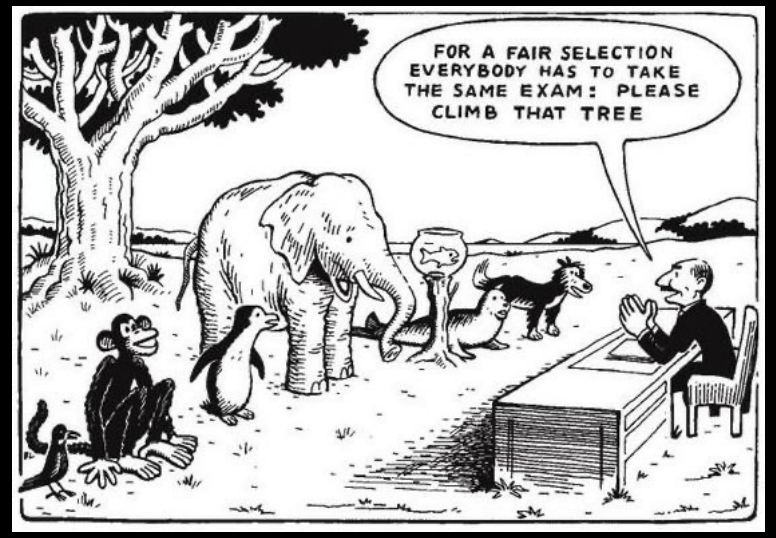
TÜV AI.LAB

Fairness bei algorithmischen Entscheidungsprozessen

Dr. Christoph Poetsch
Senior Advisor AI Ethics and Quality

Online-Vortrag | ZERTIFIZIERTE KI
Technische Prüfung von Fairnessanforderungen
19.06.24

Systematische Vorüberlegungen



Begriffliches

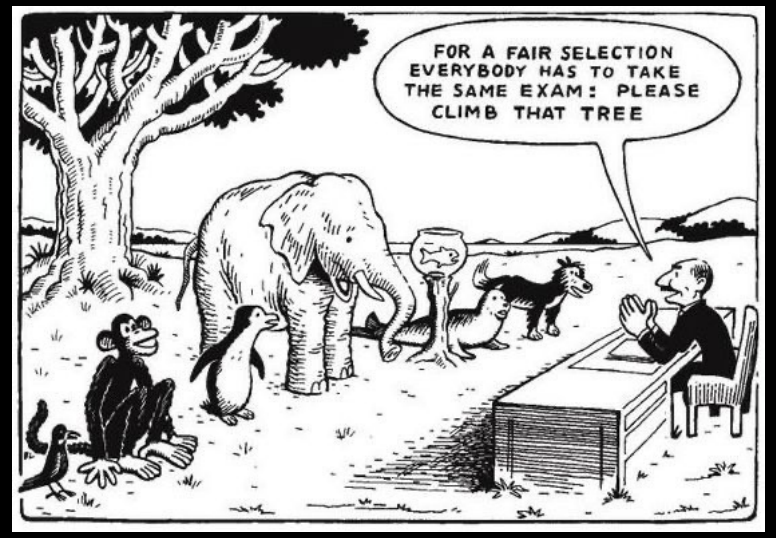
- Fairness vs. Was ist fair?
  Grundbegriff vs. einzelne, konkrete Fairnesskonzepte
- Fairness = *gesetzliche* Nicht-Diskriminierung? (vgl. EU AI Act)
- Oder Fairness als explizit *nicht gesetzlich festgeschriebene* Gerechtigkeitsvorstellung? (und dann ggfs. dem Gesetz widersprechend)

Zwei Dimensionen

- Adäquate Methode: Gleichbehandlung oder Ungleichbehandlung? (jdG, jdS; Equality, Equity)
- Kriterium : Ausgangsbedingungen/Durchführung oder Resultat? (zzgl. Korrelation/Kausalität)

TÜV AI.LAB

Systematische Vorüberlegungen



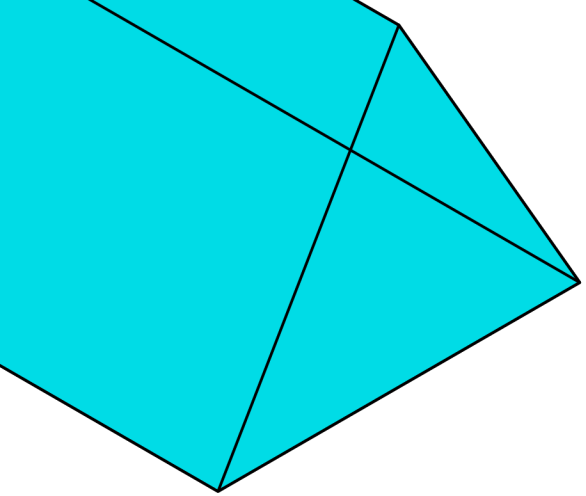
Systematische Differenzierungen

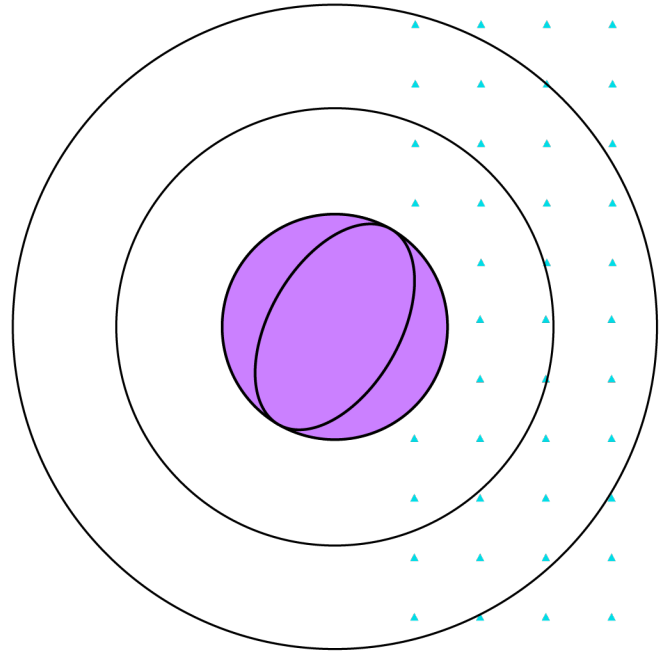
- Fokus auf Individuen oder auf Gruppen?
- Einzelfall- oder Vergleichsbetrachtung?
- Bei Vergleich: Ergebnisse in Konkurrenz zueinander oder nicht?
- Kriterium bei konkurrierenden konkreten Fairness-Konzepten?

Für die Bewertung / Messung von KI-Systemen u.a.

- Ideale oder reale Sollwerte? Falls Ersteres: welche? Falls Letzteres: welcher Bezugsrahmen?
- Einbezug der human baseline für die gegebene Aufgabe

TÜV AI.LAB





Wie lernt ein System von Daten?

Support Vector Machine

Support Vector Machine

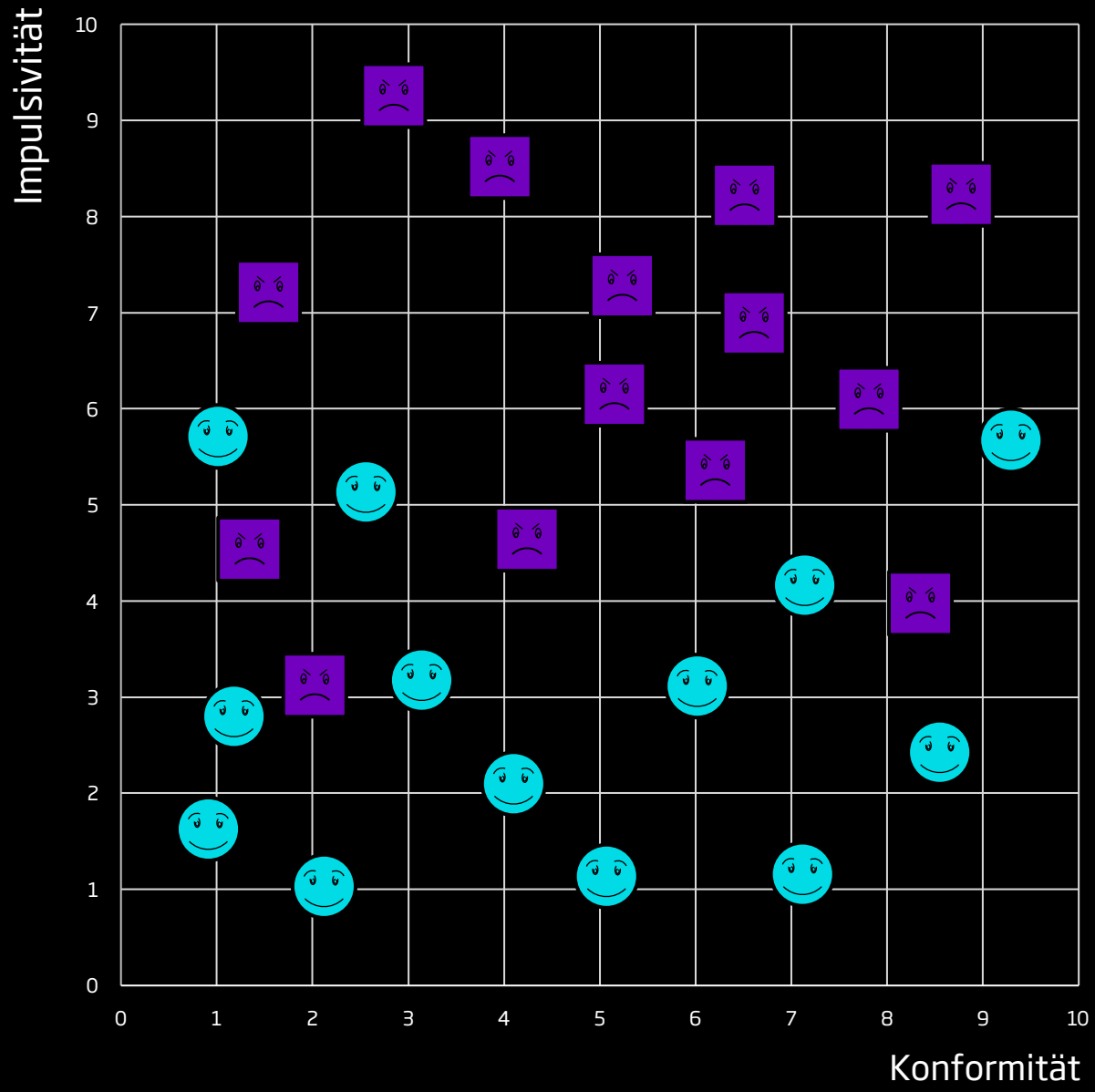
Straftäter

Unschuldige Bürger

Zeichnen Sie – in Gedanken – eine gerade Linie so zwischen die Smileys, dass die violetten möglichst gut von den türkisen getrennt sind.

Gratulation: Sie haben eine Support Vector Machine trainiert!

Die Linie dient nun als Entscheidungsregel, ob eine Person als Straftäter oder als unschuldig gilt.



TÜV AI.LAB

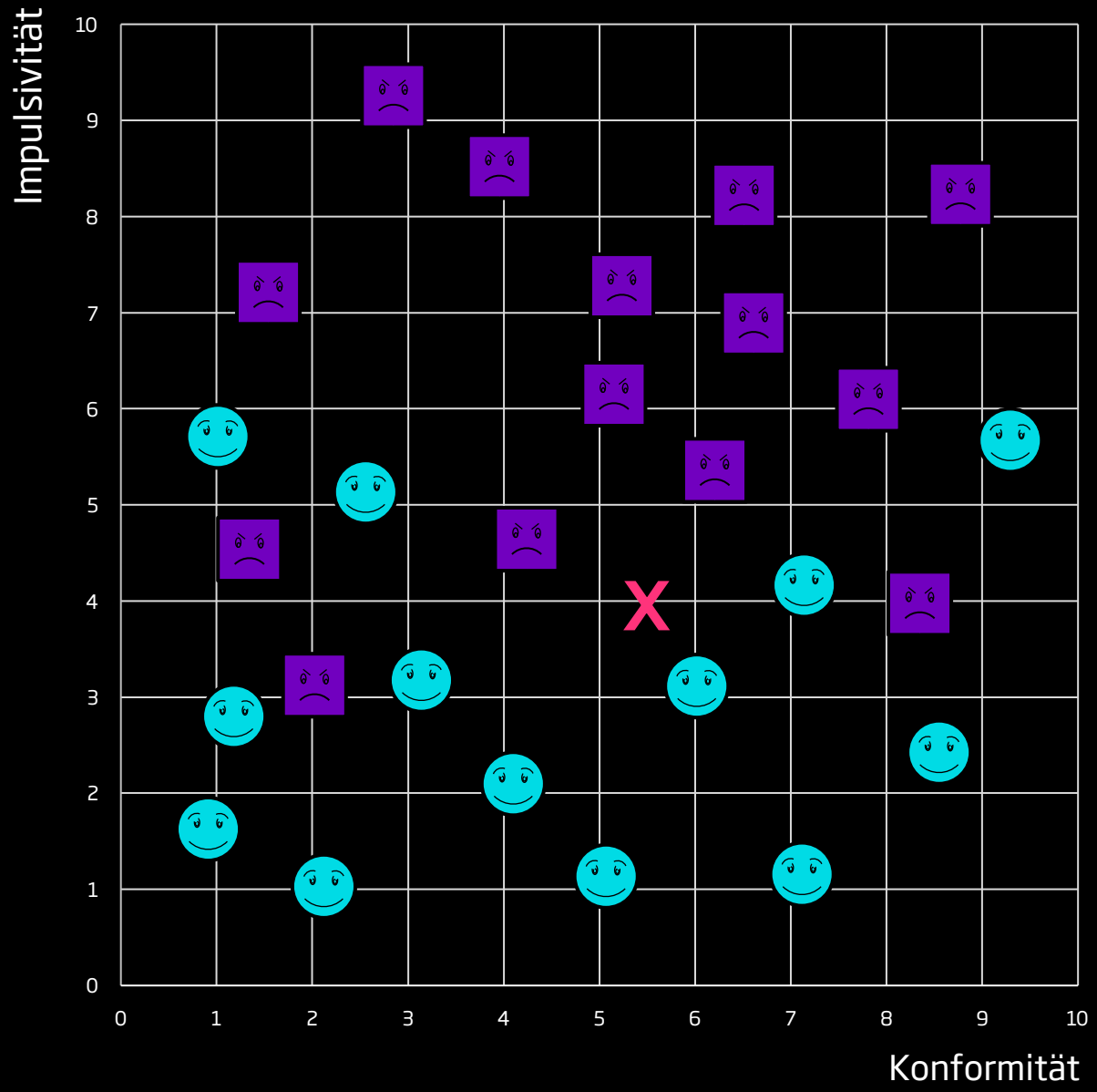
Support Vector Machine

Straftäter

Unschuldige Bürger

Bewerten Sie Manuela Mustermann:

5.5 Konformität
4.0 Impulsivität



TÜV AI.LAB

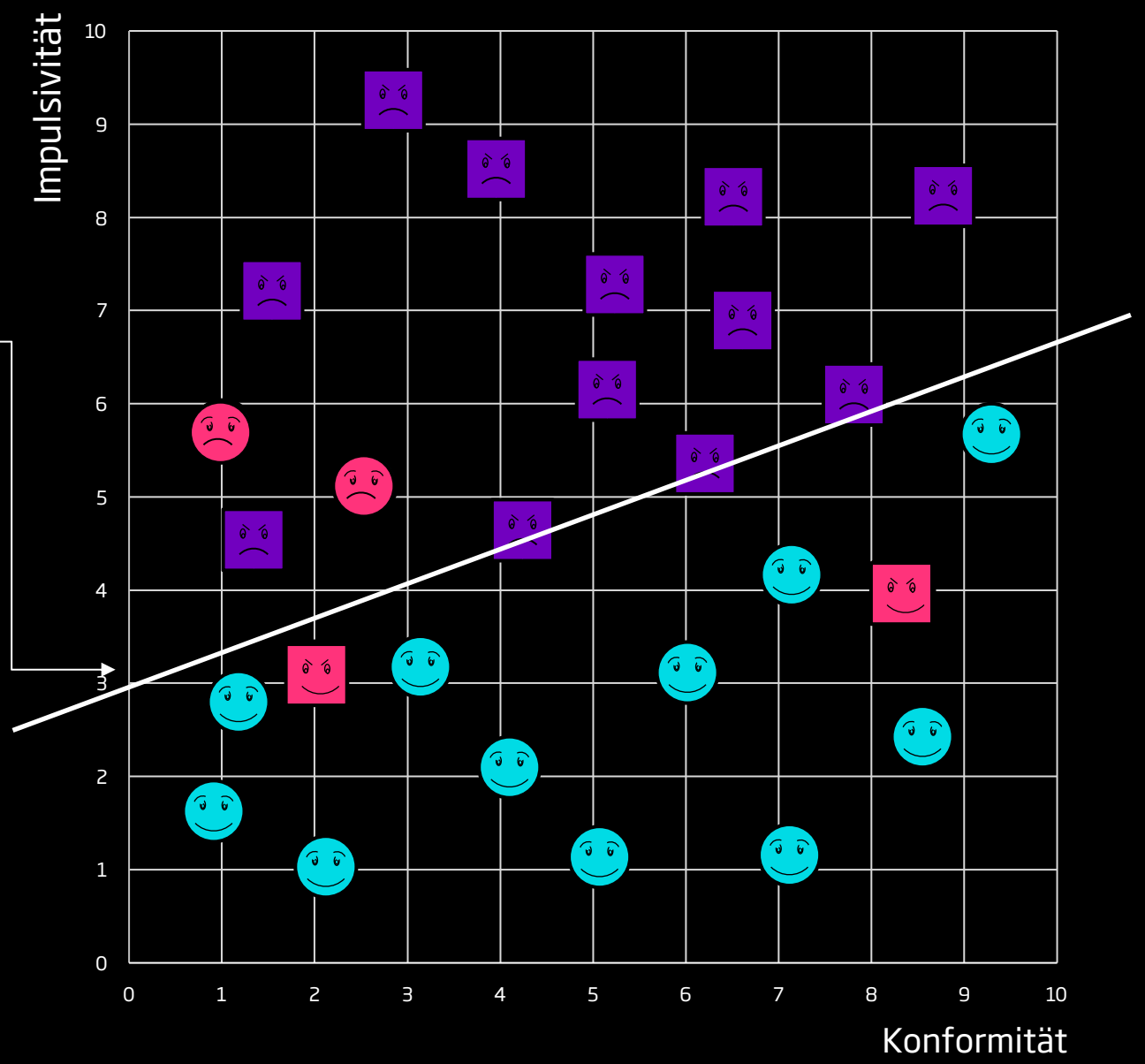
Support Vector Machine

Eine der möglichen Trennlinien

Alle möglichen Trennlinien erzeugen Fehler

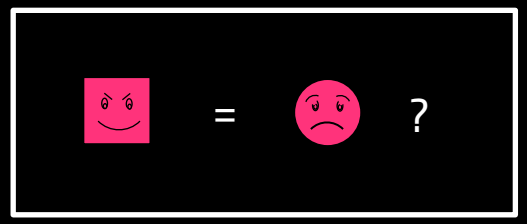
Straftäter, die unentdeckt bleiben

Unschuldige Bürger, die für Straftäter gehalten werden



TÜV AI.LAB

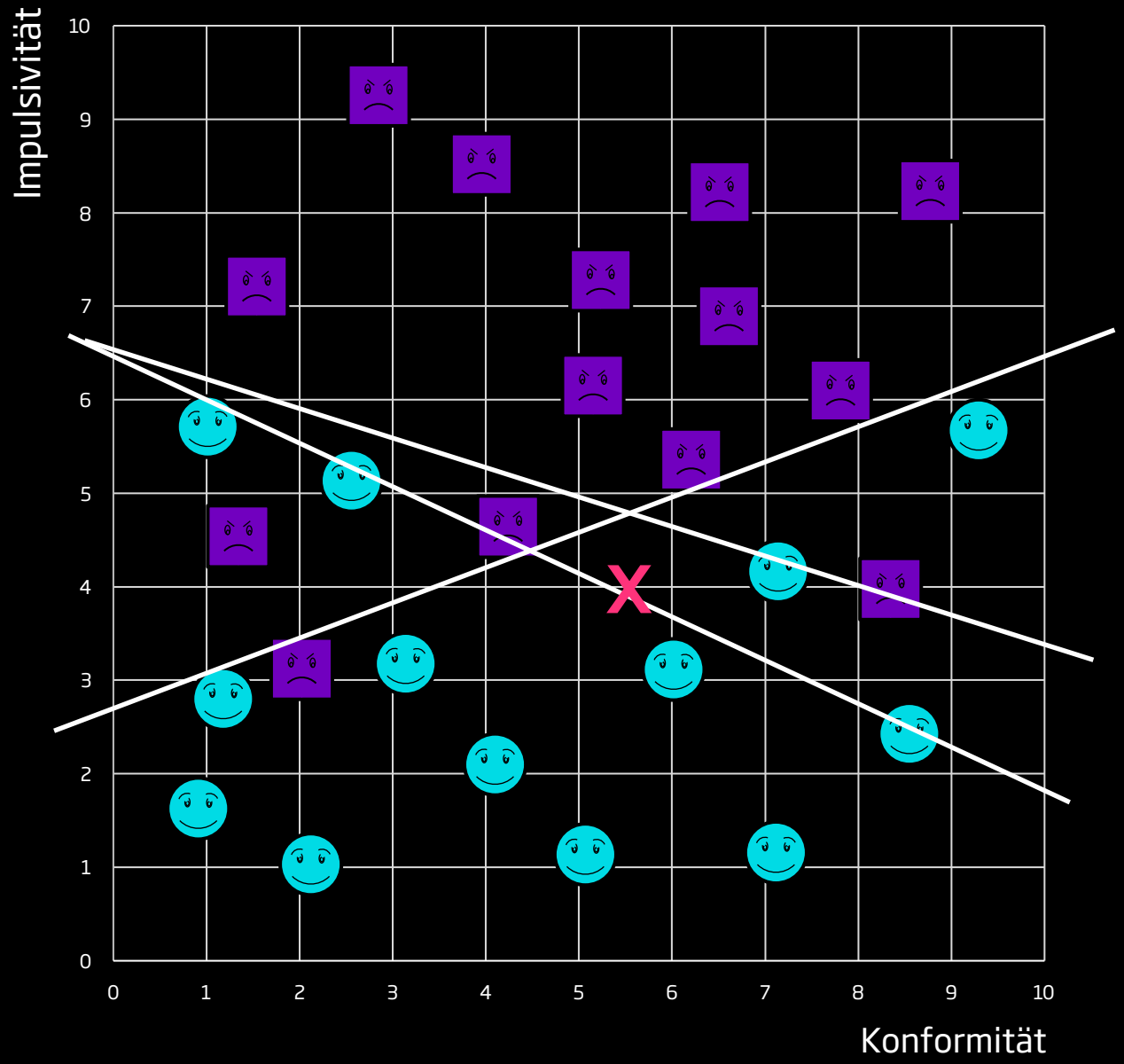
Support Vector Machine



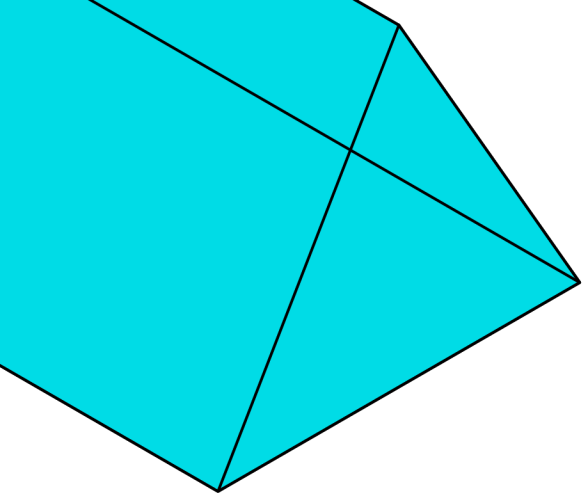
Wichtige Explikation:
Was ist „Positive“?
Was ist „Negative“?
In diesem Fall:
<u>positive Behandlung</u>
(das Gewollte/Gesollte)
oder Positiv = „Ja“ =
Verbrecher?

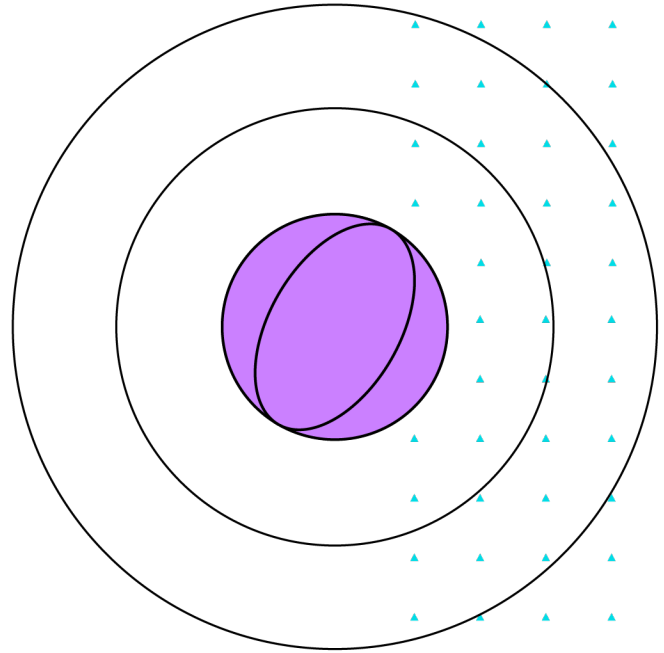
FP = FN ?

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



TÜV AI.LAB

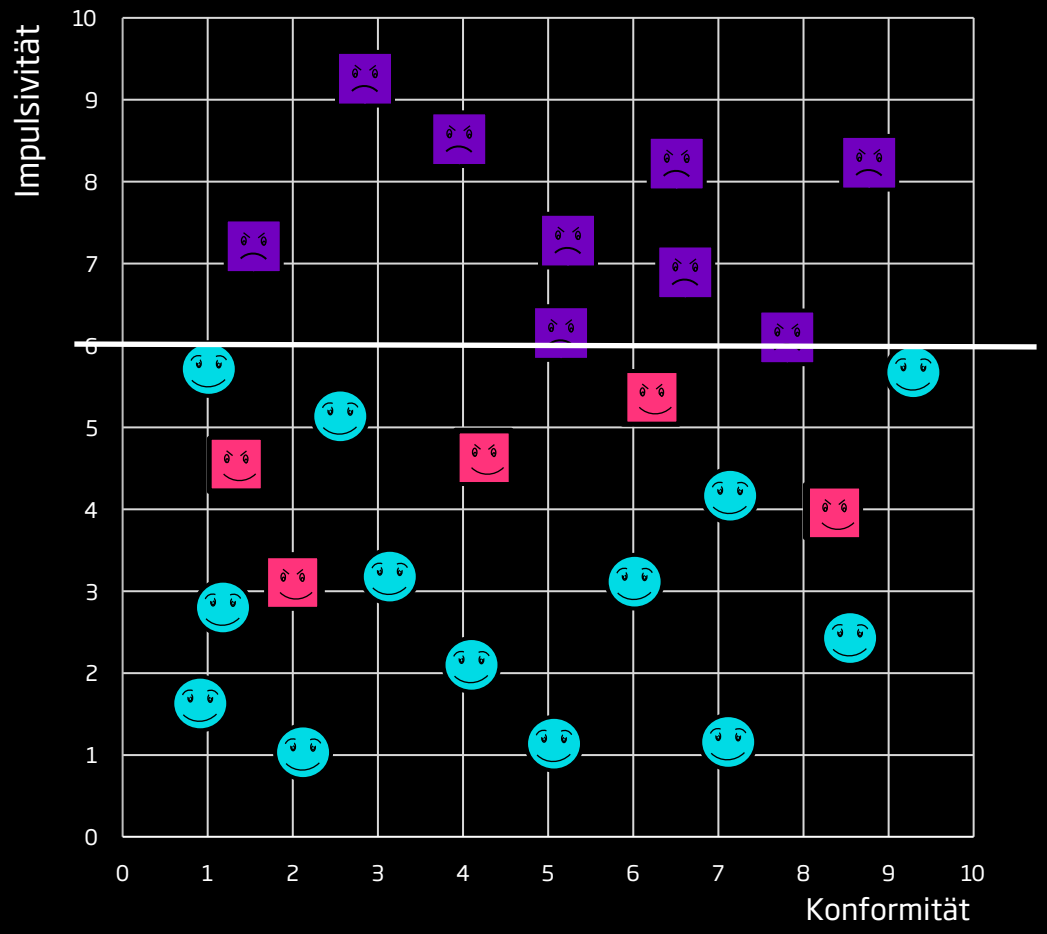




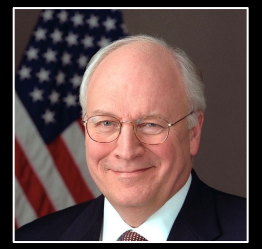
Sind beide Fehlerarten gleich zu bewerten?

„It is better that ten guilty persons escape than that **one** innocent suffers."

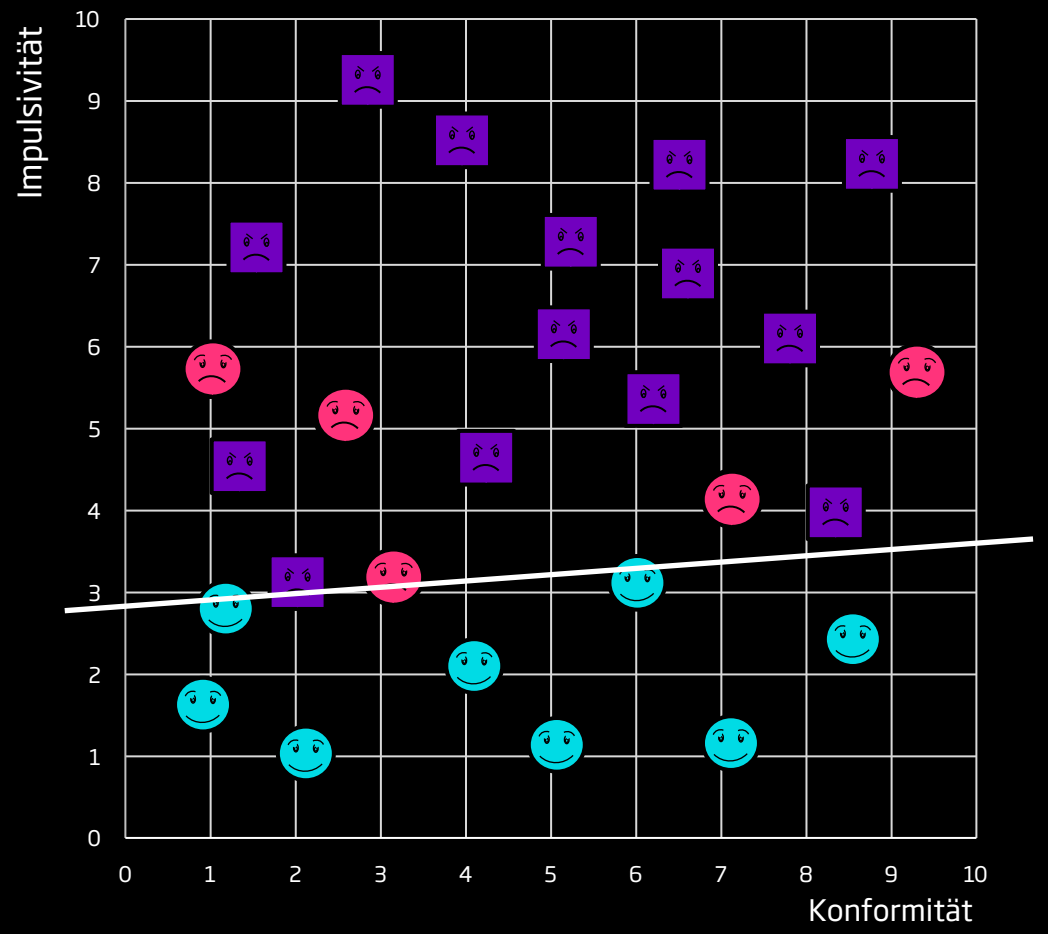
William Blackstone, Rechtsphilosoph, 1760



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



Dick Cheney, ehemaliger Vizepräsident der USA



Kurzfassung Leistungsmetriken

Klassifikation

› Auf Basis der Grundwahrheit: Konfusionsmatrix

P+N	PP	PN
P	TP	FN
N	FP	TN

TPR = TP/P	FNR = FN/P
FPR = FP/N	TNR = TN/N

PPV = TP/PP	FOR = FN/PN
FDR = FP/PP	NPV = TN/PN

TP = true positives
FP = false positives
TN = true negatives
FN = false negatives

P = Gesamtzahl der positiven Instanzen = TP + FN
N = Gesamtzahl der negativen Instanzen = FP + TN
PP = Gesamtzahl der als positiv klassifizierten Instanzen = TP + FP
PN = Gesamtzahl der als negativ klassifizierten Instanzen = TP + FP

TÜV AI.LAB

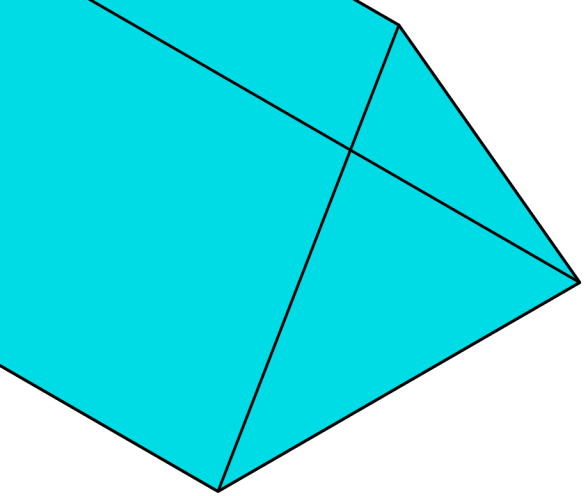
Kurzfassung Leistungsmetriken

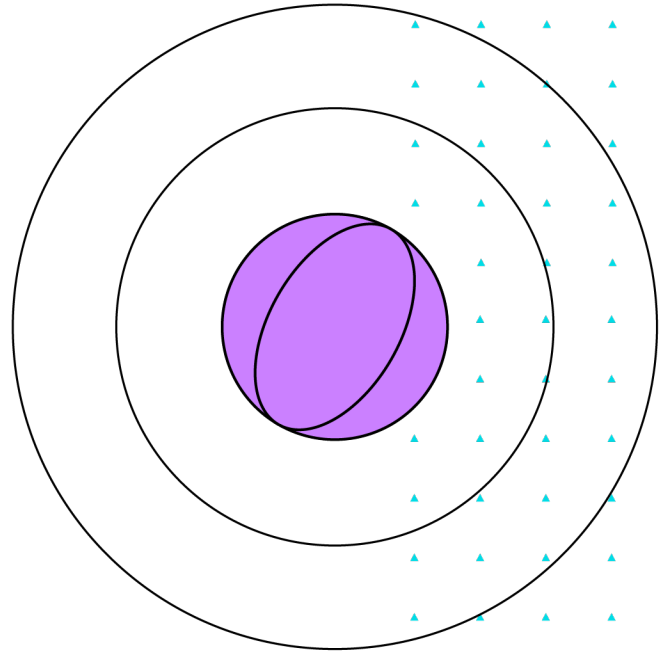
Klassifikation

- Auf Basis der Grundwahrheit: Konfusionsmatrix
- Qualität

Quality measure	Formula	
Accuracy (ACC)	$\frac{TP+TN}{TP+TN+FP+FN}$	
True positive rate (TPR), Recall or sensitivity	$\frac{TP}{TP+FN}$	<- Blackstone
True negative rate (TNR) or specifity	$\frac{TN}{TN+FP}$	<- Cheney
False positive rate (FPR) or fall-out	$\frac{FP}{FP+TN}$	
False negative rate (FNR) or miss rate	$\frac{FN}{FN+TP}$	
Positive predictive value (PPV) or precision	$\frac{TP}{TP+FP}$	
False discovery rate (FDR)	$\frac{FP}{FP+TP}$	
Negative Predictive Value (NPV)	$\frac{TN}{TN+FN}$	
False omission rate (FOR)	$\frac{FN}{FN+TN}$	
Matthews correlation coefficient (MCC)	$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	
F1-Score	$2 * \frac{PPV * TPR}{PPV+TPR}$	
ROC-AUC	komplizierter	

TÜV AI.LAB





1. Beobachtung

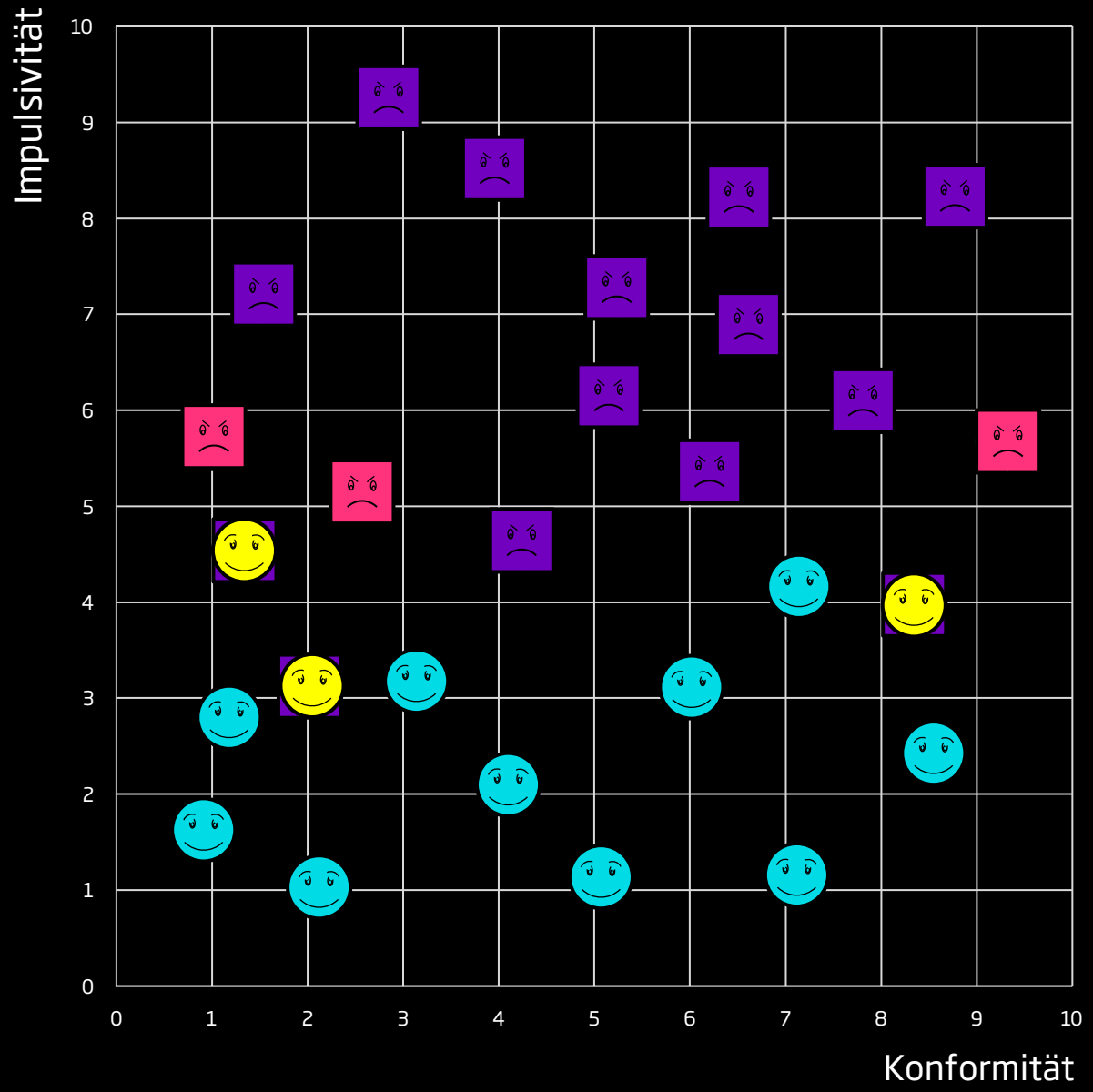
Was durch eine KI optimiert werden soll, ist eine gesellschaftliche bzw. systematische Entscheidung.

Datenqualität

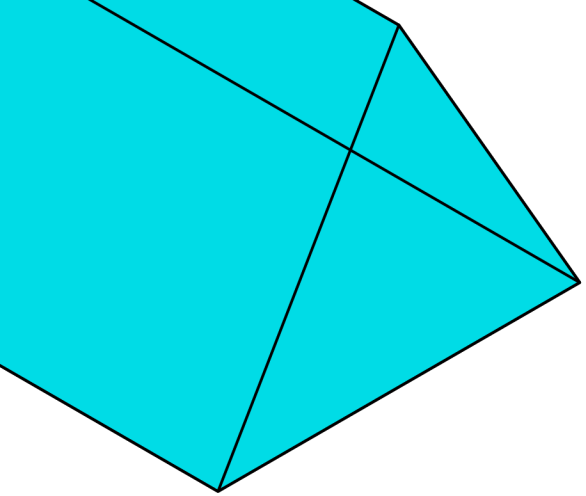
Noch nicht entdeckte Straftäter

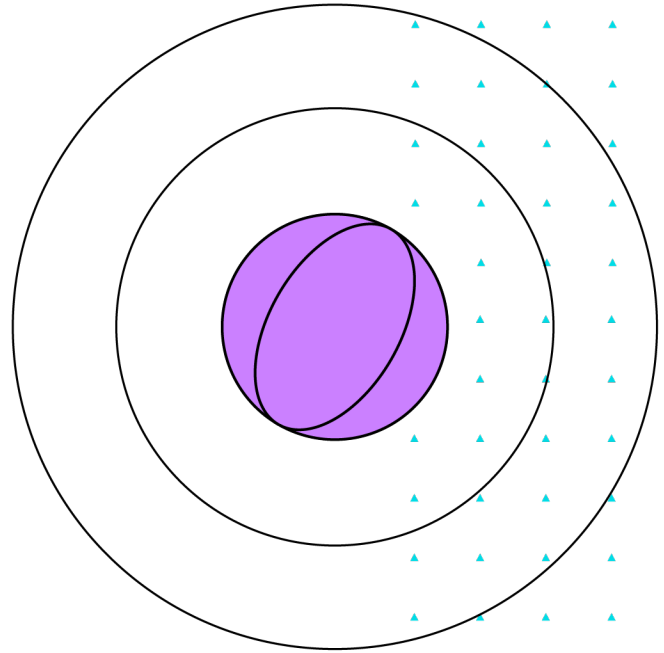
Unschuldig im Gefängnis

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



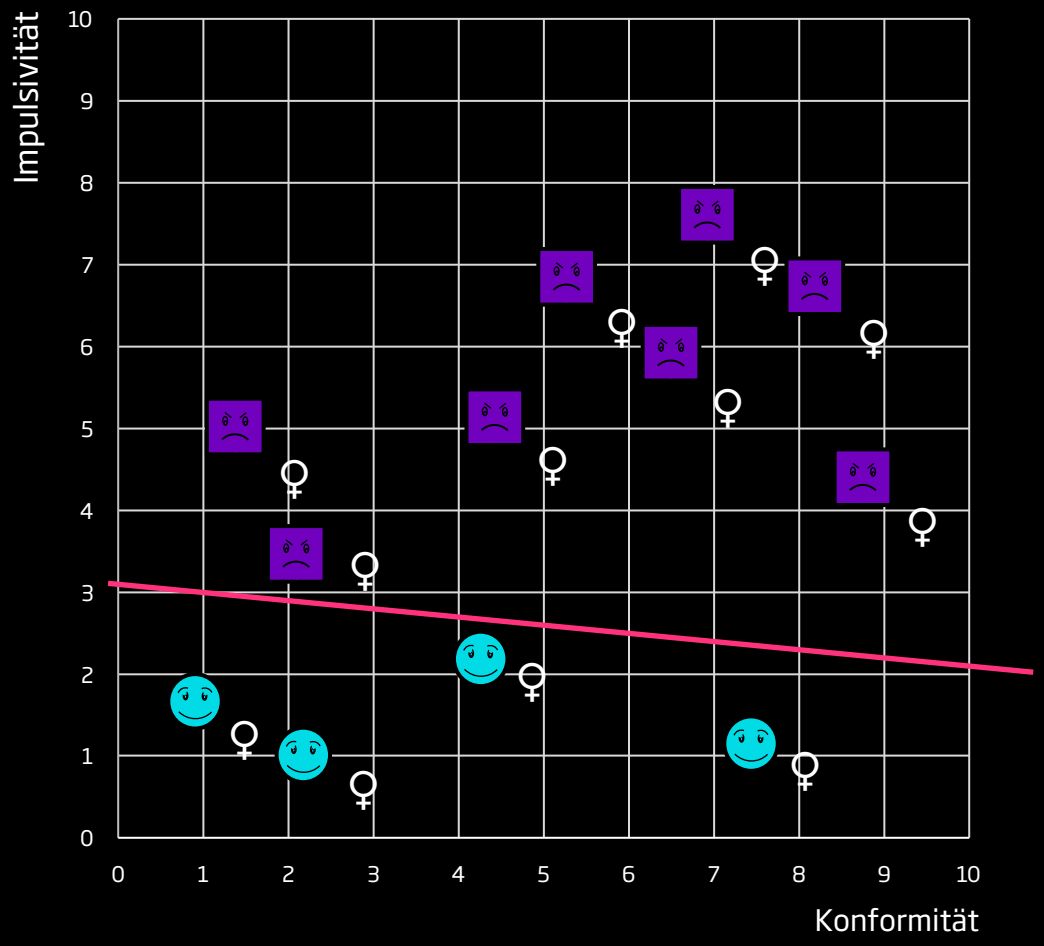
TÜV AI.LAB

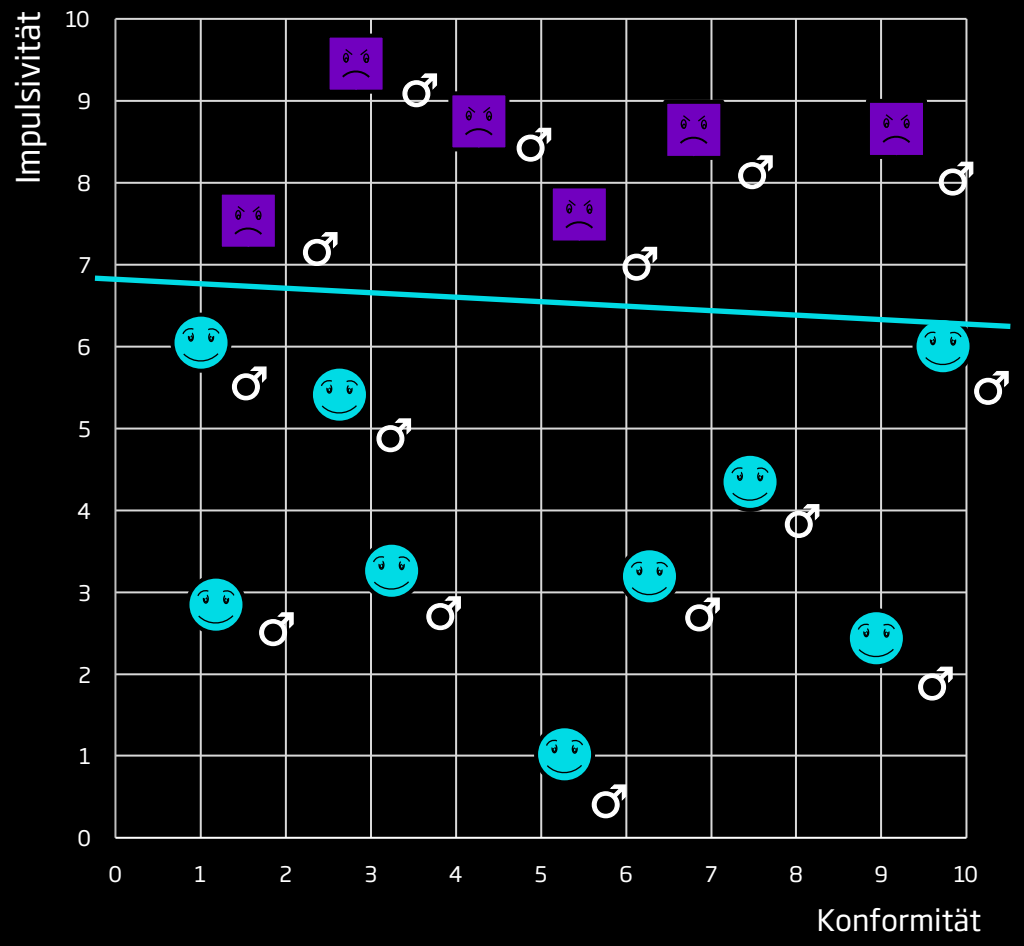




2. Beobachtung

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



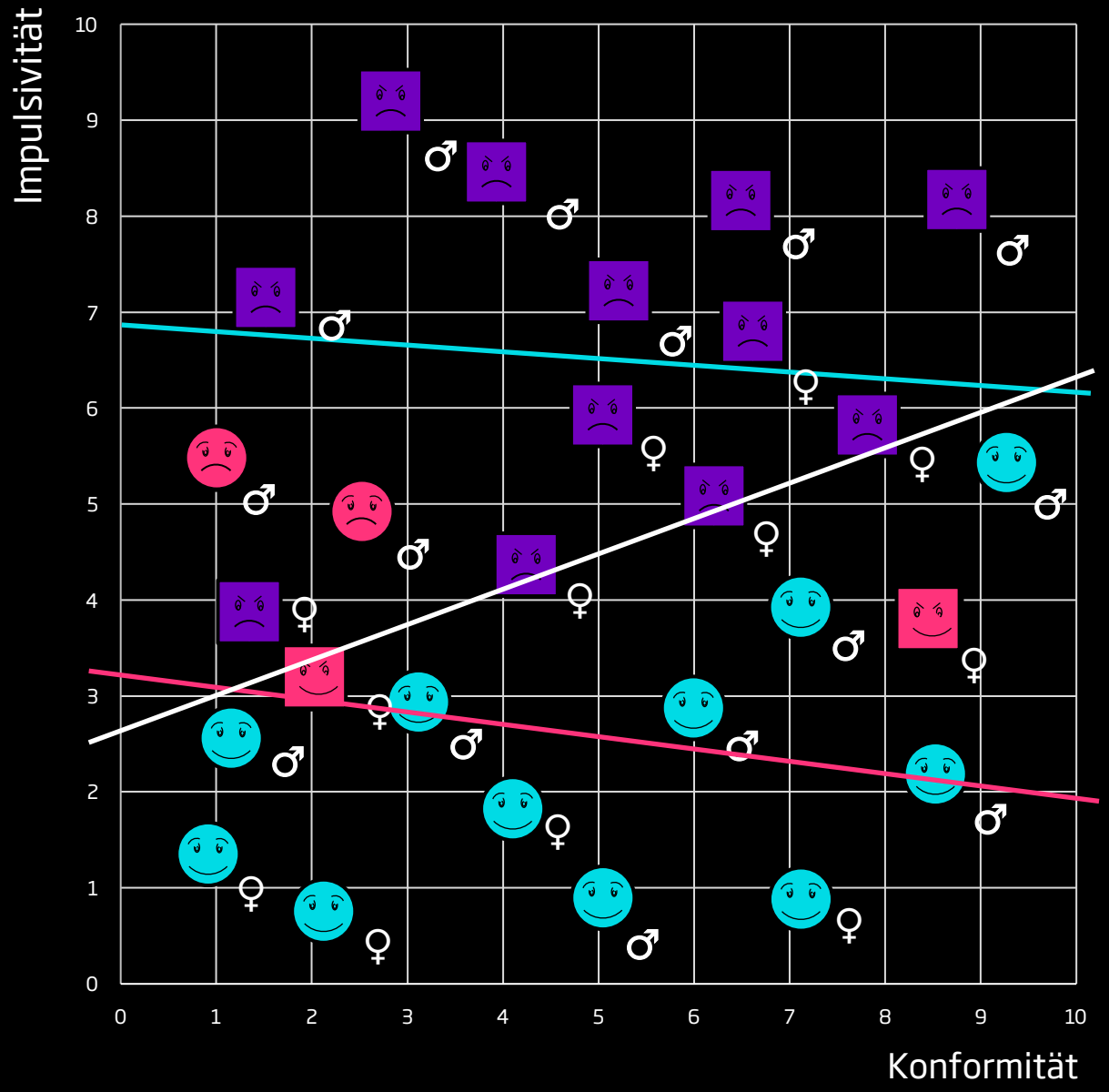


Ergebnis: In diesem fiktiven Beispiel wird für jede Teilgruppe eine optimale Entscheidungsregel ohne Fehler gefunden.

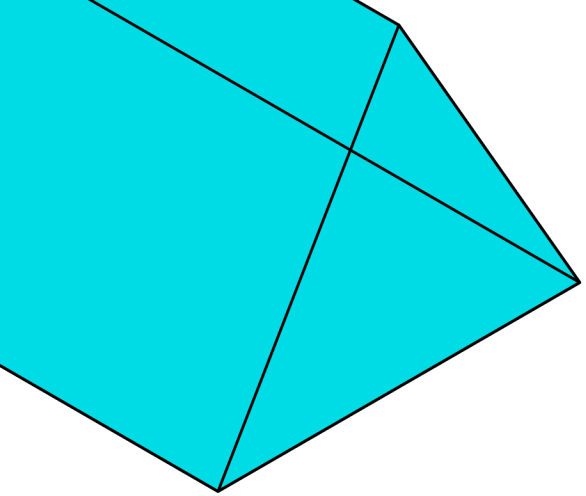
Support Vector Machine

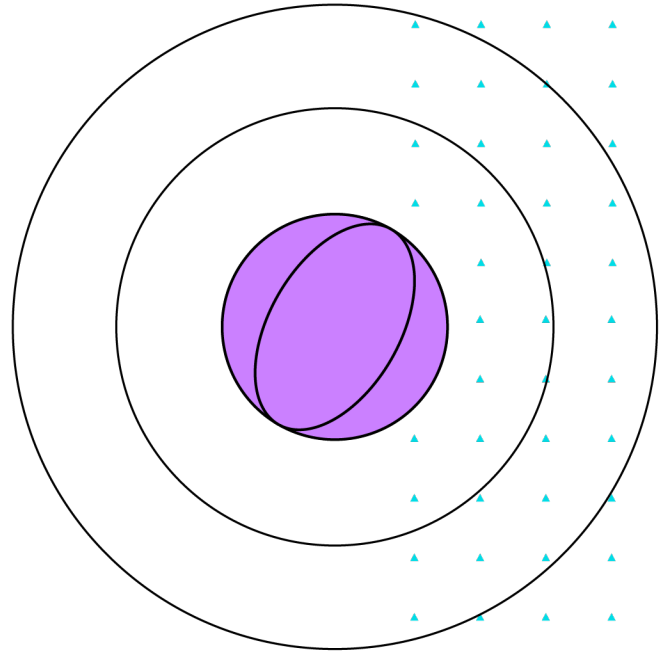
Legt man dagegen beide Gruppen zusammen, diskriminiert die (neu) trainierte Support Vector Machine **Männer:**

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



TÜV AI.LAB



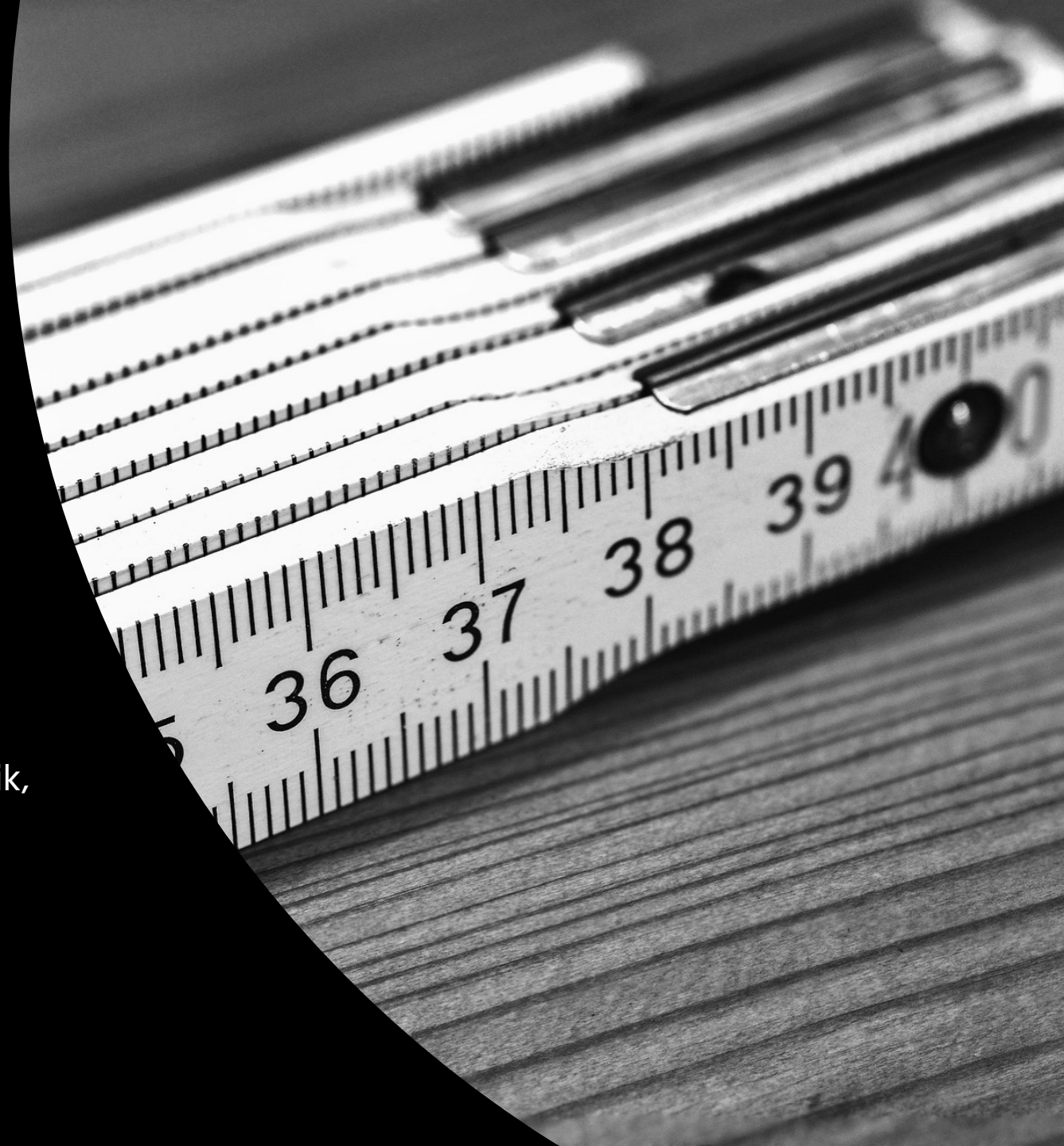


3. Beobachtung

Eine geschützte Information kann wichtig sein.
Diskriminierung wird nicht *per se* dadurch vermieden, dass die Information vorenthalten wird.

Diskriminierung messen

- Qualitätsmaß(e) wählen, z.B. False-Positive-Rate?
- (Statistische) Gleichheit der Teilgruppen im Qualitätsmaß fordern?
  Ansatz: Teilgruppe sollte min. 80% des maximalen Wertes haben (Buolamwini 2017: 49).
- Vorsicht: Manche Fairnessmaße widersprechen einander! (Zweig/Krafft 2018)
- Gesellschaftlicher sowie systematischer Diskurs (Ethik, Rechtswissenschaften, usw.) zur Wahl der Metrik(en) notwendig



TÜV
AI.LAB

Kurzfassung Fairnessmetriken

Klassifikation

- Auf Basis der Grundwahrheit (Konfusionsmatrix)
- Qualität
- Fairness

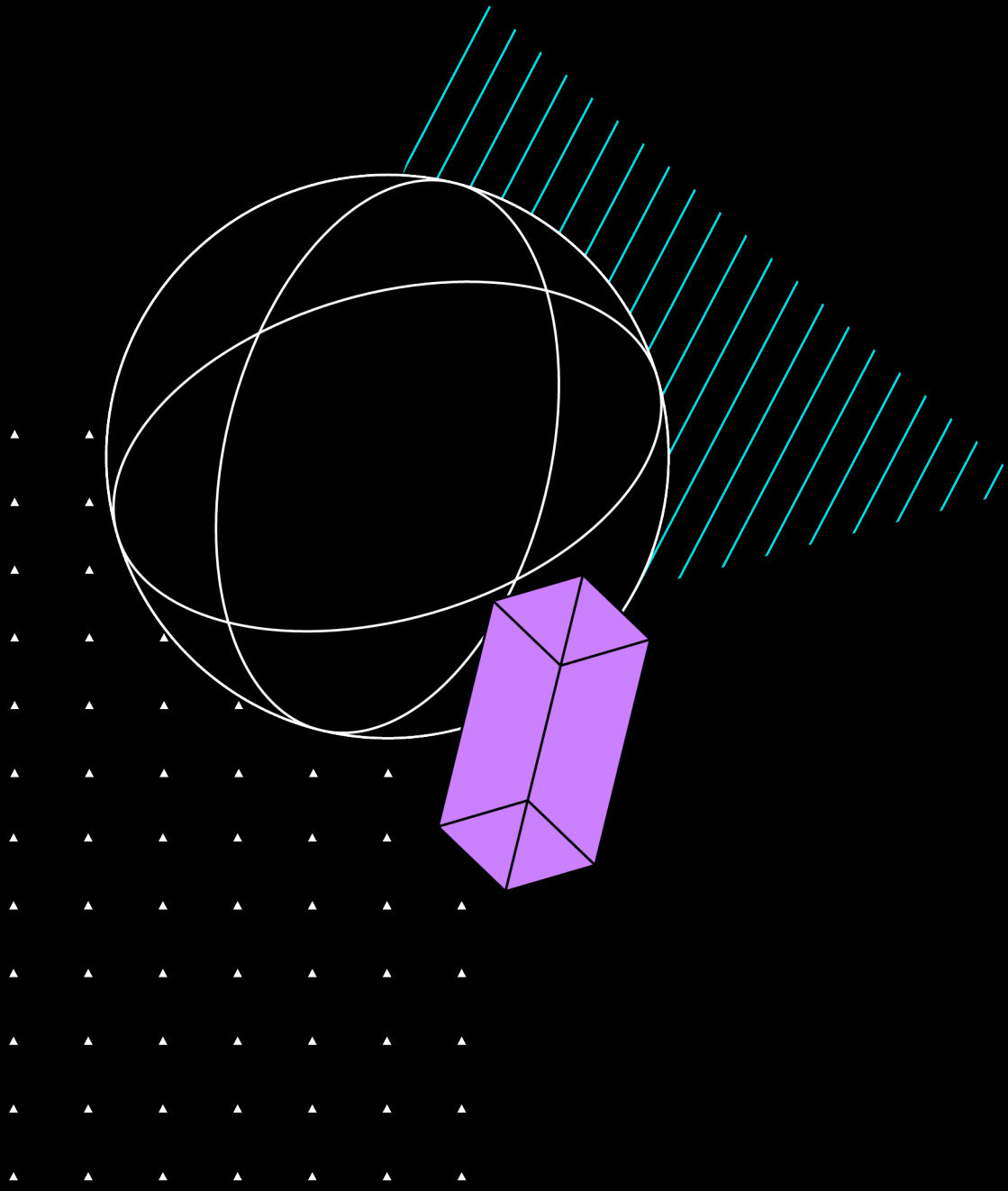
Fairness measure	Requires equality of
Overall accuracy equality	ACC
Separation Conditional procedure accuracy Equalized odds	TPR and FPR
Equal opportunity	TPR
Error rate balance	FPR and FNR
Sufficiency	PPV and FOR
Conditional use accuracy	PPV and NPV

TÜV AI.LAB

Vielen Dank

Dr. Christoph Poetsch
Senior Advisor AI Ethics & Quality

christoph@tuev-lab.ai



TÜV
AI.LAB