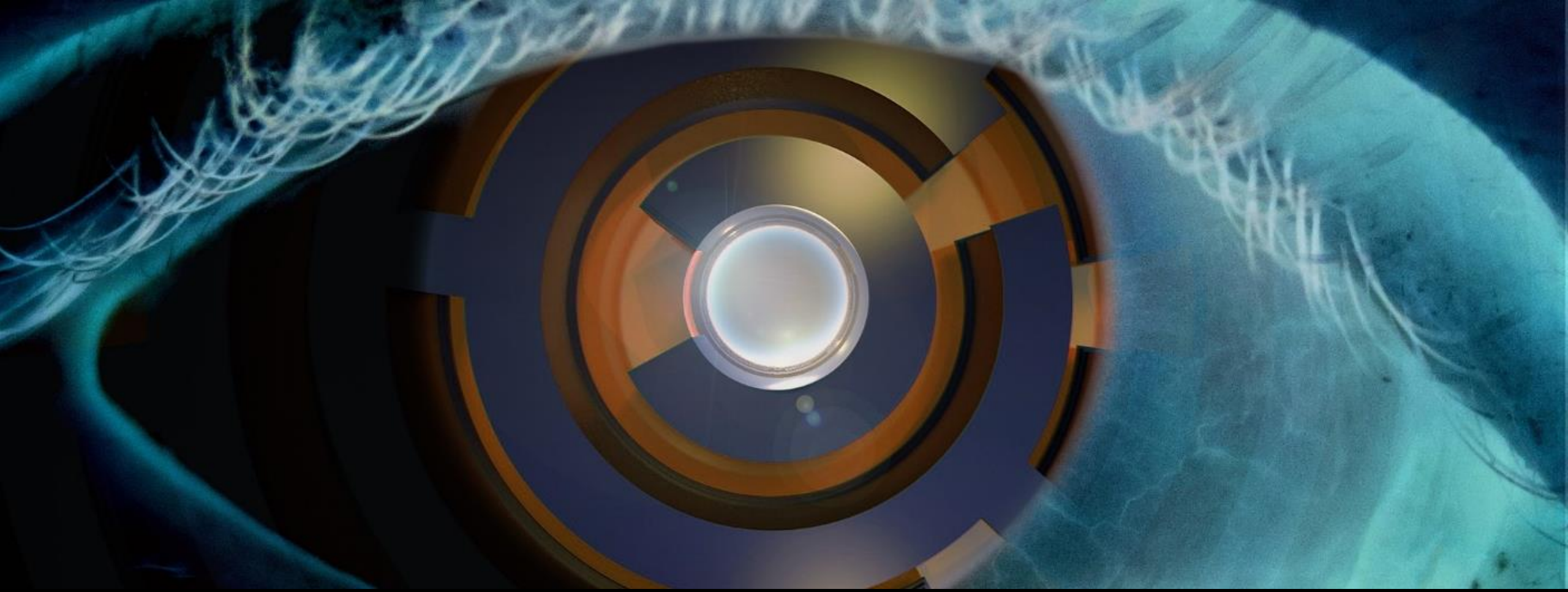


Algorithm Accountability

Auto-Uni Wolfsburg, 7.6.2018

Prof. Dr. Katharina A. Zweig

Leiterin
Algorithm Accountability Lab,
TU Kaiserslautern

@nettwerkerin

Ethik

Griechisch für
“sittliches Verständnis”

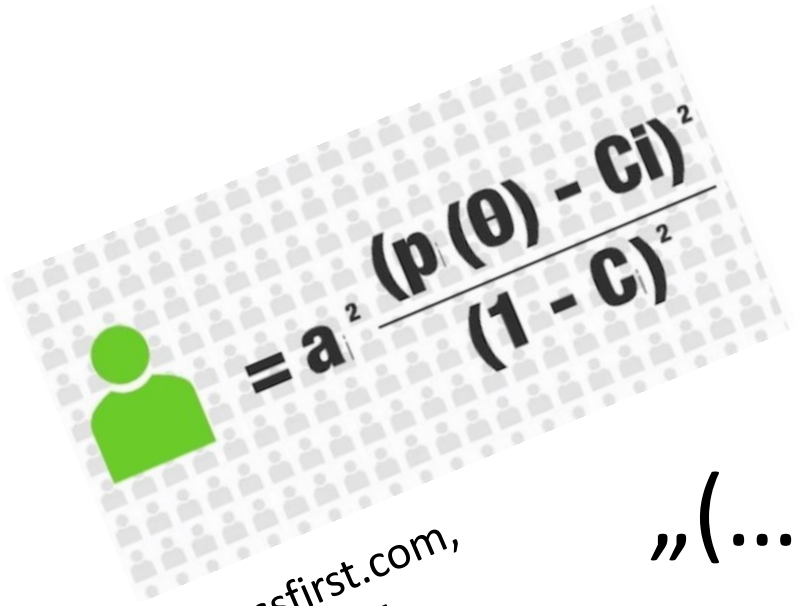
Gibt es eine Datenethik?

Bedingungen und Bewertung menschlicher Taten

Nun…

…Menschen handeln oft irrational…

Kann Data Science und maschinelles Lernen helfen?



Assessfirst.com, 16.11.2017

Let's take the emotion out of the process and replace it with a data-driven approach…"

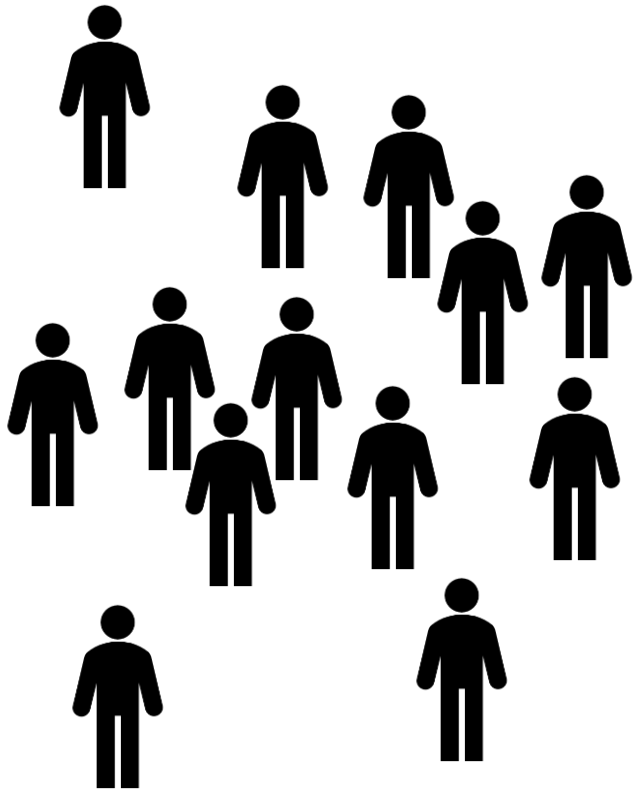
iNostix (by Deloitte), 16.11.2017

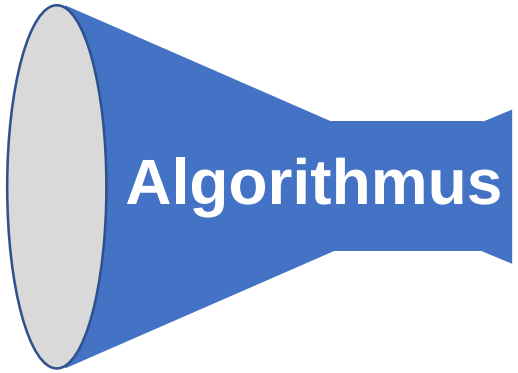
„(…) with the availability of good data, the predictive possibilities are virtually unlimited (…)"

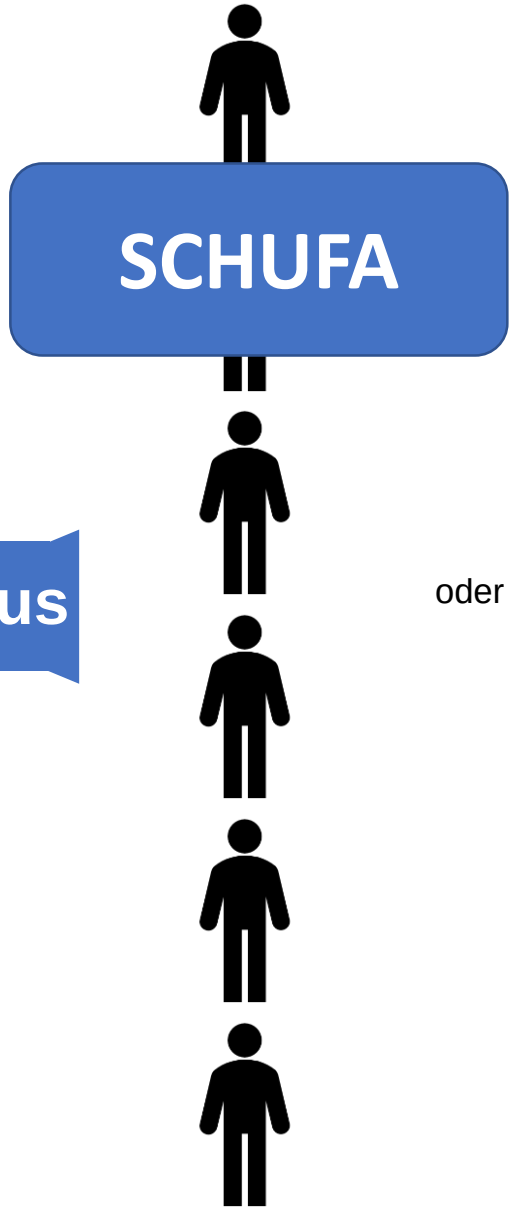
https://www.inostix.com/predict-hiring-success/
16.11.2017

@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

Algorithmische Entscheidungssysteme

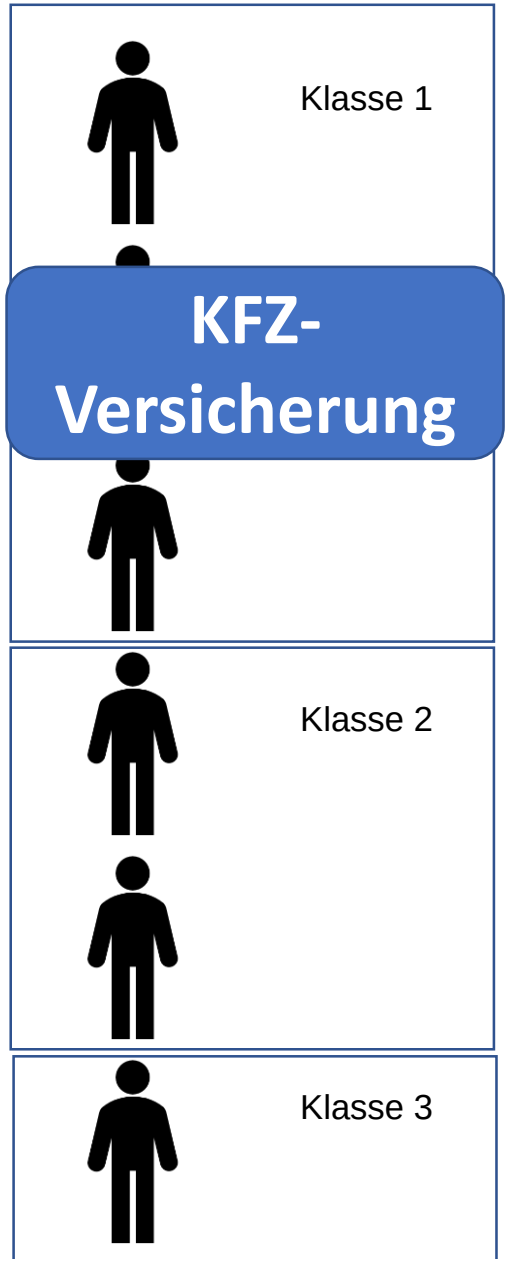






Scoring-Verfahren

oder



Klassifikation

Forschung

Vorhersage des Rückfälligkeitsrisiko von Kriminellen

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

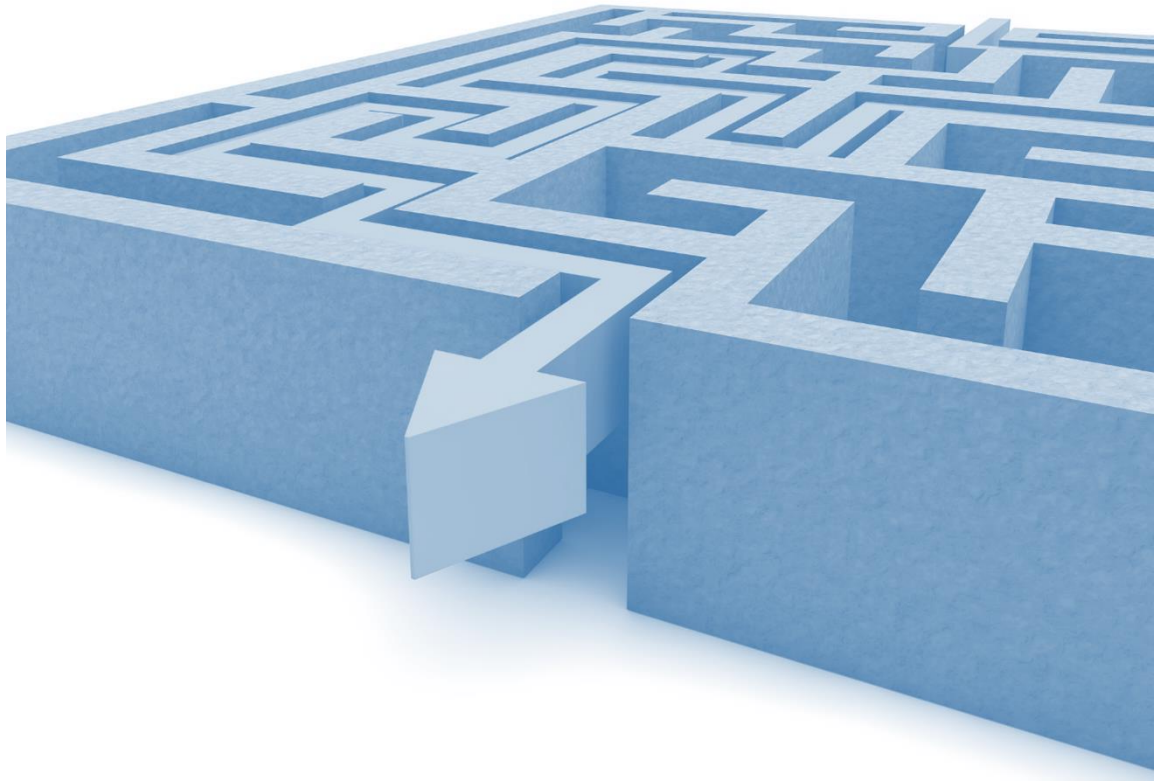
Der PUTPUT der uns sagt, wie Input mit Output zusammenhängt.

INPUT

OUTPUT

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 für ein Problem: Navigation

Navigation

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

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

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

Der **Algorithmus** berechnet die **Lösung –** mathematisch bewiesen.

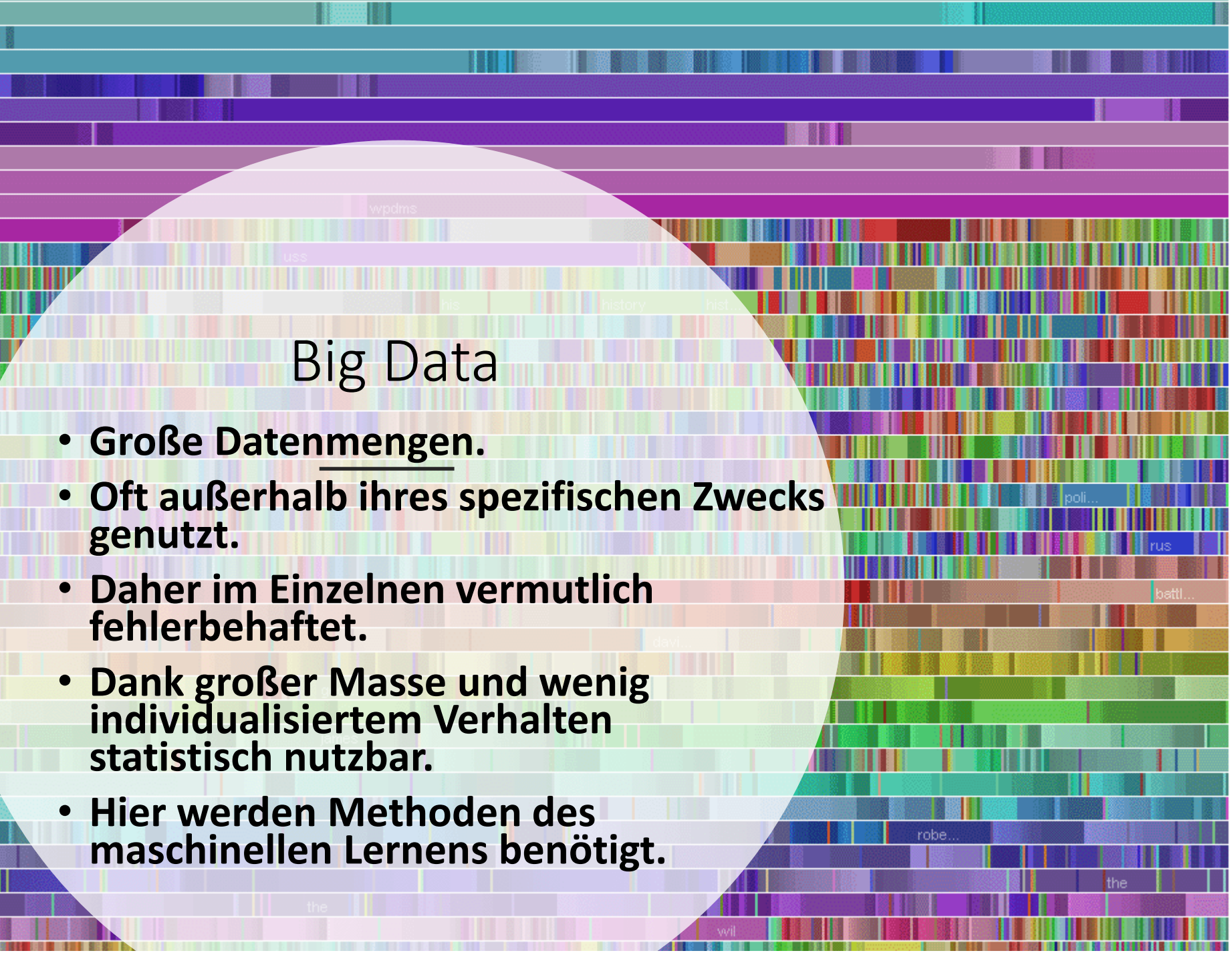
Output: optimale Route

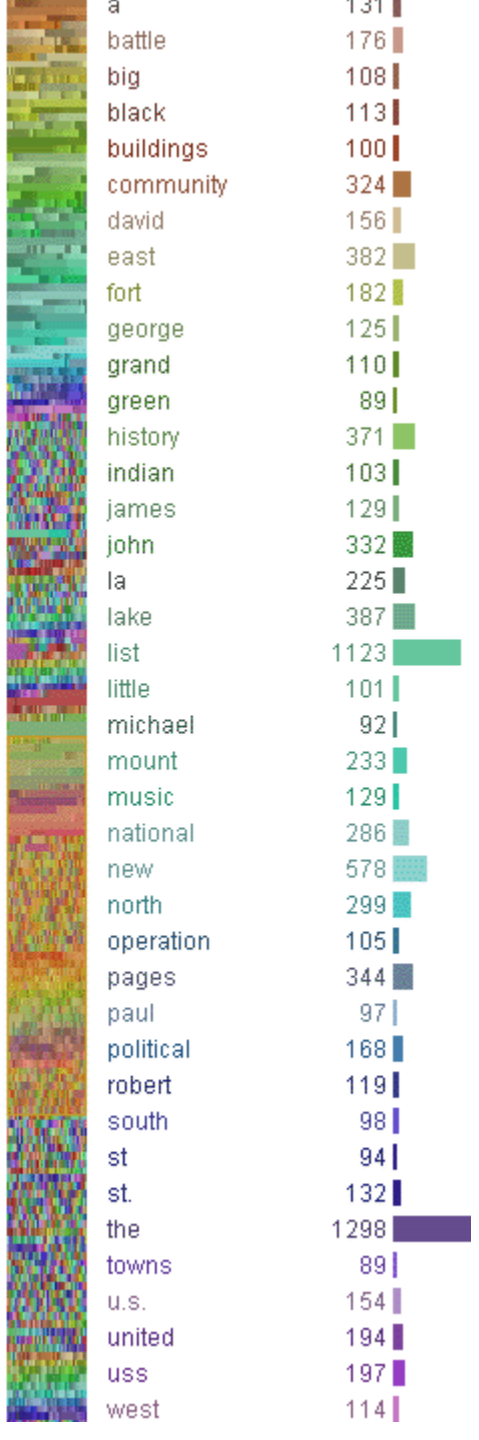
B wie Big Data

Daten als Grundlage

Big Data

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





Dazu gehört auch ein Drowsiness Measuring Decision Making System

C wie Computerintelligenz

Sebastian lernt „heiss“ und „warm“

Juli **März** **September**

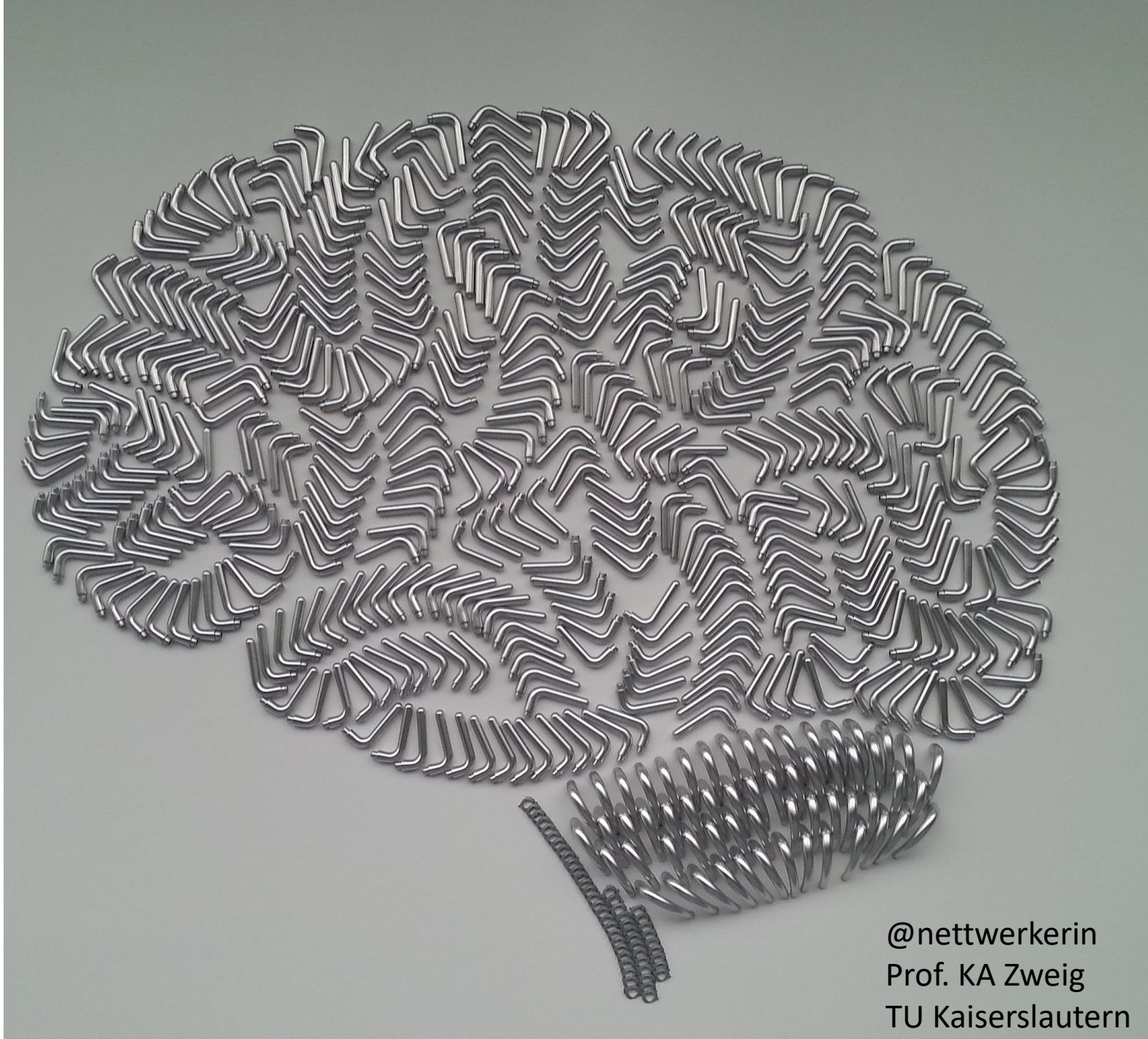
Zu vorsichtig: Darf nicht dampfen Zu mutig geworden

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**.



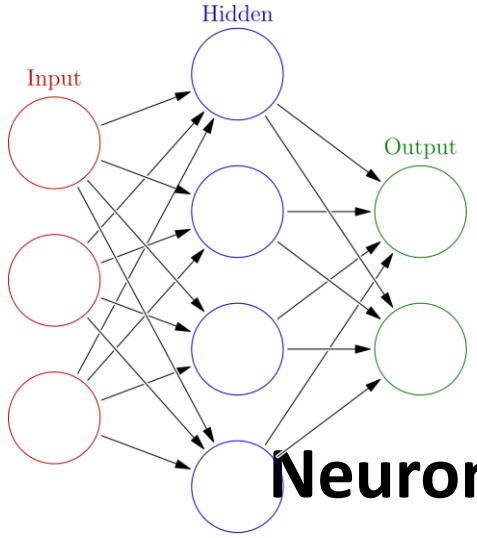
@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

Computer lernen

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

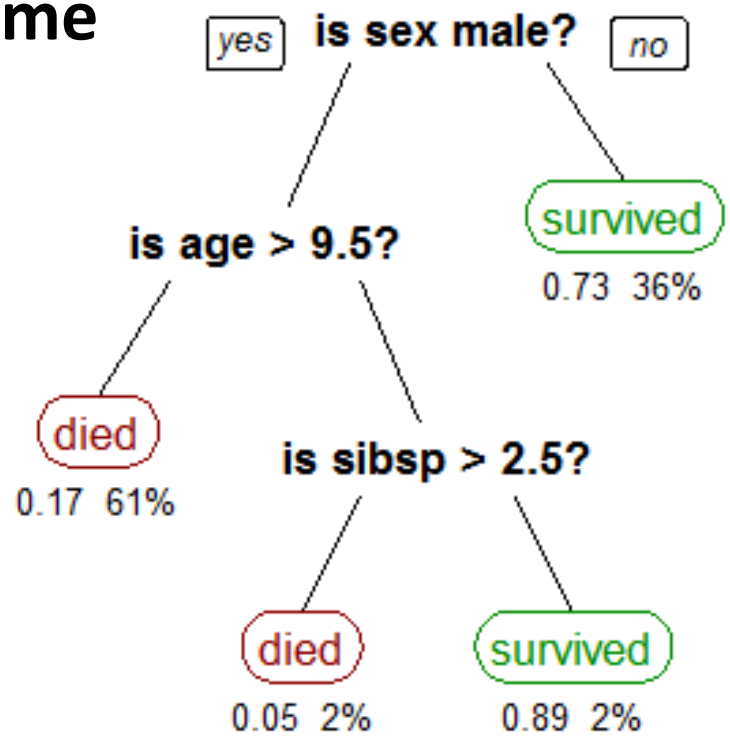
Optimal auch **Rückkopplung**.

Er lernt **generelle Regeln**.



Neuronales Netz

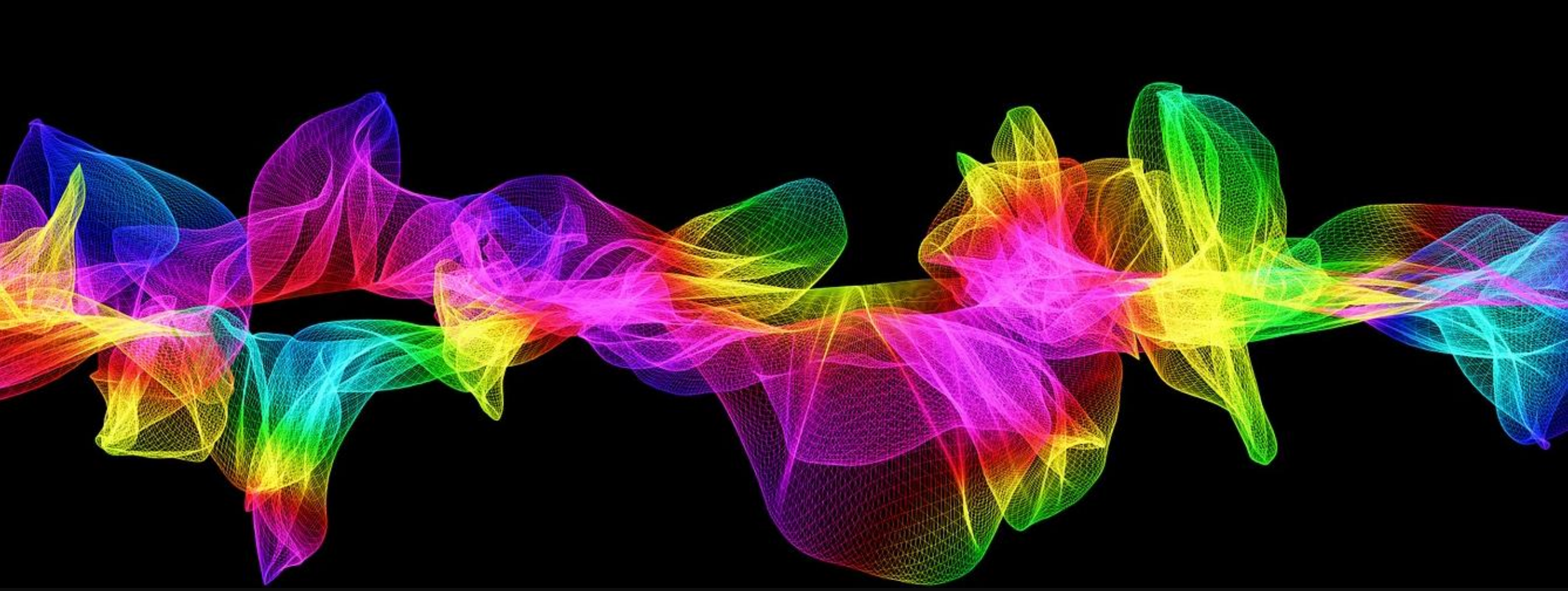
Entscheidungs-bäume



Formel

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

@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

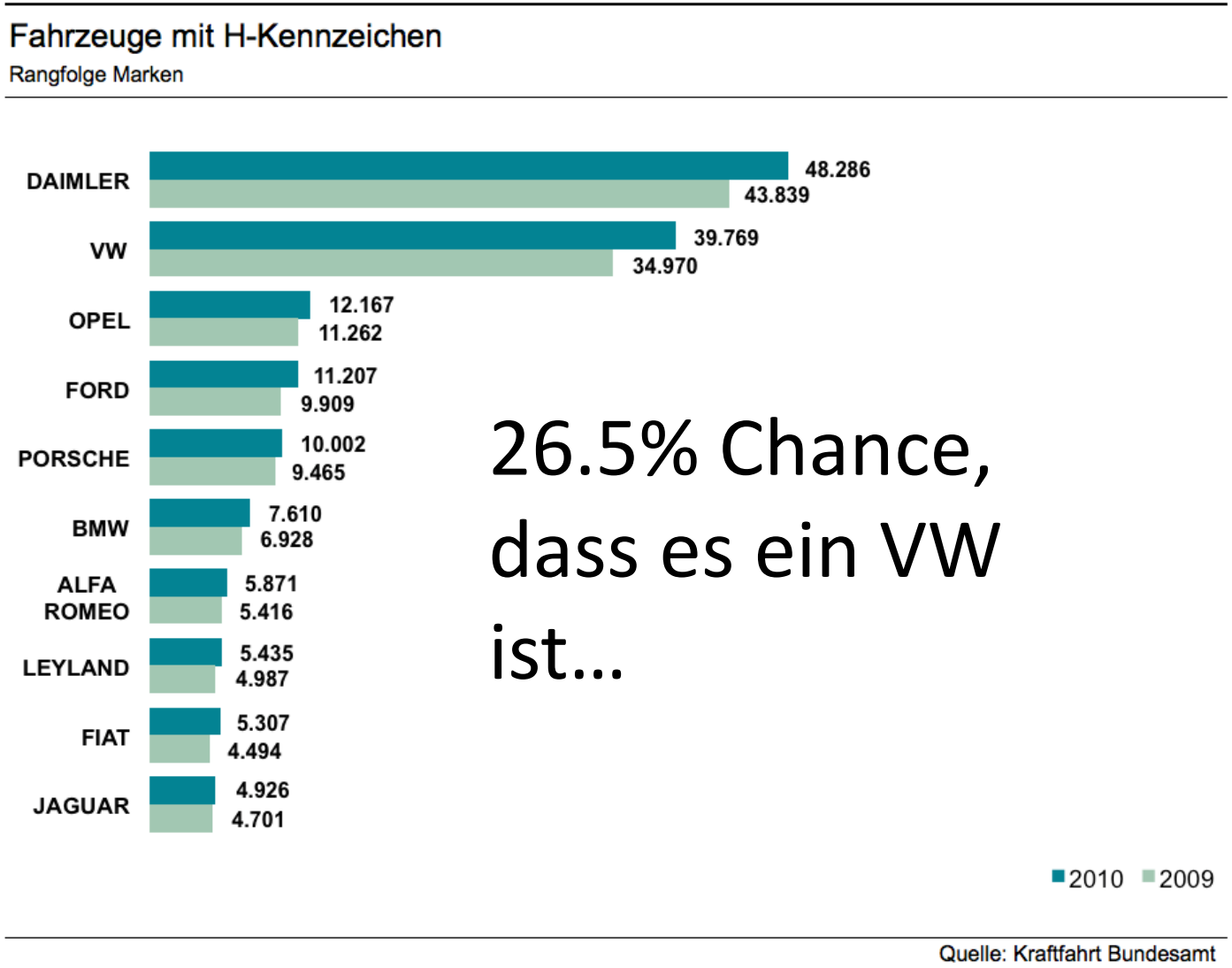


“Lernen” mit Korrelationen

VW Museum

Sollten Sie nachfragen?

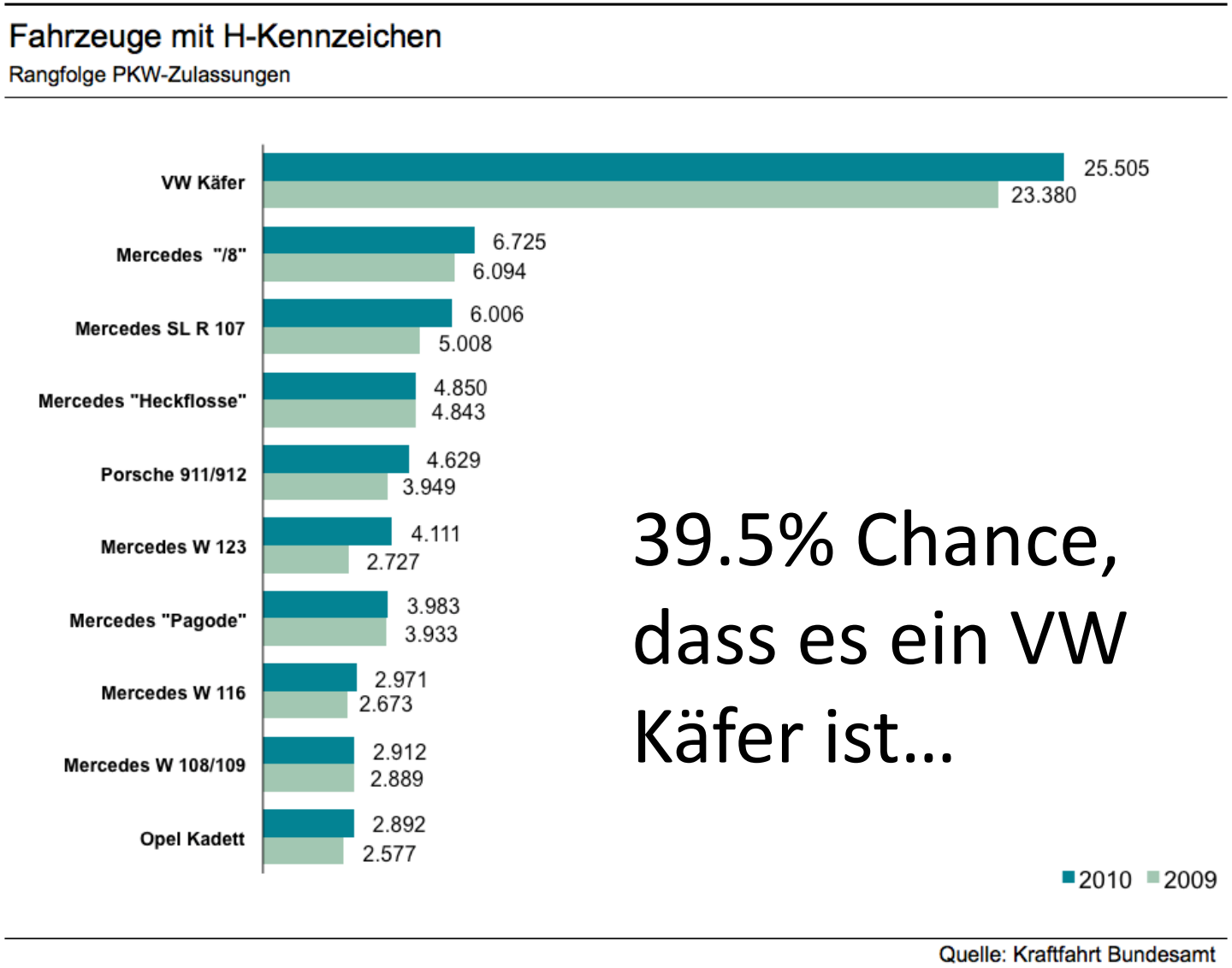
Ein vielleicht interessantes Angebot liegt vor…



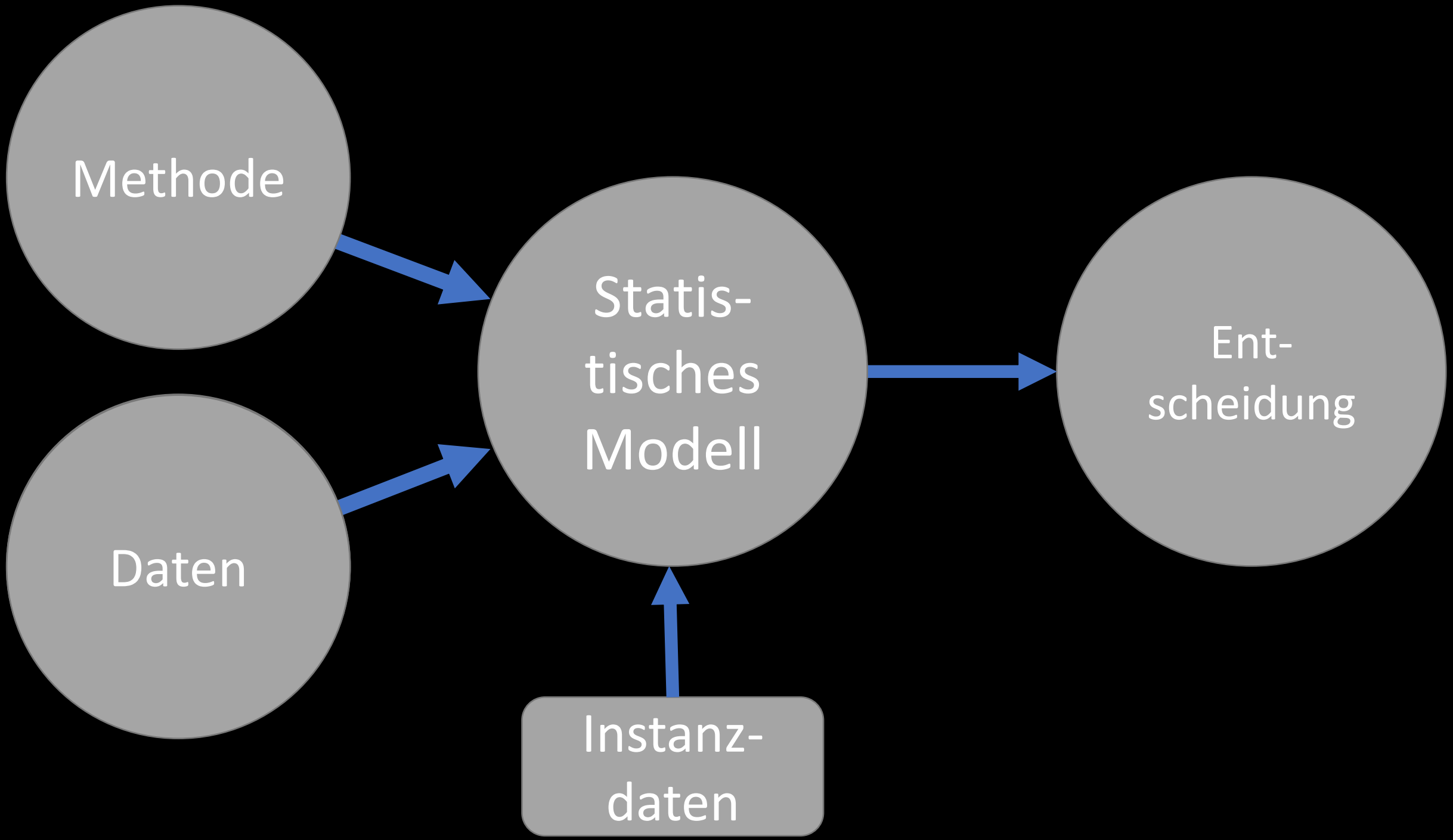
26.5% Chance, dass es ein VW ist…

Sollten Sie nachfragen?

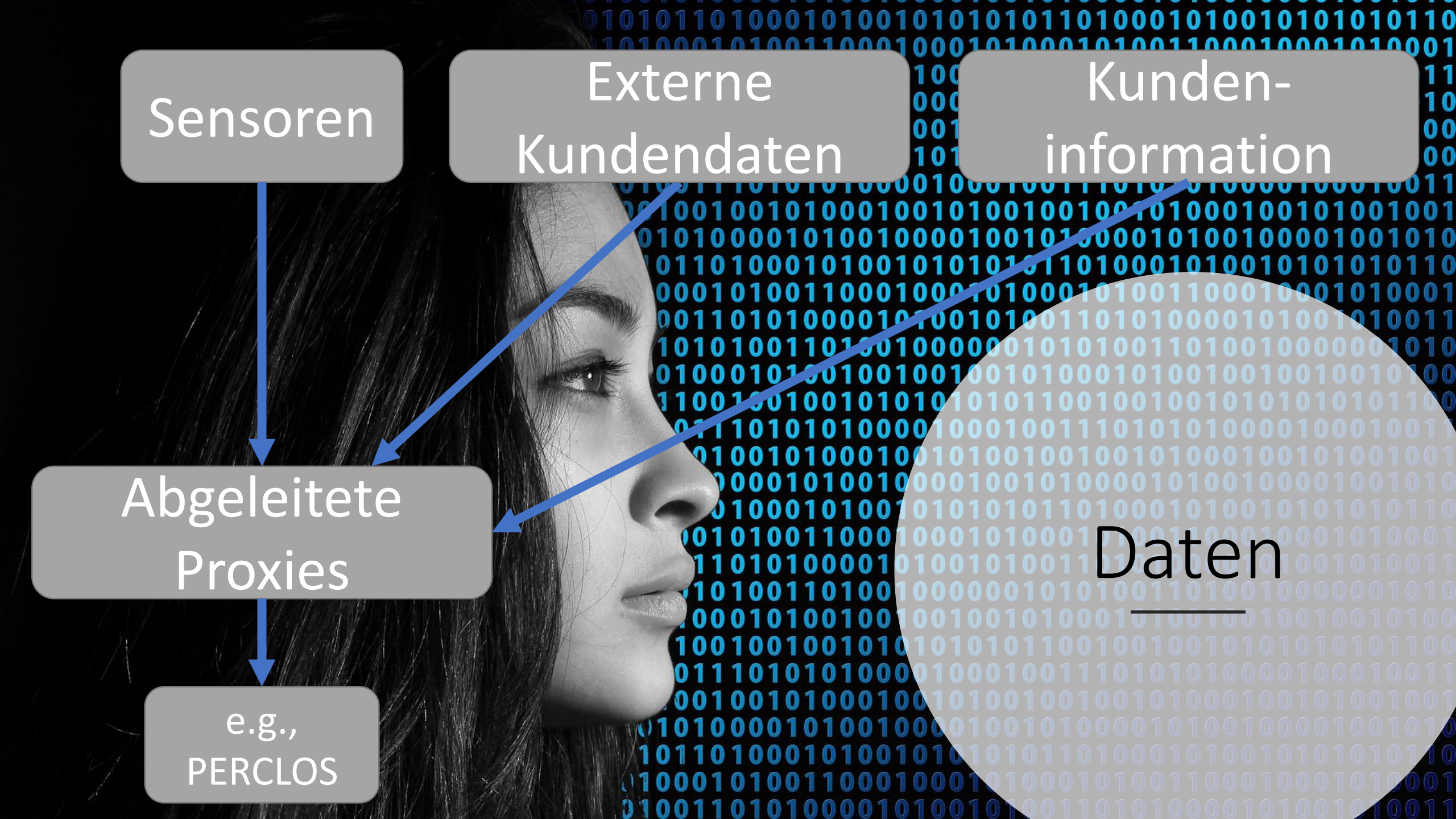
Ein vielleicht interessantes Angebot liegt vor…



39.5% Chance, dass es ein VW Käfer ist…


Methode
Daten
Statis-
tisches
Modell
Ent-
scheidung
Instanz-
daten

Daten ⇨ Datenauswahl


Sensoren
Externe Kundendaten
Kunden-information
Abgeleitete Proxies
e.g., PERCLOS
Daten

Google Patents

Drowsiness detection system

Abstract

This invention describes a non-intrusive system used to dete
and at risk of falling asleep at the wheel due to drowsiness.
drowsiness detection systems and a control unit. This redur
drowsiness assessment. The first subsystem consists of an
headliner and seat, which detects head movements that are
driver. The second subsystem consists of heart rate monitor
wheel. The control unit is used to analyze the sensory data a
state and therefore corresponding risk of falling asleep while
intelligent software algorithms, and the data provided by the
characteristics that may indicate a drowsy driver. If the drive
outputted which may be used to activate a response system
automobiles; this system may be used in any type of vehicle

Kopfbewegungen

Herzschlag

Augenlider (PERCLOS)

B2

OF Find Prior Art Similar

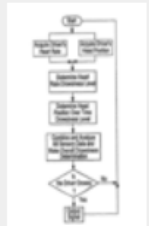
dam Basir, Jean Pierre Bhavnani, Fakhreddine Desrochers

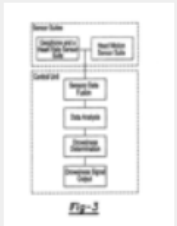
: Intelligent Mechatronic Systems Inc

: Intelligent Mechatronic Systems Inc

2-01-18

Images (3)





Classifications

G08B21/06 Alarms for ensuring the safety of persons indicating a condition of sleep, e.g. anti-dozing alarms

View 1 more classifications

Family: US (1)

Date	App/Pub Number	Status
2003-01-21	US10348037	Active
2003-08-14	US20030151516A1	Application
2004-11-23	US6822573B2	Grant

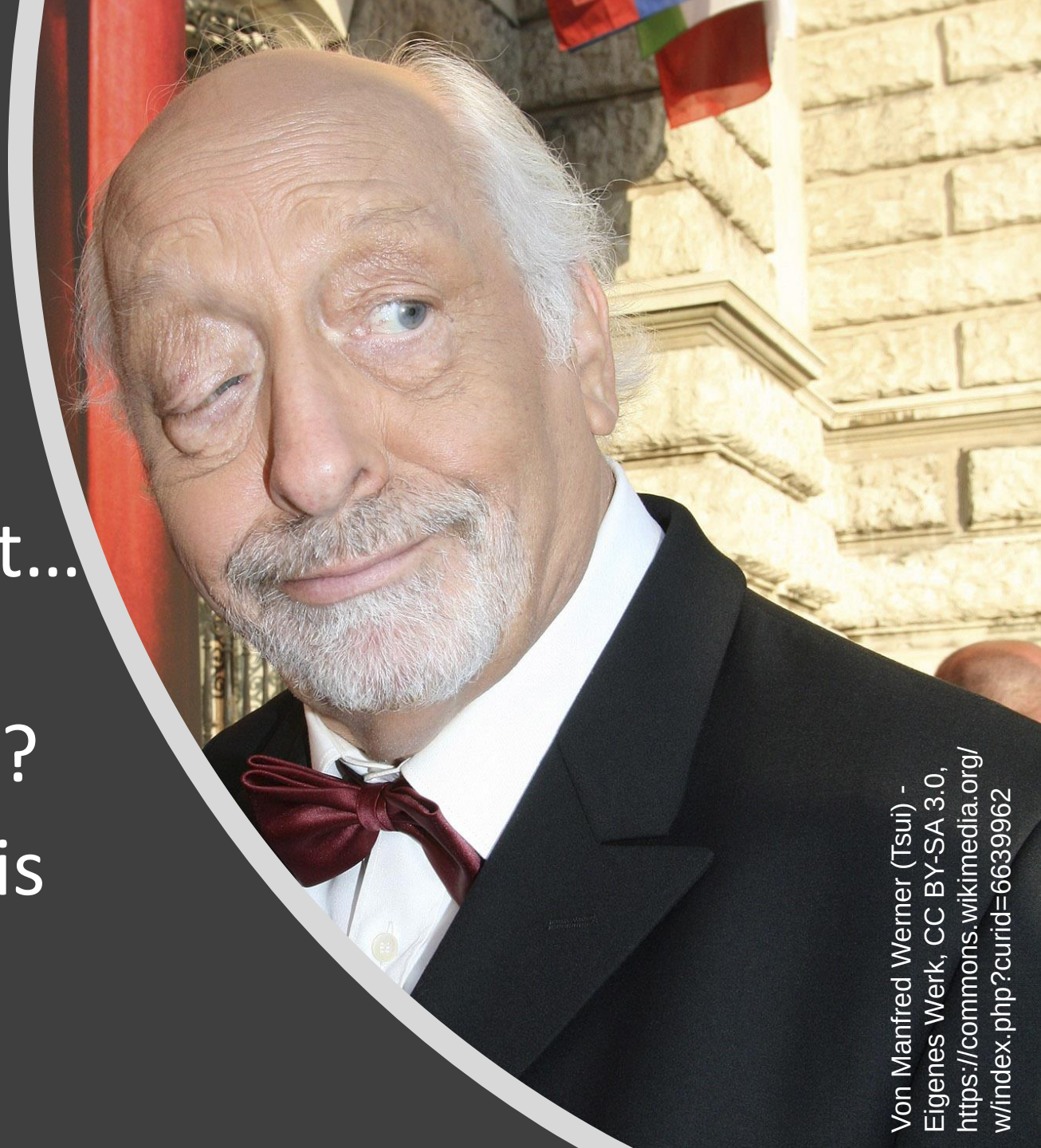
Info: Patent citations (14), Cited by (47), Legal events, Similar documents, Priority and Related Applications

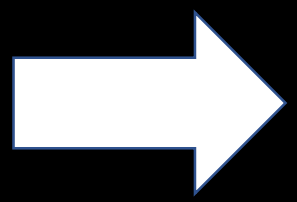
External links: USPTO, USPTO Assignment, Espacenet, Global Dossier, Discuss

Ethische Überlegungen bei der Datenauswahl

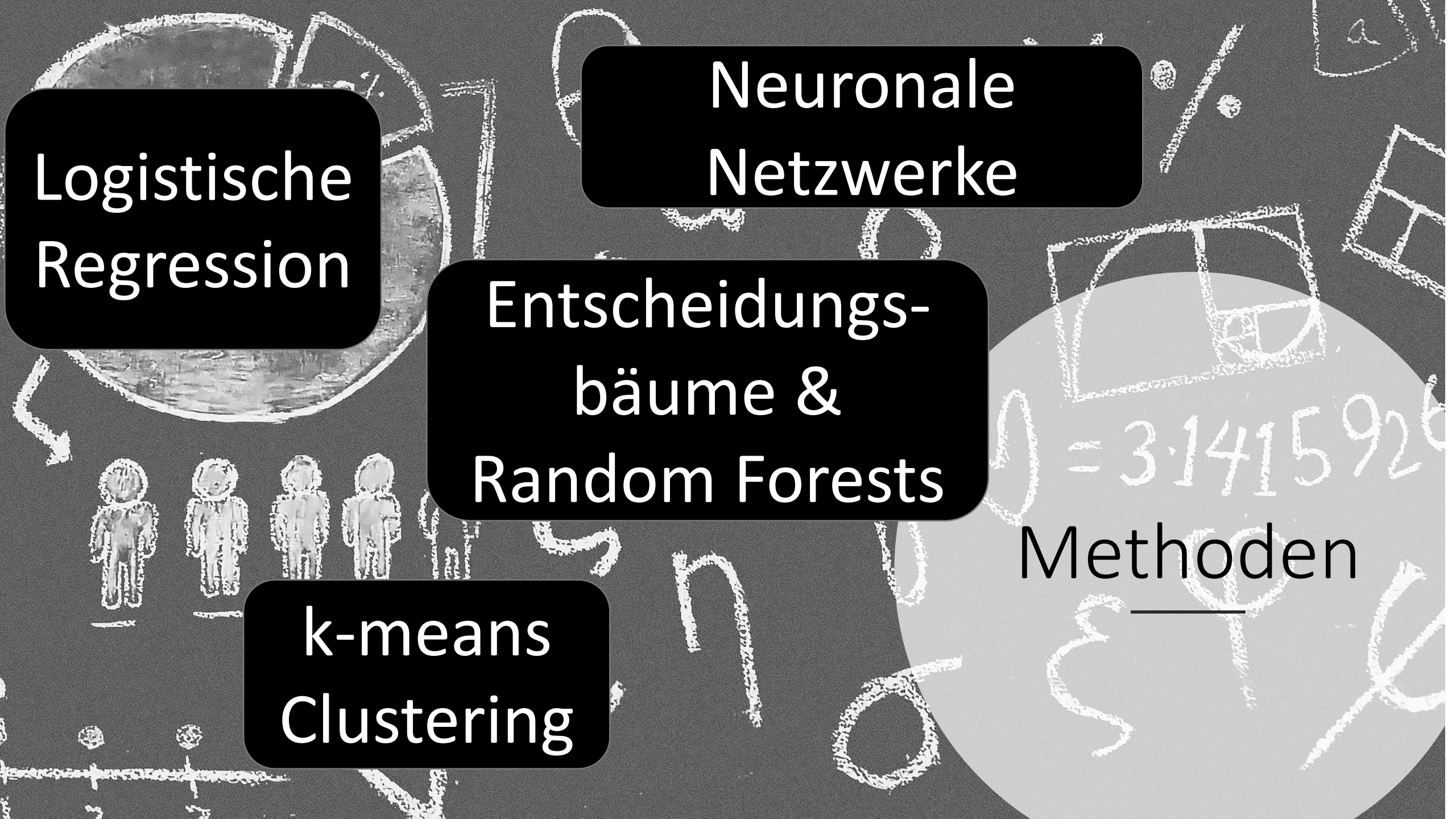
Durch die Videosensorik diskriminieren wir vielleicht…

- …Personen mit trockenen Augen oder Kontaktlinsen?
- …Personen mit einer Ptosis wie Karl Dall?





Methodenauswahl


Logistische Regression
Neuronale Netzwerke
Entscheidungs-bäume & Random Forests
k-means Clustering
Methoden

Regressionsansätze

- Die Algorithmen-designer entscheiden, welche Methode verwendet wird.
- Die Software sollte eine einzige Zahl ausgeben.
- Je höher die Zahl, desto höher die Erfolgs-wahrscheinlichkeit.
- Beispiel logistische Regression:

$$
\begin{aligned}
& 3 * (\text{PerClos}) \\
+ & 3 * (\text{no of jerky head movements}) \\
- & 2 * (\text{heart rate}) \\
+ & \ldots
\end{aligned}
$$

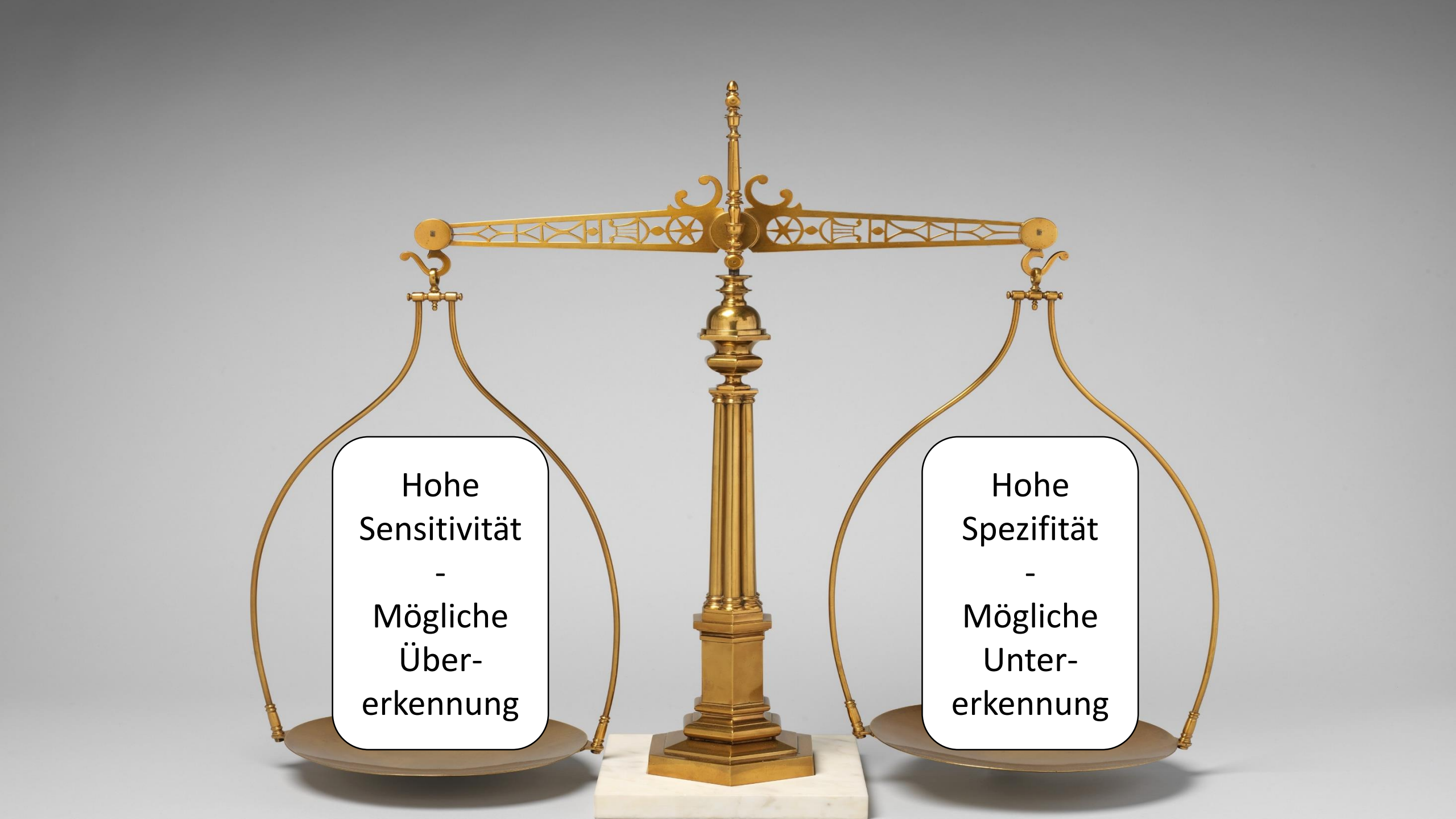
Allgemein

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

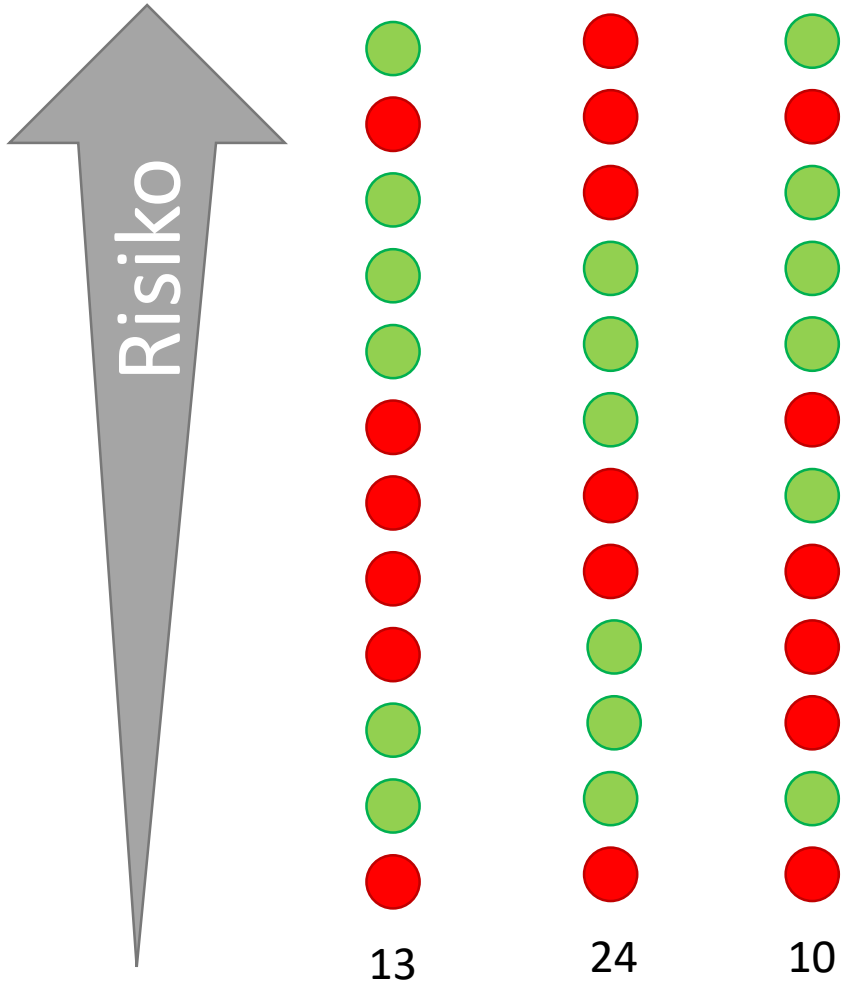
$$
\begin{aligned}
& w_1 * (\text{PerClos}) \\
+\ & w_2 * (\text{no of jerky head movements}) \\
-\ & w_3 * (\text{heart rate}) \\
+\ & \ldots
\end{aligned}
$$

Qualität eines Algorithmus


ROC AUC
...und 20 mehr
Positive Predictive Value
Sensitivität
Qualitäts-Maße
Accuracy


Hohe Sensitivität - Mögliche Über-erkennung
Hohe Spezifität - Mögliche Unter-erkennung

- Rote Kugeln symbolisieren müde, grüne wache FahrerInnen.
- Optimale Sortierung: Alle roten oben, alle grünen darunter.
- Qualitätsmaß: Paare von rot und grün, bei denen die rote Kugel über der grünen einsortiert ist. (**ROC AUC**)



@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

Ist das Qualitätsmaß sinnvoll?

- Wenn zwei Fahrer zur Verfügung stellen, von denen einer jetzt sofort fahren soll: **ja**
- Wenn es um die Erkennung der müdesten Fahrer geht: **nein**
- Hier müssen andere Qualitätsmaße benutzt werden.

@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

einen Jagdhund zu trainieren,

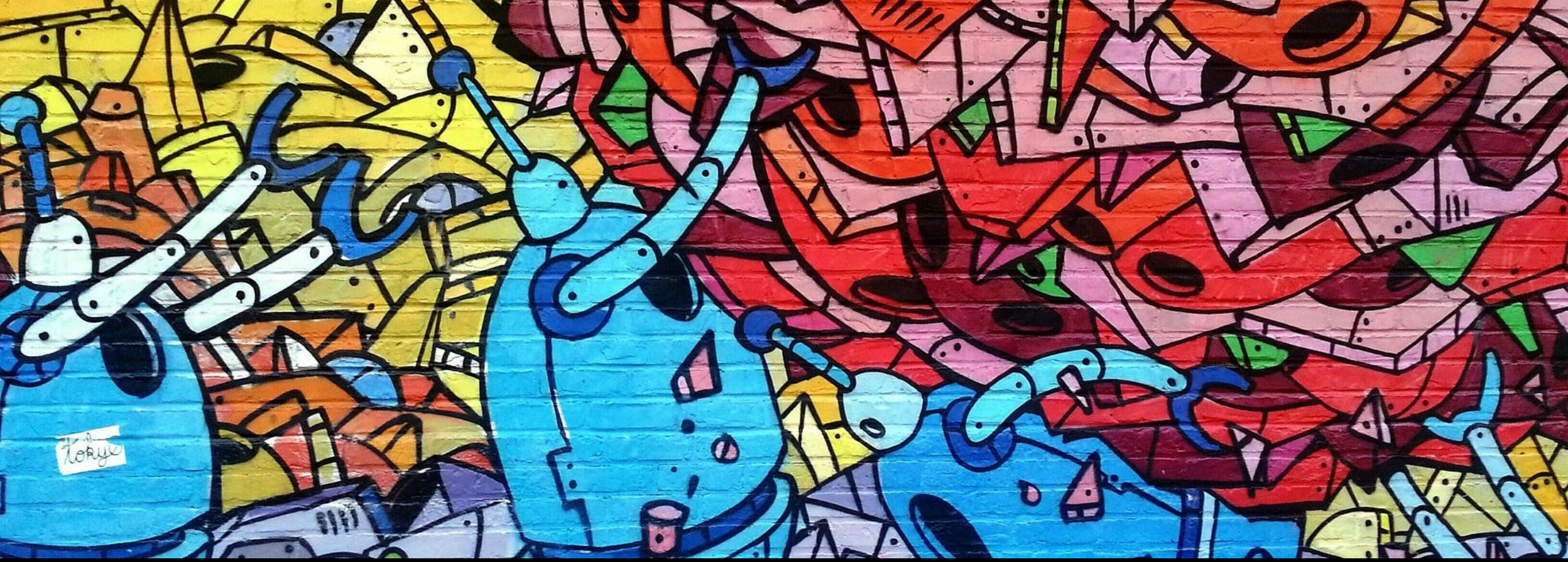
um mit ihm Schafe zu hüten.

Das ist wie...

Probleme von algorithmischen Entscheidungssystemen (ADM Systemen) im People Assessment

1. **Wer entscheidet, wann ein ADM System „gut“ 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…

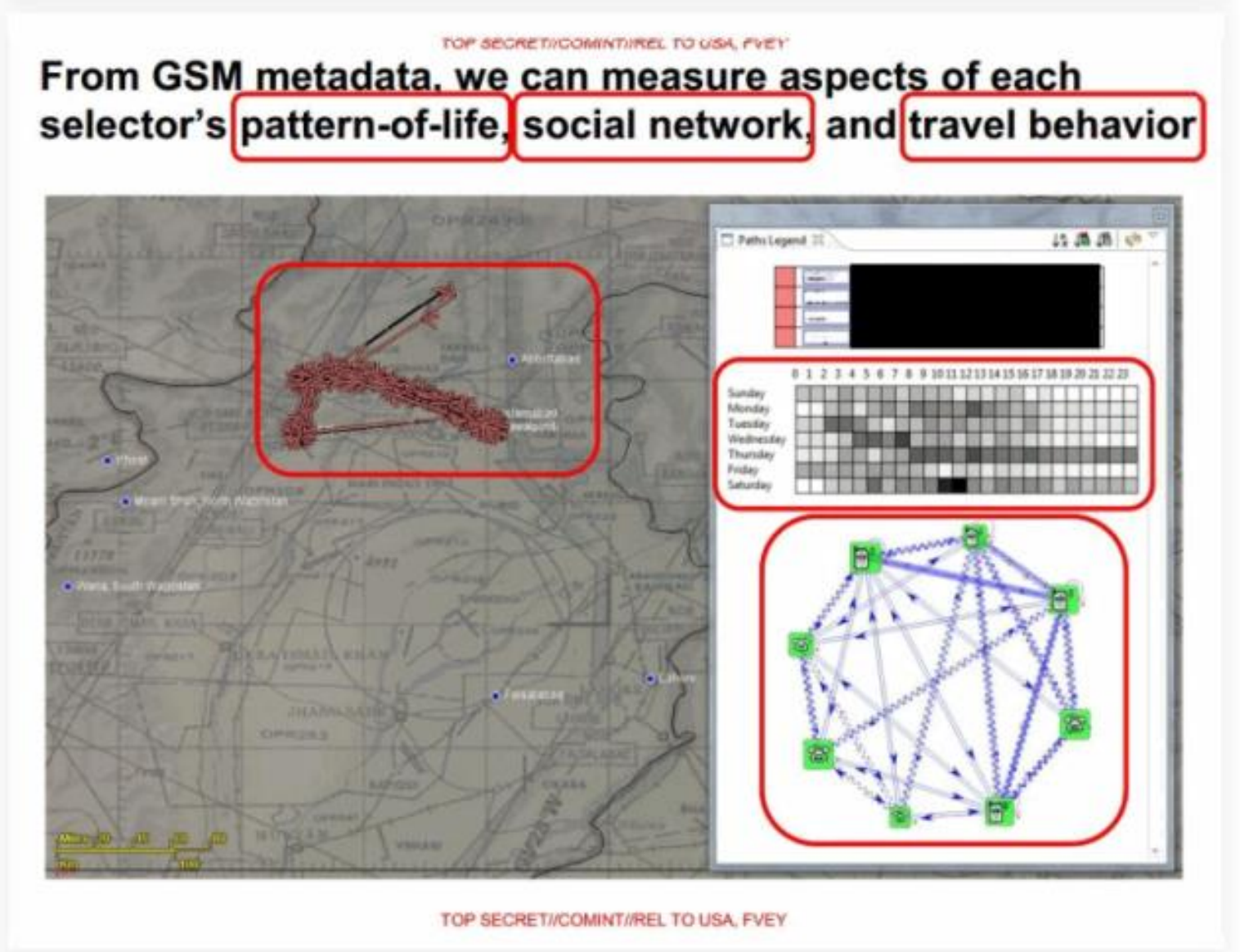
- … basieren auf Korrelationen von Eigenschaften mit gewünschtem Verhalten.
- Quasi **algorithmisch legitimierte Vorurteile**:
  - Zu 70% müde heißt:
  - Von 100 Personen, die „sich genau so verhalten dieser Mensch“, sind 70 zu müde, um zu fahren.

Generell

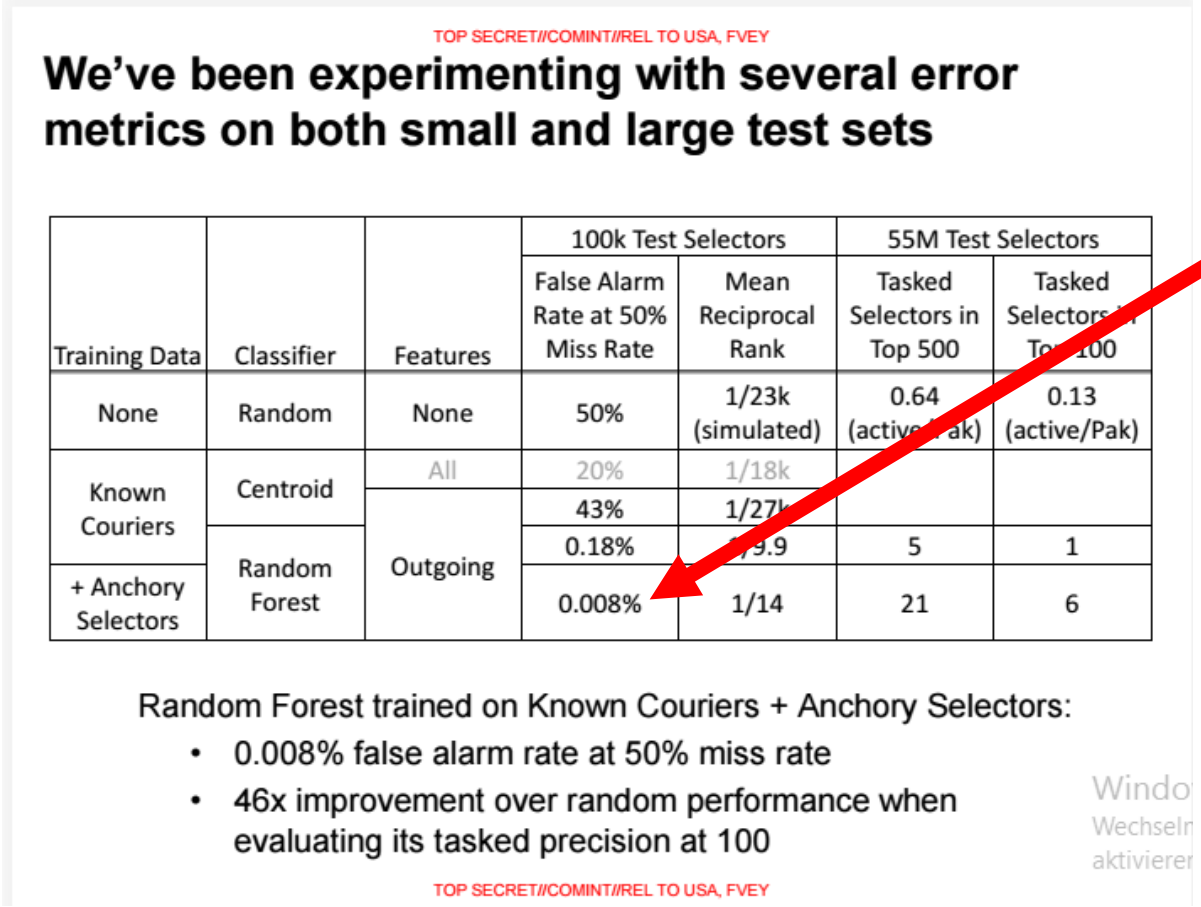
Prinzipiell können ADM Systeme vieles entscheiden:

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

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:

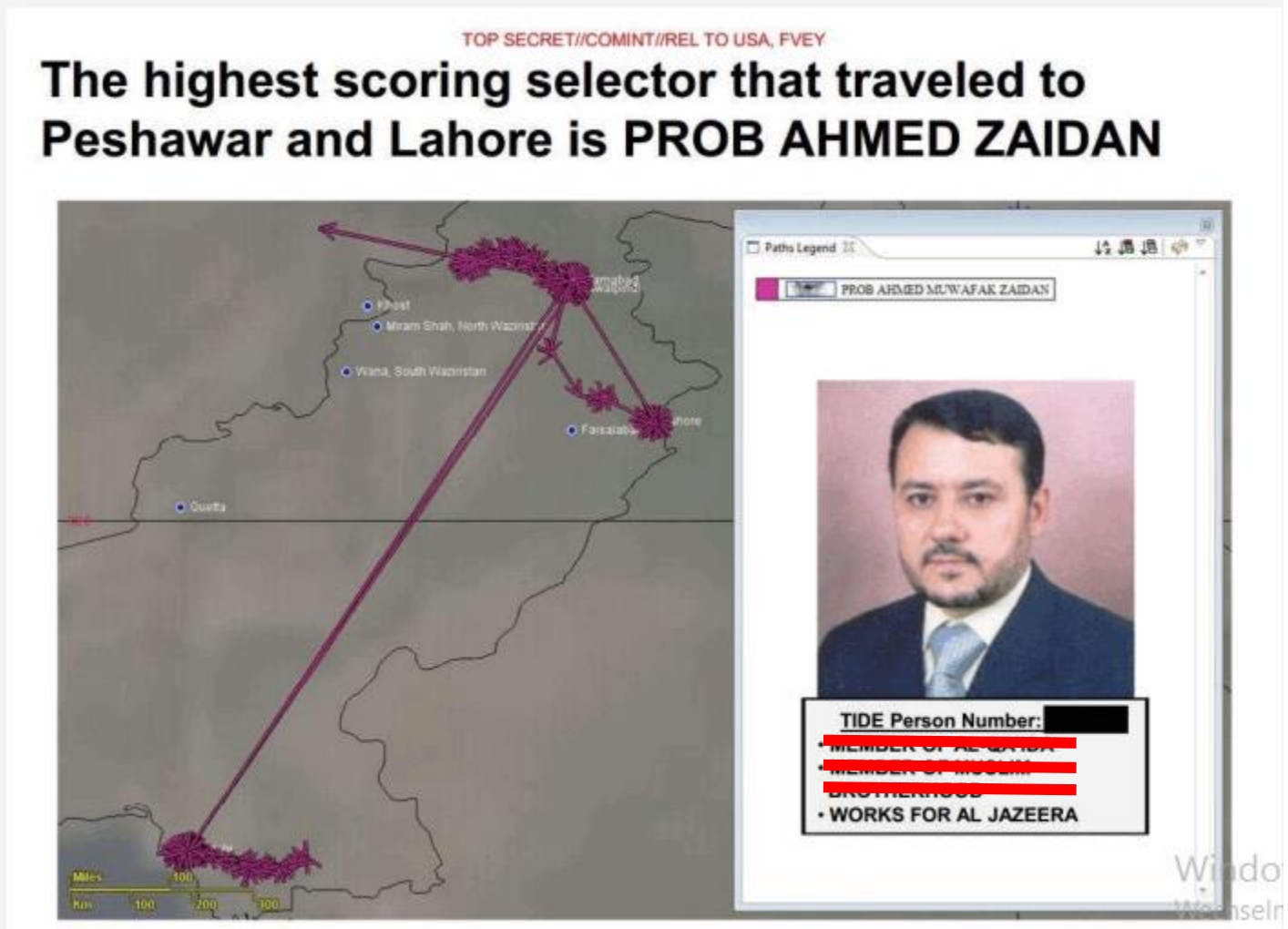
- 0.008% false alarm rate at 50% miss rate
- 46x improvement over random performance when evaluating its tasked precision at 100

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…



Methoden-
auswahl

Daten-
auswahl

Qualitäts-
maße

Statistisches
Modell

Ent-
scheidung

Instanz-
daten

Methoden-
auswahl

Daten-
auswahl

Qualitäts-
maße

Statistisches
Modell

Inter-
pretation

Instanz-
daten

Probleme von algorithmischen Entscheidungssystemen (ADM Systemen) im People Assessment

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

@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

Methoden-
auswahl

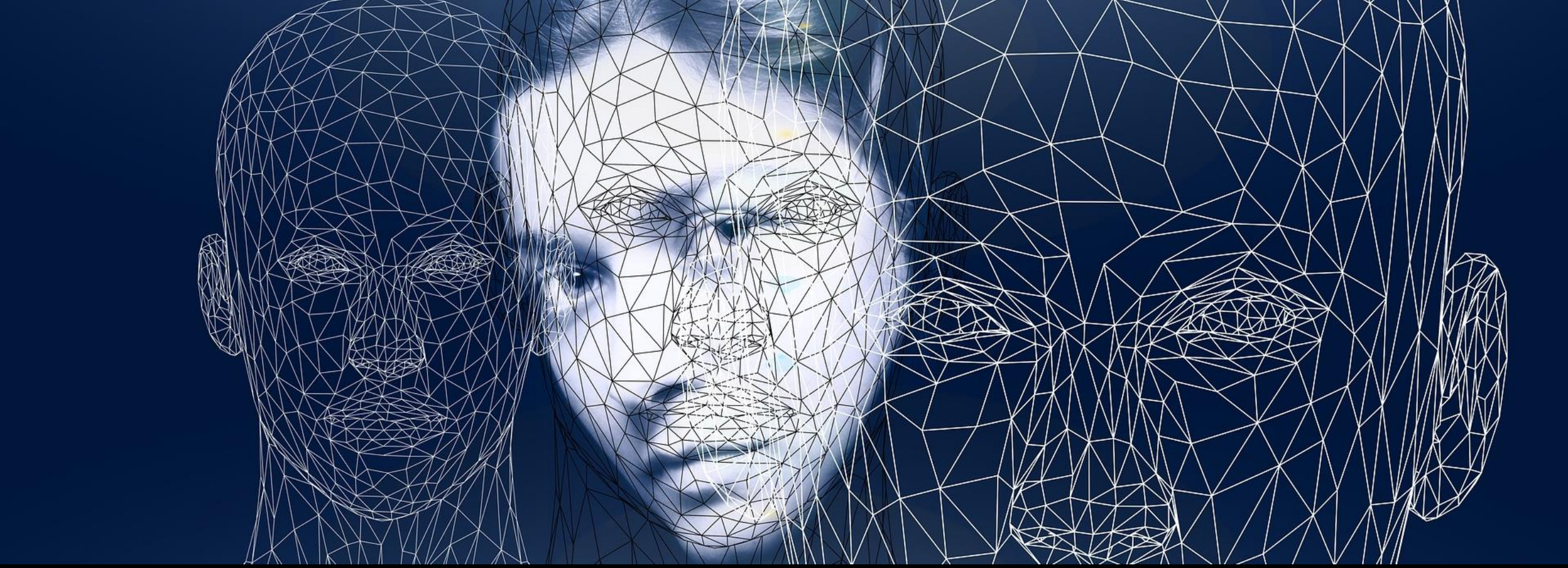
Daten-
auswahl

Qualitäts-
maße

Statistisches
Modell

Inter-
pretation

Instanz-
Daten



Sozio-informatische Gesamtbetrachtung

Soziale Einbettung des Systems

- Wird sich der Fahrer oder die Fahrerin auf das System verlassen?
- Wollen Speditionen Zugriff auf Fahrerdaten?
  - Oder die FahrerInnen?
- Könnte das System auch für Büros nützlich sein?
  - Welche Verantwortlichkeiten erwachsen daraus?

Probleme von algorithmischen Entscheidungssystemen (ADM Systemen) im People Assessment

1. Wer entscheidet, wann ein ADM System „gut“ ist?
2. ADM Systeme ergeben nur Wahrscheinlichkeiten, keine Wahrheiten.
3. **ADM Systeme können nicht einfach in einen neuen sozialen Kontext verpflanzt werden.**

@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

Methoden-
auswahl

Daten-
auswahl

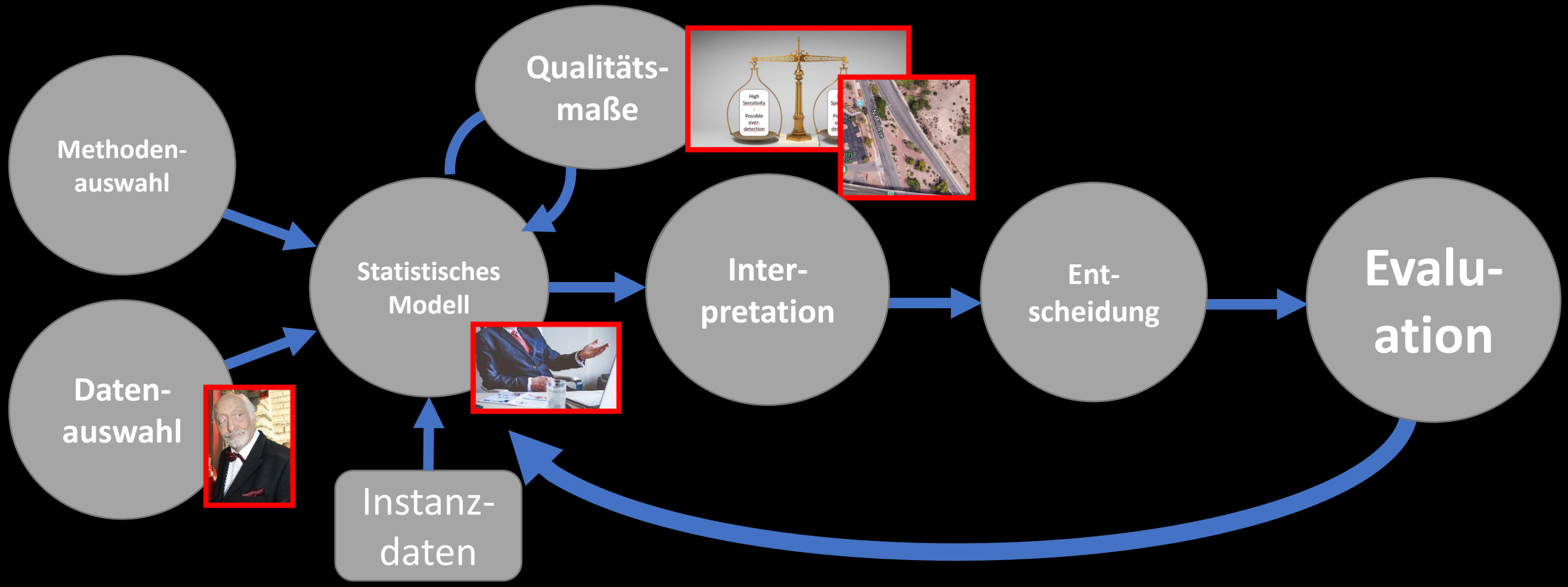
Qualitäts-
maße

Statistisch
es Modell

Instanz-
daten

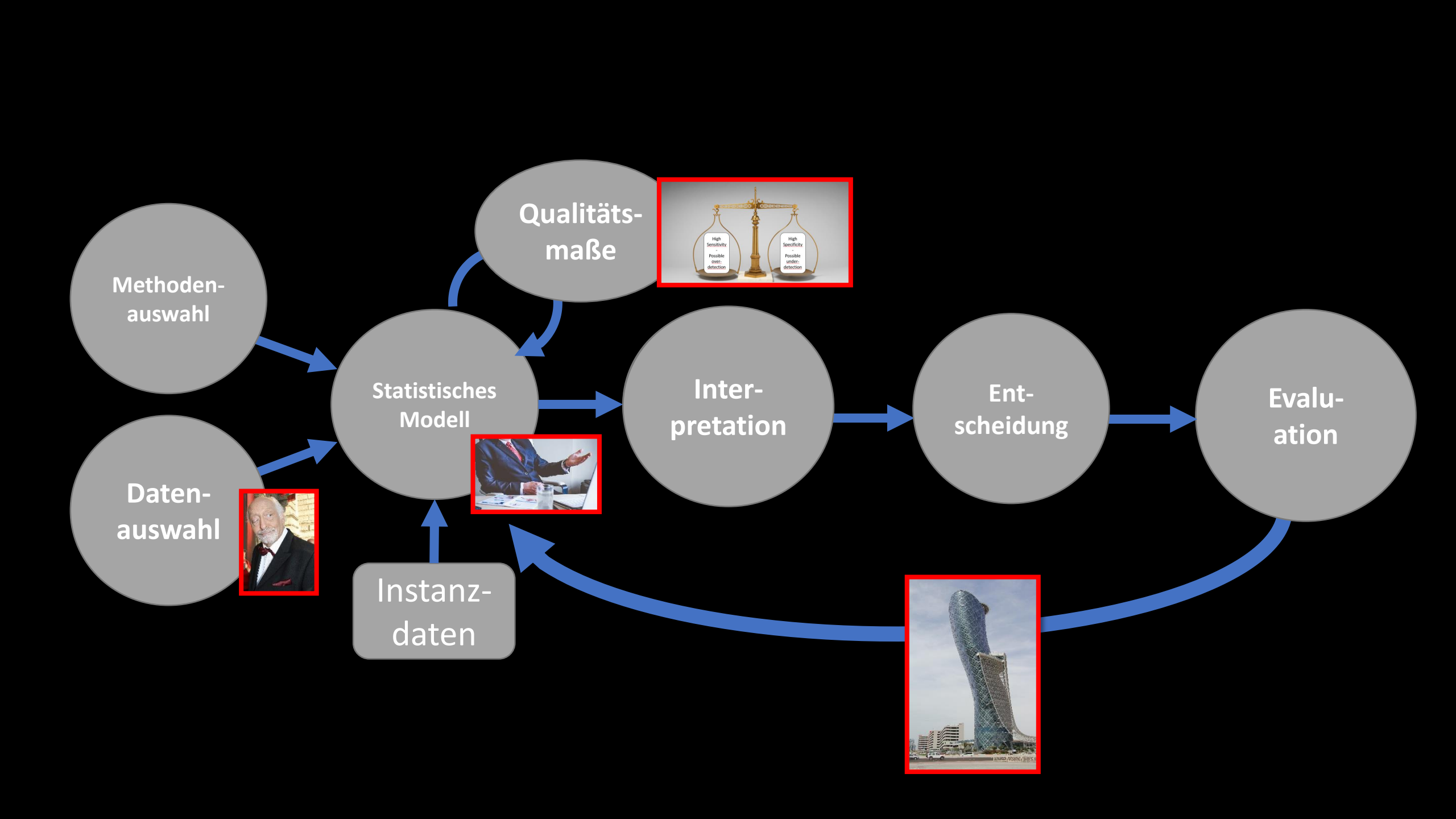
Inter-
pretation

Ent-
scheidung

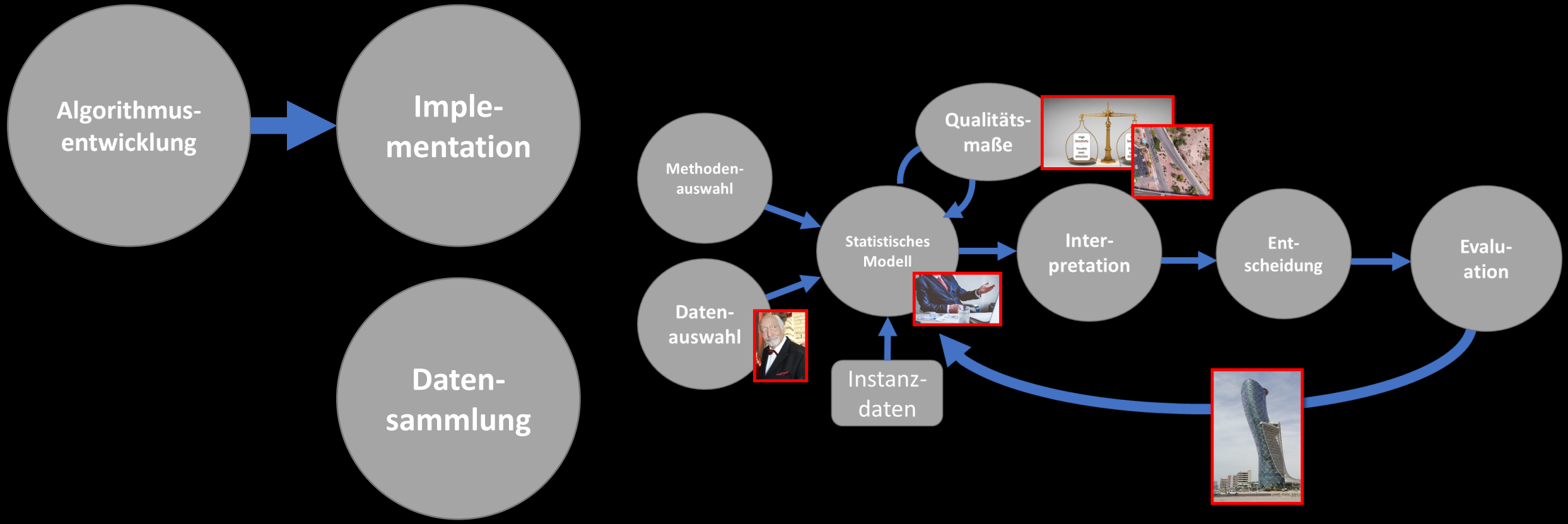

Methoden-
auswahl
Daten-
auswahl
Qualitäts-
maße
Statistisches
Modell
Inter-
pretation
Ent-
scheidung
Evalu-
ation
Instanz-
daten

Feedback-Probleme

Personen, die wegen der Entscheidung des Sensors nicht fahren dürfen, können nicht nachweisen, dass sie sicher gefahren wären.

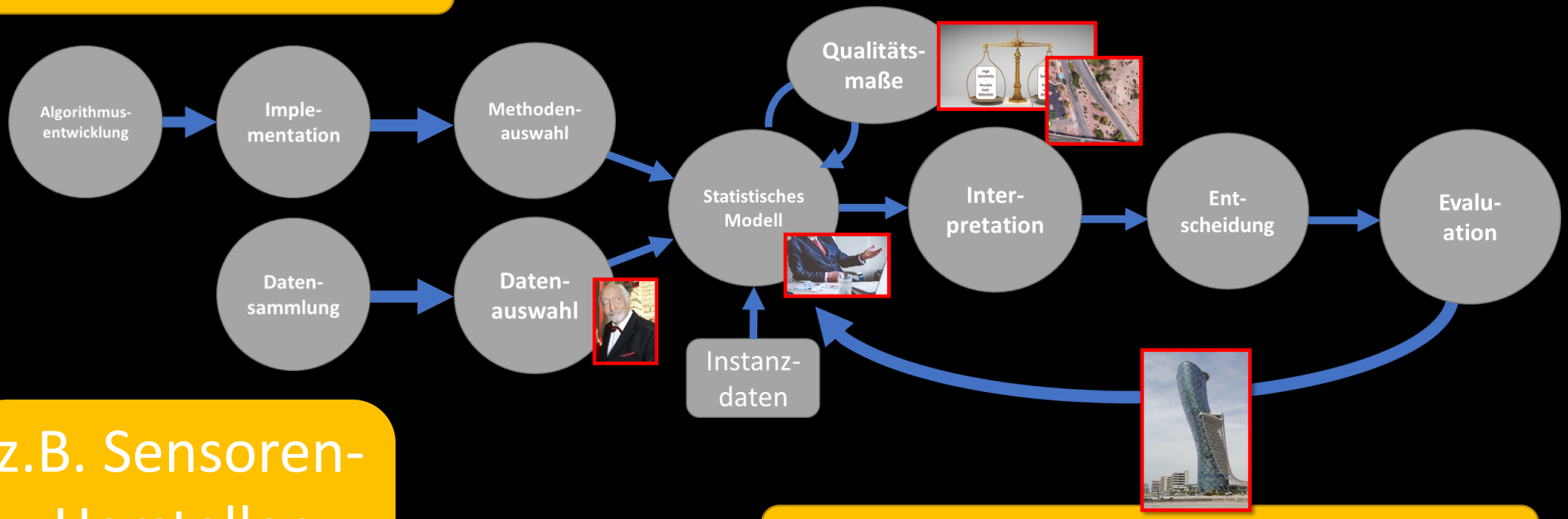

Methoden-
auswahl
Daten-
auswahl
Qualitäts-
maße
Statistisches
Modell
Instanz-
daten
Inter-
pretation
Ent-
scheidung
Evalu-
ation

Wer ist verantwortlich?



Wer ist verantwortlich?

Informatiker



z.B. Sensoren-Hersteller

Kunden Ihres ADM Systems

Data Scientists / Ingenieure

Kann ein fehlendes ADM System unethisch sein?

Und darüber hinaus?

Der Uber-Unfall in Tempe

Zwischen-bericht

“According to Uber, emergency braking maneuvers are not enabled while the vehicle is under computer control, to reduce the potential for erratic vehicle behavior.”

“The vehicle operator is relied on to intervene and take action.”

“The system is not designed to alert the operators.”

https://www.ntsb.gov/investigations/AccidentReports/Reports/HWY18MH010-prelim.pdf

https://www.google.de/maps/place/N+Mill+Ave,+Tempe,+AZ+85281,+USA/@33.4364084,-111.9436953,335m/data=!3m1!1e3!4m5!3m4!1s0x872b09338c84a6a3:0x4eb48ca97885c3f7!8m2!3d33.4379952!4d-111.9435544

Welche ethischen Entscheidungen traf Uber?

- Wurde die Wahrscheinlichkeit für Fußgängerübergänge von Straßenkarten und Satellitenbildern abgeleitet?
- Balance zwischen “Sicherheit (anderer)” und “Komfort der Passagiere”.
- Es fehlte ein Alarmsystem!
- Warum kein Alarmsystem, wenn der “operator” unaufmerksam ist?

Probleme von algorithmischen Entscheidungssystemen (ADM Systemen) im People Assessment

1. Wer entscheidet, wann ein ADM System „gut“ ist?
2. ADM Systeme ergeben nur Wahrscheinlichkeiten, keine Wahrheiten.
3. ADM Systeme können nicht einfach in einen neuen sozialen Kontext verpflanzt werden.
4. **Das Fehlen eines ADM / IT-Systems kann auch unethisch sein.**

@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

Gibt es eine Datenethik?

Ethik der Data Science

Drowsiness Measuring Decision Making System

Fig. 1 from AG Reece and Danforth: "Instagram Photos reveal Predictive Markers of Depression", EPJ Data Science 2017, 6:15

Depressionserkennung auf Instagram

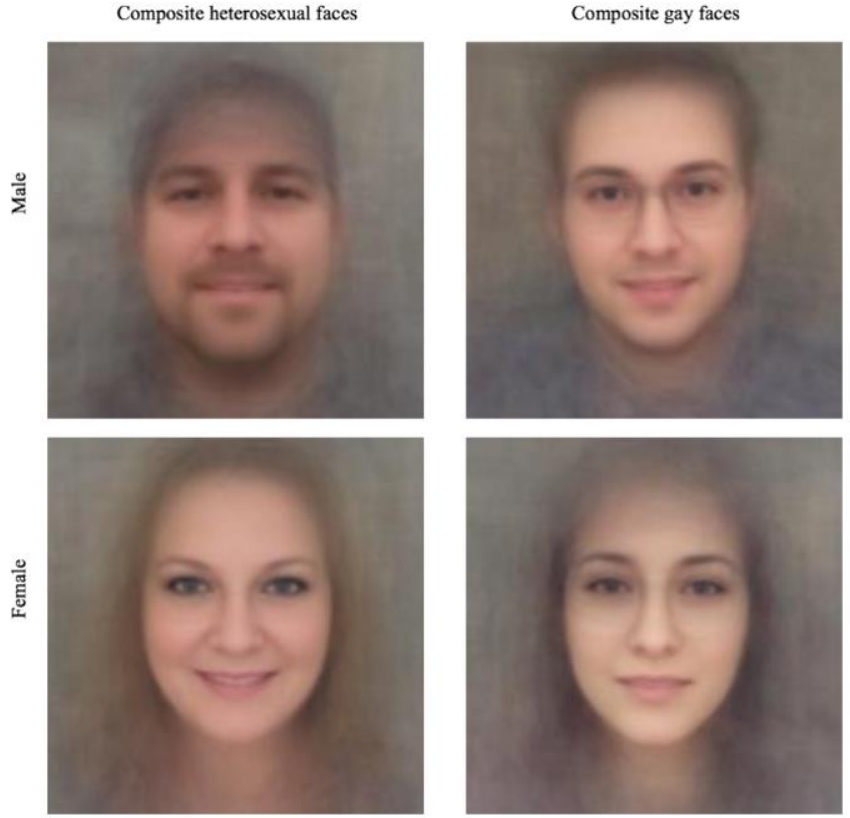
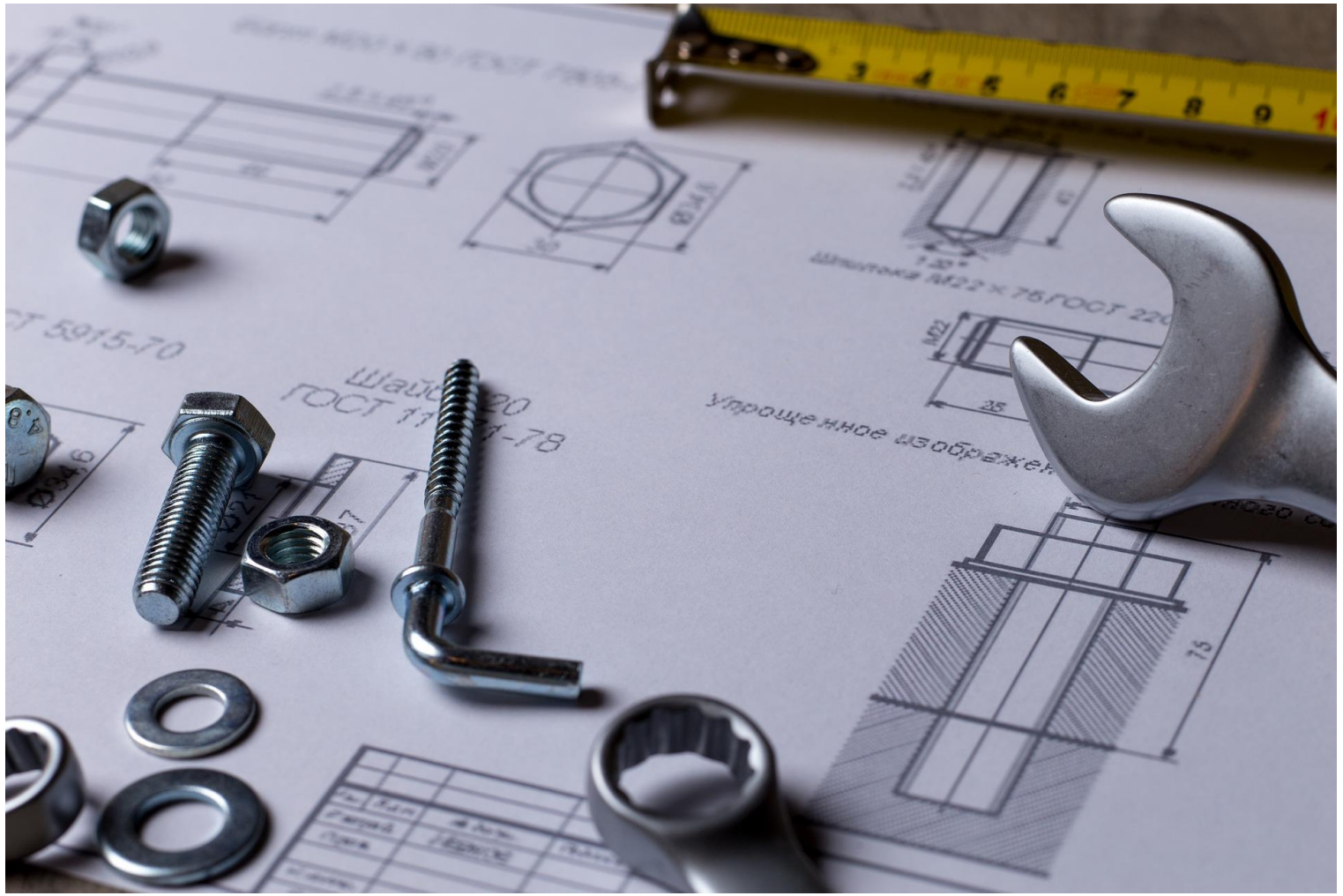


Fig. 4 from Y Wang and M Kosinski: "Deep Neural Networks can detect sexual orientation from faces", to appear in Journal of Personality and Social Psychology", preprint at: "https://osf.io/fk3xr/", used with permission from the last author

“Gaydar”

Wie können wir bessere Systeme bauen?



Best Practice für die Entwicklung und Nutzung von ADM Systemen, die Menschen bewerten

- „Normale“ Algorithmen sind gut überprüfbar.
- Fokus auf **lernende Systeme,** deren Entscheidungen Menschen betreffen.
- Wichtig sind Entwicklungsteams **mit hoher Diversität**, sowohl was die Personen als auch Ausbildungen angeht.
- Weiterbildung im Bereich „Data Science Literacy“ und „Ethik für Data Scientists“.
- **Klare Kommunikationsprozesse** entlang der „langen Kette der Verantwortlichkeiten“:
  - **Z.B. Fehlerraten der Daten von Sensoren**.
  - Über alle Designentscheidungen, die getroffen wurden, und in welchem Kontext sie gültig sind.
- **Klare Abgrenzungen der jeweiligen Verantwortlichkeiten.**

„There are no simple solutions for complex problems. Whoever promises them, is fooling himself and others.“

Volkswagen at the IAA 2017