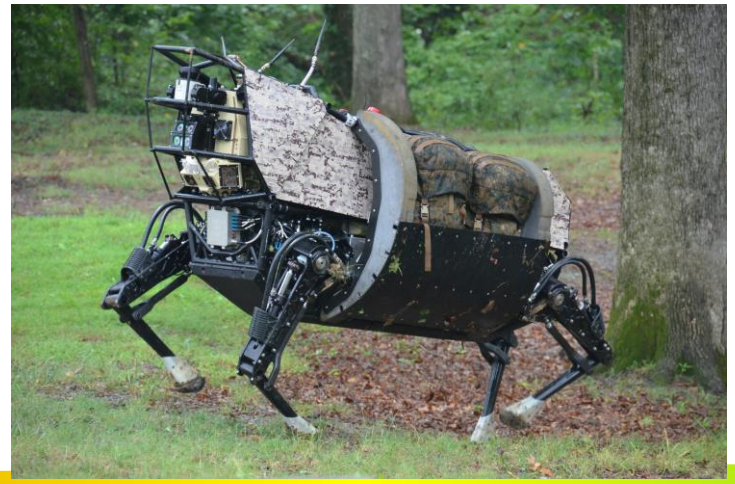
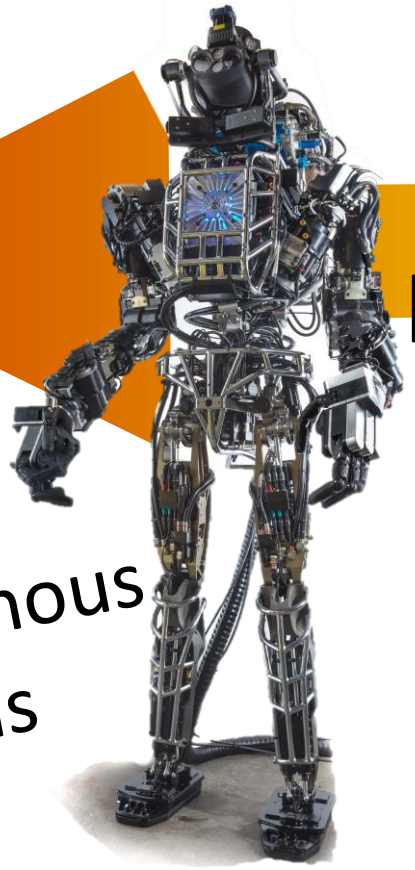
Mehr Sicherheit durch Algorithmen?

Gewerkschaft

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

DeepFake





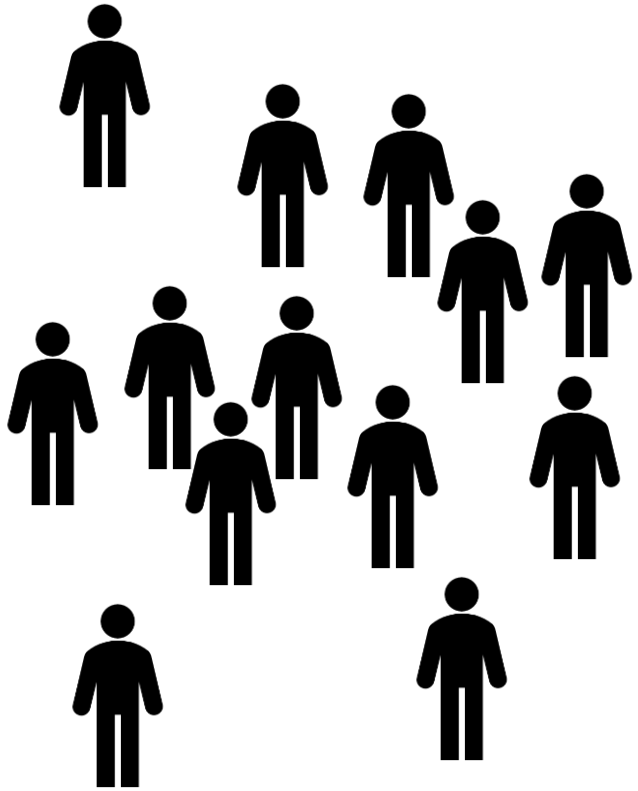
LAWs

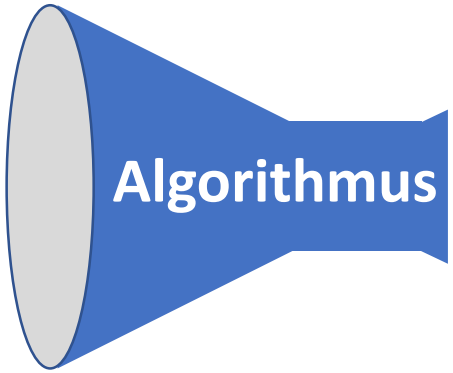
Lethal autonomous weapons

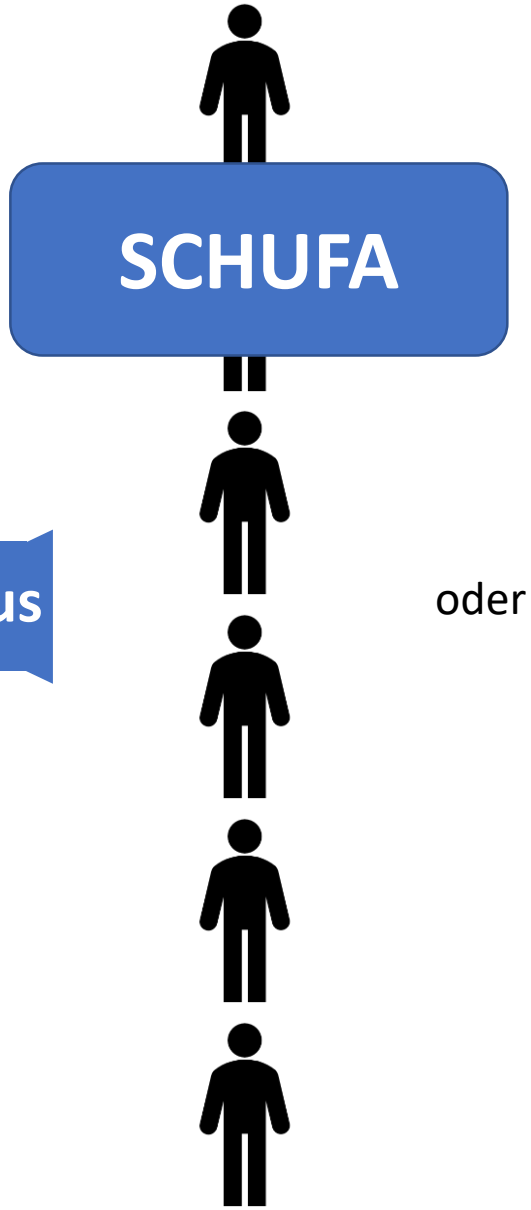
Künstliche Intelligenz und Sicherheit

Risk Assessment von Personen

Algorithmische Entscheidungssysteme







Scoring-Verfahren

oder

Klassifikation

Wer soll richten?

Unser Menschenbild

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

Menschen – so irrational!

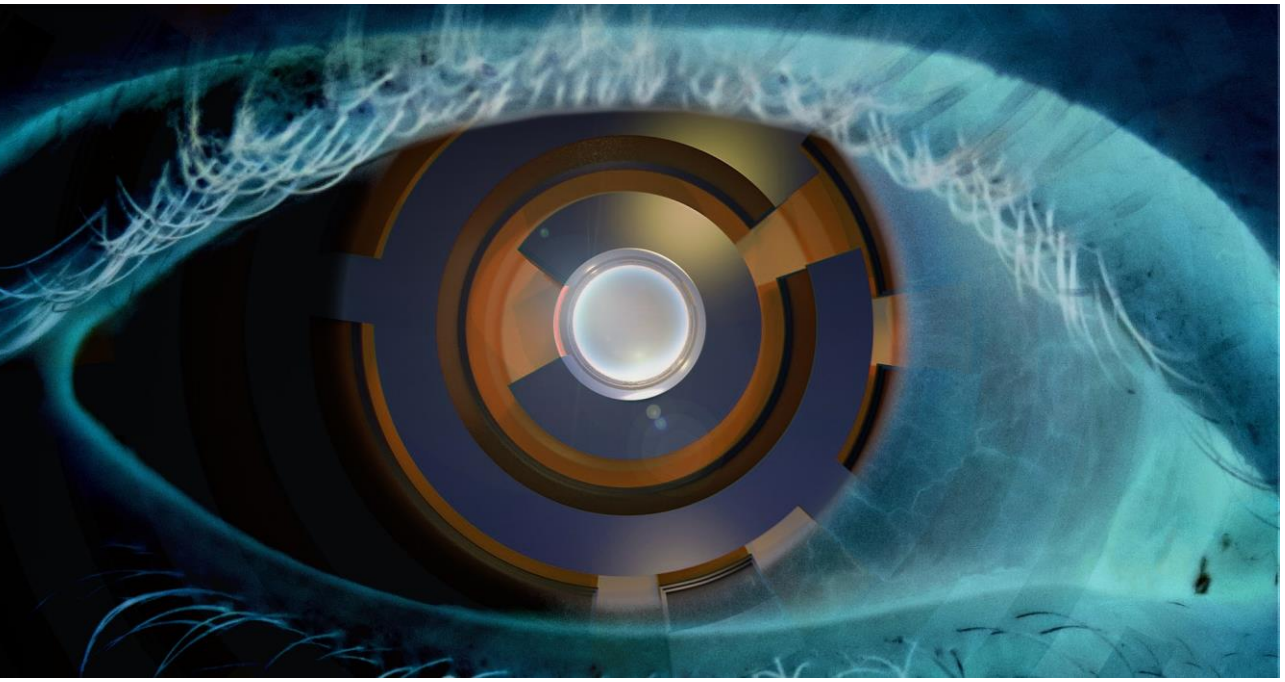
- Richter müssen vorzeitige Haftentlassungsanträge begutachten.
- Studie: je weiter von der letzten Pause weg, desto weniger risikoreiche Entscheidungen[1].
- Eine Vielzahl solcher Studien scheint zu beweisen:
  - Menschen sind irrational und vorurteilsbeladen.

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

Problemfall USA

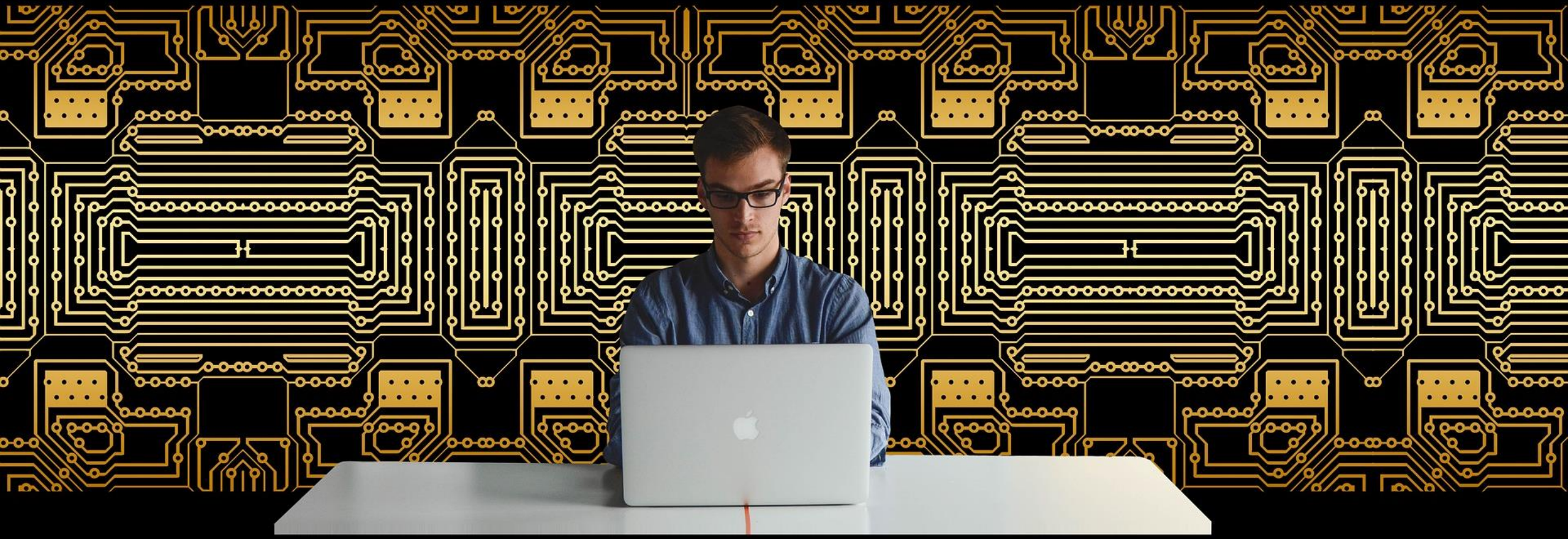
- Zweithöchste Inhaftierungssrate weltweit.
- 6x höhere Rate von Afroamerikanern und 2x höhere Rate von Latinos als von Weißen.
- Prognose: jeder dritte afroamerikanische Junge im Alter von 10 Jahren wird eine Gefängnisstrafe absitzen müssen.

American Civil Liberties Union

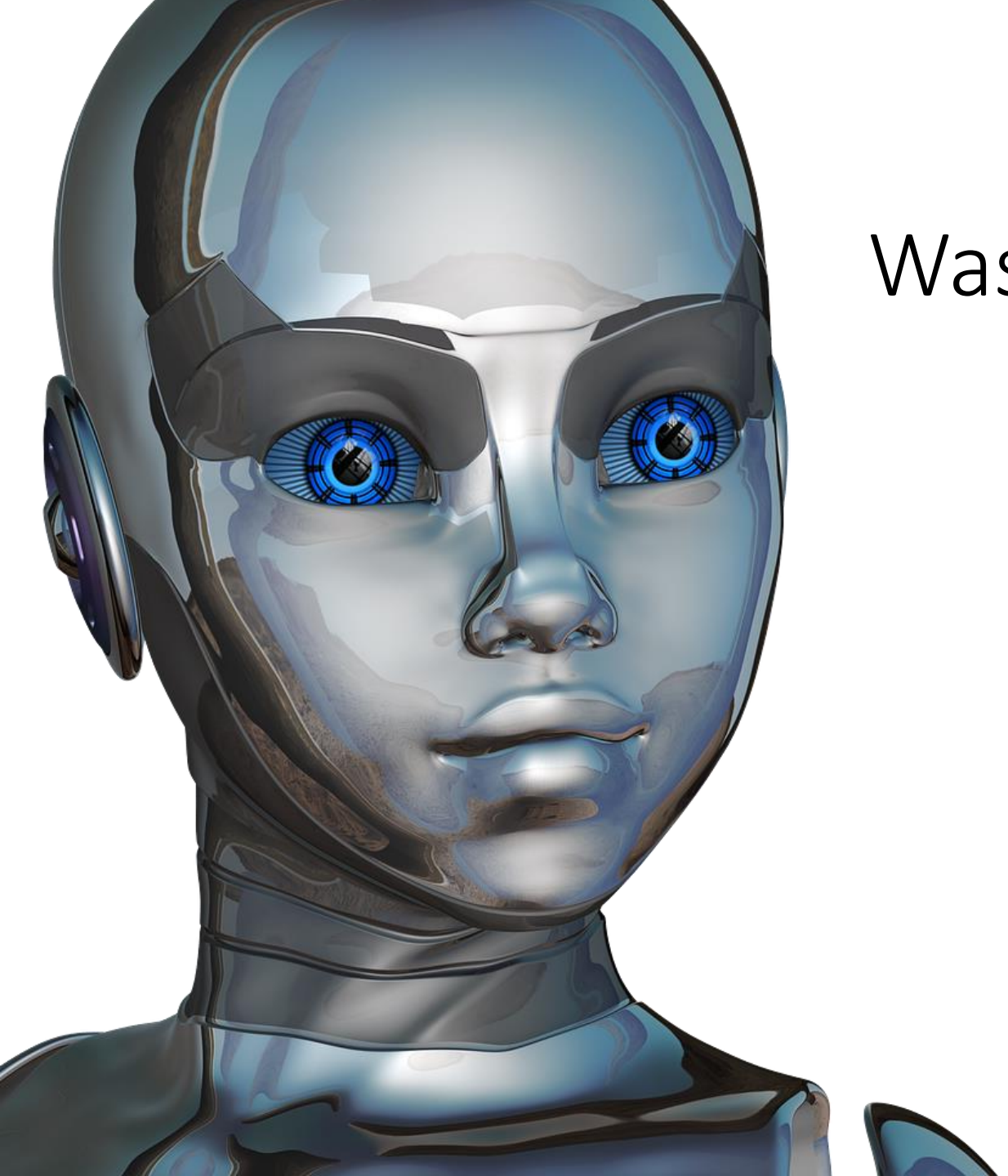


- Amerikanische Bürgerrechtsunion (seit 1920) fordert:
- Algorithmische Entscheidungssysteme sollten überall im Prozess eingesetzt werden, …
- … um Fairness und Objektivität zu sichern.
- Dazu sollen Computer aus Daten Entscheidungsregeln lernen.

SMART REFORM IS POSSIBLE - States Reducing Incarceration Rates and Costs While Protecting Communities, Report from August 2011, p. 9



Können Computer lernen?



Was heißt Lernen?

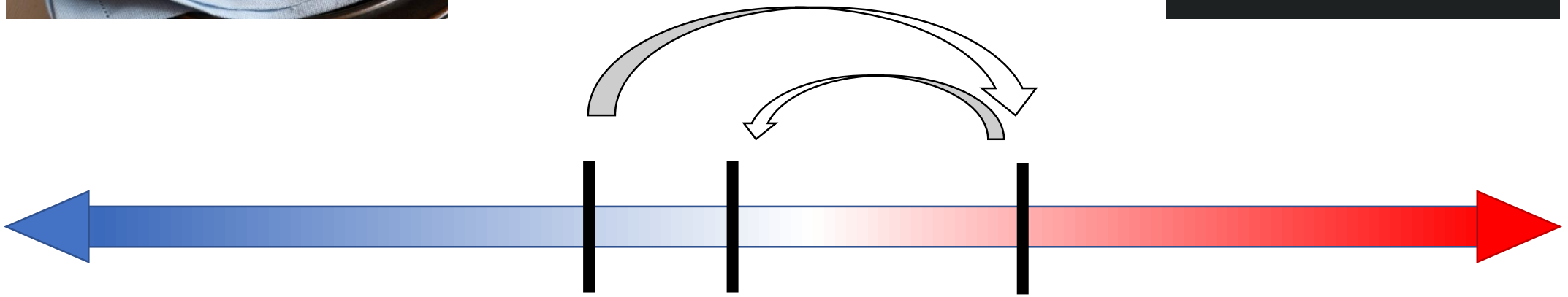
Einfach:

In derselben Situation ein vorher gezeigtes Verhalten wiederholen.

Generalisiert:

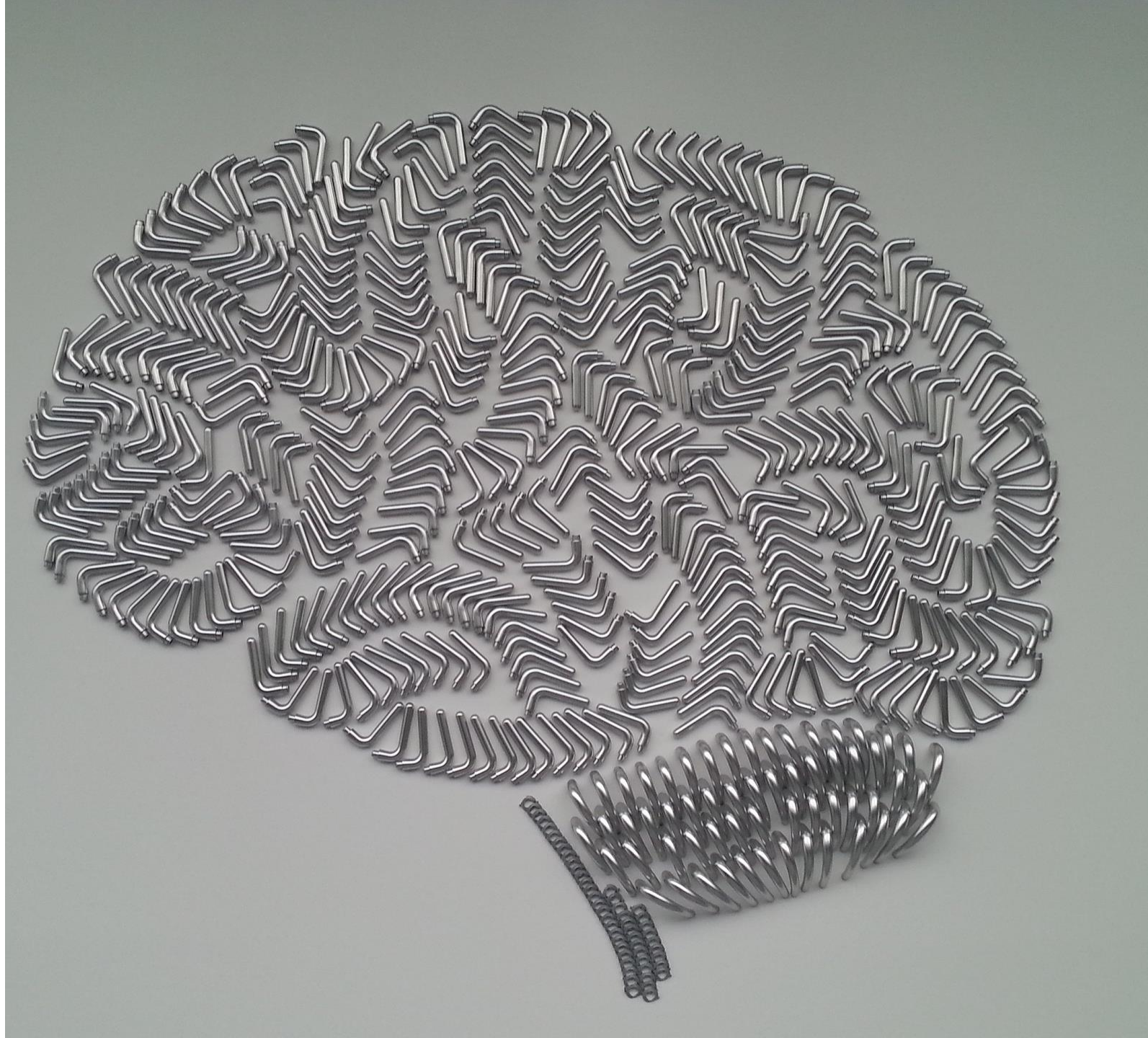
In derselben Art von Situation das richtige Verhalten aus einer Reihe von Möglichkeiten auswählen.

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**.

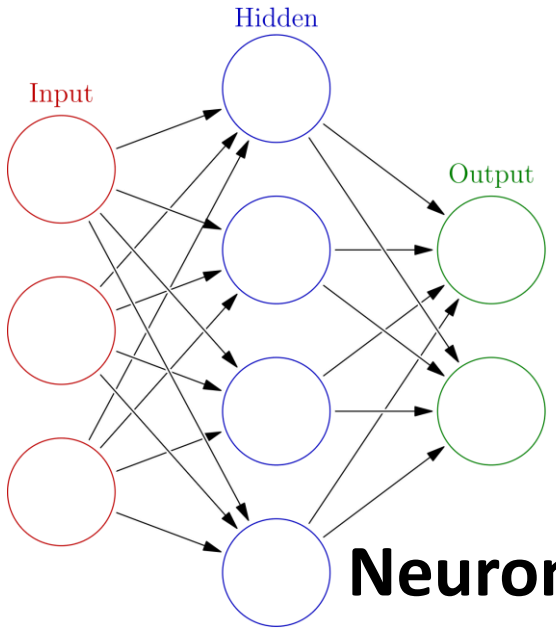


Computer lernen

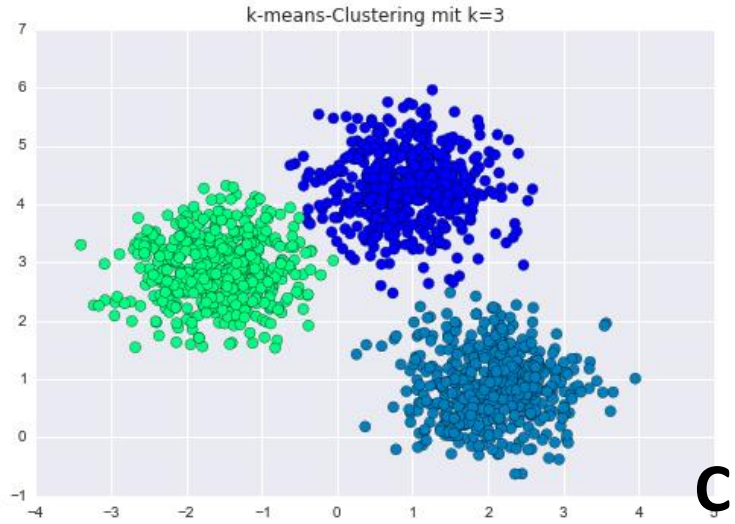
Damit ein Computer lernen kann, benötigt er ebenfalls eine **Struktur,** um Gelerntes abzuspeichern.

Optimal auch **Rückkopplung**.

Er lernt **generelle Regeln**.



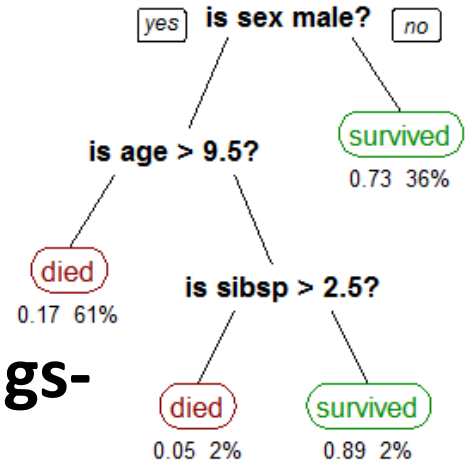
Neuronales Netz



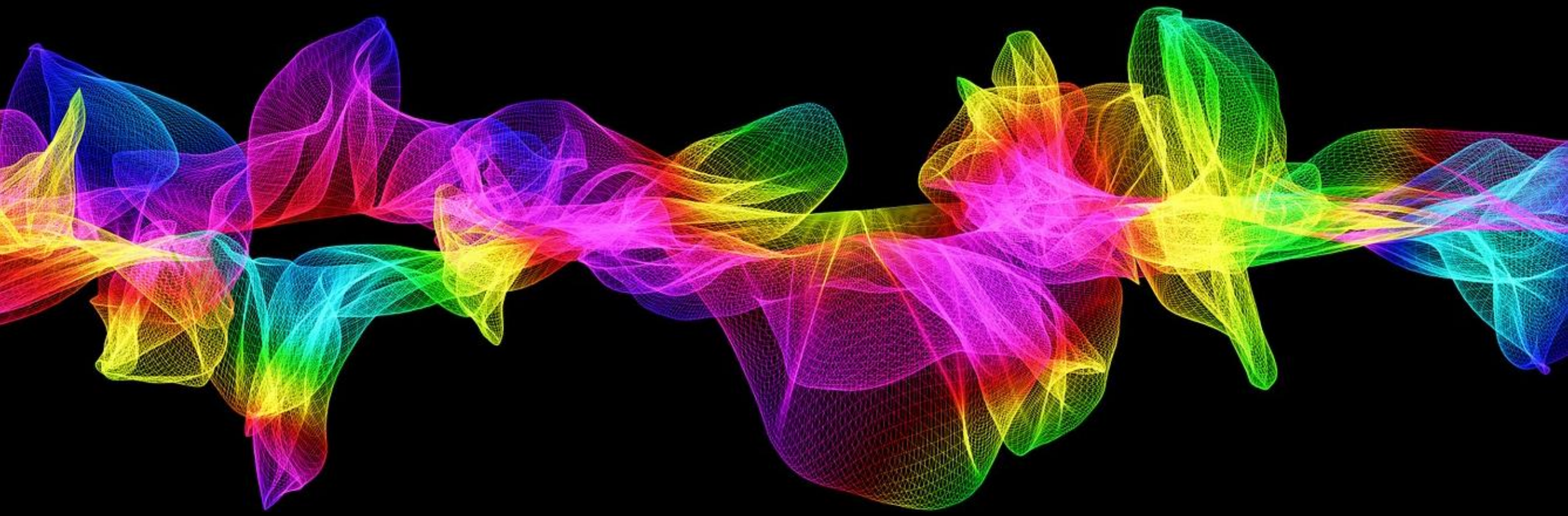
Clustering

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

Formel



Entscheidungs-bäume

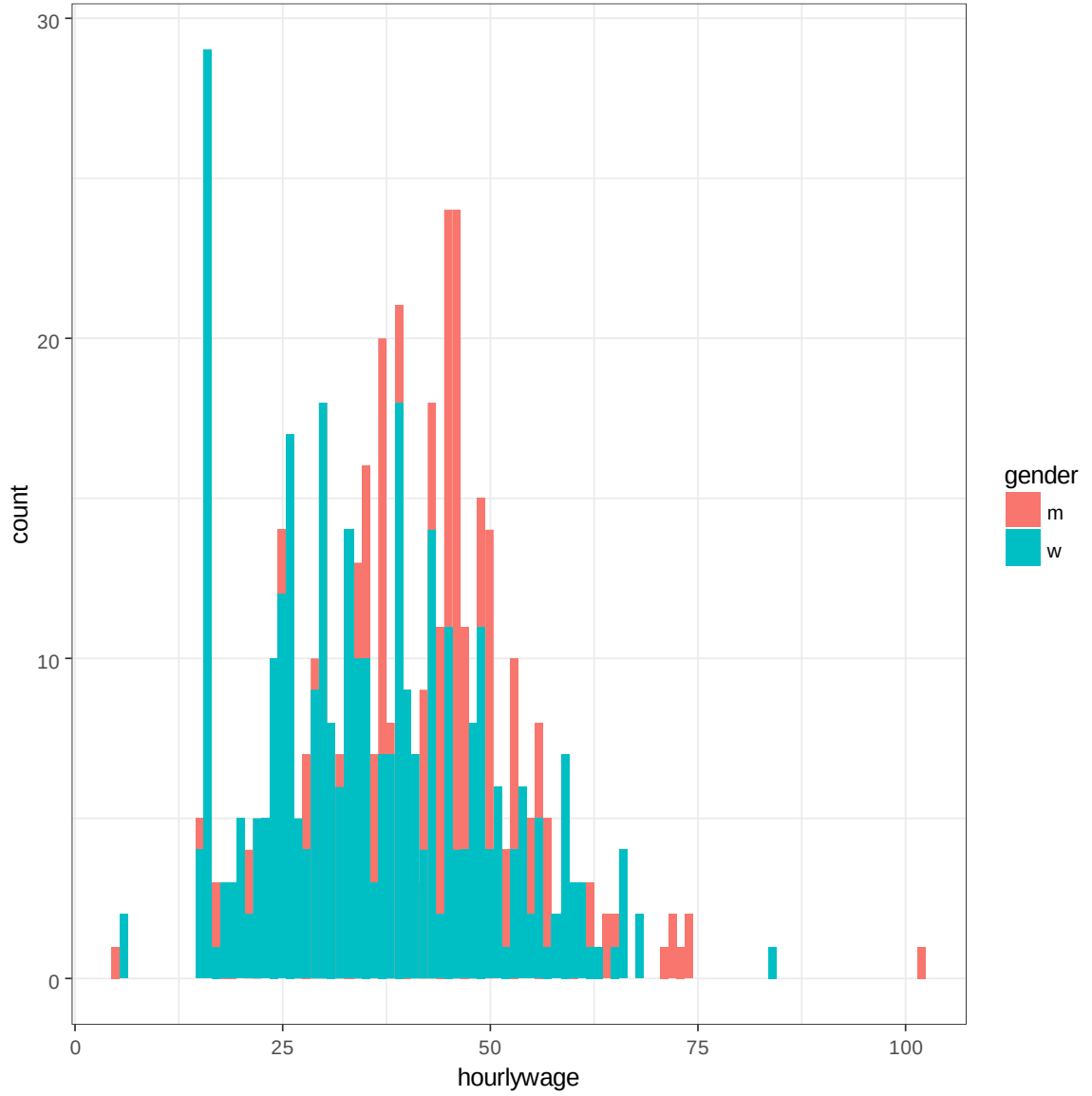


“Lernen” mit Korrelationen

Gehälter in Seattle

Sie bekommen Daten von einer Person – diese verdient weniger als $25 pro Stunde.

Basierend auf den Daten, ist die Person weiblich oder männlich?



@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

Lernen mit Formeln

Individuelle Risikobewertung der Rückfälligkeit von Kriminellen

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.

Regressionsansätze

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

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

Allgemein

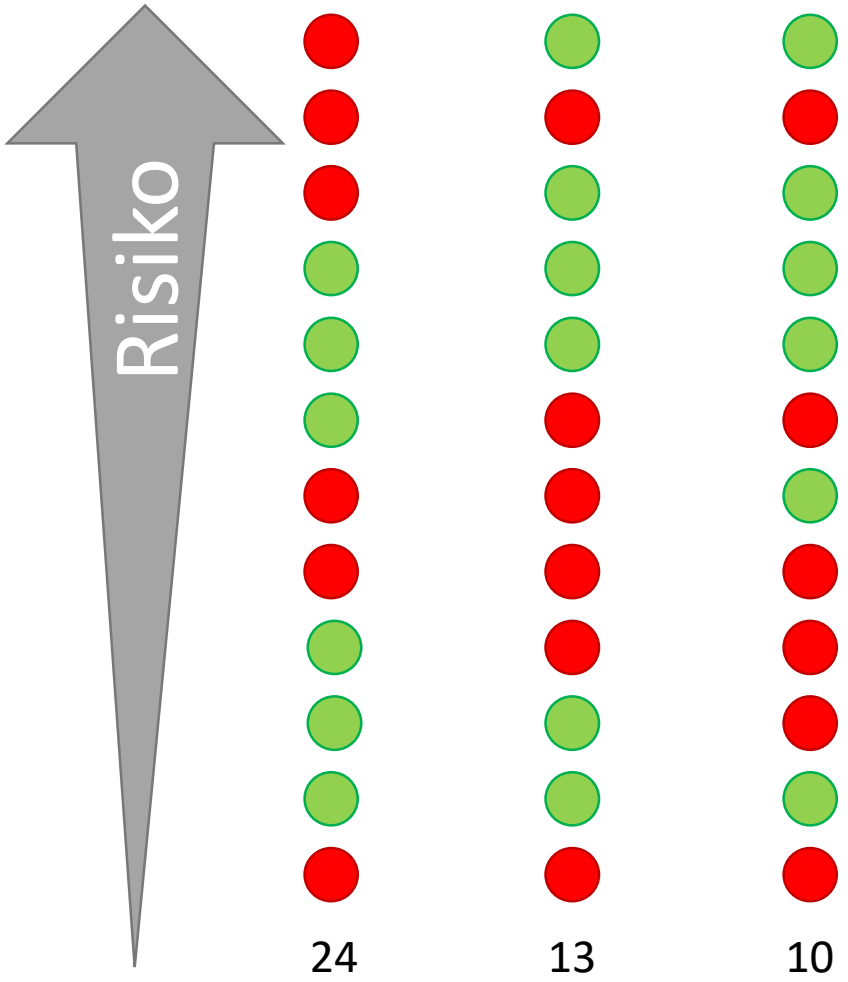
$$
\begin{aligned}
& w_1 * \text{ bisherige Verhaftungen} \\
- & w_2 * \text{ Anzahl Tage seit letzter Verhaftung} \\
+ & w_3 * \text{ (Wenn Mann, dann 1, sonst 0)} \\
+ & w_4 * \text{ (Wenn Raubüberfall, dann 1, sonst 0)} + ...
\end{aligned}
$$

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.

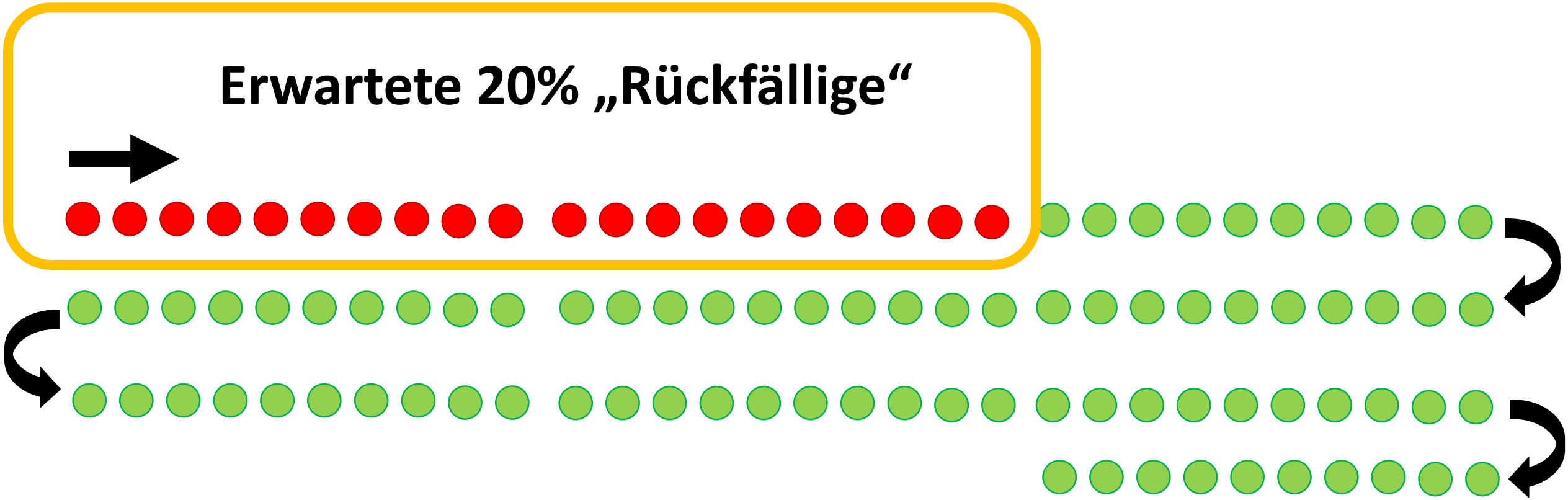
Optimale Sortierung: Alle roten oben, alle grünen darunter.

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

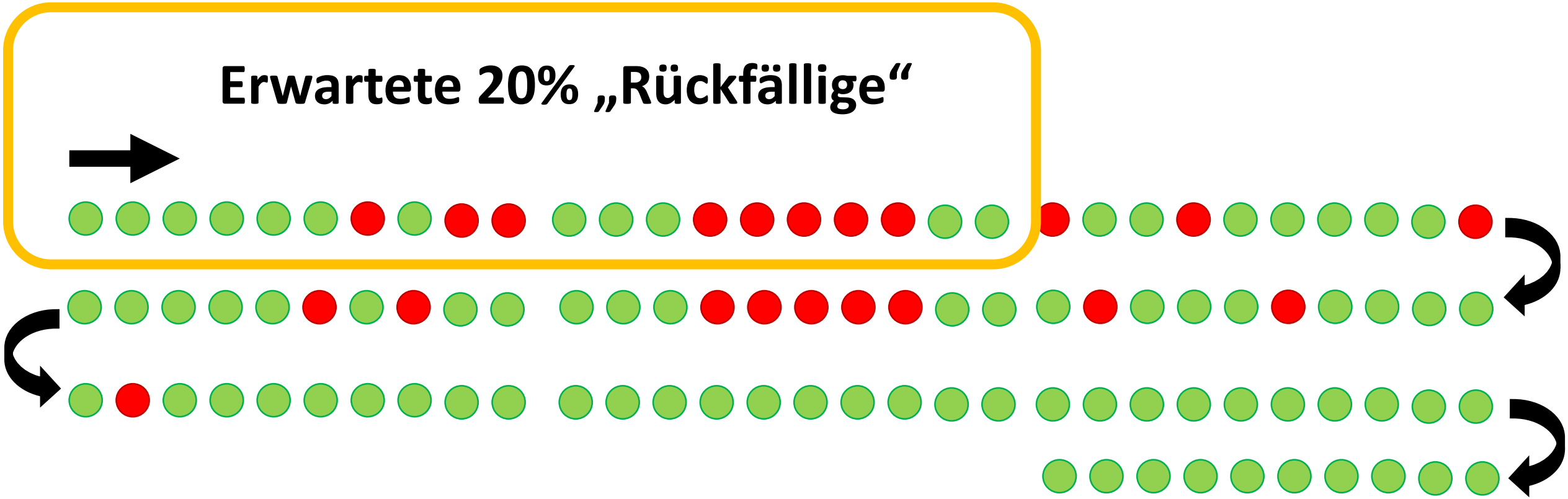
Oregon Recidivism Rate Algorithm

- 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



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



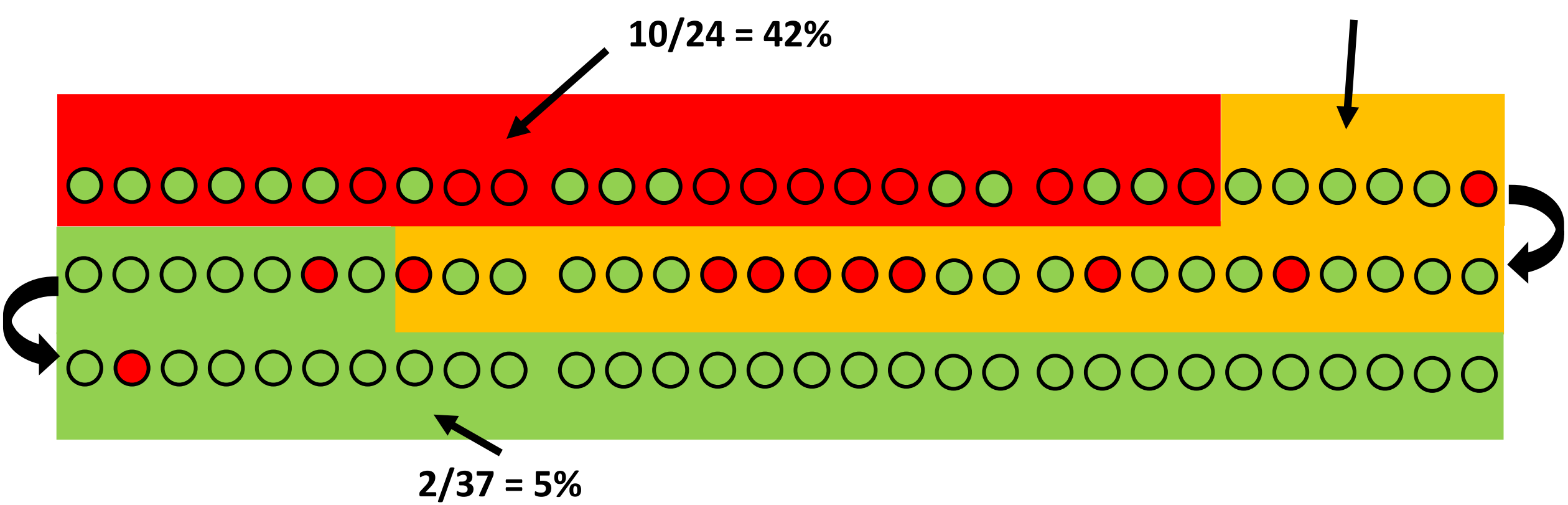
einen Jagdhund zu kaufen,

um Schafe zu hüten.

Das ist wie...

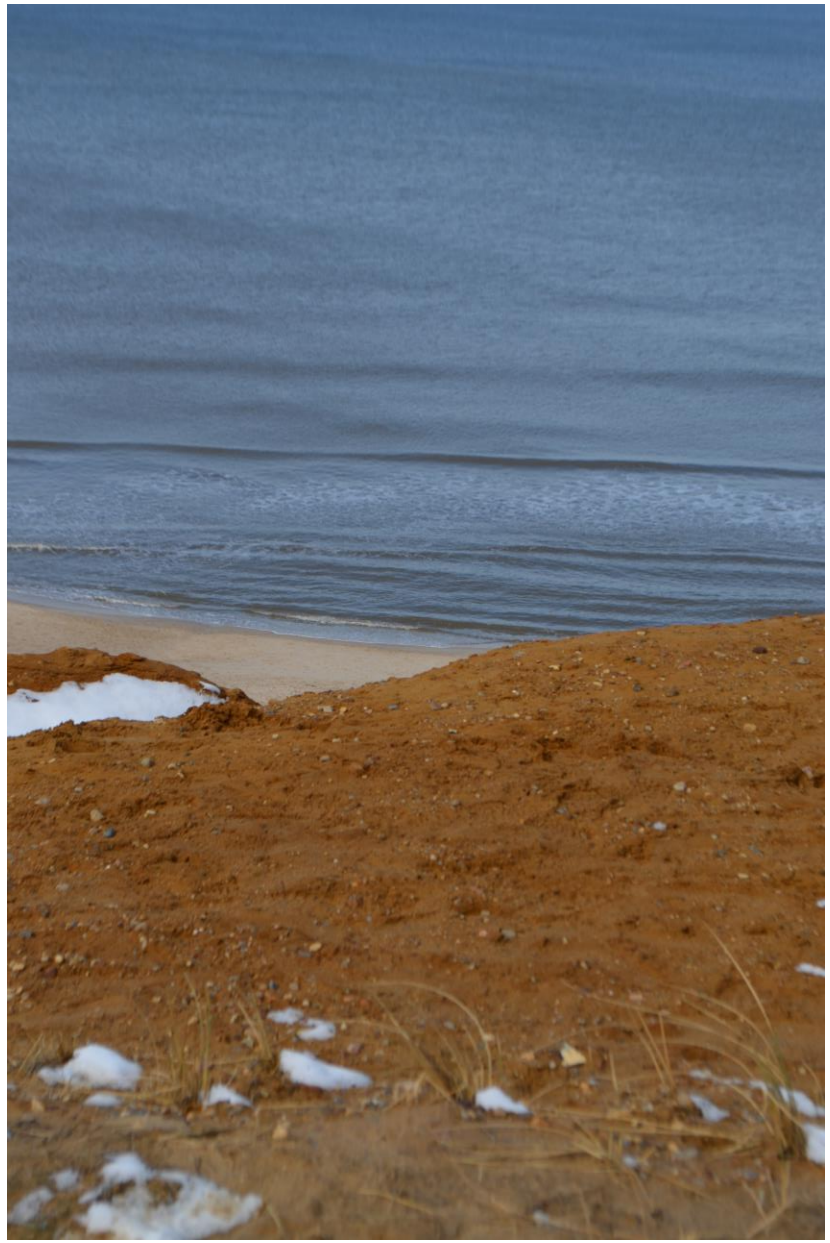
Vom Scoring zur Klassifikation

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



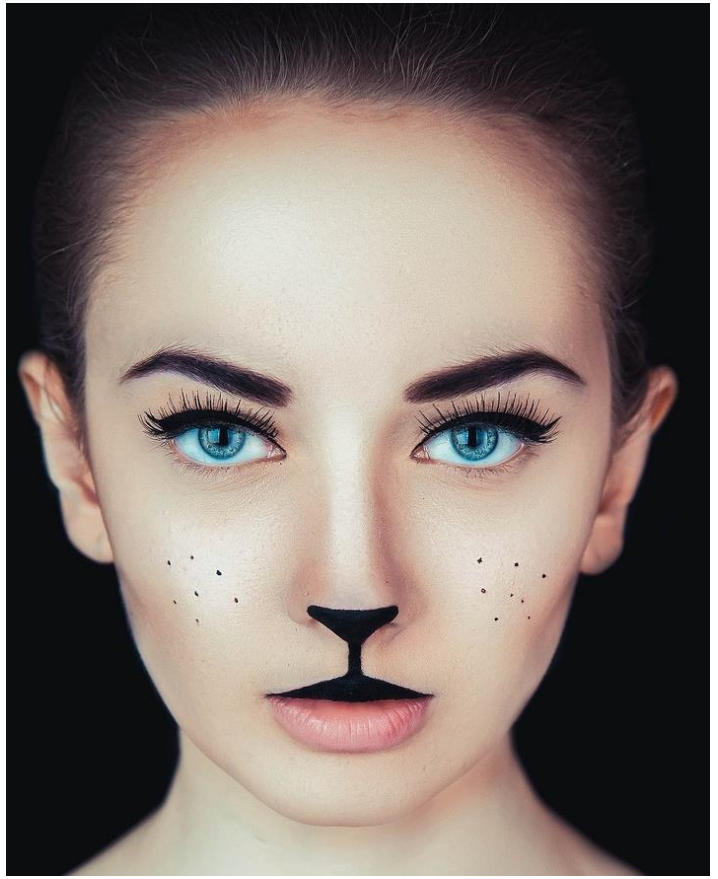
Statistische Vorhersagen über Menschen

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/

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**)

Können uns Algorithmen vor “racial profiling” bewahren?

Können Algorithmen diskriminieren?

Gleichberechtigung

Wenn man auf Google nach „CEO“ sucht…

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

Diskriminierung

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

Algorithmen in einer demokratischen Gesellschaft

Generell

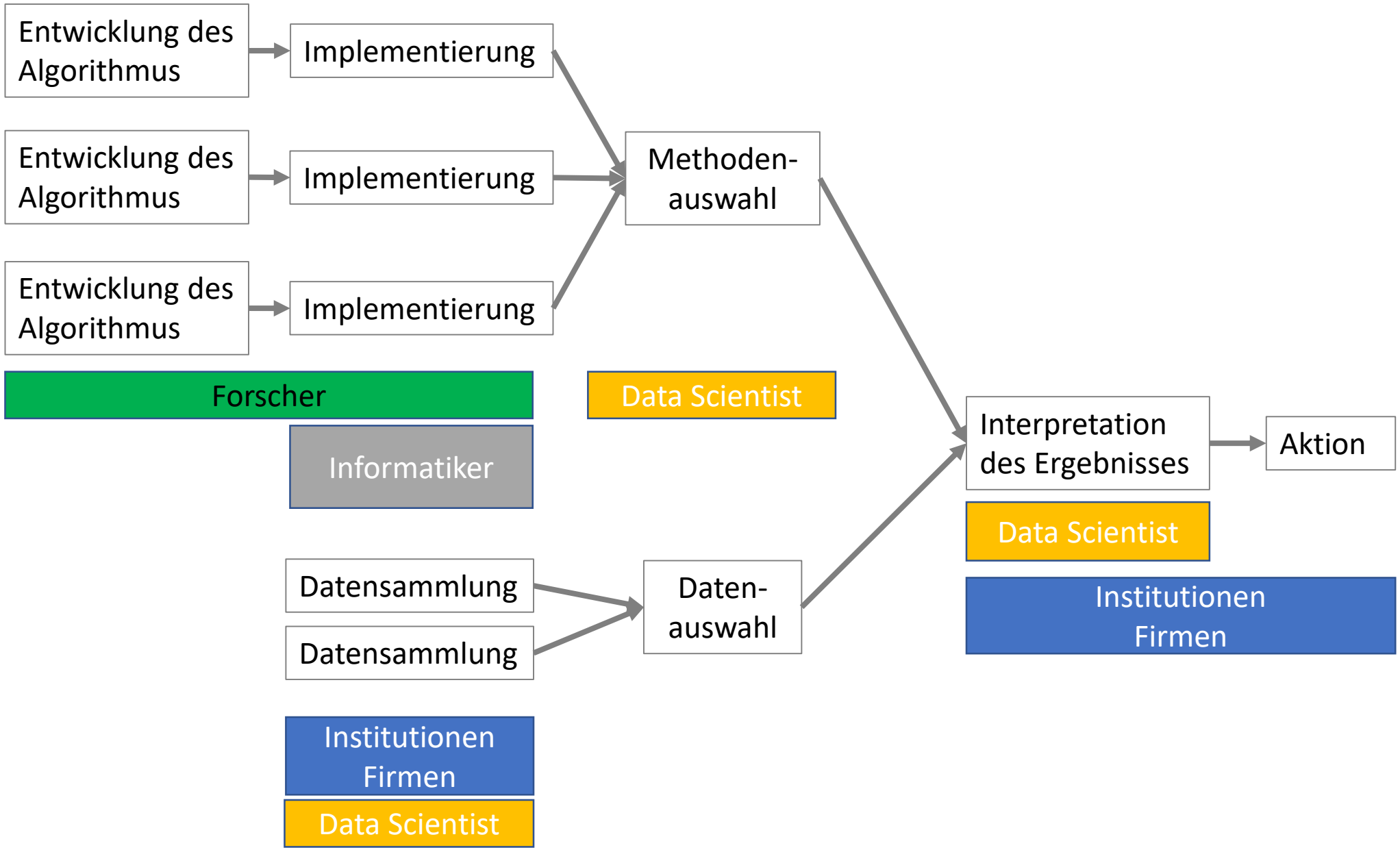
Prinzipiell können algorithmische Entscheidungssysteme für sehr viele, schwierige Fragestellungen in derselben Art gebaut werden:

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

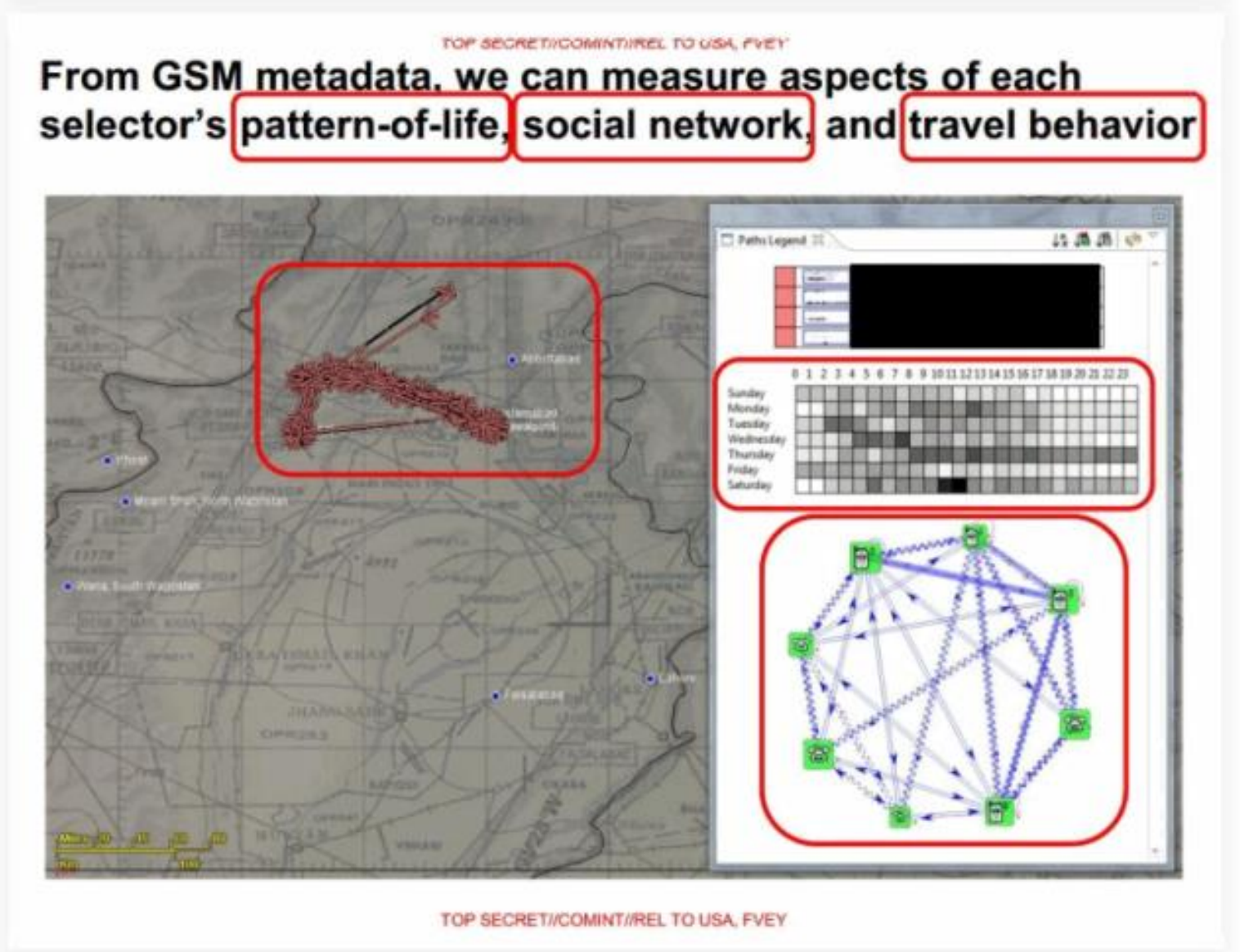
Ihre Aufgabe heute….

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

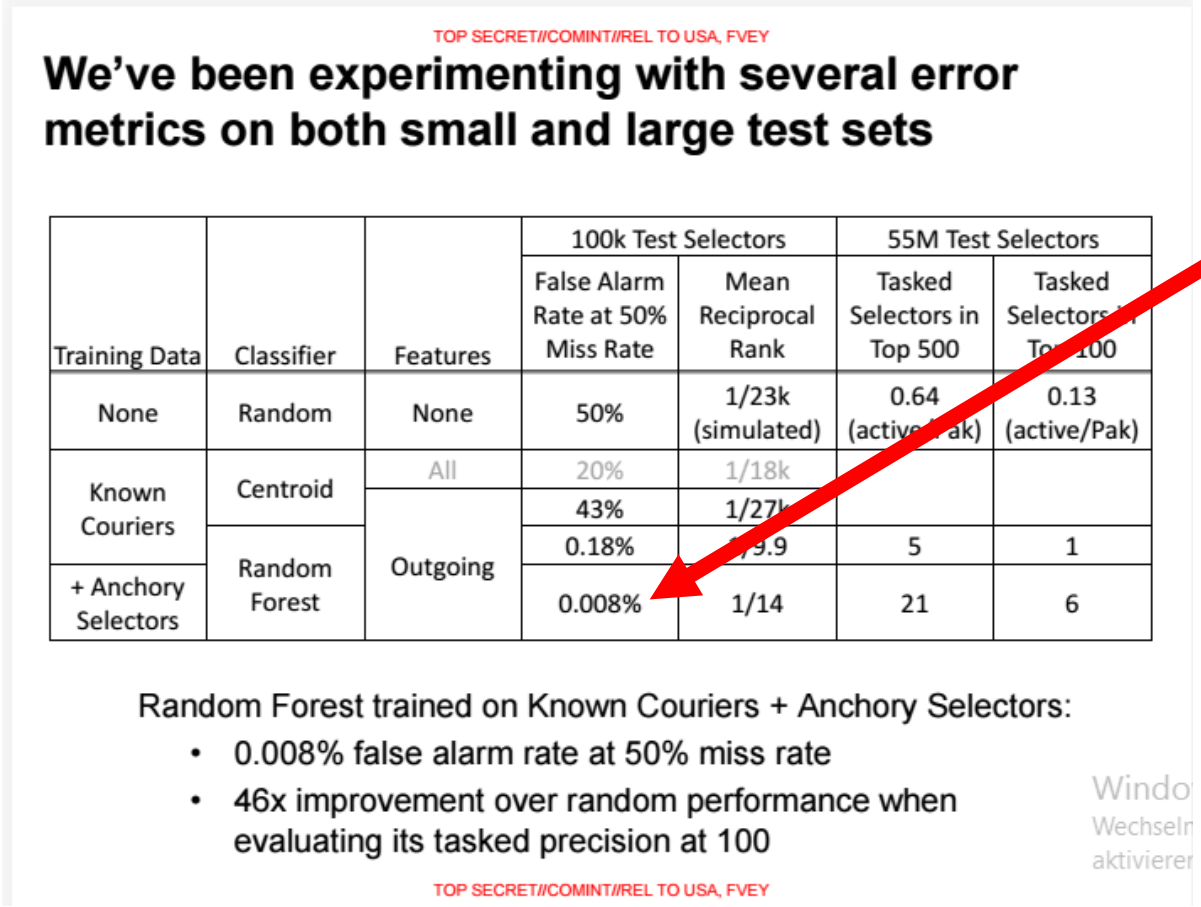
Designprozess



Capturing terrorists with network analysis



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		
	Random Forest		0.18%	1/9.9	5	1
+ Anchory Selectors			0.008%	1/14	21	6

Random Forest trained on Known Couriers + Anchory Selectors:

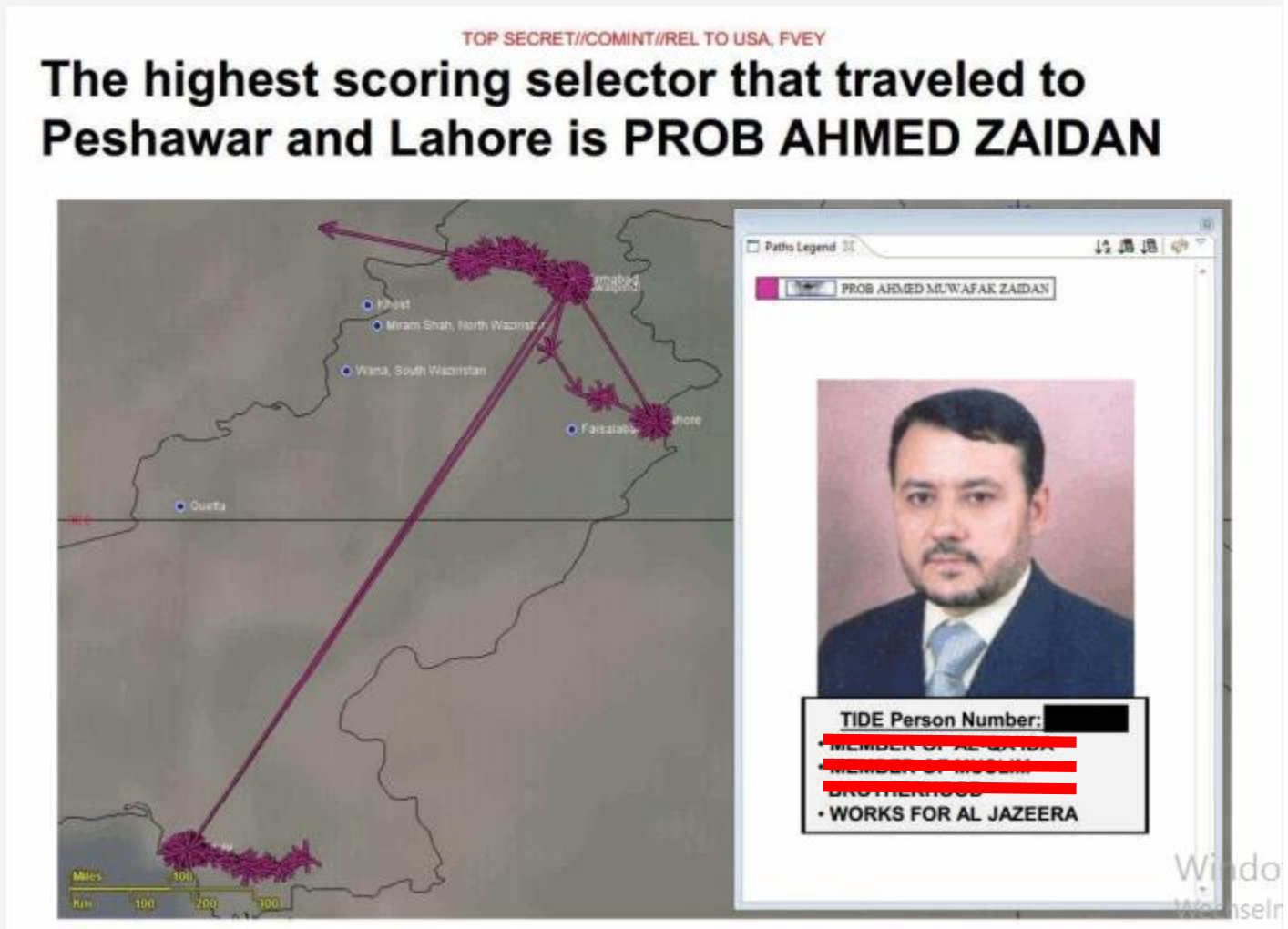
- 0.008% false alarm rate at 50% miss rate
- 46x improvement over random performance when evaluating its tasked precision at 100

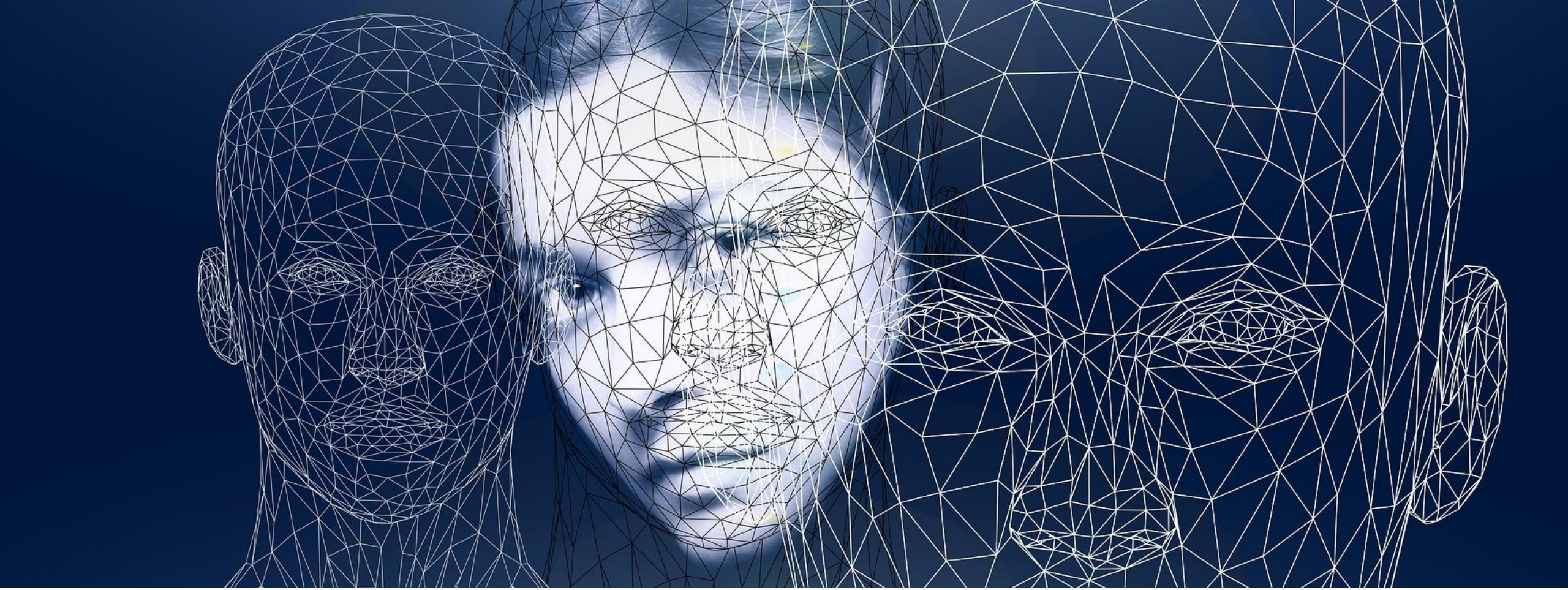
TOP SECRET//COMINT//REL TO USA, FVEY

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 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!**
  - Abgelehnte Bewerber und Bewerberinnen,
  - in Lager verschleppte Verdächtige.

Sozioinformatische Gesamtbetrachtung

Ein Beispiel

- Glaube: Softwarefirmen vermitteln den Eindruck, „echte“ AI wäre quasi nur noch eine Frage von Jahren:
  - Z.B. Sophia, der erste Roboter, der eine Staatsbürgerschaft erhielt.
  - Aber auch Elon Musk, der warnt vor LAWs, die automatisch ihr menschliches Ziel erkennen und gezielt töten.
- Fakt: Die Bilderkennung ist sehr gut bei Standardfotos, weniger gut im Video und in Menschenmengen.
- Reaktion:
  - Camouflage-Makeup, veränderter Gang, Haltung, …. machen es noch wahrscheinlicher, dass die falsche Person attackiert wird.
  - „Adversarial AI“: Gegenlernen von Beispielen, die von der ursprünglichen AI falsch interpretiert werden.

Einschätzung

- Algorithmisch Verfahren **könnten** dabei helfen, bessere Entscheidungen zu treffen.
  - Sie können riese Datenmengen durchsuchen.
  - Sie könnten neue “Muster” entdecken.
  - Könnten Diskriminierung vermeiden.
- Allerdings sind sie heute oft noch nicht gut genug, insbesondere da, wo sehr wenige Menschen identifiziert werden müssen unter vielen Unschuldigen.

Probleme von algorithmischen Entscheidungssystemen (ADM Systemen) im People und 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. **Sie helfen kleine Gruppen zu identifizieren, aber mit vielen „falsch Positiven“ (falsch Verdächtigen).**
5. **ADM Systeme können soziale Prozesse verändern.**
6. **Die Reaktionen können das Problem enorm verschärfen.**

@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

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)