Können Algorithmen diskriminieren?

ALGORITHM ACCOUNTABILITY LAB

Wie funktioniert sie, wo ist sie hilfreich, wo sollte sie nicht eingesetzt werden?

Prof. Dr. Katharina A. Zweig,
TU Kaiserslautern

Konstituierende Sitzung der Enquete-Kommission „Künstliche Intelligenz“ am 27.9.

Aus der Rede von Bundestagspräsidenten Dr. Schäuble:

- „Die künstliche Intelligenz gilt Vielen als neue Zauberformel des technischen Fortschritts, …
- … sie wird dichten, …
- … sie wird belohnen und bestrafen …“

https://www.bundestag.de/gebaerdensprache/mediathek?videoid=7272804#url=bWVkaWF0aGVrb3ZlcmxheT92aWRlb2lkPTcyNzI4MDQ=&mod=mediathek

Die zwei Ängste

Sie wird dichten

Sie wird richten

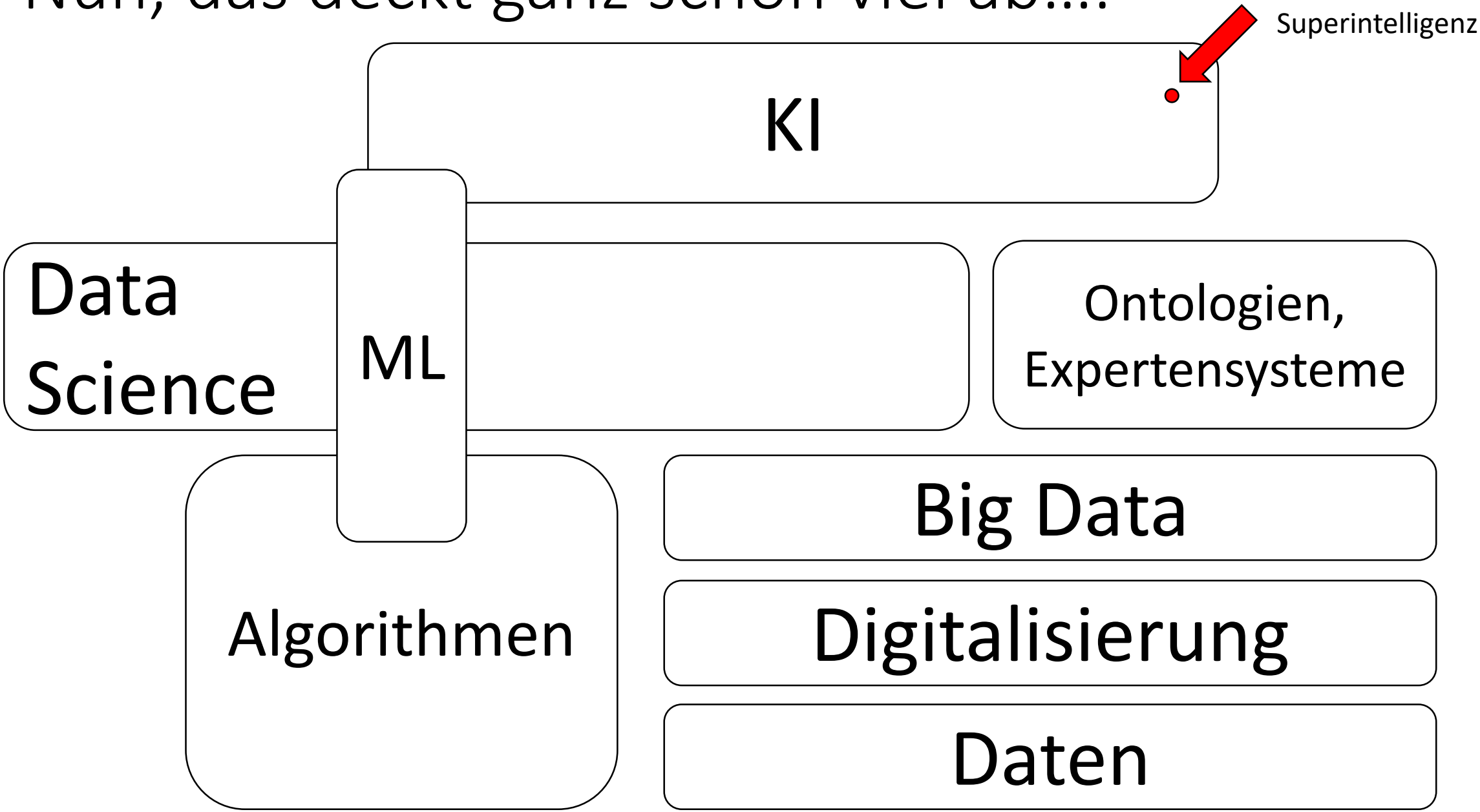
Aber was ist überhaupt KI?

SNV, „Eckpunkte einer nationalen Strategie für Künstliche Intelligenz“:

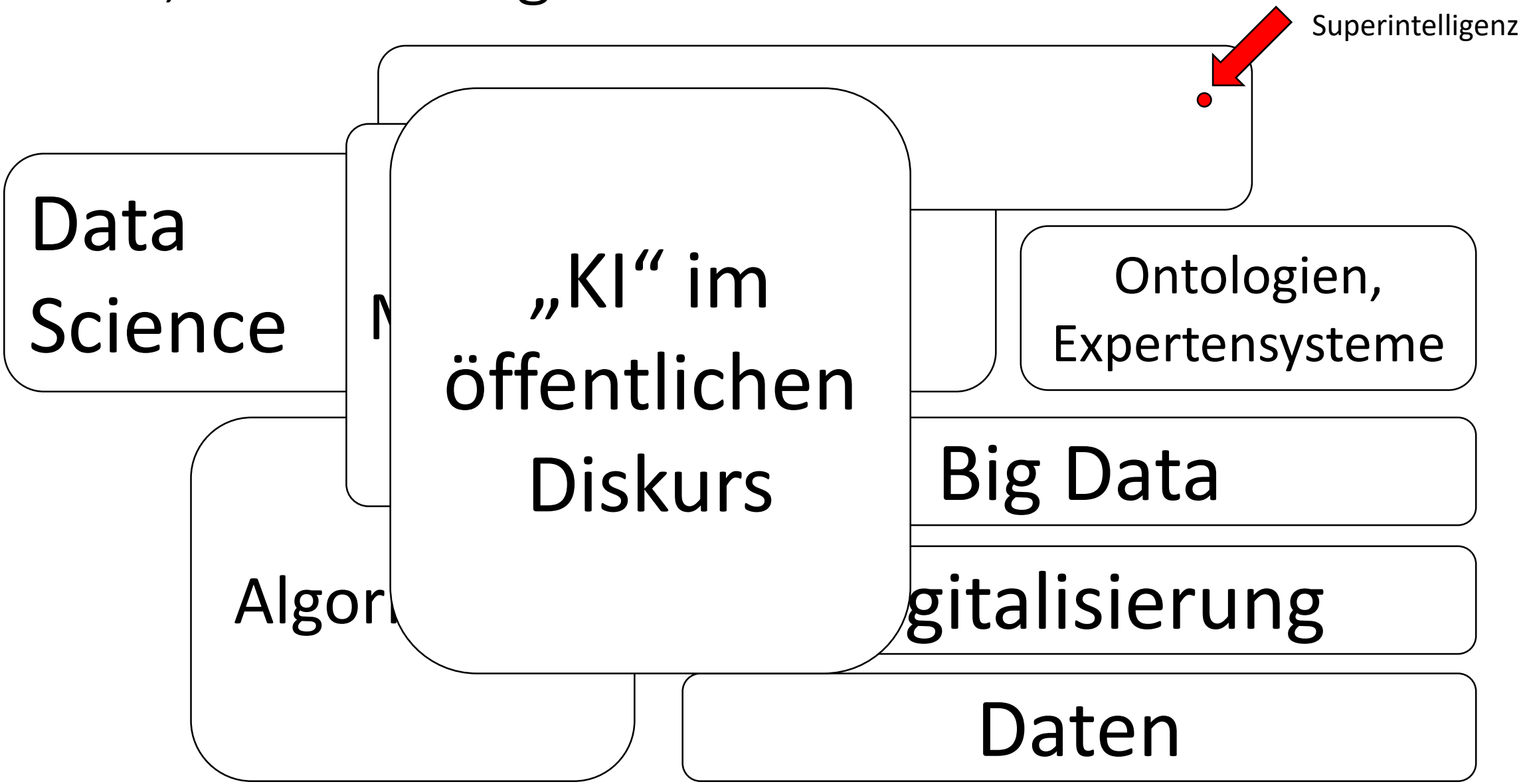
„(…) Verfahren (…), welche die Übertragung von bislang menschlich getroffenen Entscheidungen, Bewertungen und Handlungen auf Computer und Maschinen erlauben.

https://www.stiftung-nv.de/sites/default/files/ki_strategie.pdf

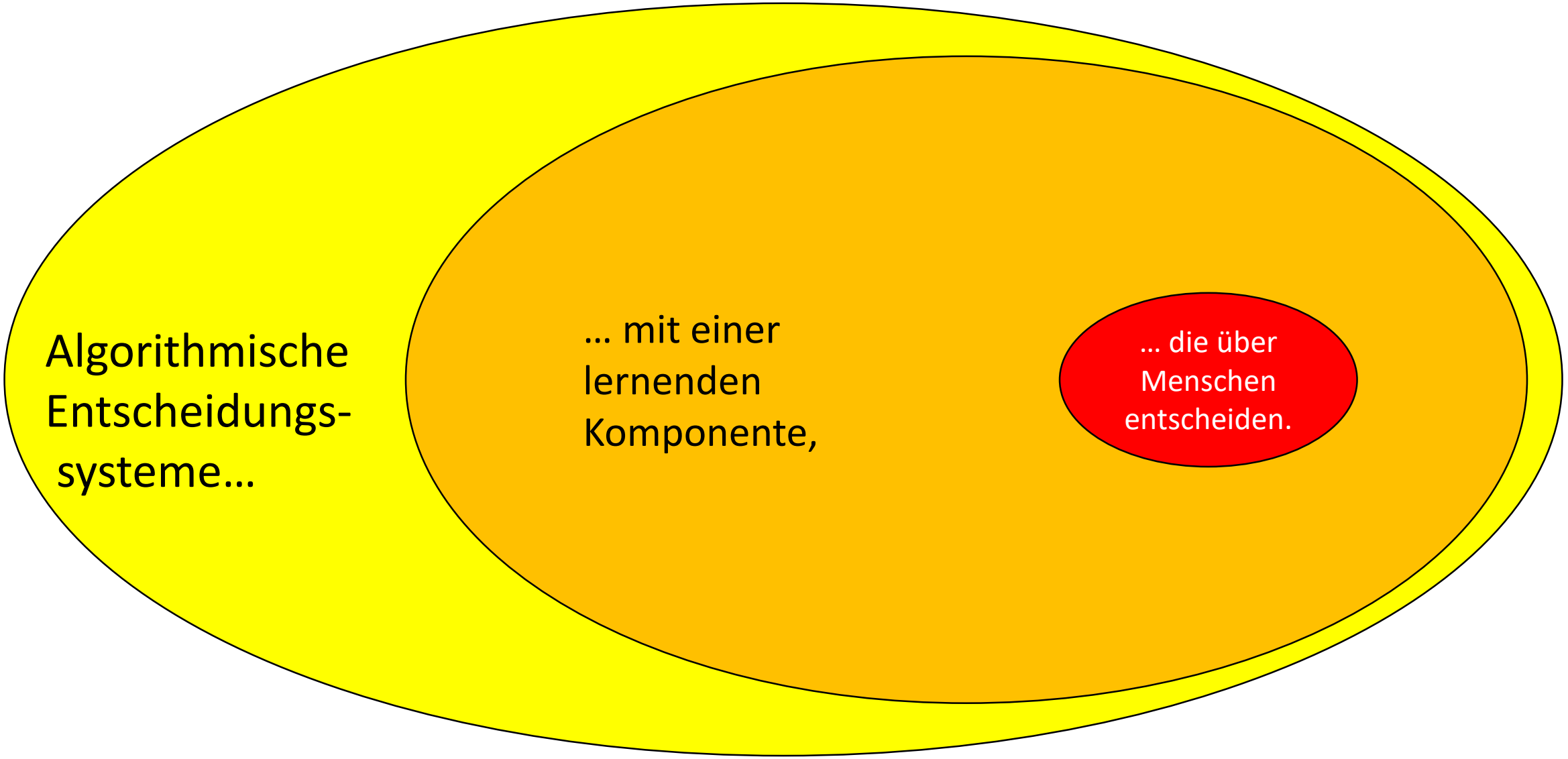
Nun, das deckt ganz schön viel ab....



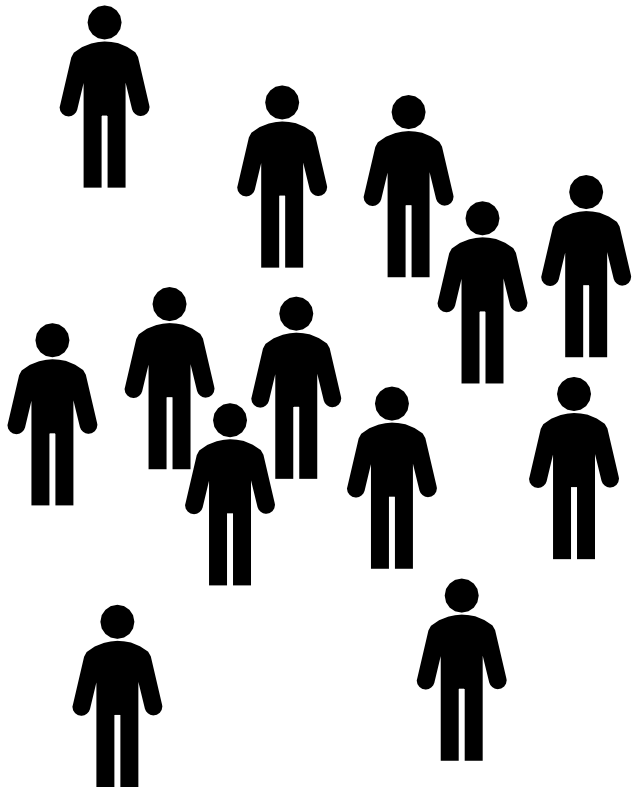
Nun, das deckt ganz schön viel ab….

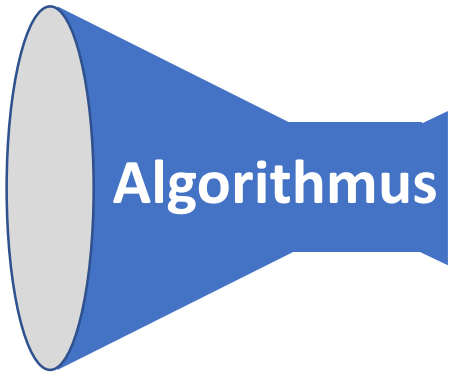


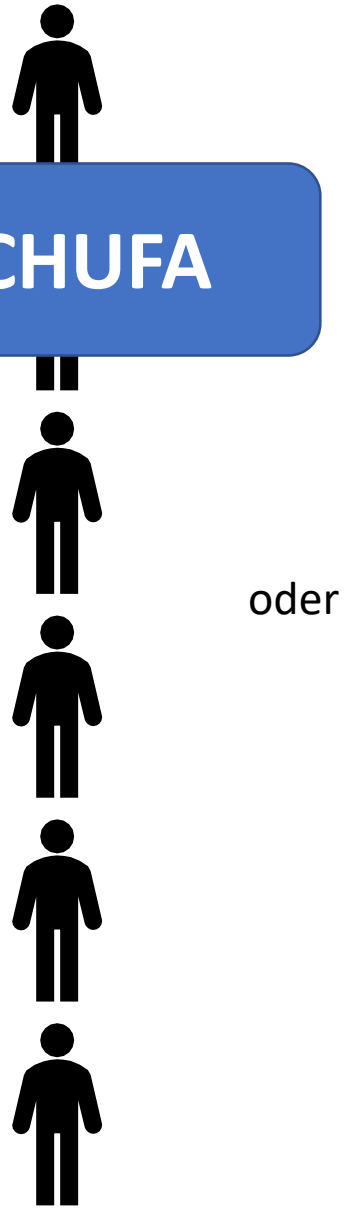
Welche algorithmischen Entscheidungssysteme (ADM Systeme) sind problematisch?



Algorithmische Entscheidungssysteme (ADM Systeme)





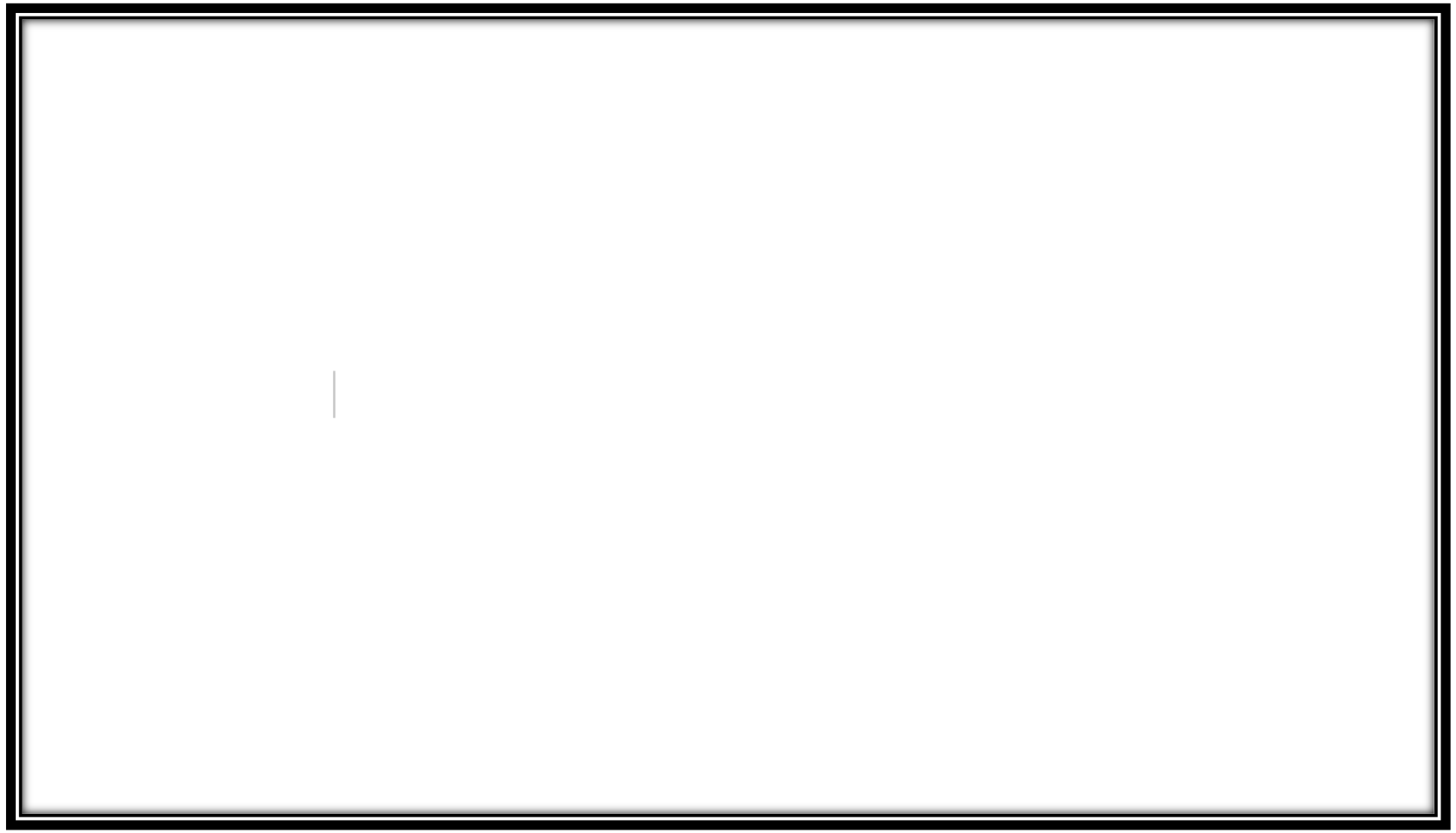


Scoring-Verfahren

oder

Klassifikation

Kann KI HR?



Diskriminierung bei Bewerbungen

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

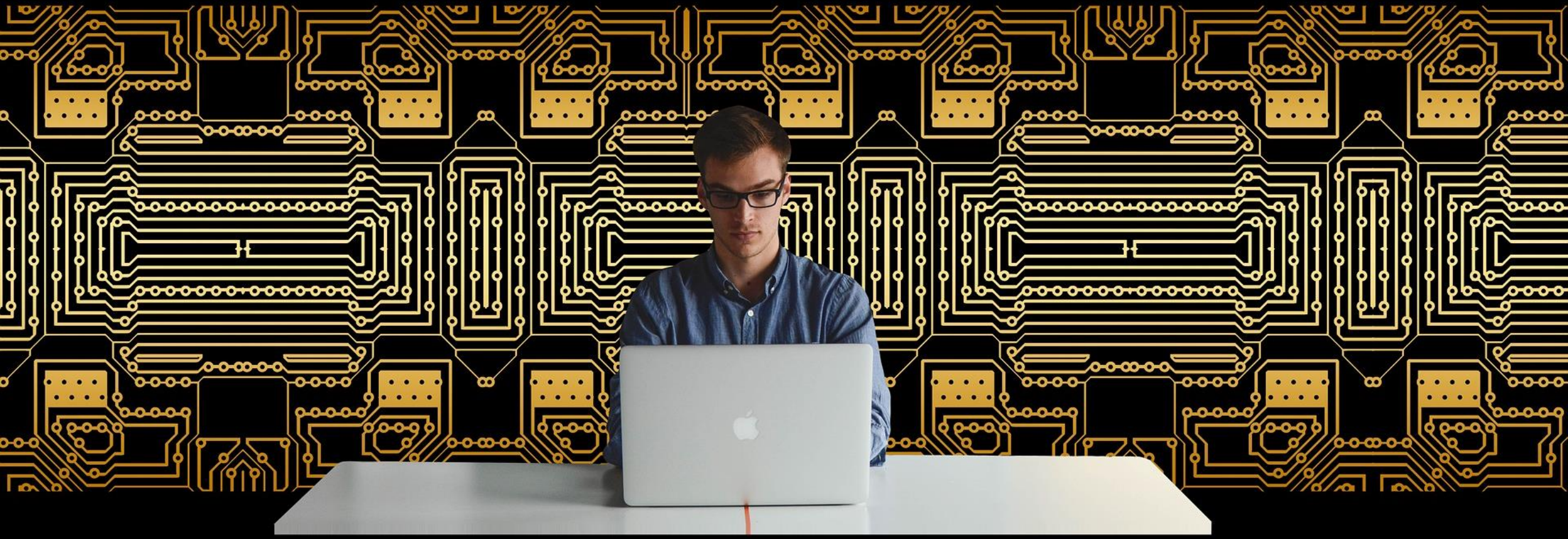
1 Kaas, L. & Manger, C.: “Ethnic Discrimination in Germany's Labour Market: A Field Experiment”, German Economic Review, 2011 , 13 , 1-20

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

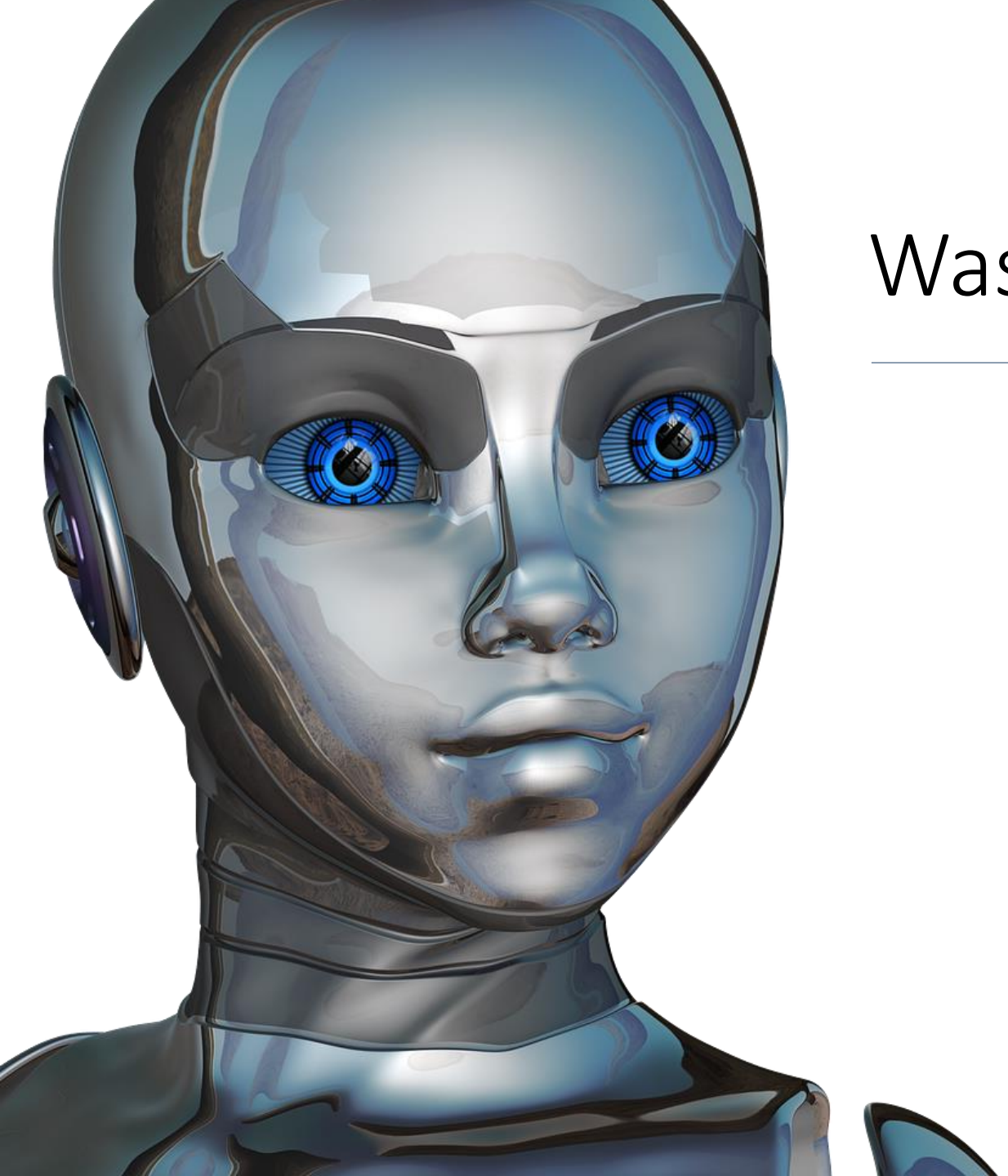
Könnten Computer das besser?

- Die ersten Firmen testen *algorithmische Entscheidungssysteme* (oder Entscheidungs-unterstützungssysteme)[1].
- Eigenschaften, nach denen nicht diskriminiert werden darf, können vor ihnen besser verborgen werden.
- Sie sind objektiv und arbeiten nahezu fehlerfrei.
- (objektiv := „reproduzierbar dieselbe Entscheidung bei derselben Eingabe von Daten“)

1 Claire Miller: “Can an Algorithm hire Better than a Human?”, The New York Times, June 25, 2015, https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html



Können Computer lernen?



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 „heiss“ und „warm“

Juli **März** **September**

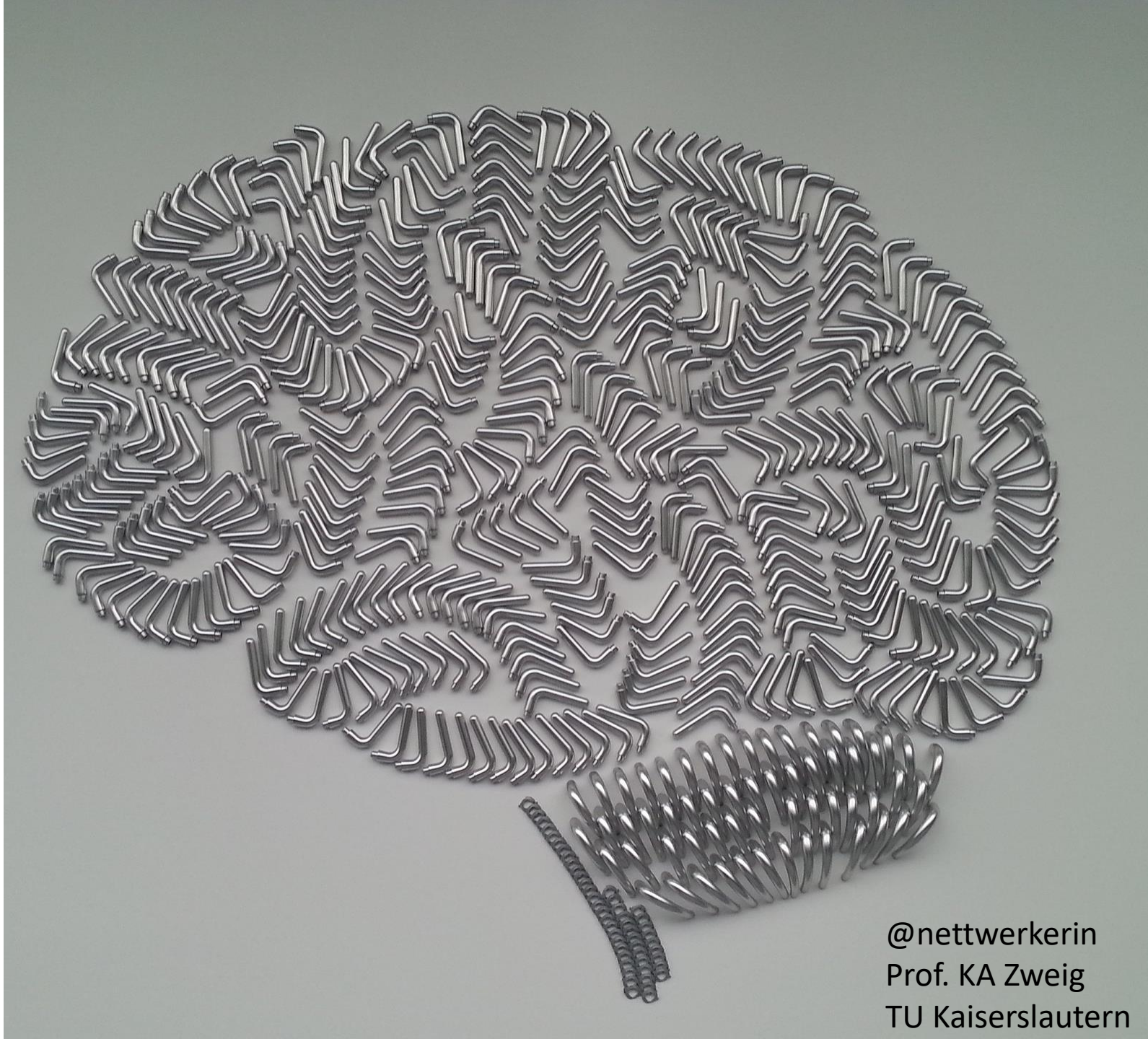
Zu vorsichtig: Darf nicht dampfen Zu mutig geworden

Alles muss kalt sein

@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

Sebastian lernt…

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



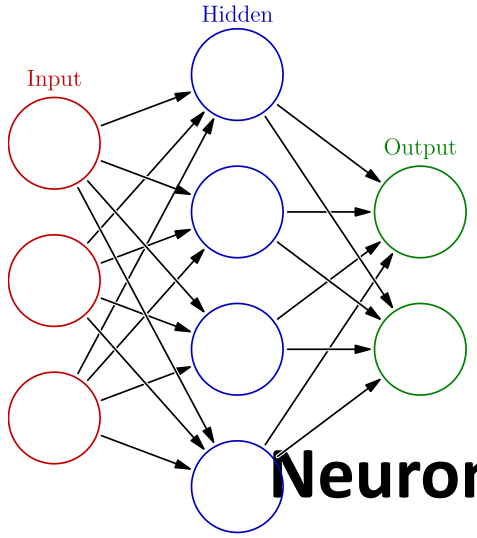
@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

Computer lernen

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

Optimal auch **Rückkopplung**.

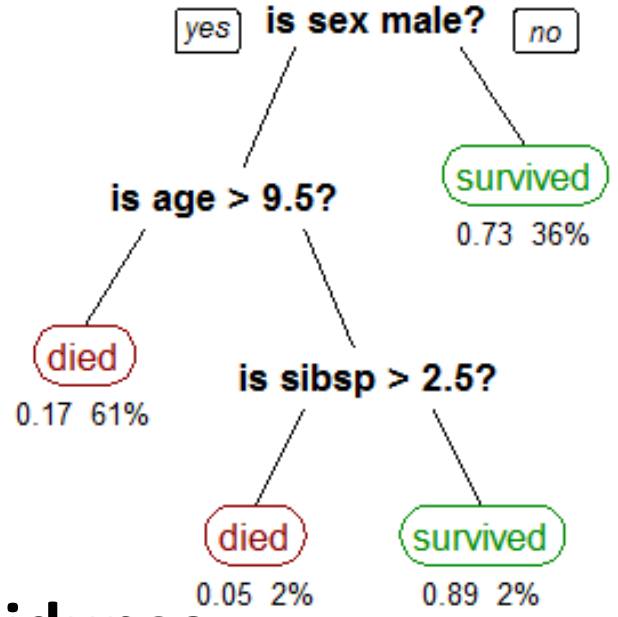
Er lernt **generelle Regeln**.



Neuronales Netz

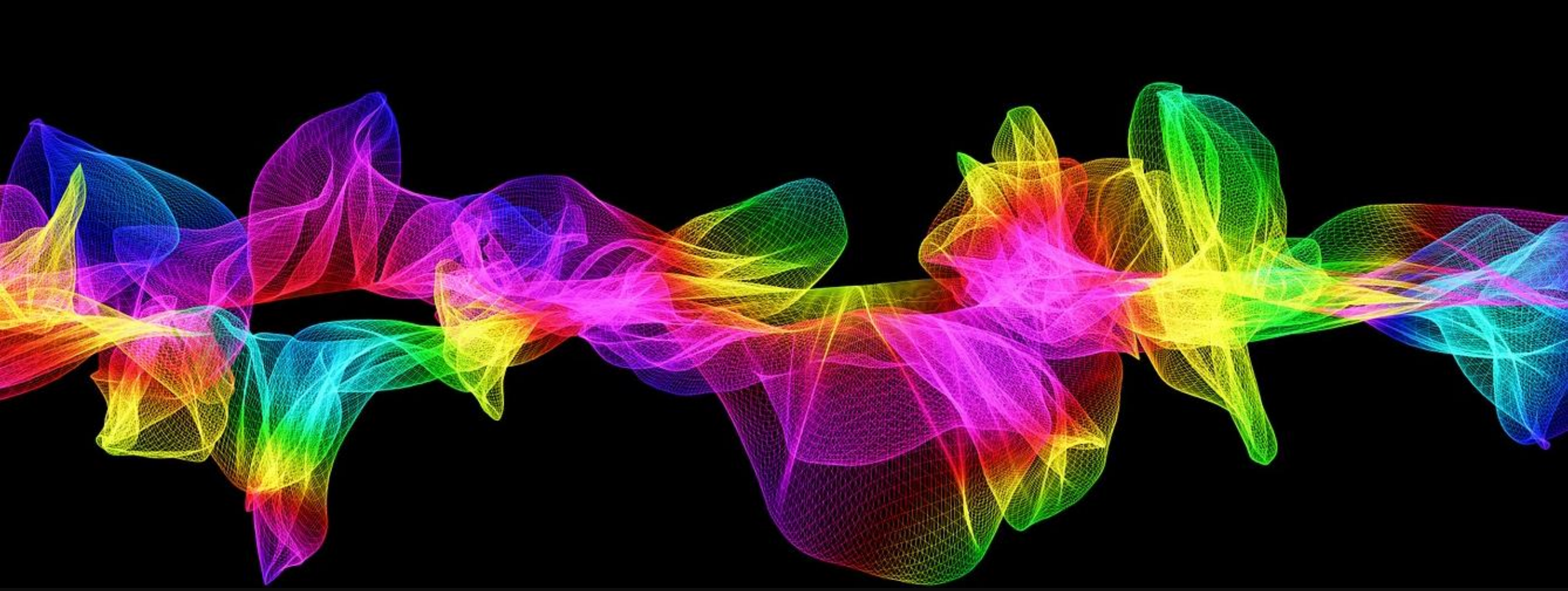
Formel

$$w_1 * \#Vh - w_2 * \#day_1 Vh + w_3 * I[g = male] * 1 + w_4 * I[T = R] * 1.0 + \cdots$$



Entscheidungs-bäume

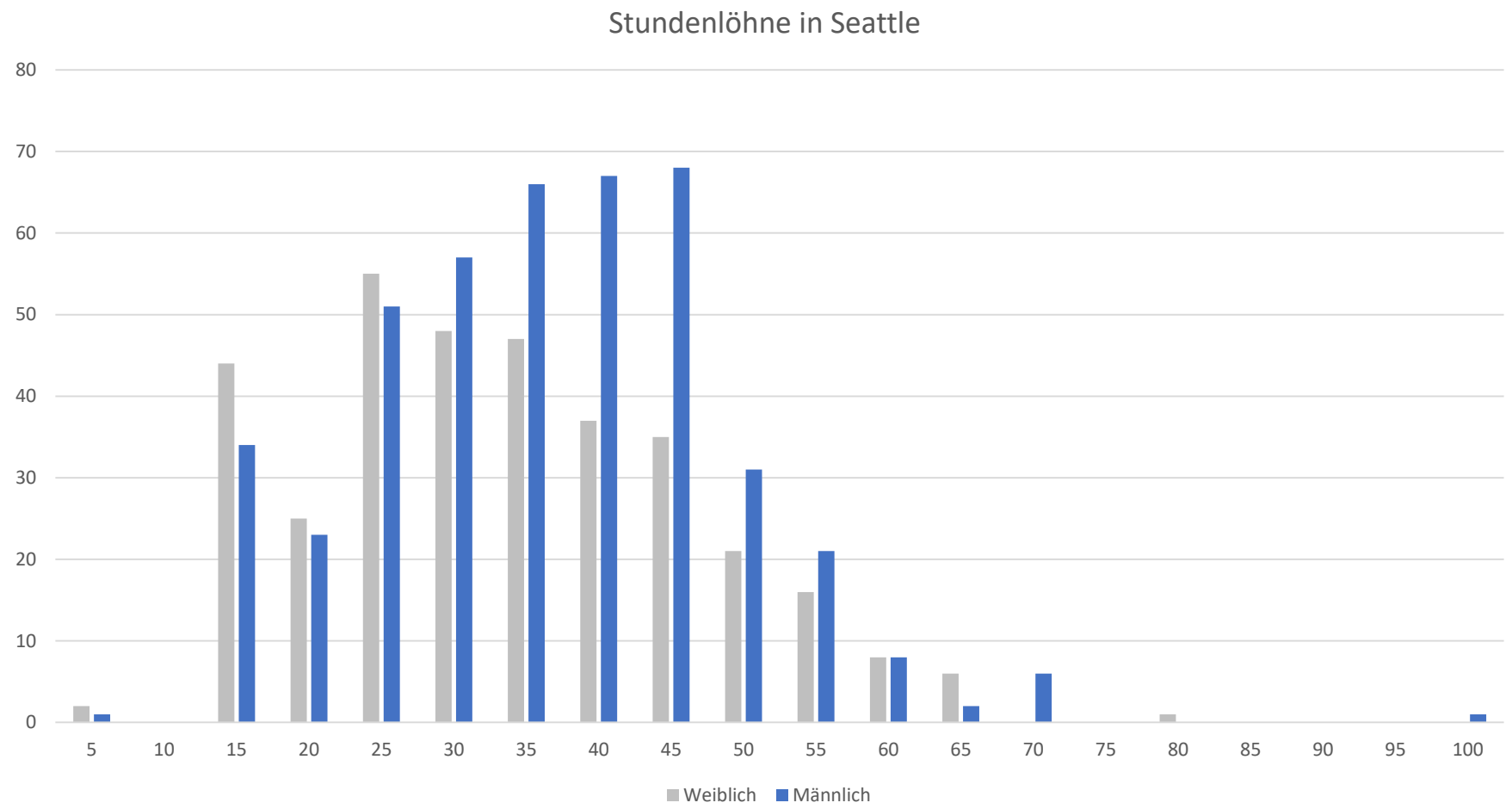
@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern



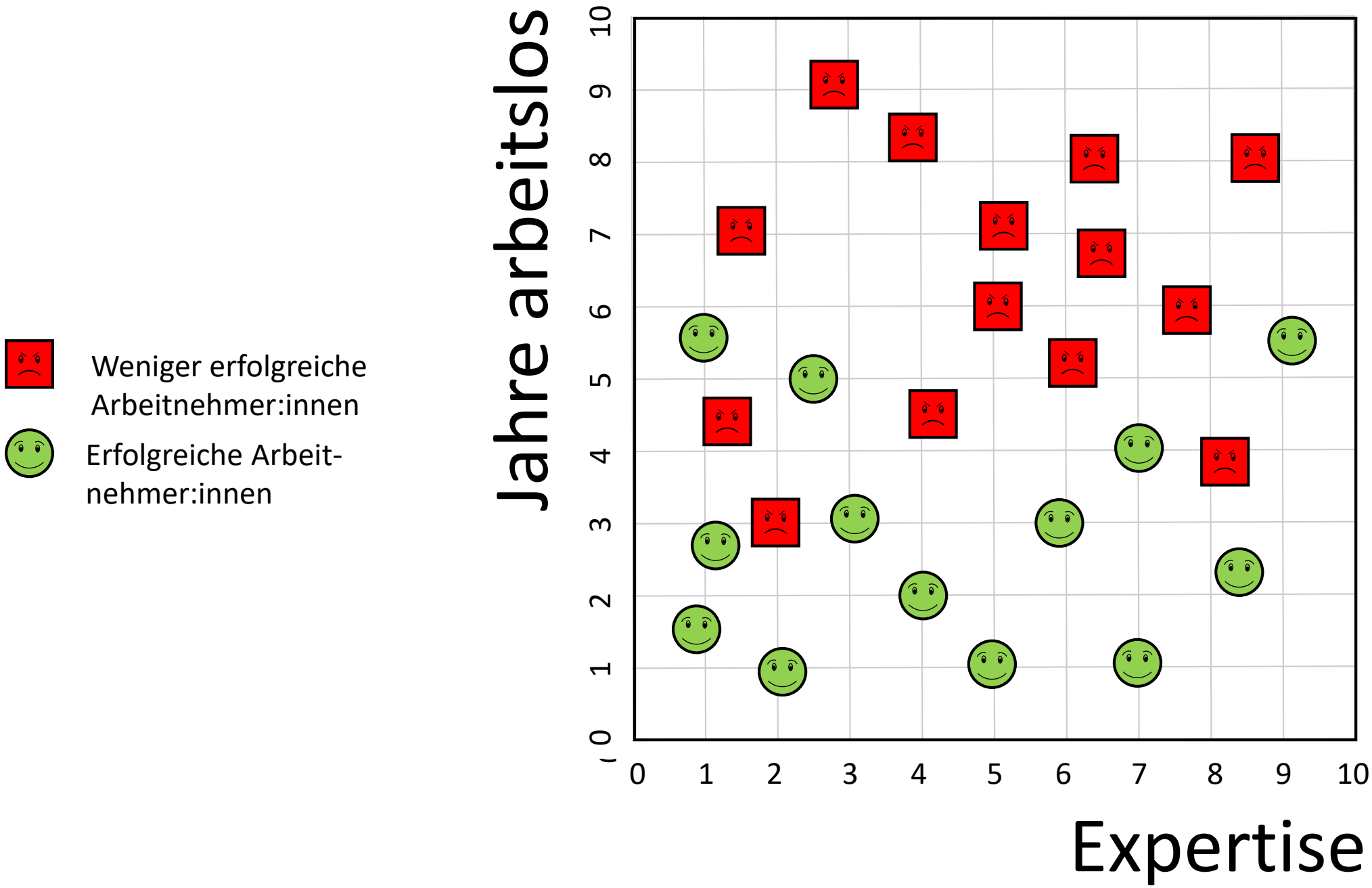
“Lernen” mit Korrelationen

Heißen Sie unsere(n) neue(n) Mitarbeiter(in) willkommen!

- Anteil weiblicher Angestellter?
  - 44%
- Anteil weiblicher Angestellter mit Lohn unter $25?
  - 55%



“Lernen” mit SVMs


Weniger erfolgreiche Arbeitnehmer:innen
Erfolgreiche Arbeit-nehmer:innen
Jahre arbeitslos
0
1
2
3
4
5
6
7
8
9
10
Expertise

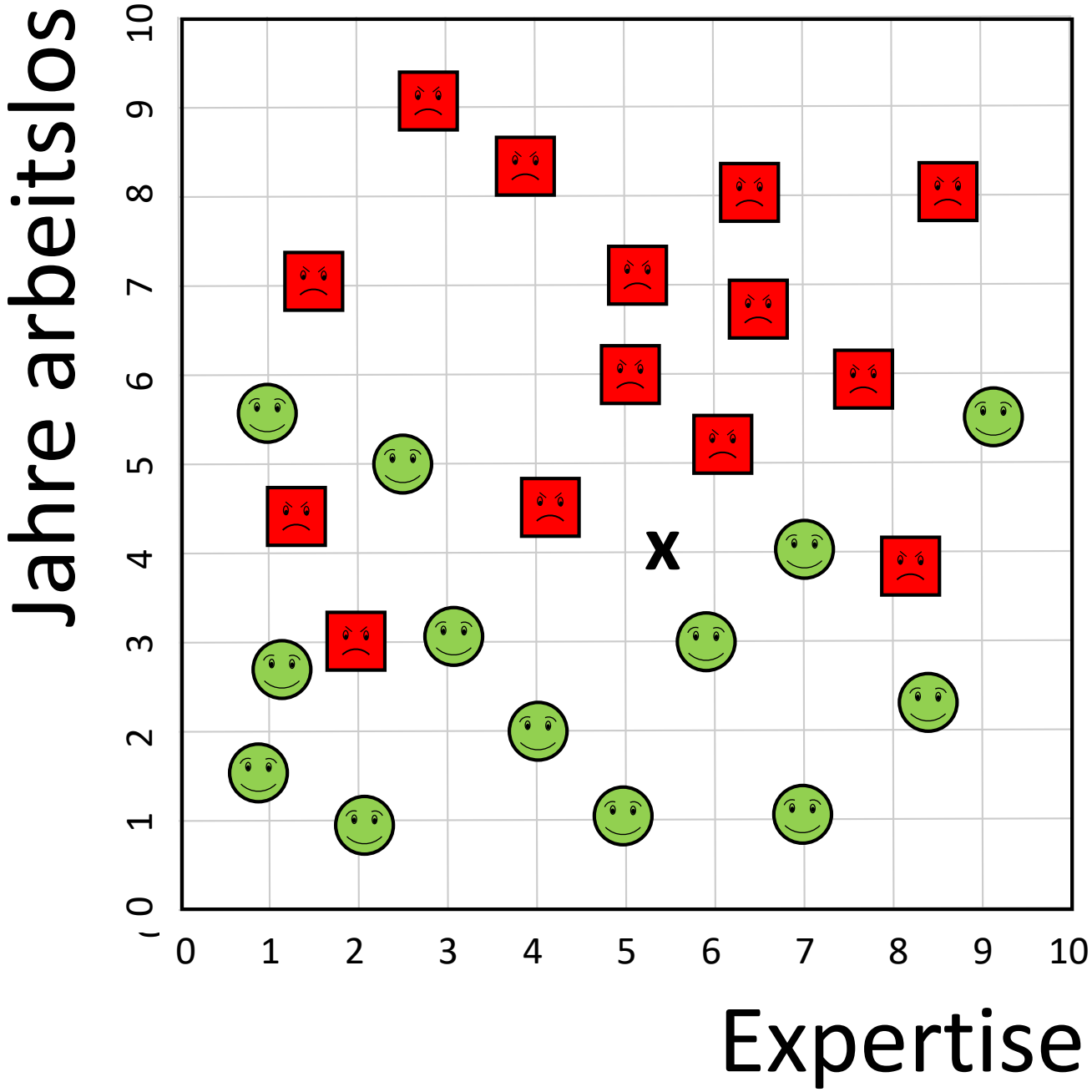
Weniger erfolgreiche Arbeitnehmer:innen

Erfolgreiche Arbeit-nehmer:innen

Bewerten Sie Frau Müller:

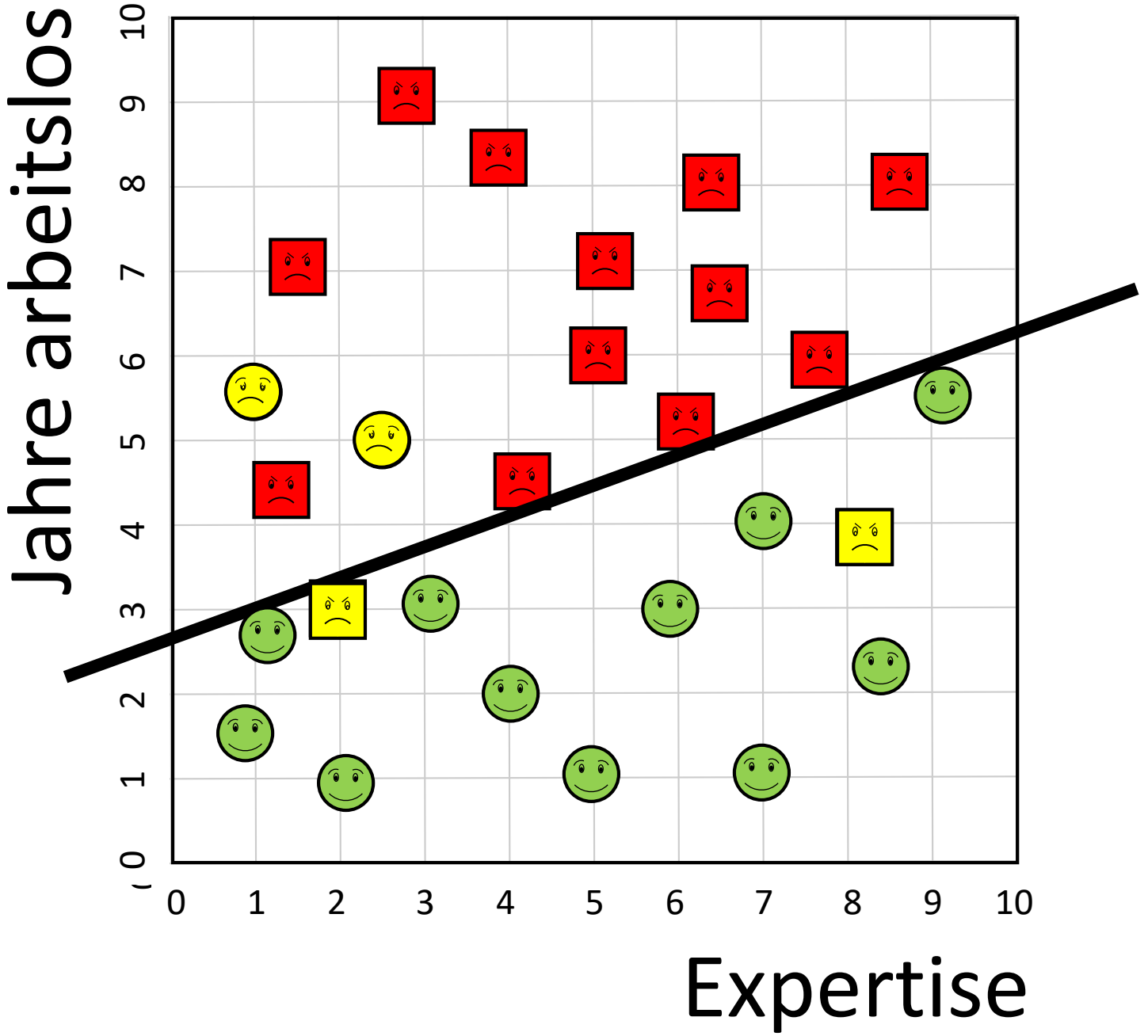
5.5 Jahre Erfahrung

4 Jahre arbeitslos



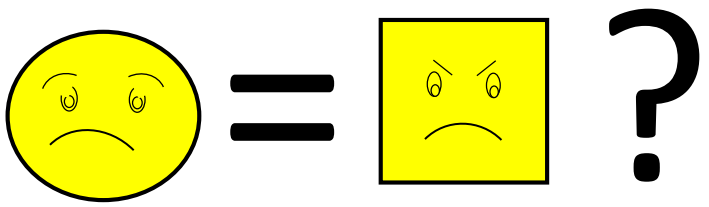
Weniger erfolgreiche Arbeitnehmer:innen

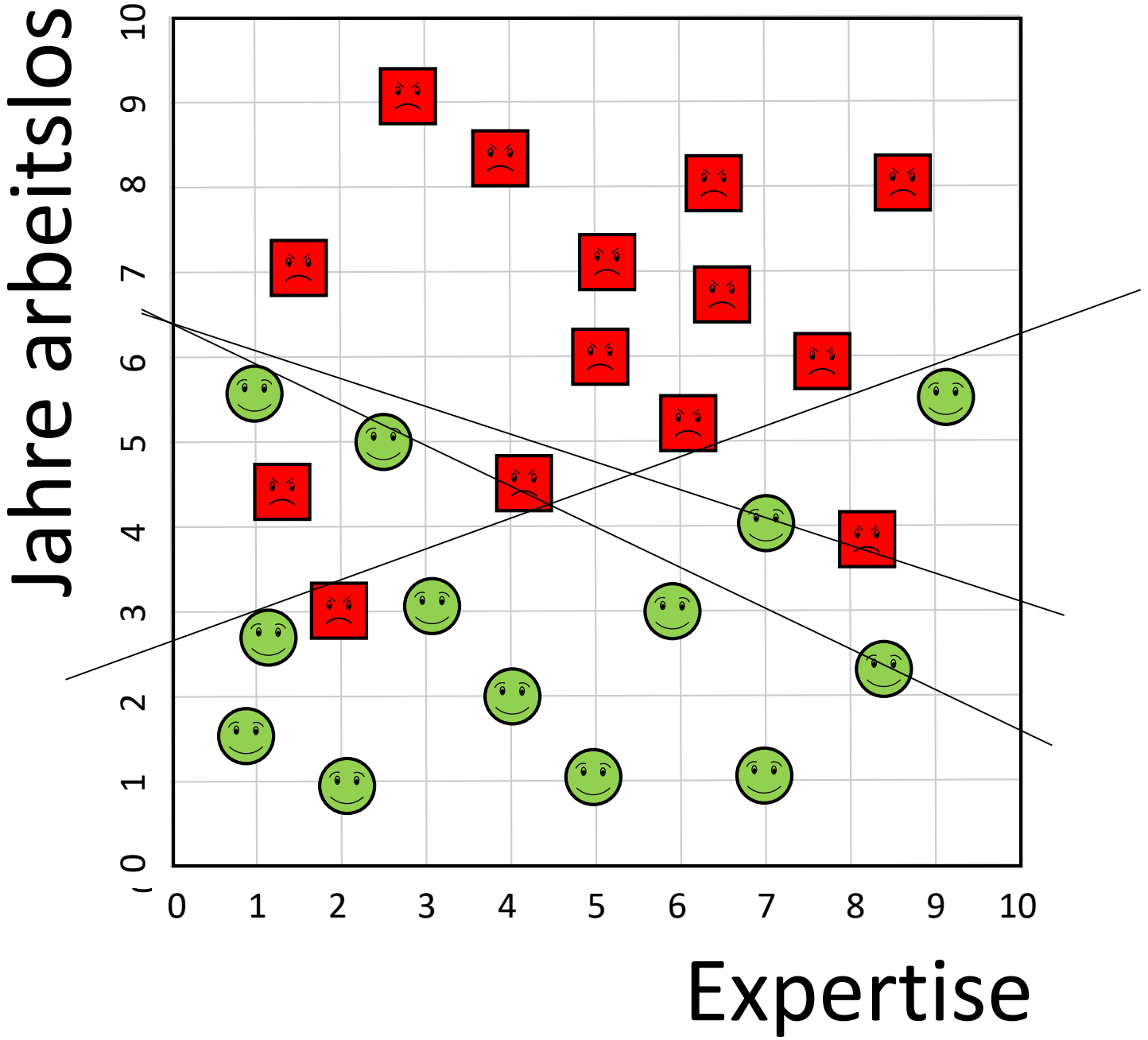
Erfolgreiche Arbeit-nehmer:innen



Weniger erfolgreiche Arbeitnehmer:innen

Erfolgreiche Arbeit-nehmer:innen





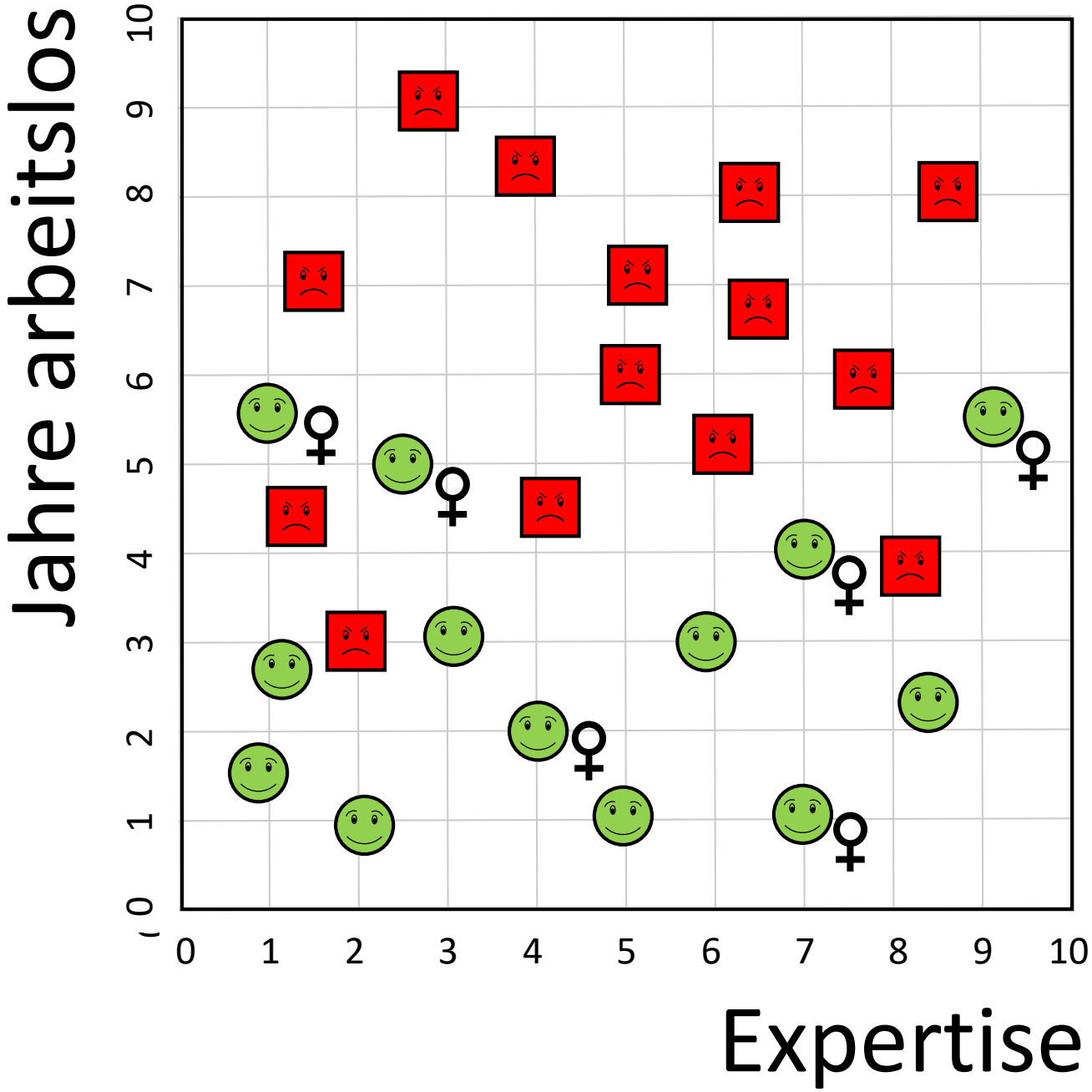
Datengrundlagen

- Meistens mehr als zwei Eigenschaften.
- Am wichtigsten:
  - **War Einstellung erfolgreich?**

Weniger erfolgreiche Arbeitnehmer:innen

Erfolgreiche Arbeit-nehmer:innen

Daten-grundlage

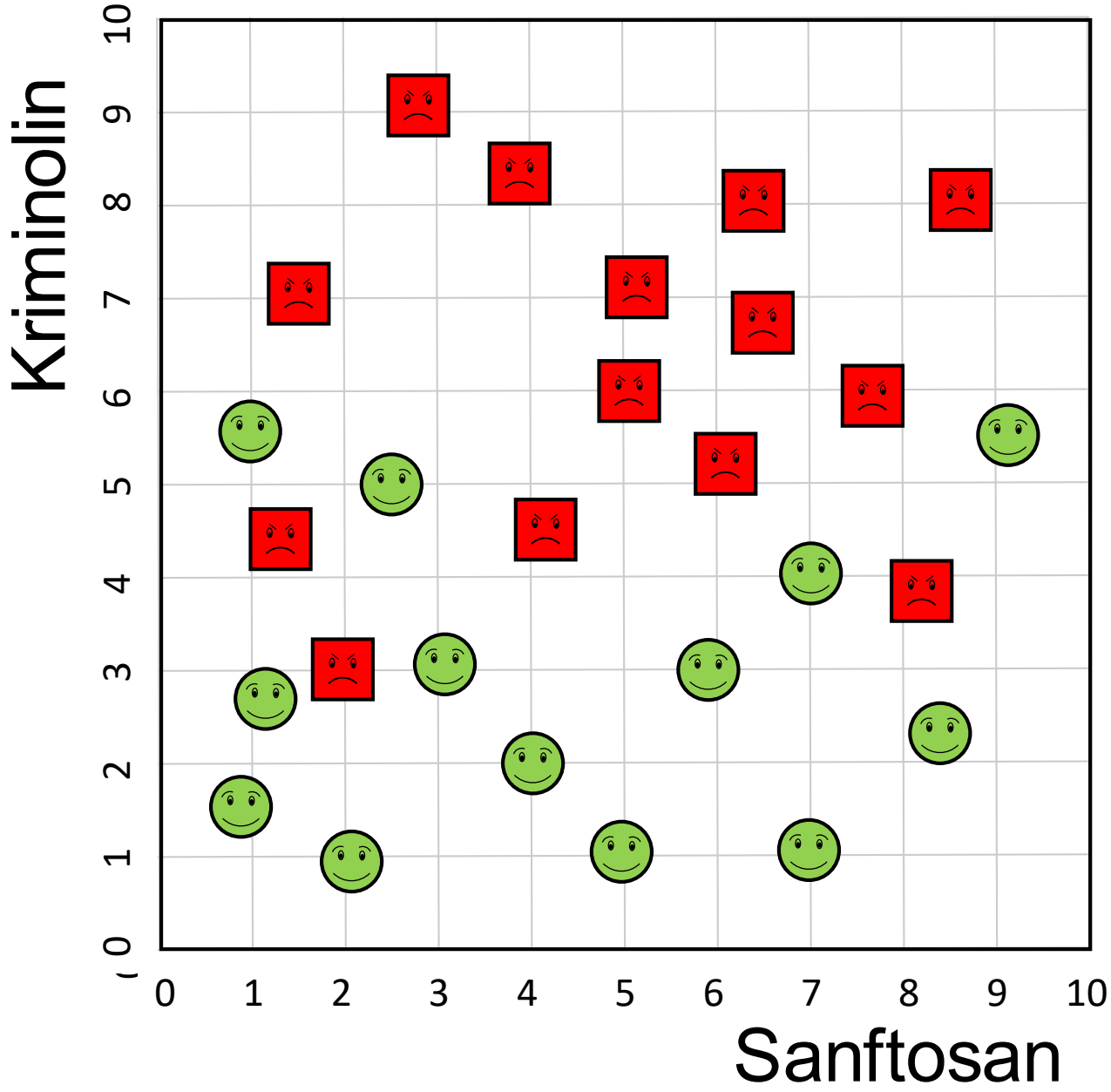



Ausbildung
Lebenszeiten
Alter
Arbeitgeber
-wechsel
Bewerbungs-
schreiben
Rechtschreibung
Wortvielfalt
Ton
Social Media?

Qualität eines Algorithmus

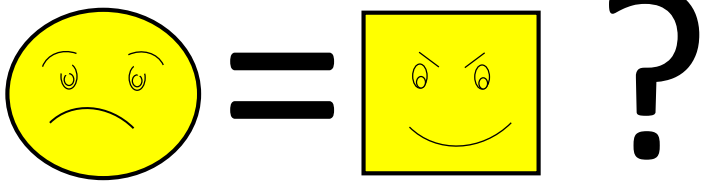
Bösartige Kriminelle

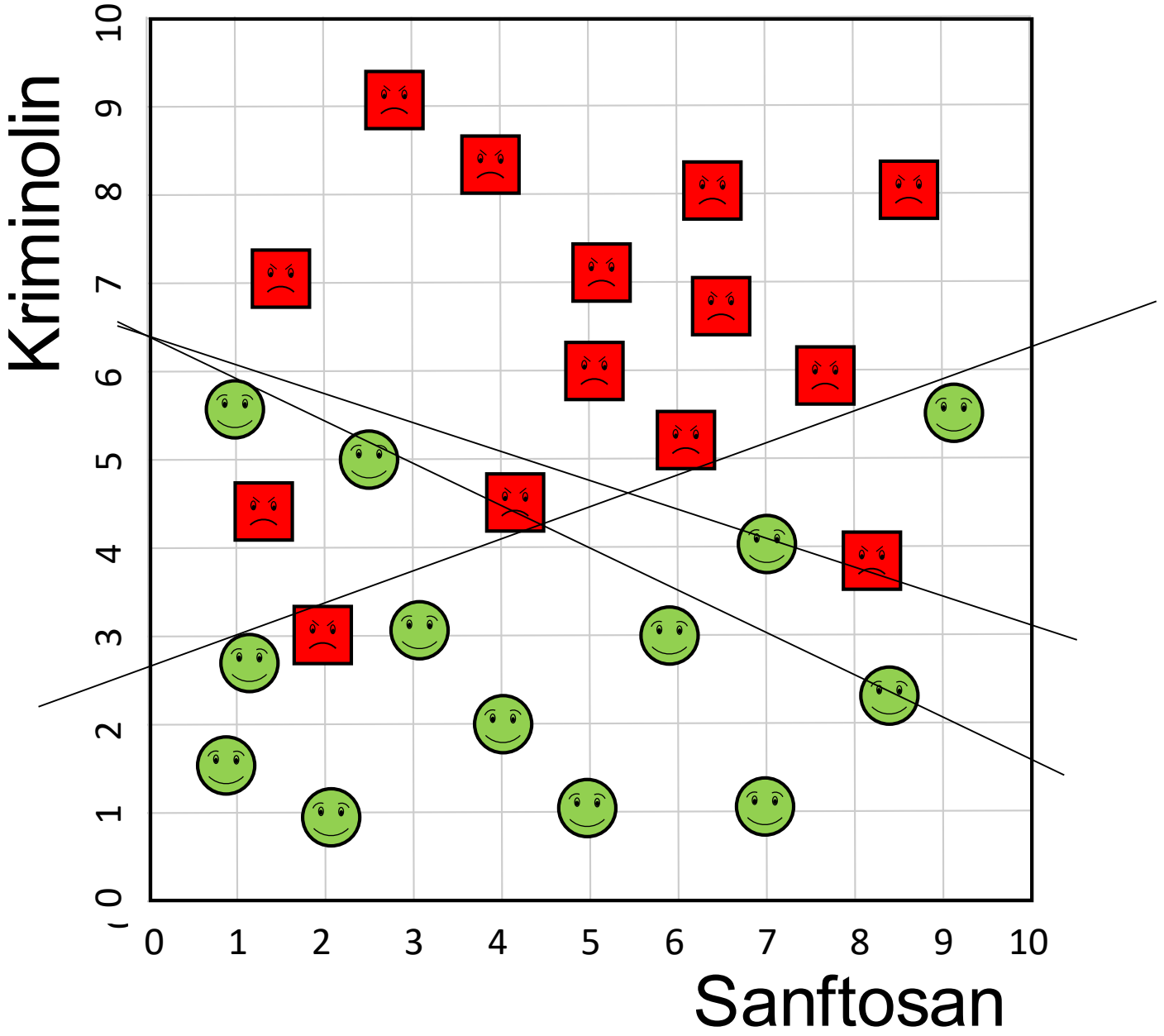
Unschuldige Bürger



Bösartige Kriminelle

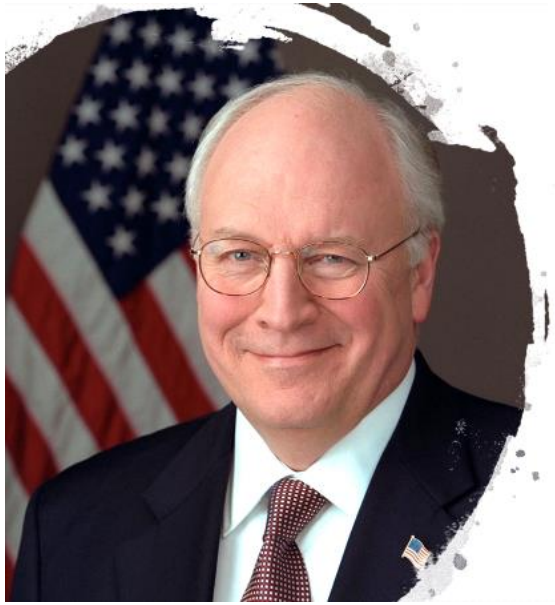
Unschuldige Bürger





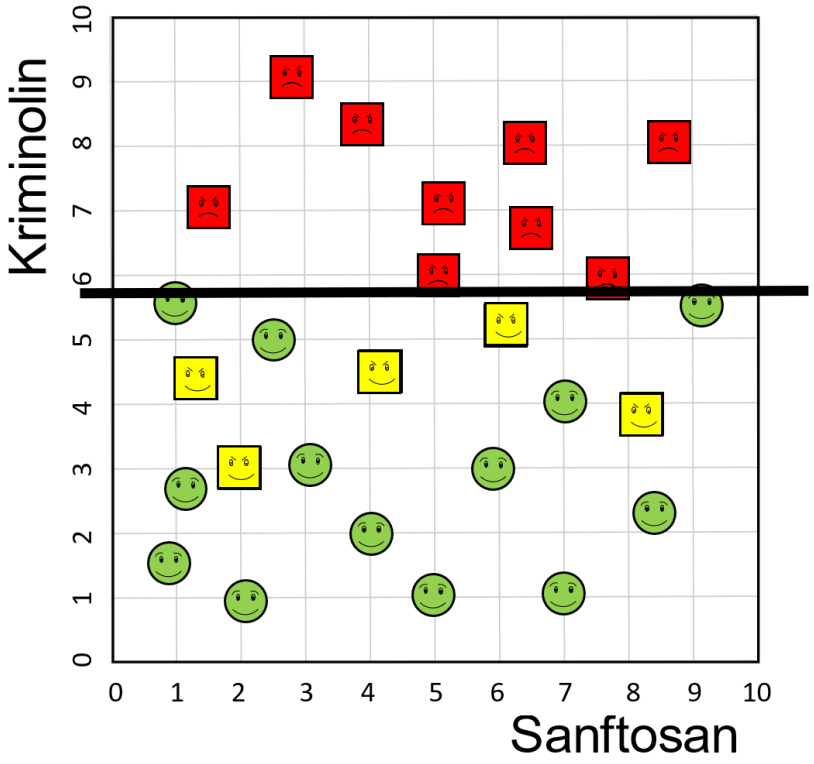
„It is better that ten guilty persons escape than that **one** innocent suffer.“

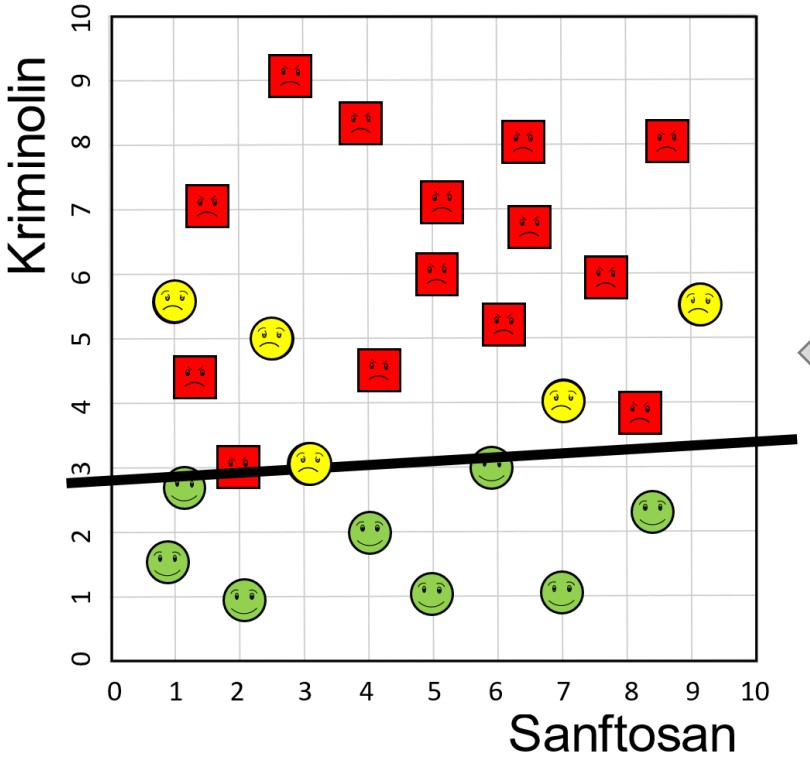
William Blackstone,
Rechtsphilosoph, 1760



"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,

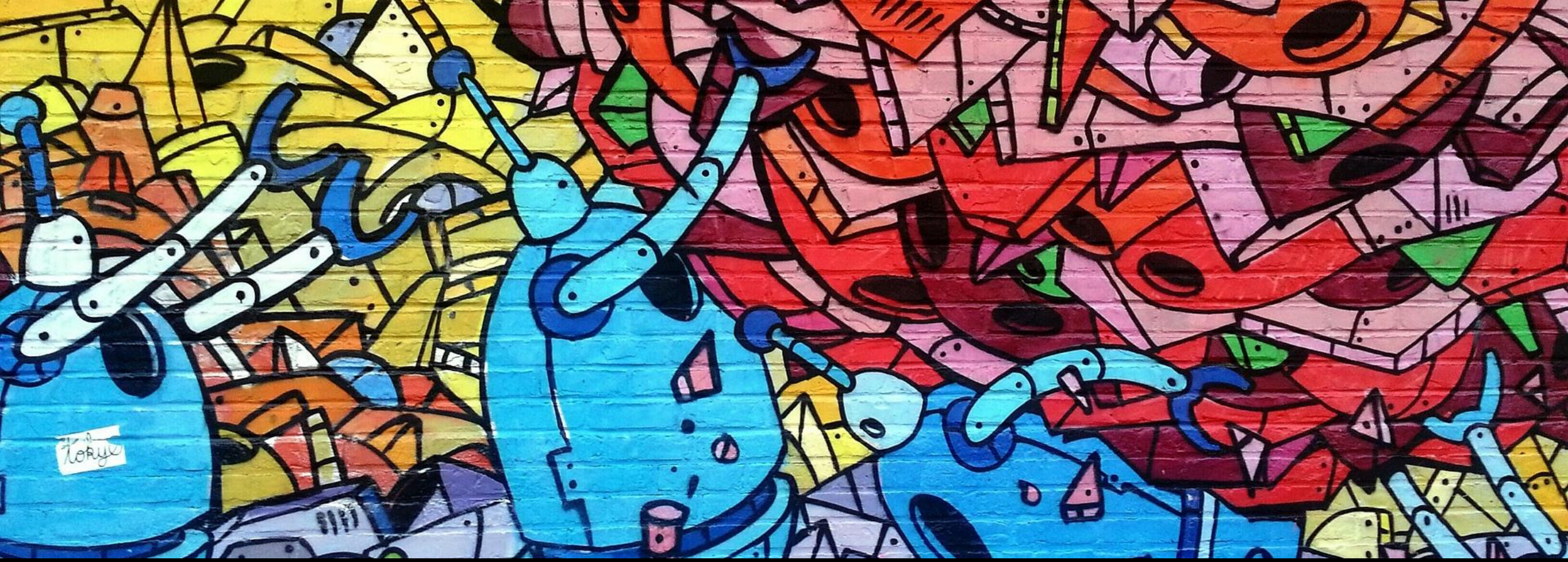




Qualität von ADM Systemen

1. **Wer entscheidet, wann ein ADM System „gut“ ist? Wer, wann es „fair“ ist?**

@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern



Wahrscheinlichkeit & Wahrheit

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**)

@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

Algorithmen…

- … basieren auf Korrelationen von Eigenschaften mit gewünschtem Verhalten.
- Quasi **algorithmisch legitimierte Vorurteile**:
  - Zu 70% erfolgreich heißt:
  - Von 100 Personen, die „genau so sind wie dieser Mensch“, sind 70 nachher erfolgreich.

Qualität von ADM Systemen

1. Wer entscheidet, wann ein ADM System „gut“ ist? Wer, wann es „fair“ ist?
2. **ADM Systeme ergeben nur Wahrscheinlichkeiten, keine Wahrheiten.**

@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

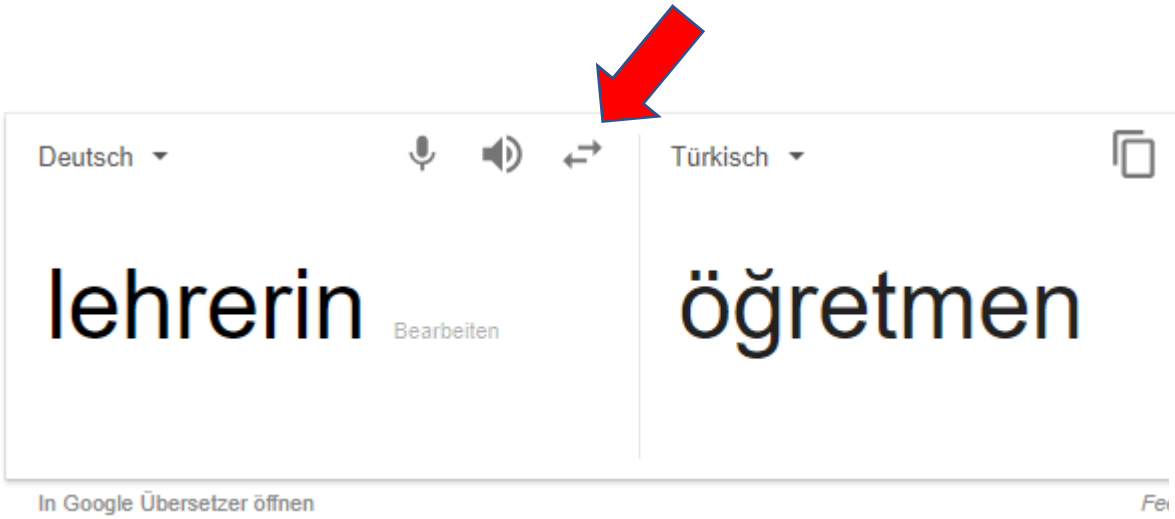
Computer: Objektiv und fehlerfrei?

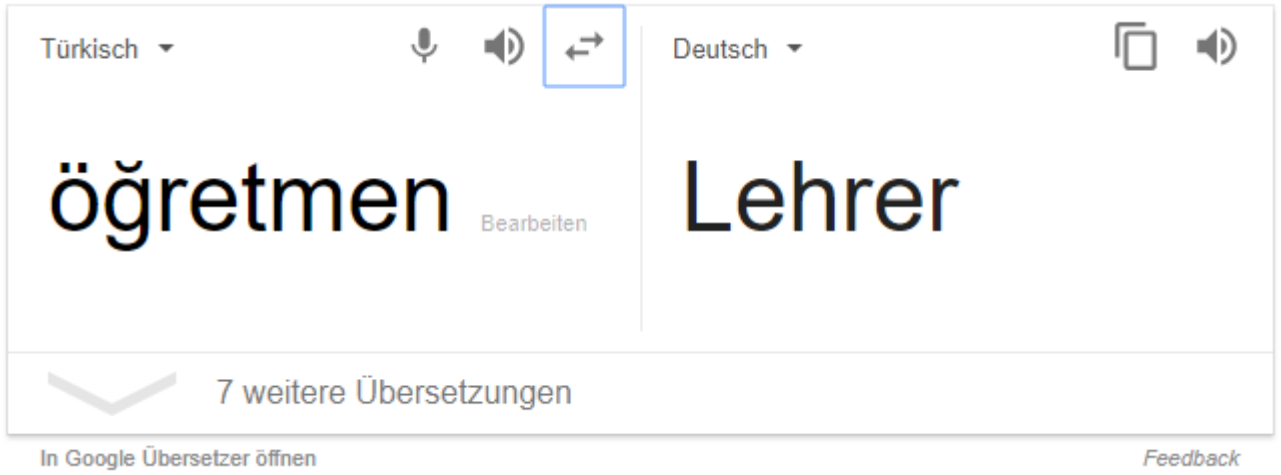
Gleichberechtigung

Wenn man auf Google nach „CEO“ sucht…

Passt das einigermaßen: ca. 8% der (deutschen) CEOs sind Frauen

Übersetzungen

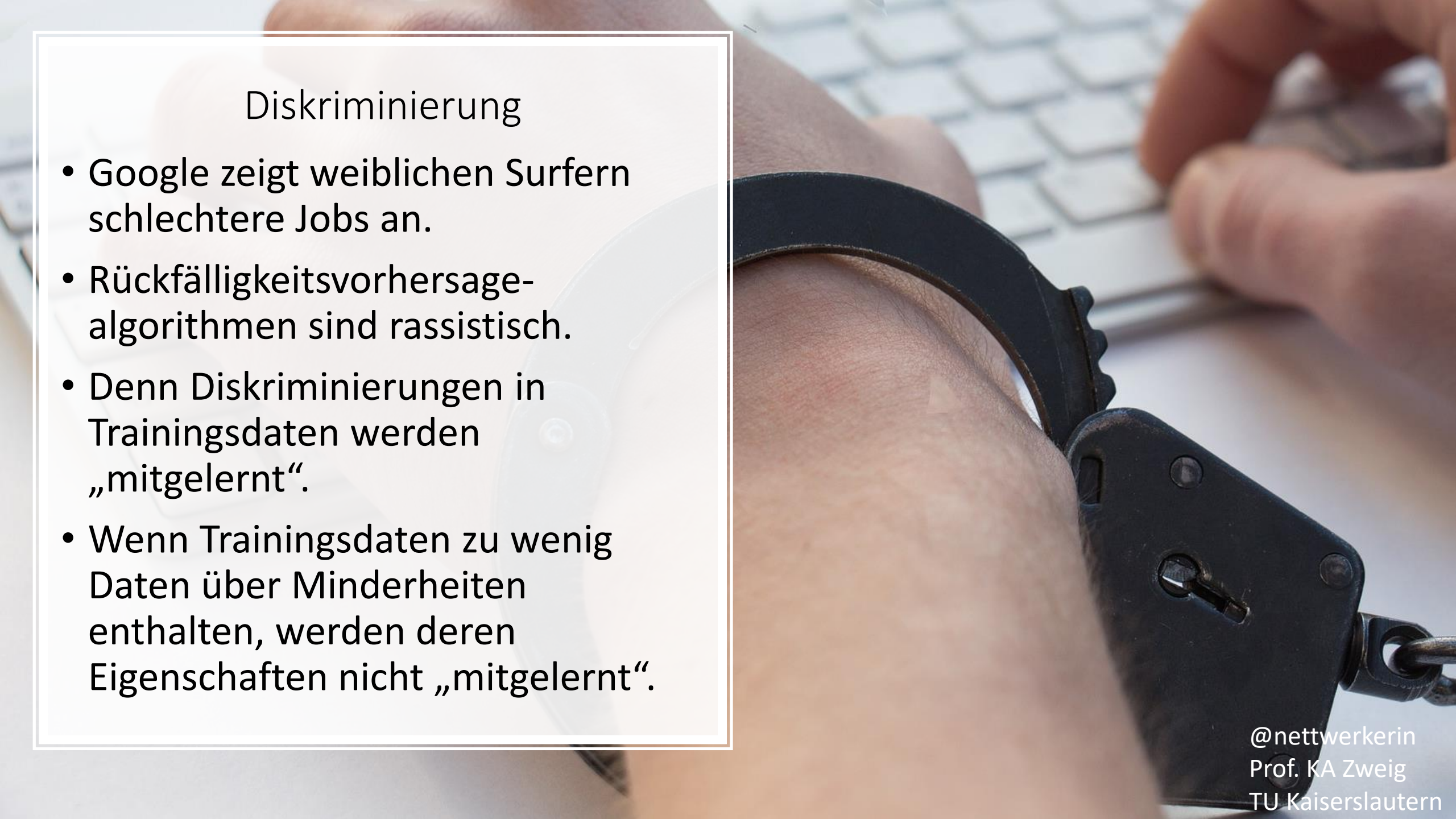




Und das passiert, wenn ich auf Pixabay nach „Chef“ suche…

Diskriminierung

- Google zeigt weiblichen Surfern schlechtere Jobs an.
- Rückfälligkeitsvorhersage-algorithmen sind rassistisch.
- Denn Diskriminierungen in Trainingsdaten werden „mitgelernt“.
- Wenn Trainingsdaten zu wenig Daten über Minderheiten enthalten, werden deren Eigenschaften nicht „mitgelernt“.

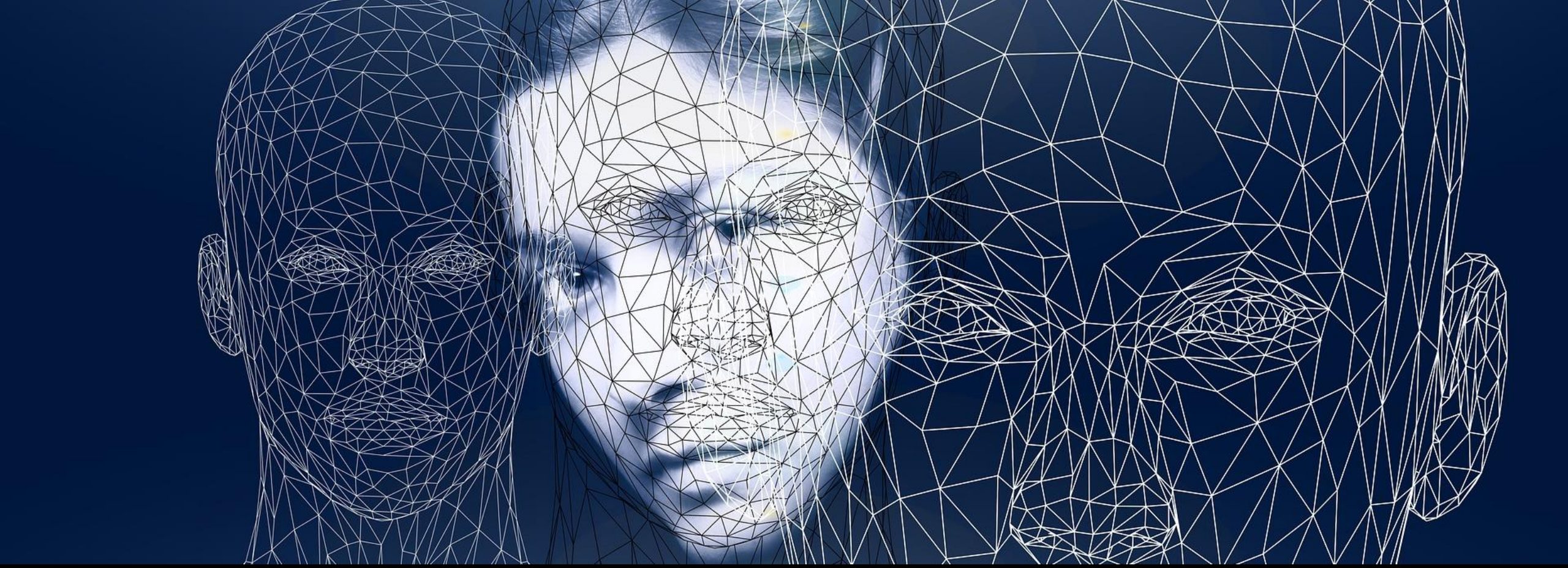


@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

Qualität von ADM Systemen

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

@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern



Sozio-informatische Gesamtbetrachtung

Sozio-informatische Betrachtungsweise

- KI schafft neue Anreize für menschliche Akteure.
- Diese reagieren auf die Anreize und wirken auf die KI ein.

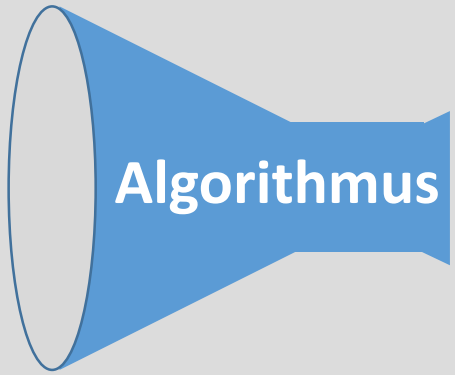
Probleme der Einbettung der ADM in den sozialen Prozess

- **Aufmerksamkeitsökonomie** von Entscheiderinnen und Entscheidern.
- **„Best practice“** erfordert Nutzung der Software.
- **Delegation von Verantwortung!**
- Manchmal kann ein falsch Beurteilter **die Vorhersage prinzipiell nicht entkräften!**
  - Z.B. abgelehnte Bewerberin, inhaftierter Krimineller

@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

Algorithmische Entscheidungs**systeme** (ADM Systeme)

Daten



Bewertete

Nutzer des ADM Systems

Soziales System

Qualität von ADM Systemen

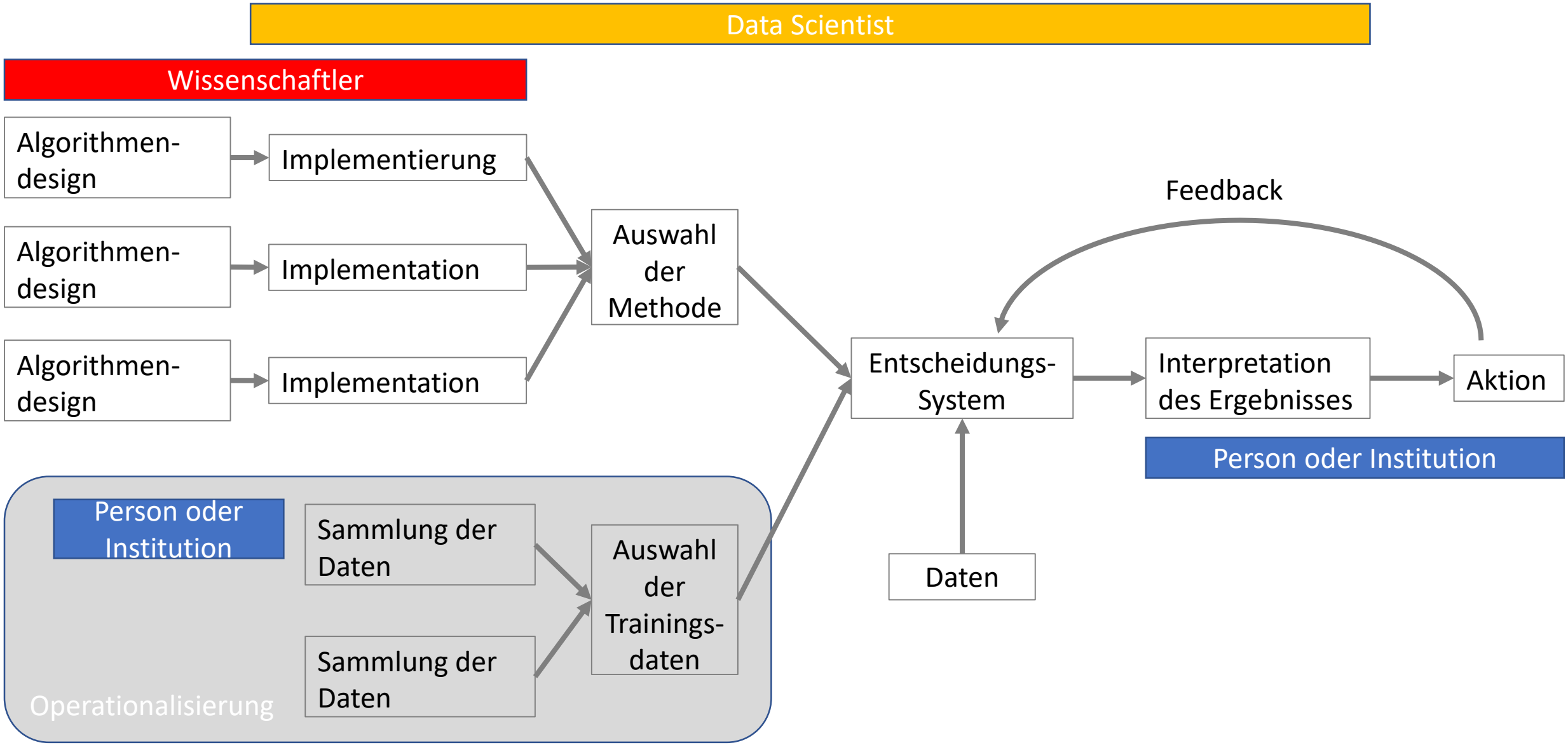
1. **Wer entscheidet, wann ein ADM System „gut“ ist?**
2. **ADM Systeme ergeben nur Wahrscheinlichkeiten, keine Wahrheiten.**
3. **ADM Systeme können diskriminieren.**
4. **ADM Systeme bedürfen einer sozio-informatischen Gesamtanalyse.**

@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

Wie gut sind die Robo-Richter?

- Ganz schön schlecht: COMPAS
  - Hochrisiko-Kategorie:
    - Gewöhnliche Kriminaltaten: nur zu 50% richtig!
    - Schwere Straftaten: nur zu 20% richtig!
- Ein amerikanisches Terroristenidentifikationssystem tönt:
  - „Nur 0.008% falsch Positive!“
  - Bei 55 Millionen Einwohner sind das 4.400 Unschuldige, um wenige Hundert zu identifizieren.
  - Von den „Hochrisikopersonen“ also vermutlich unter 20%!
- Im medizinischen Bereich teilweise besser als Doktoren!

Lange Kette der Verantwortlichkeiten



Zusammenfassung

- Algorithmen können diskriminieren
- Algorithmen können trotzdem Entscheidungsprozesse verbessern
- Es gibt Situationen, die keine algorithmischen Entscheidungen erlauben.

Weitere Literatur

- Katharina Zweig: „Auch Maschinen können diskriminieren“, 16.1.2018, MERTON Magazin, https://merton-magazin.de/auch-algorithmen-koennen-diskriminieren
- Studie für die Bertelsmann-Stiftung:
  Zweig, Fischer & Lischka: „Wo Maschinen irren können“ (Serie AlgoEthik, No. 4, 2018)
- Broschüre der Bayerischen Landesmedienanstalt (Zweig, Krafft & Hauer, 2016): „Dein Algorithmus - meine Meinung“