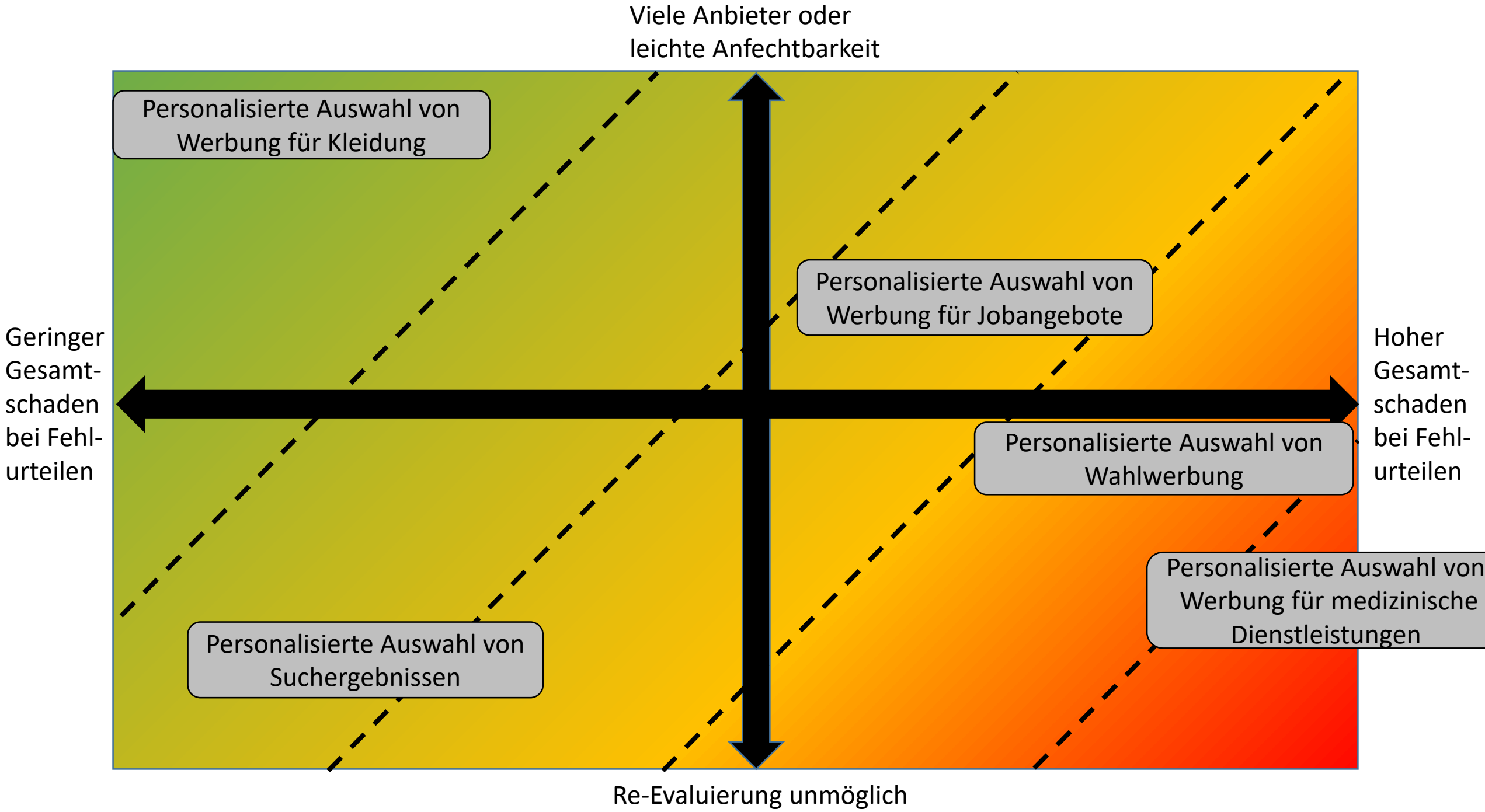
Der Fall YouTube

Der Fall YouTube


Viele Anbieter,
einfacher Wechsel
Geringer Gesamt-schaden bei Fehl-urteilen
Videovorschläge für Erwachsene
Videovorschläge für Kinder
Hoher Gesamt-schaden bei Fehl-urteilen
Monopol

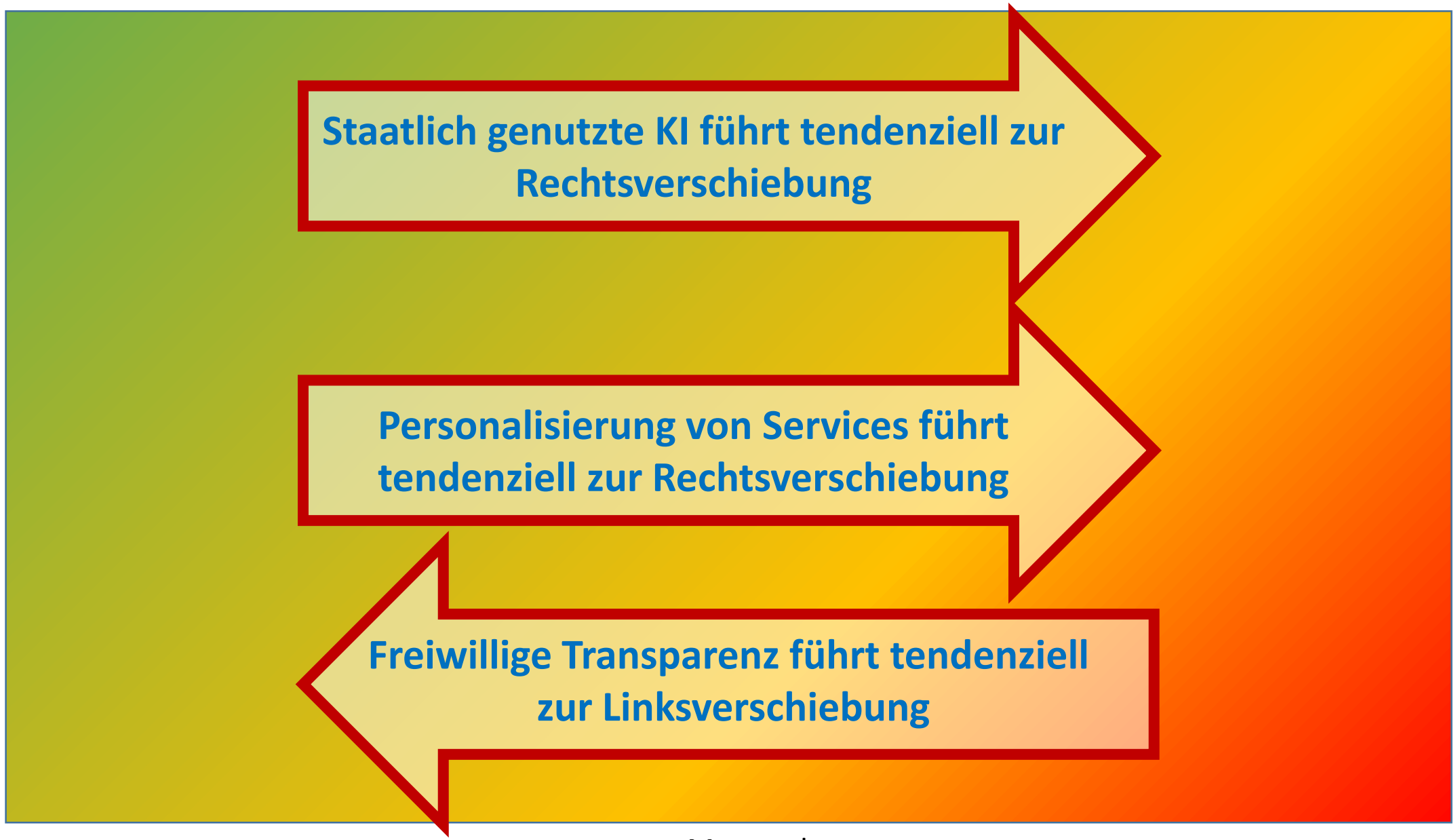
Der Fall YouTube




Viele Anbieter oder
leichte Anfechtbarkeit
Personalisierte Auswahl von Werbung für Kleidung
Personalisierte Auswahl von Werbung für Jobangebote
Geringer Gesamt-schaden bei Fehl-urteilen
Hoher Gesamt-schaden bei Fehl-urteilen
Personalisierte Auswahl von Wahlwerbung
Personalisierte Auswahl von Werbung für medizinische Dienstleistungen
Personalisierte Auswahl von Suchergebnissen
Re-Evaluierung unmöglich

Merke:

Viele Anbieter,
einfacher Wechsel



Geringer Gesamtschaden bei Fehlurteilen

Hoher Gesamtschaden bei Fehlurteilen

Monopol

Wie kommt die Ethik in den Rechner?

Über Sie,
über mich,
über uns.

Weitere Informationen

- Studie für die Bertelsmann-Stiftung: Zweig, Fischer & Lischka: „Wo Maschinen irren können“ (Serie AlgoEthik, No. 4, 2018)
- Zwei Kapitel im Sammelband (Un)Berechenbar? des Fraunhofer FOKUS, Kompetenzzentrum ÖFIT, 2018
  - Zweig & Krafft: „Fairness und Qualität algorithmischer Entscheidungen“
  - Krafft & Zweig: „Wie Gesellschaft algorithmischen Entscheidungen auf den Zahn fühlen kann“
- Studie für den Bundesverband der Verbraucherzentralen und Verbraucherverbände: Krafft & Zweig: “Transparenz und Nachvollziehbarkeit algorithmenbasierter Entschiedungsprozesse” eine
- Studie vom Fraunhofer FOKUS, Kompetenzzentrum Öffentliche IT (ÖFIT): Opiela, Mohabbat Kar, Thapa & Weber: “Exekutive KI 2030 – Vier Zukunftsszenarien für Künstliche Intelligenz in der öffentlichen Verwaltung”

verbraucherzentrale
Bundesverband