

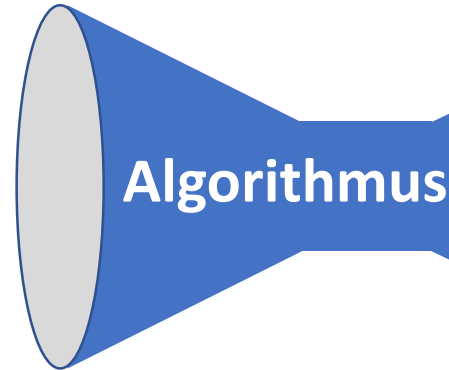
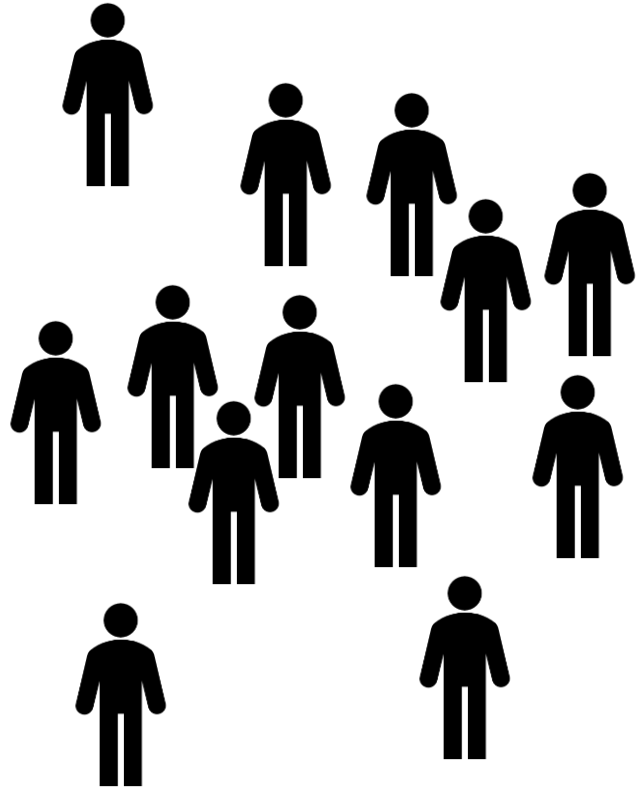


Entseelte Entscheidungen

BMBF Verbundprojekt INTEGRAM
9.4.2018 in Tübingen

Prof. Dr. Katharina A. Zweig
Algorithm Accountability Lab
TU Kaiserslautern

Algorithmische Entscheidungssysteme



Scoring-Verfahren

oder



Klassifikation

Wer soll richten?





Unser Menschenbild

Sind Menschen eigentlich dazu geeignet, über andere Menschen zu entscheiden?

Menschen – so irrational!

- Richter müssen vorzeitige Haftentlassungsanträge begutachten.
- Studie: je weiter von der letzten Pause weg, desto weniger risikoreiche Entscheidungen¹.
- 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

Problemfall USA

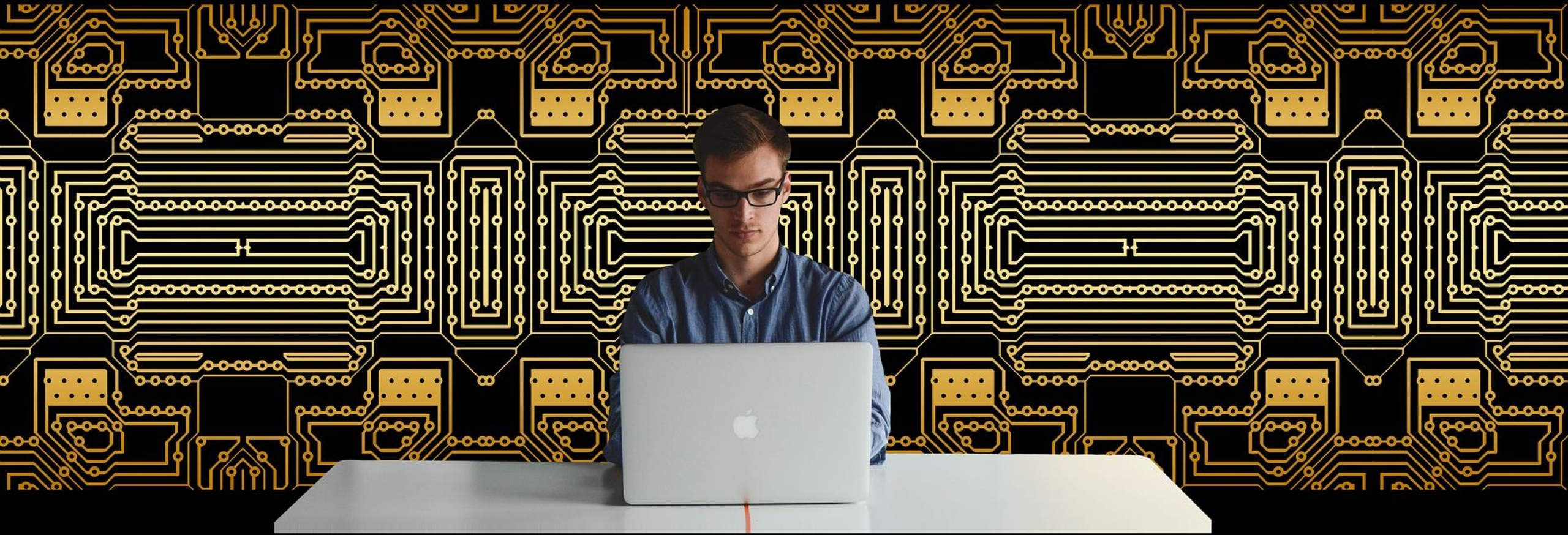
- Zweithöchste Inhaftierungsrate weltweit.
- 6x höhere Rate von Afroamerikanern und 2x höhere Rate von Latinos als von Weißen.
- Prognose: jeder dritte afroamerikanische Junge im Alter von 10 Jahren wird eine Gefängnisstrafe absitzen müssen.



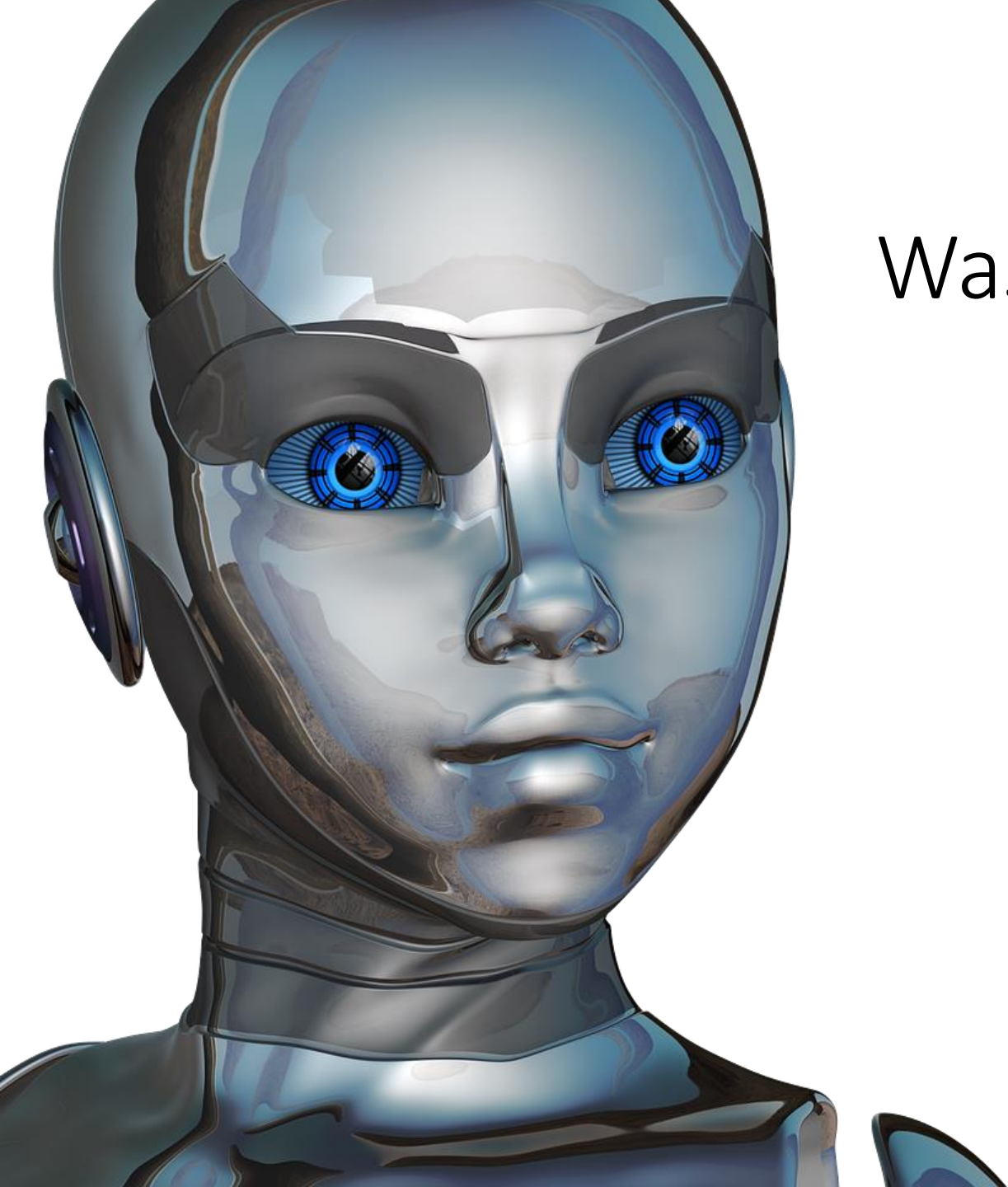
American Civil Liberties Union



- Amerikanische Bürgerrechtsunion (seit 1920) fordert:
- Algorithmische Entscheidungssysteme sollten überall im Prozess eingesetzt werden, ...
- ... um Fairness und Objektivität zu sichern.
- Dazu sollen Computer aus Daten Entscheidungsregeln lernen.



Können Computer lernen?



Was heißt Lernen?

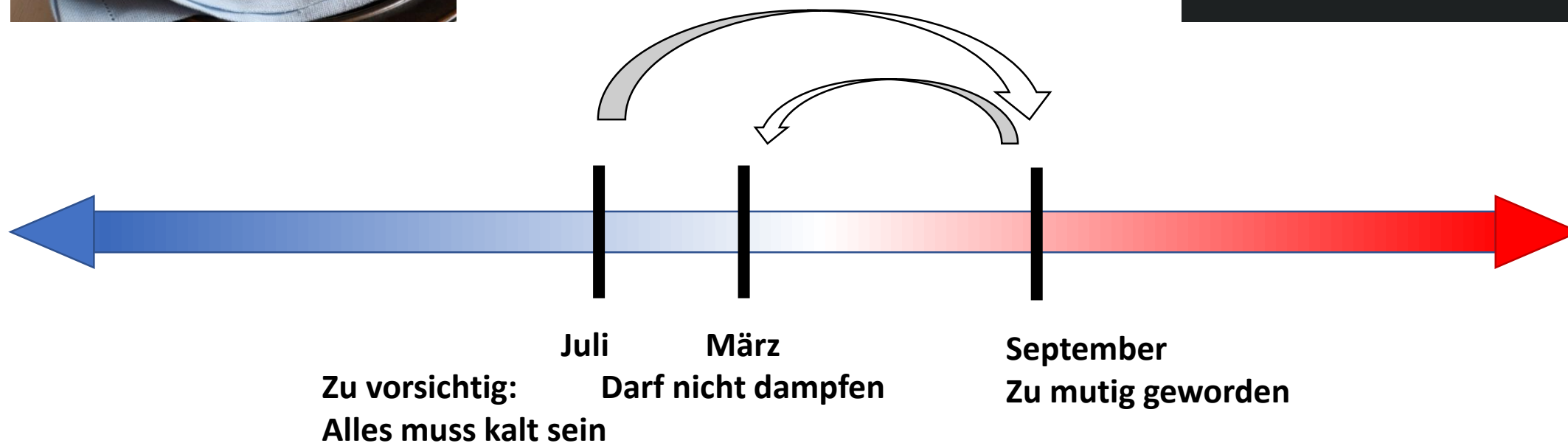
Einfach:

In derselben Situation ein vorher gezeigtes Verhalten wiederholen.

Generalisiert:

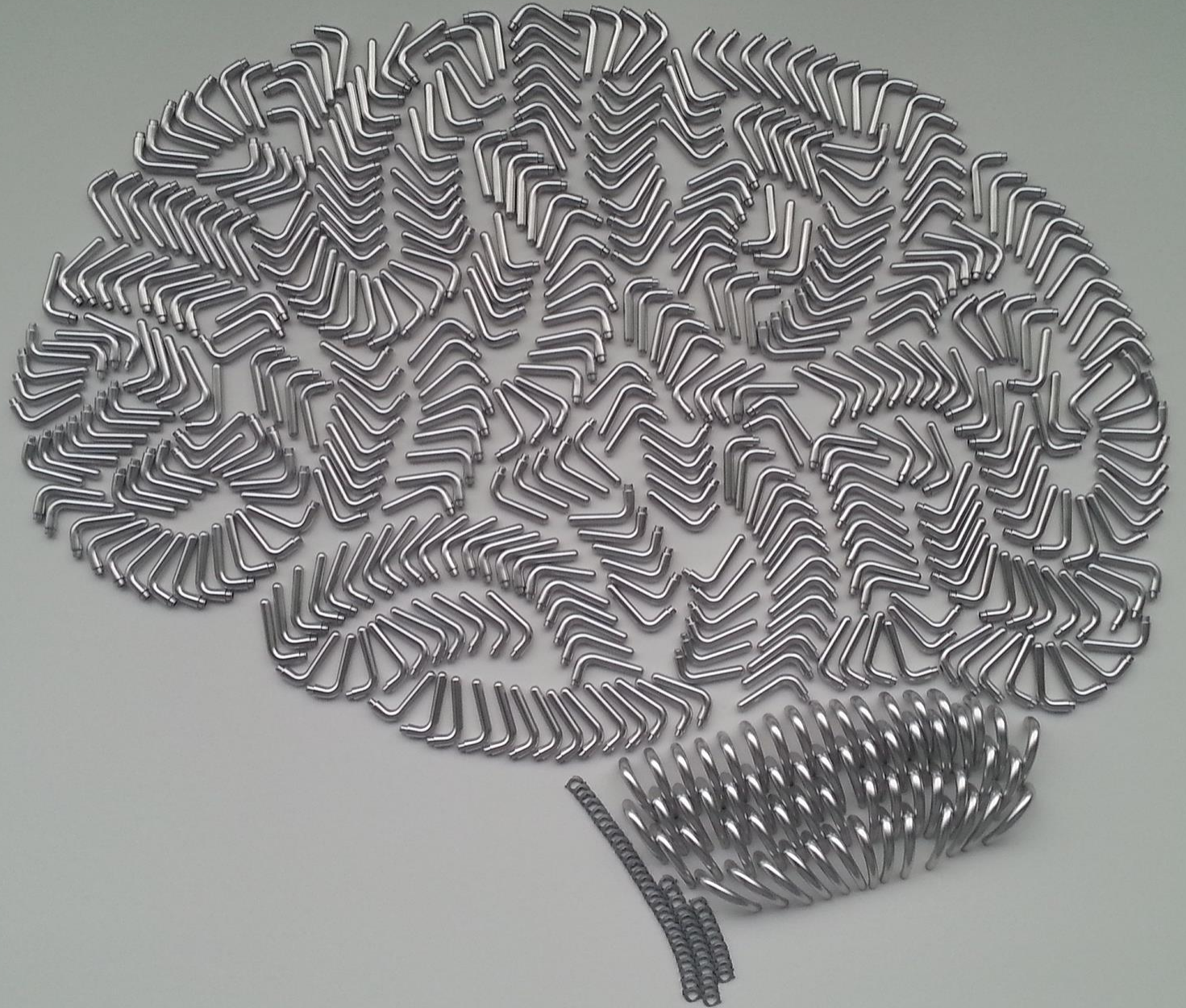
In derselben Art von Situation das richtige Verhalten aus einer Reihe von Möglichkeiten auswählen.

Sebastian lernt „heiss“ und „warm“



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**.

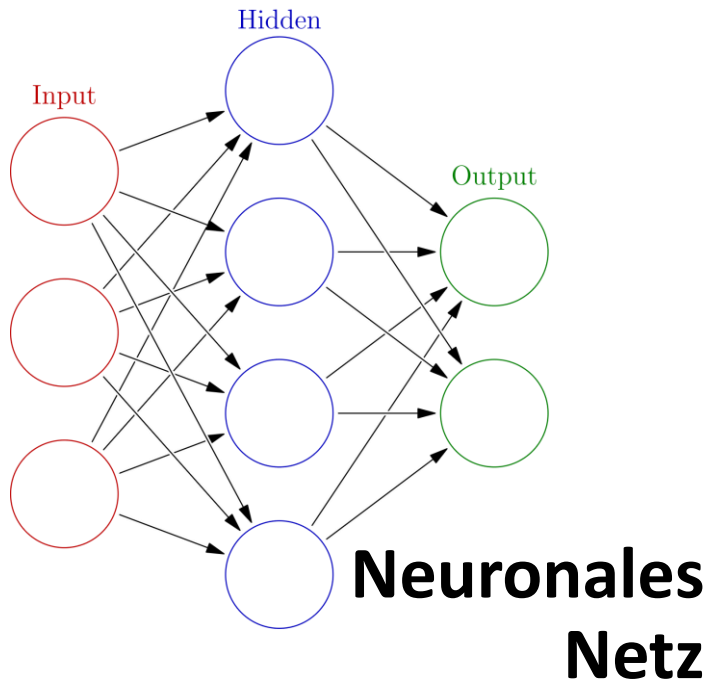


Computer lernen

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

Optimal auch **Rückkopplung**.

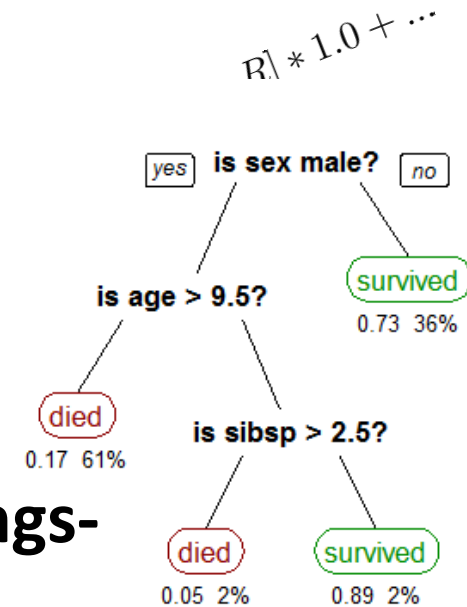
Er lernt **generelle Regeln**.



Formel

$$w_1 * \#V_h - w_2 * \#day_i V_h + w_3 * I[g = male]$$

Entscheidungs- bäume





Lernen mit Formeln

**Individuelle
Risikobewertung der
Rückfälligkeit von
Kriminellen**

Regressionsansätze

- Algorithmdesigner entscheiden, welche der Daten vermutlich mit „Rückfallwahrscheinlichkeit“ korrelieren.
- Resultat sollte eine einzige Zahl sein.
- Je höher die Zahl, desto höher die Rückfallwahrscheinlichkeit.
- Beispiel Formel:

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

Allgemein

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

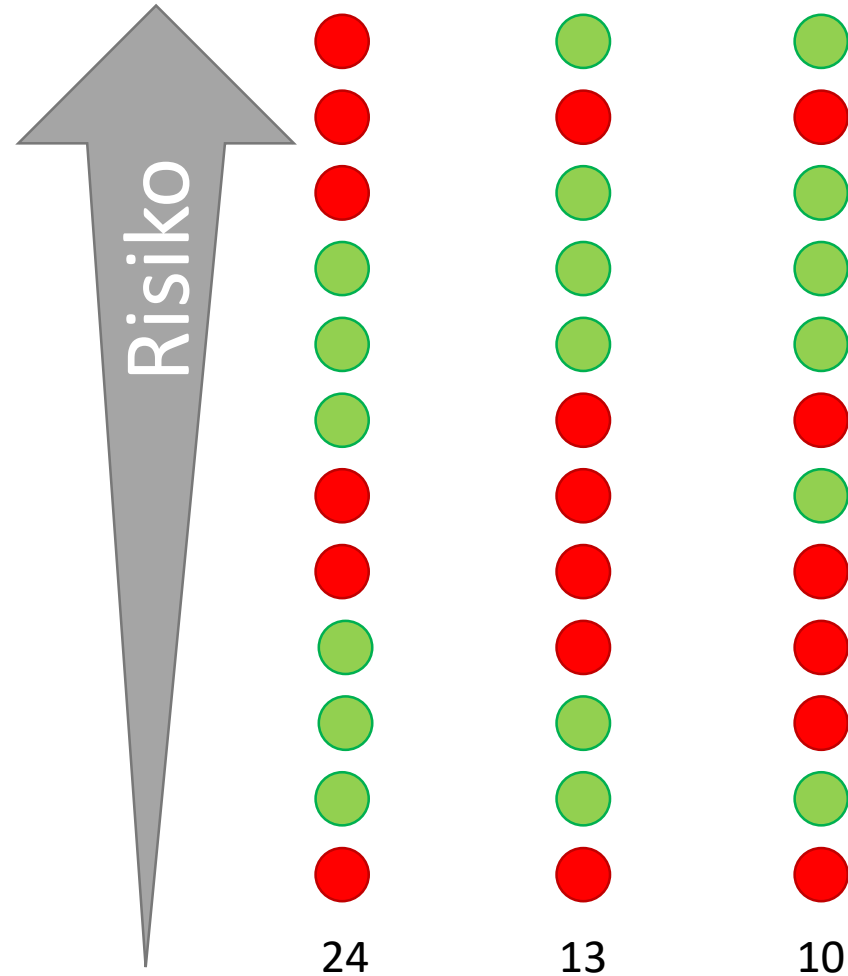
Der Computer bestimmt die Gewichte und bekommt ein Feedback (Rückkopplung), inwieweit die damit resultierende Bewertung tatsächlich mit dem (beobachteten) Verhalten übereinstimmt.



Qualität eines Algorithmus |

„Lernen“ von Gewichten

- Algorithmus probiert Gewichte und berechnet Risiko für alle Personen im Datenset.
- Bewertet jeweils, wie viele erwiesenermaßen Rückfällige möglichst weit oben stehen.
- Die Gewichtung, die das maximiert, wird für weitere Daten genommen.



Grüne Kugeln symbolisieren resozialisierte, rote rückfällige Kriminelle.

Optimale Sortierung: Alle roten oben, alle grünen darunter.

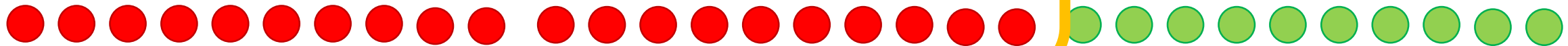
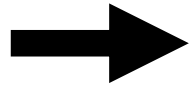
Qualitätsmaß: Paare von rot und grün, bei denen die rote Kugel über der grünen einsortiert ist.

Oregon Recidivism Rate Algorithm

- 72 von 100 Paaren werden korrekt sortiert.
- So werden aber keine Urteile gefällt!
- Sondern: Reihe von Angeklagten, von denen diejenigen mit dem höchsten Rückfallrisiko benannt werden sollen.
- Rückfallquote bei jugendlichen Kriminellen liegt z.B. bei 20%.

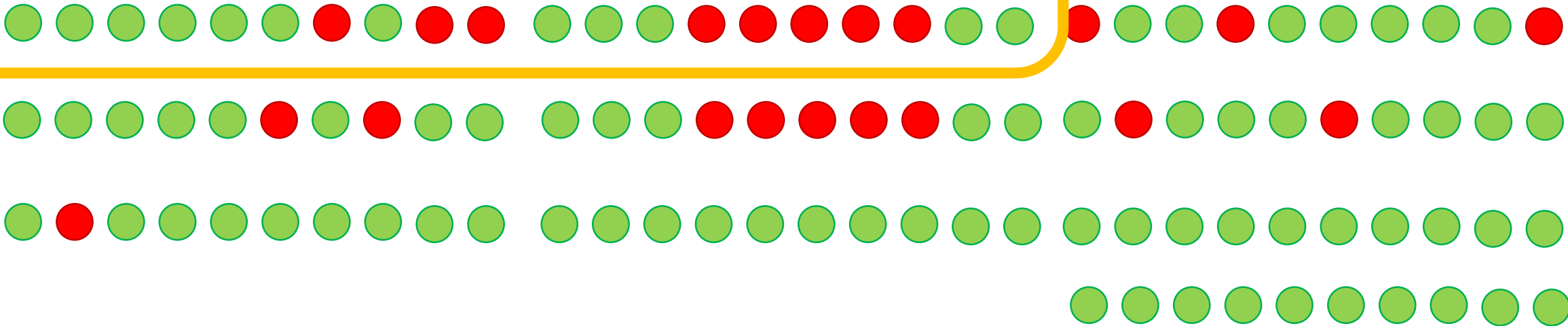
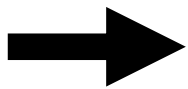
Optimale Sortierung

Erwartete 20% „Rückfällige“



Mögliche Sortierung eines Algorithmus mit dieser „Güte“ (75/100 Paaren)

Erwartete 20% „Rückfällige“



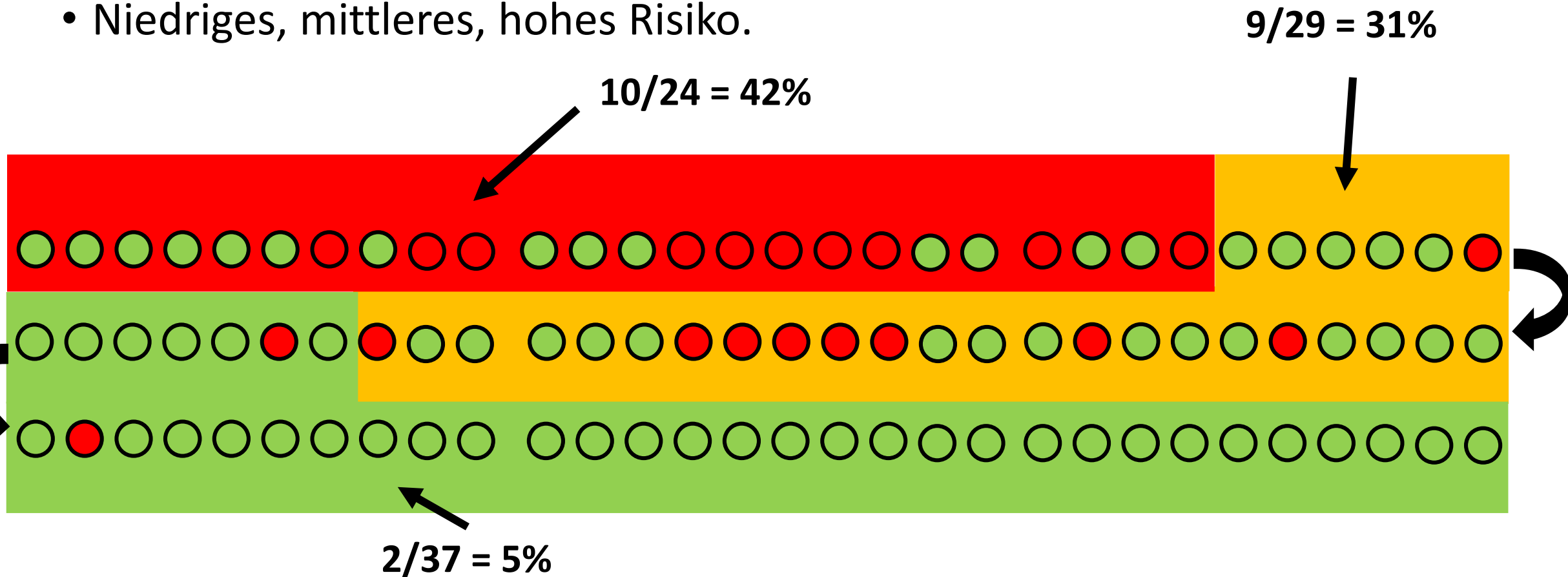
Das ist wie...

„Kaufen Sie diesen wunderbaren Wagen. TÜV? Brauchen Sie nicht! Und sehen Sie nur, die unglaublich gut erhaltenen Sommerreifen. Das ist noch Qualität!“



Vom Scoring zur Klassifikation

- ACLU fordert: Es soll drei Klassen geben.
- Niedriges, mittleres, hohes Risiko.

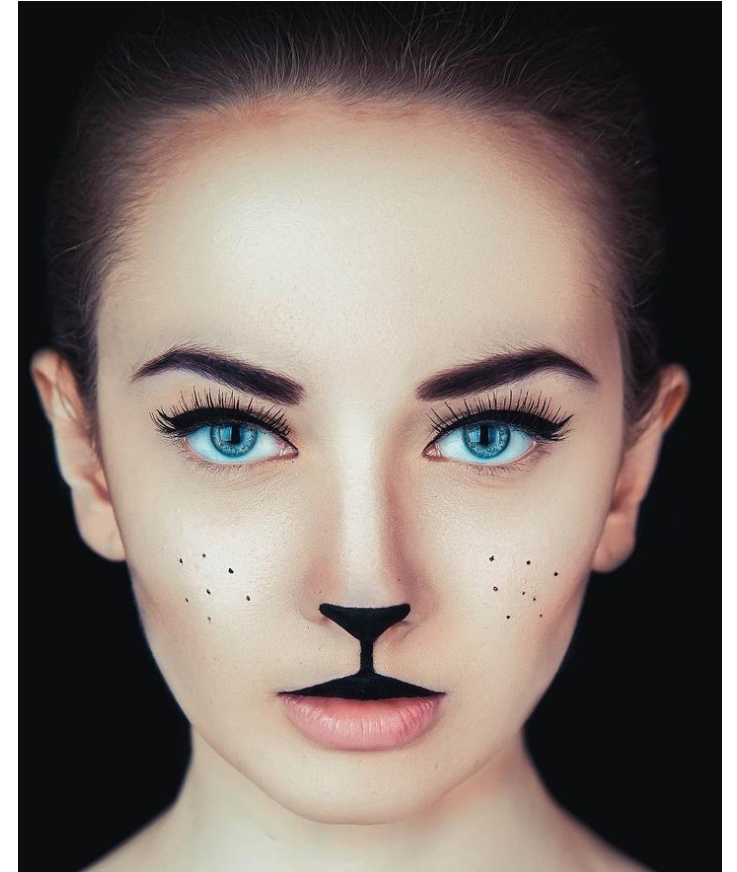




Statistische Vorhersagen |
über Menschen |

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*.



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**)



Algorithmen in einer demokratischen Gesellschaft

Generell

Prinzipiell können algorithmische Entscheidungssysteme für sehr viele, schwierige Fragestellungen in derselben Art gebaut werden:

- Automatische Leistungsbewertung
- Kreditvergabe
- Schulische und universitäre Ausbildungen, die durch algorithmische Entscheidungssysteme unterstützt werden
- Algorithmen, die das Sterberisiko von Kranken bewerten
- Gefährder-, Terroristenidentifikation
- ...



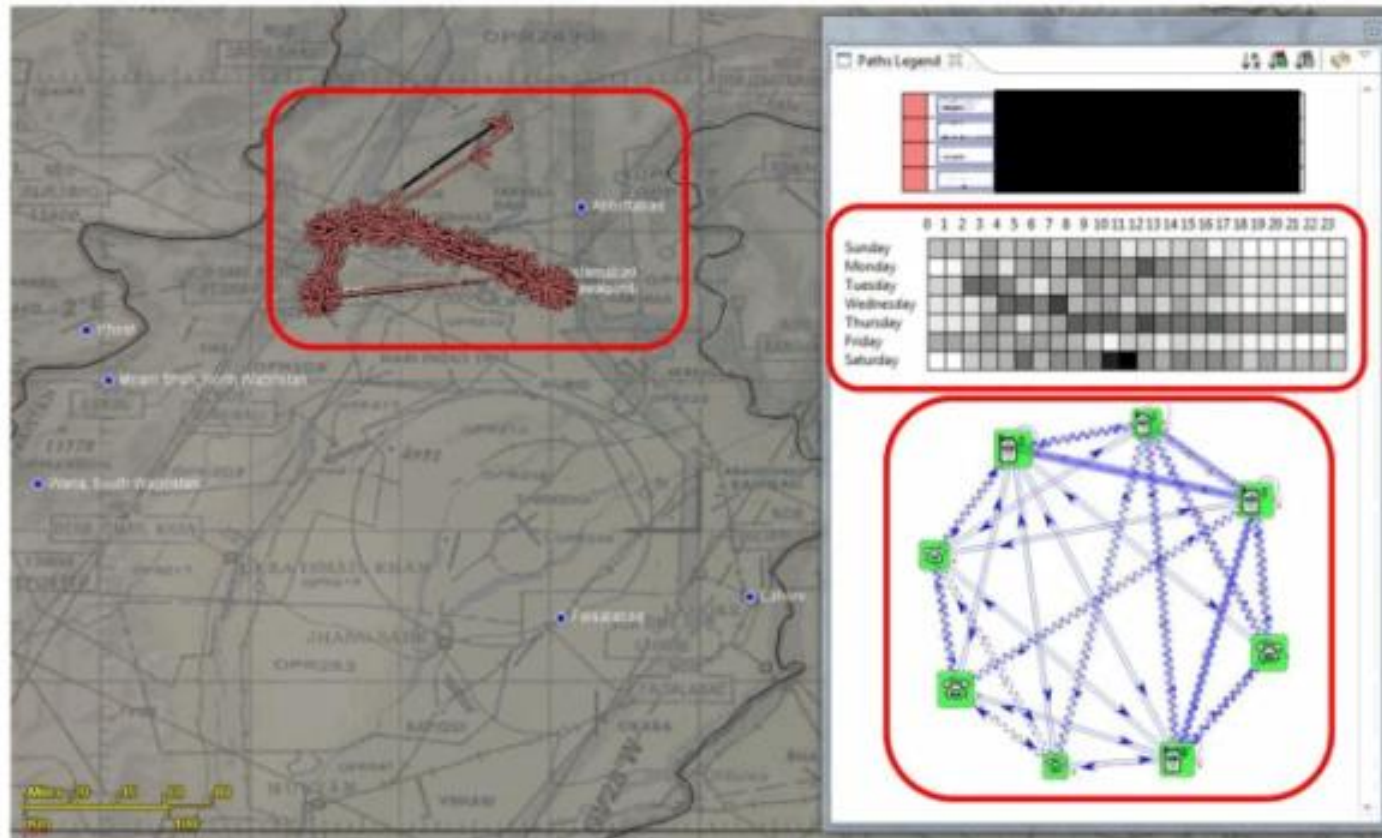
Ihre Aufgabe heute....

Entwickeln Sie ein
algorithmisches Entscheidungssystem,
dass **gewaltbereite Extremisten**
frühzeitig identifiziert!

Capturing terrorists with network analysis

TOP SECRET//COMINT//REL TO USA, FVEY

From GSM metadata, we can measure aspects of each selector's **pattern-of-life**, **social network**, and **travel behavior**



Terroristenidentifikation SKYNET

TOP SECRET//COMINT//REL TO USA, FVEY

We've been experimenting with several error metrics on both small and large test sets

Training Data	Classifier	Features	100k Test Selectors		55M Test Selectors	
			False Alarm Rate at 50% Miss Rate	Mean Reciprocal Rank	Tasked Selectors in Top 500	Tasked Selectors in Top 100
None	Random	None	50%	1/23k (simulated)	0.64 (active/Pak)	0.13 (active/Pak)
Known Couriers	Centroid	All	20%	1/18k		
		Outgoing	43%	1/27k		
+ Anchory Selectors	Random Forest		0.18%	1/9.9	5	1
		0.008%	1/14	21	6	

Random Forest trained on Known Couriers + Anchory Selectors:

- 0.008% false alarm rate at 50% miss rate
- 46x improvement over random performance when evaluating its tasked precision at 100

Windows
Wechseln
aktivieren

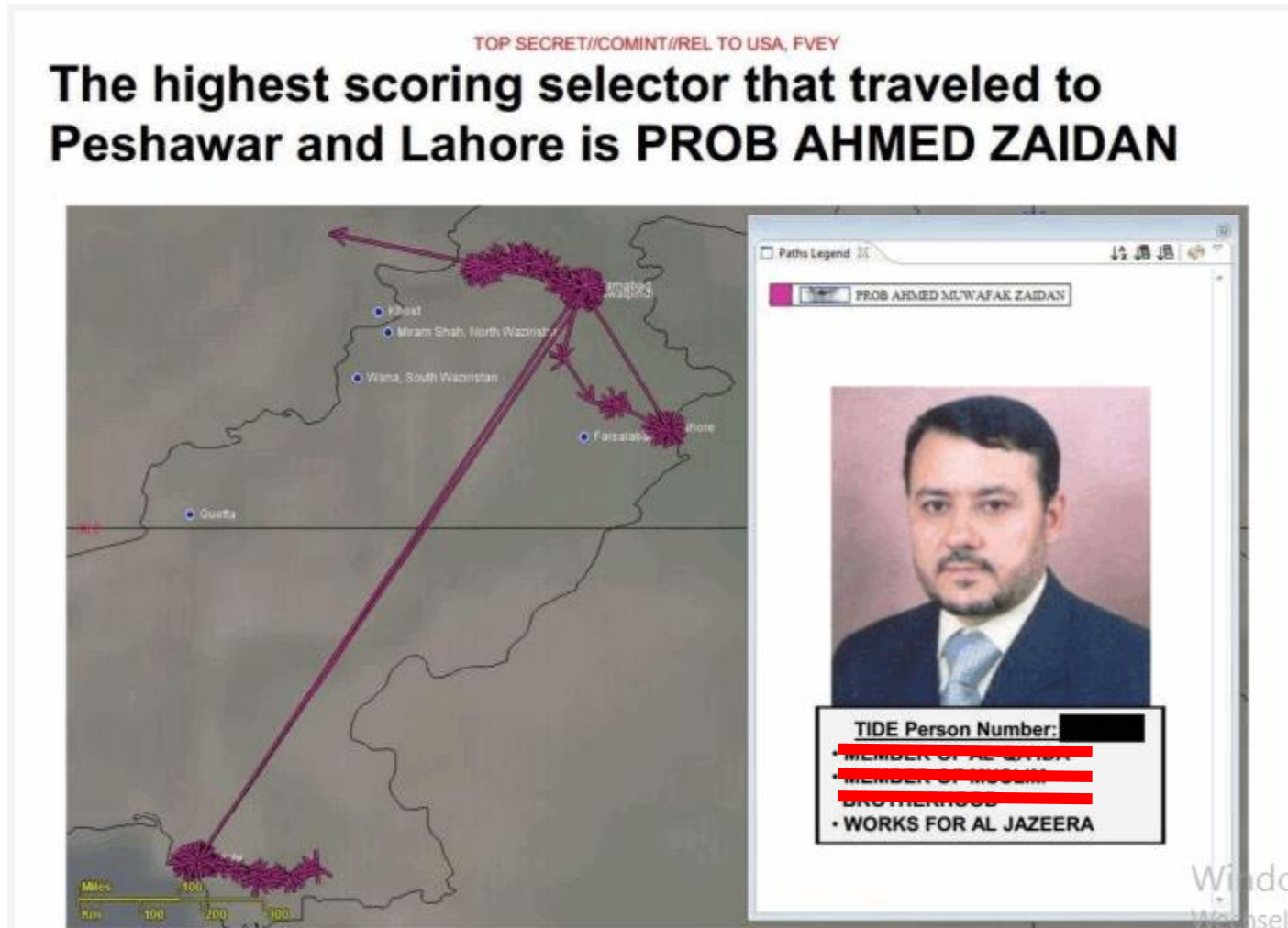
TOP SECRET//COMINT//REL TO USA, FVEY

Das sind 4.400 Unschuldige, um die Hälfte der vermeintlichen Terroristen zu identifizieren!

<https://theintercept.com/document/2015/05/08/skynet-courier/>

<https://theintercept.com/2015/05/08/u-s-government-designated-prominent-al-jazeera-journalist-al-qaeda-member-put-watch-list/>

Top-“Kurier“ der Terroristen laut Algorithmus ist...

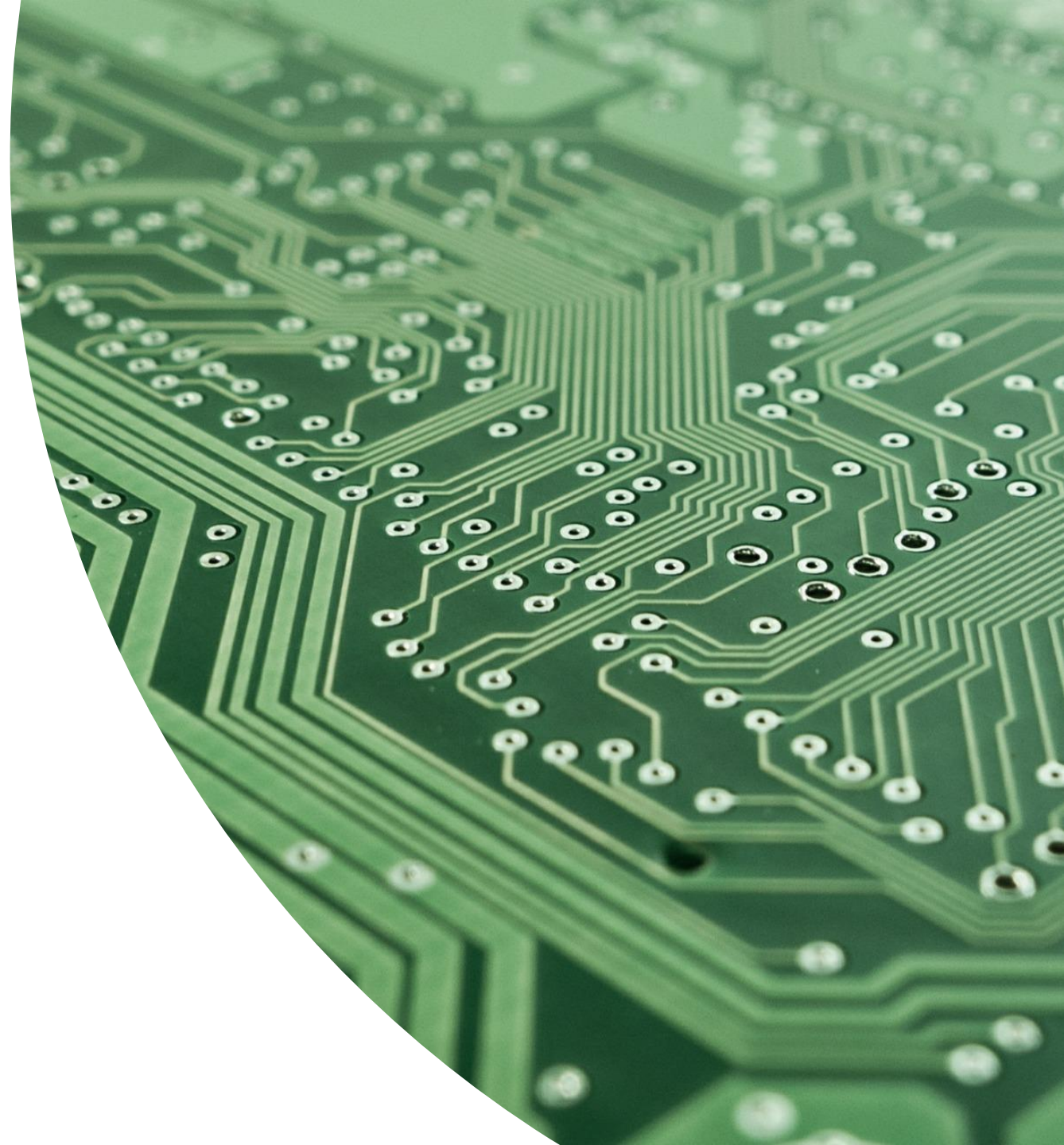




Sozio-informatische Gesamtbetrachtung

Probleme des technischen Systems

- Sind die **Daten** diskriminierungs- und fehlerfrei bzw. gut genug?
 - Häufiges Problem: “Entity Recognition”.
- Was heißt das:
 - „zu 40% rückfällig“ und wie wird das dem Nutzer dargestellt?
- **Wie gut ist das ADM System**, welche Annahmen trifft es?
- Wer **entscheidet**, was “gut” ist?
- **Ist das System vorurteilsfrei?** Garantiert?
- **Bleibt** es vorurteilsfrei? Garantiert?
- Wie geht es mit **Minderheiten** um?
- Wie werden **Ermessensspielräume** umgesetzt?



Probleme der Einbettung der ADM in den sozialen Prozess

- **Aufmerksamkeitsökonomie** von Entscheiderinnen und Entscheidern.
- „**Best practice**“ erfordert Nutzung der Software.
- **Delegierung von Verantwortung!**
- Manchmal kann ein(e) falsch-negativ Beurteilte(r) **die Vorhersage prinzipiell nicht entkräften!**
 - Z.B. abgelehnte Bewerberin oder ins Gefängnis gesteckte Kriminelle

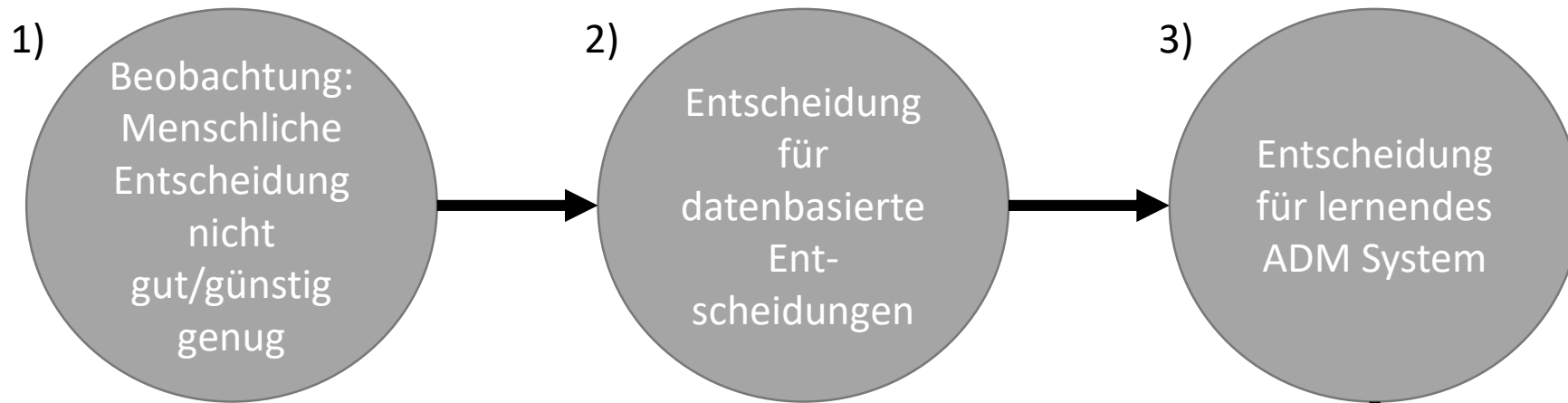
Einschätzung

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

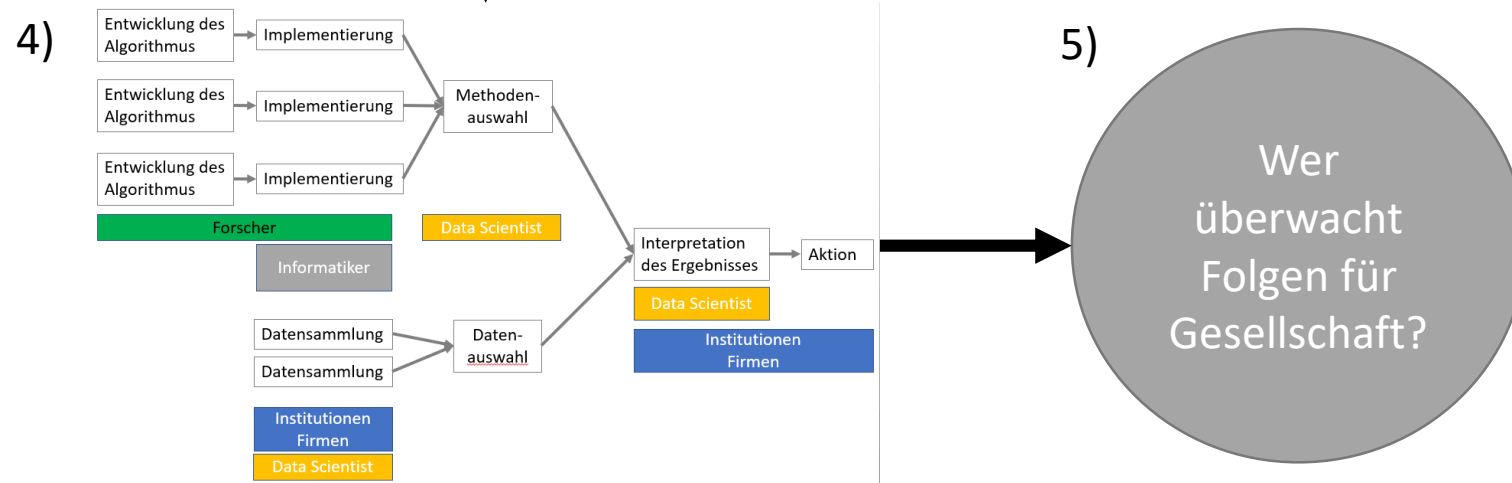


Quis custodiet ipsos algorithmos

Der „Automated Decision Making“-TÜV vulgo: „Algorithmen TÜV“ (Kenneth Cukier und Viktor Mayer-Schönberger: „Big Data“)

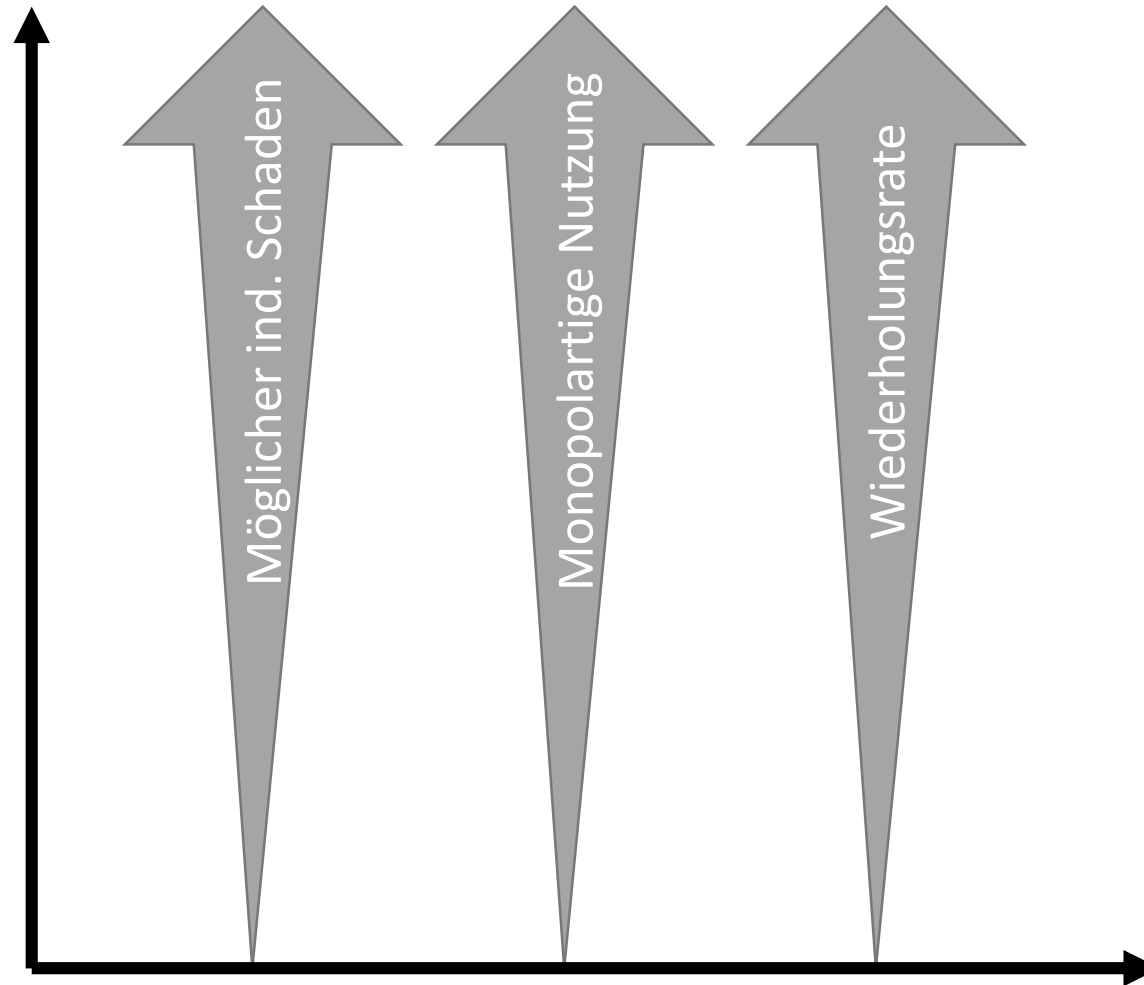


- Was ist gesellschaftliches oder ökonomisches Ziel?
- Was genau wird optimiert?
- Wie wird Datensatz zusammengestellt?
- Wie wird Qualität gemessen?
- Wer entscheidet dies?



Notwendigkeit von Technikfolgenabschätzung und Technikfolgenüberwachung

Technikfolgen-
abschätzung
und
Technikfolgen-
überwachung
notwendig



Im Nachhinein,
bei Verdachtsfall
ausreichend?

Wie könnte ein „Algorithmen-TÜV“ aussehen?

- Unabhängige Prüfstelle mit Siegelvergabe
- Möglichst auch mit Forschungsauftrag
- Identifikation der **kleinstmöglichen Menge** an zu überprüfenden Algorithmen
 - Die meisten Algorithmen sind harmlos;
 - Produkthaftung ermöglicht, dass andere, z.B. Versicherungen, Interesse an korrekten Algorithmen haben;
 - Wettbewerb ermöglicht, dass andere ‚neutralere‘ Algorithmen anbieten.
 - **Kein weiteres Innovationshemmnis!**
- **Non-Profit**

Beipackzettel für Algorithmen



Welches Problem „kuriert“ der Algorithmus?

Was ist das Einsatzgebiet des Algorithmus, was seine Modellannahmen?

Welche „Nebenwirkungen“ hat der Algorithmus durch seine Einbettung in einen sozialen Prozess?

Schlussformel

... zu Risiken und Nebenwirkungen der Digitalisierung befragen Sie bitte Ihren nächstgelegenen Data Scientist oder den deutschen Algorithmen TÜV.

Weitere Informationen



1. Broschüre der Bayerischen Landesmedienanstalt
Kostenlos zu beziehen von der BLM
Googlen nach „BLM Dein Algorithmus - meine
Meinung“

Prof. Dr. Katharina A. Zweig
zweig@cs.uni-kl.de
@nettwerkerin bei Twitter

2. Studie für die
Bertelsmann-Stiftung (2018)

