# My AI discriminates? How could this happen and who is to blame?

**Marc Hauer**

**TrustedAI GmbH**
**Algorithm Accountability Lab**
**@hauer_p**

# Workshopleitung

ALGORITHM ACCOUNTABILITY LAB

GOAL
Governance von und durch Algorithmen

Marc Hauer, M.Sc.

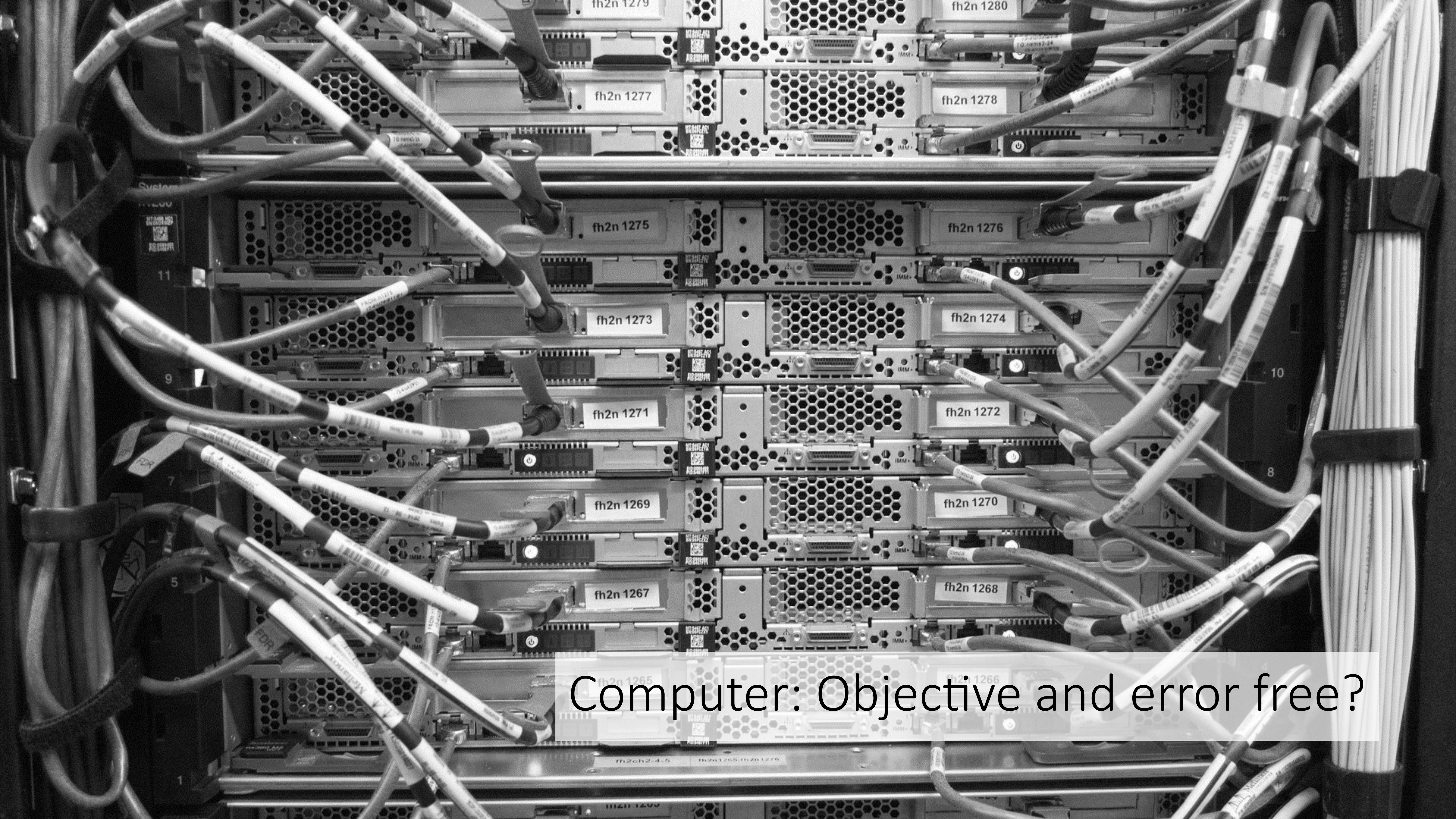
- PhD candidate on the Algorithm Accountability Lab of TU Kaiserslautern
- Ministerial project: Governance of and by algorithms
- Ministerial project: Testing, Auditing and Certification of AI
- Media education consultant of the Landesmedienzentrum Baden-Württemberg
- Consultant of the TrustedAI GmbH

ExamAI

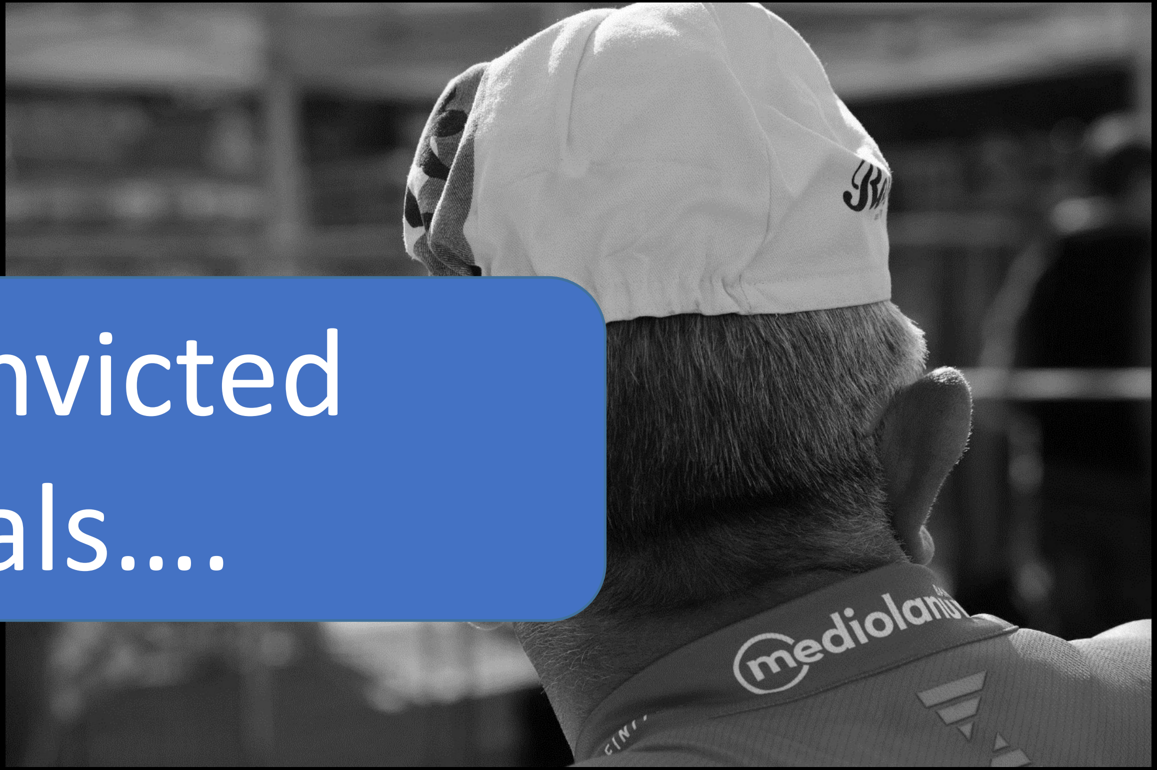
TRUSTED AI

# Goals of the Trusted AI GmbH

Guidance in the ethical development and use of AI systems.



# Computer: Objective and error free?



…two convicted criminals….

# Brisha and Vernon,….

Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner: Machine Bias, ProPublica, 23.5.2016
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Who would do it again?

# Humans – so irrational!

- Study: less risky decisions the longer it has been since the last break [1].
- A large number of such studies seem to prove:
  - Humans are irrational and prejudiced.

[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

# ACLU (American Civil Liberties Union) demans:

2011

accurate data analysis to calculate the risk of offenders actually recidivating and becoming a danger to society

Chettiar, I. M., & Gupta, V. (2011). Smart Reform is Possible: States Reducing Incarceration Rates and Costs While Protecting Communities. *Available at SSRN 1934415*.

2019

**no** accurate data analysis to calculate the risk of offenders actually recidivating and becoming a danger to society

https://civilrights.org/2018/07/30/more-than-100-civil-rights-digital-justice-and-community-based-organizations-raise-concerns-about-pretrial-risk-assessment/

# How can computers learn?

# Experience-based learning

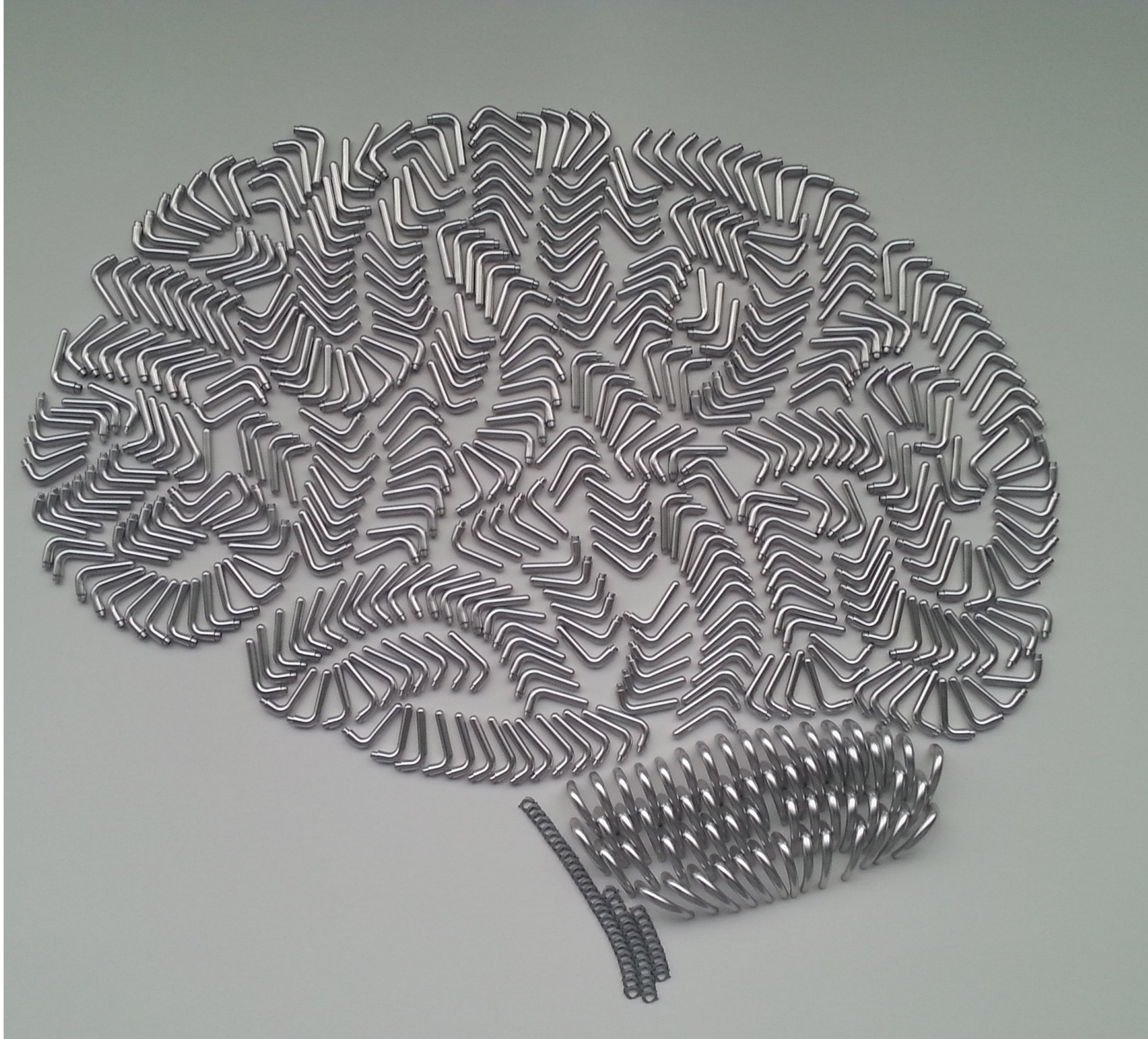
Behavior in video conferences :

- In the beginning, people often got into each other's words.

- We learned to read the facial expressions and gestures of the participants!

# Humans learn…

- through feedback
- through storing in a structure: the neurons and their connections
- through generalization of what has been learned.



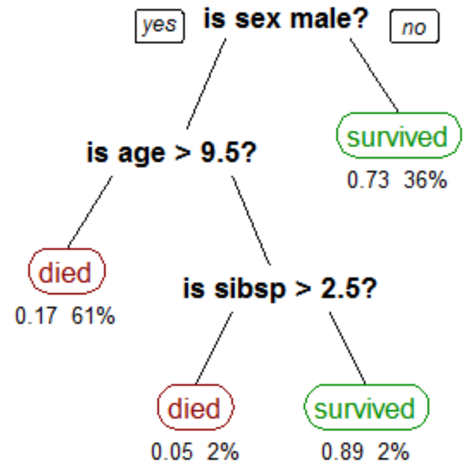
# Computers learn

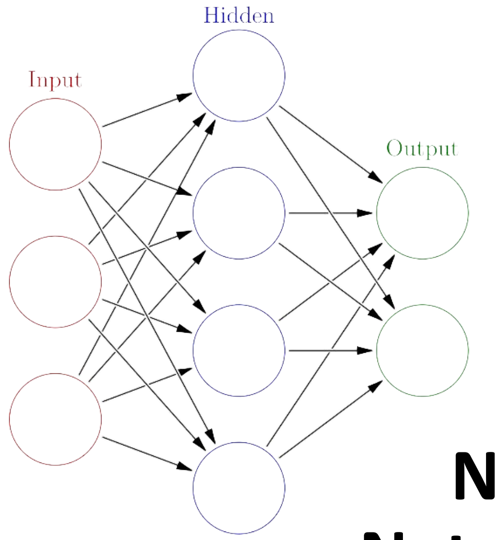
For a computer to learn, it also needs a **structure** to store what it has learned.

Optimally also through **feedback**.

It learns **general rules**.

**Decision trees**





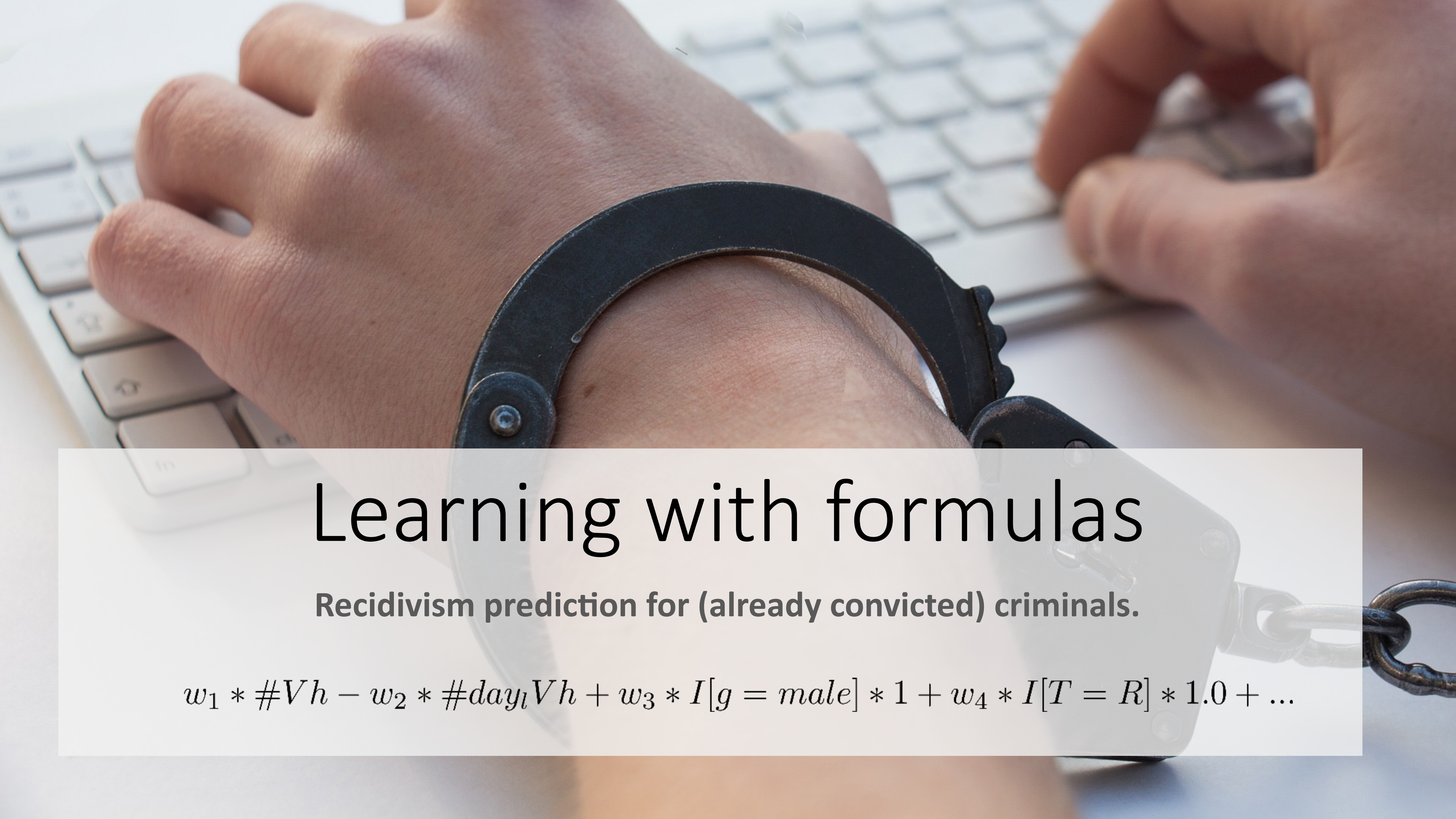
**Neural Networks**

**Formulas**

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

"CART tree titanic survivors" by Stephen Milborrow - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons - https://commons.wikimedia.org/wiki/File:CART_tree_titanic_survivors.png#/media/File:CART_tree_titanic_survivors.png

By Glosser.ca - Own work, Derivative of File:Artificial neural network.svg, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=24913461



# Learning with formulas

**Recidivism prediction for (already convicted) criminals.**

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

# Data basis

- Machine learning methods use e.g.:
  - Age at the first arrest
  - Age of the delinquent
  - Financial situation
  - Criminal relatives
  - Gender
  - Type and number of previous convictions
  - Time of the last criminal record
  - …
- Important: At the training set, it is known whether the person has recidivated or not.

# Regression

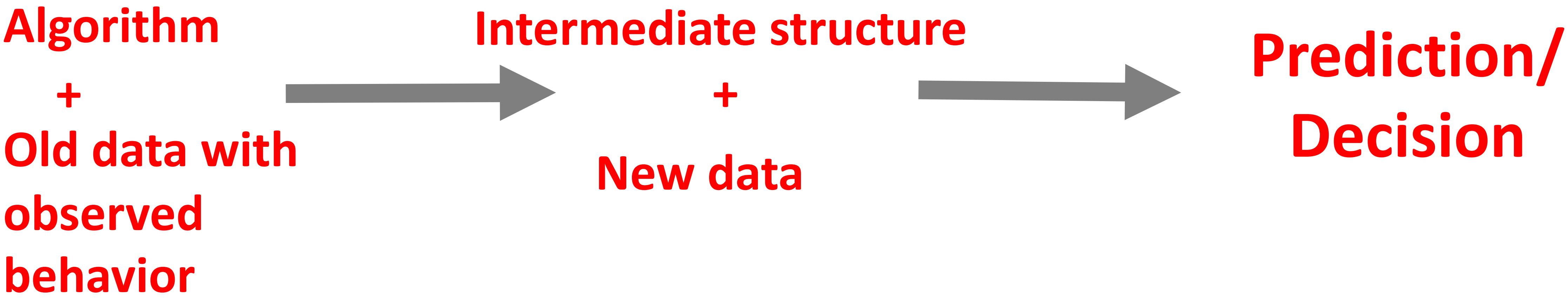
$\omega_1$ * number of previous convictions
- $\omega_2$ * days since the last arrest
+ $\omega_3$ (1 if male, 0 if not)
+ $\omega_4$ (1 if robbery, 0 if something else) + …

3 * number of previous convictions
- 2 * days since the last arrest
+ 2,5 (1 if male, 0 if not)
+ 3,5 (1 if robbery, 0 if something else) + …

The computer determines the weights and gets feedback on the extent to which the resulting score actually matches the (observed) behavior.

# Learning procedures

- **Task:** Given a set of known data, find patterns that predict how something or someone will behave on new data.
- Algorithm builds an intermediate structure - based on known data - which then generates predictions for new data.
- The algorithm is said to be "trained on the data".



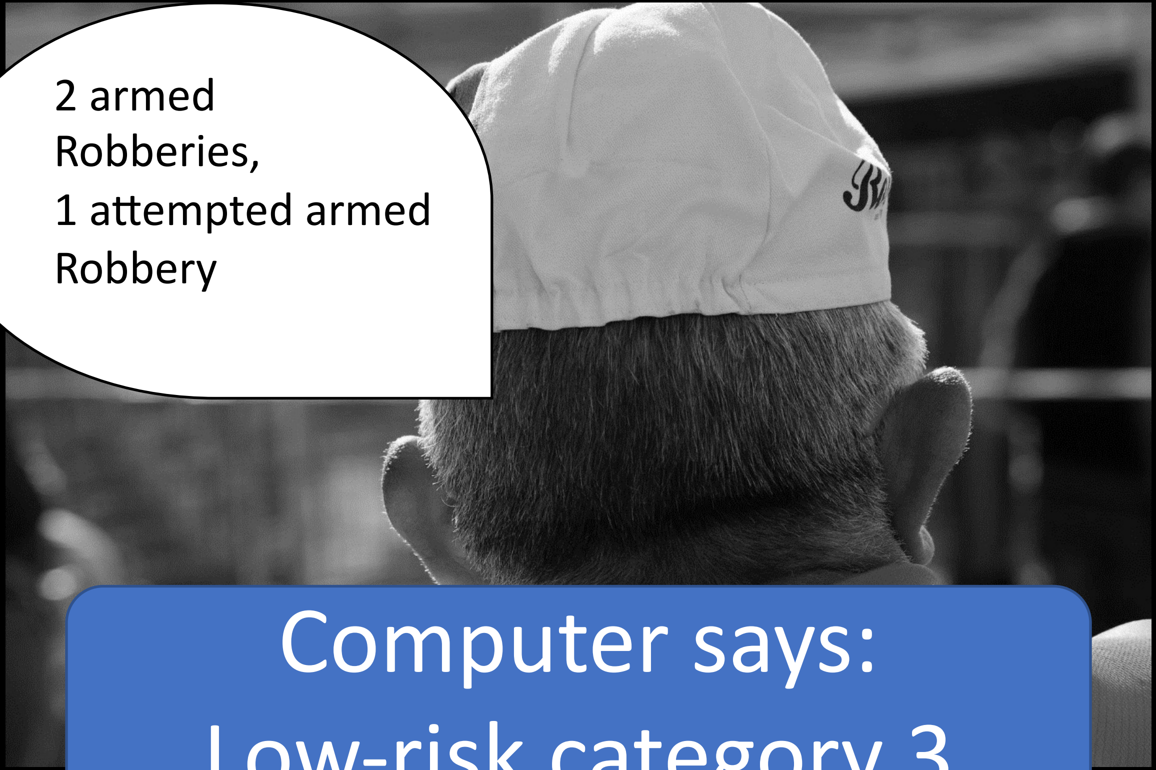
Algorithm
+
Old data with observed behavior

Intermediate structure
+
New data

Prediction/ Decision

Four times punishment according to juvenile law (minor offenses)

Computer says:
High-risk category 8



2 armed Robberies,
1 attempted armed Robbery

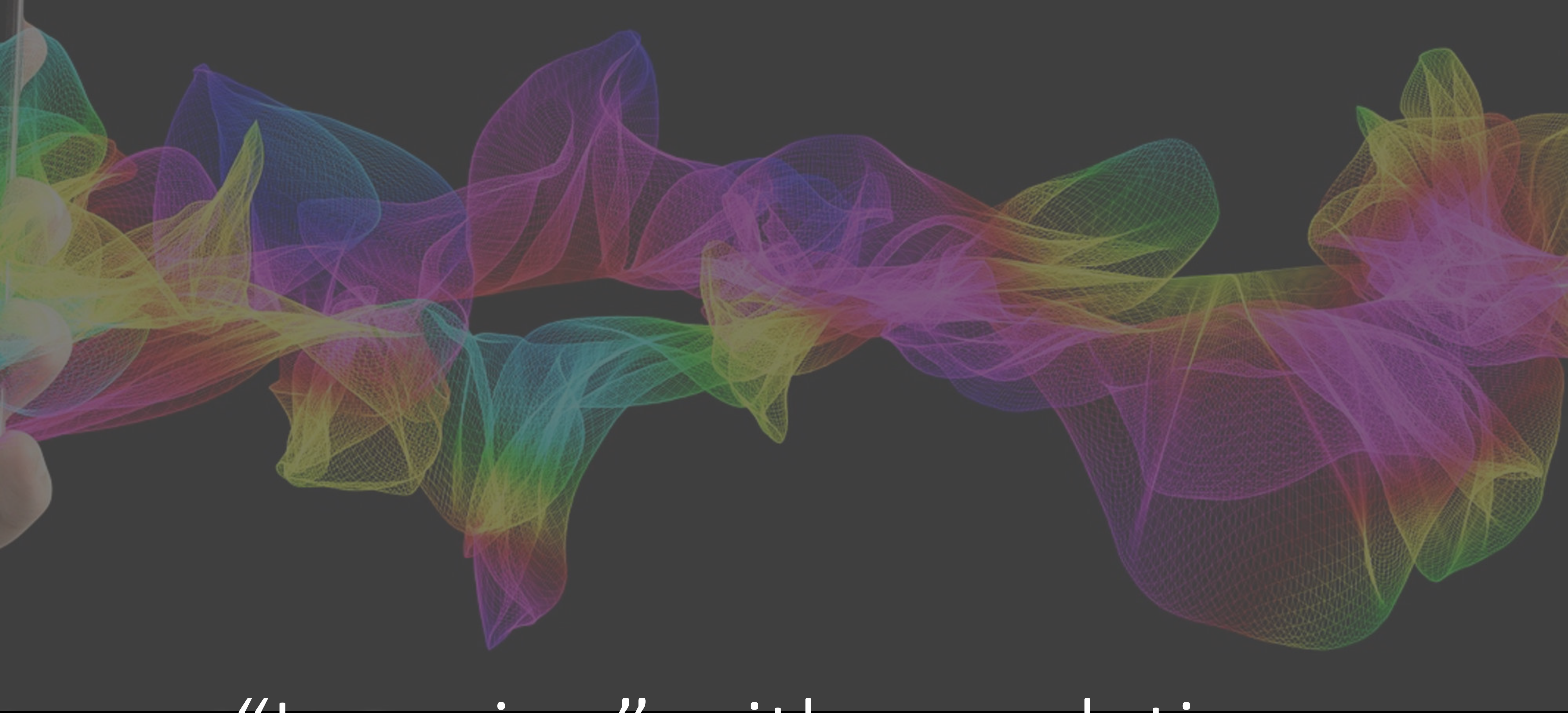
Computer says:
Low-risk category 3

# Who will do it again?

Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner: Machine Bias, ProPublica, 23.5.2016
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Who did it again?

Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner: Machine Bias, ProPublica, 23.5.2016
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing



# “Learning” with correlations

# Algorithms of artificial intelligence...

- … are based on correlations of properties with outcome to be predicted.
- Basically **algorithmically legitimized** prejudices :
  - Out of 100 offenders who are "just like this one," 70 got probation: …
  - … suspend sentence to probation

**AI systems only provide probabilities, not the truth.**

# How does a system learn from data?

**DIY:**
**Today, you are the**
**„Support Vector Machine“**

TRUSTED AI

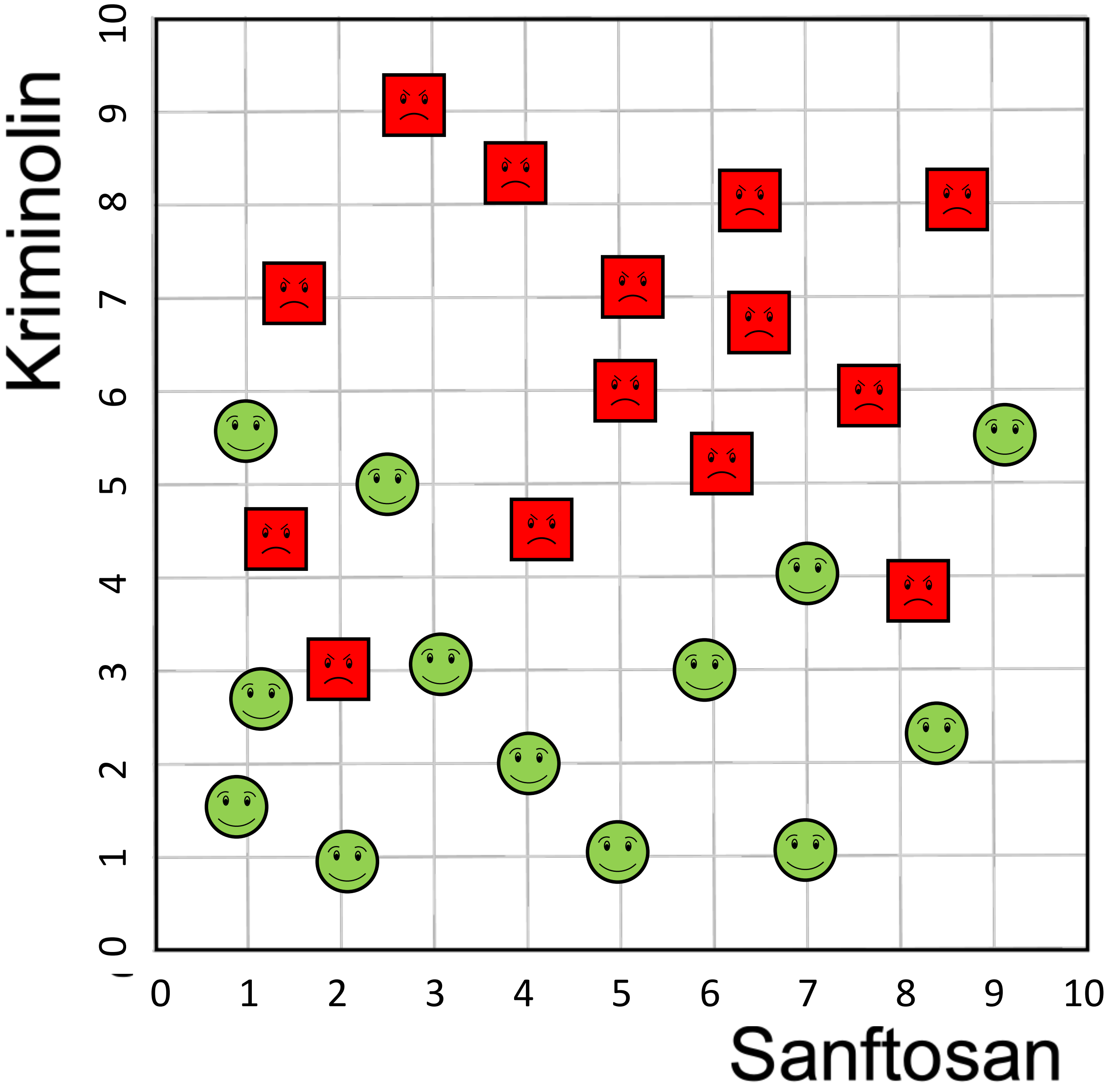
🟥 Malicious criminal

☺ Innocent citizen

Draw a dividing line between the smileys so that the red ones are separated from the green ones as good as possible.

Congratulations: You have trained a Support Vector Machine!

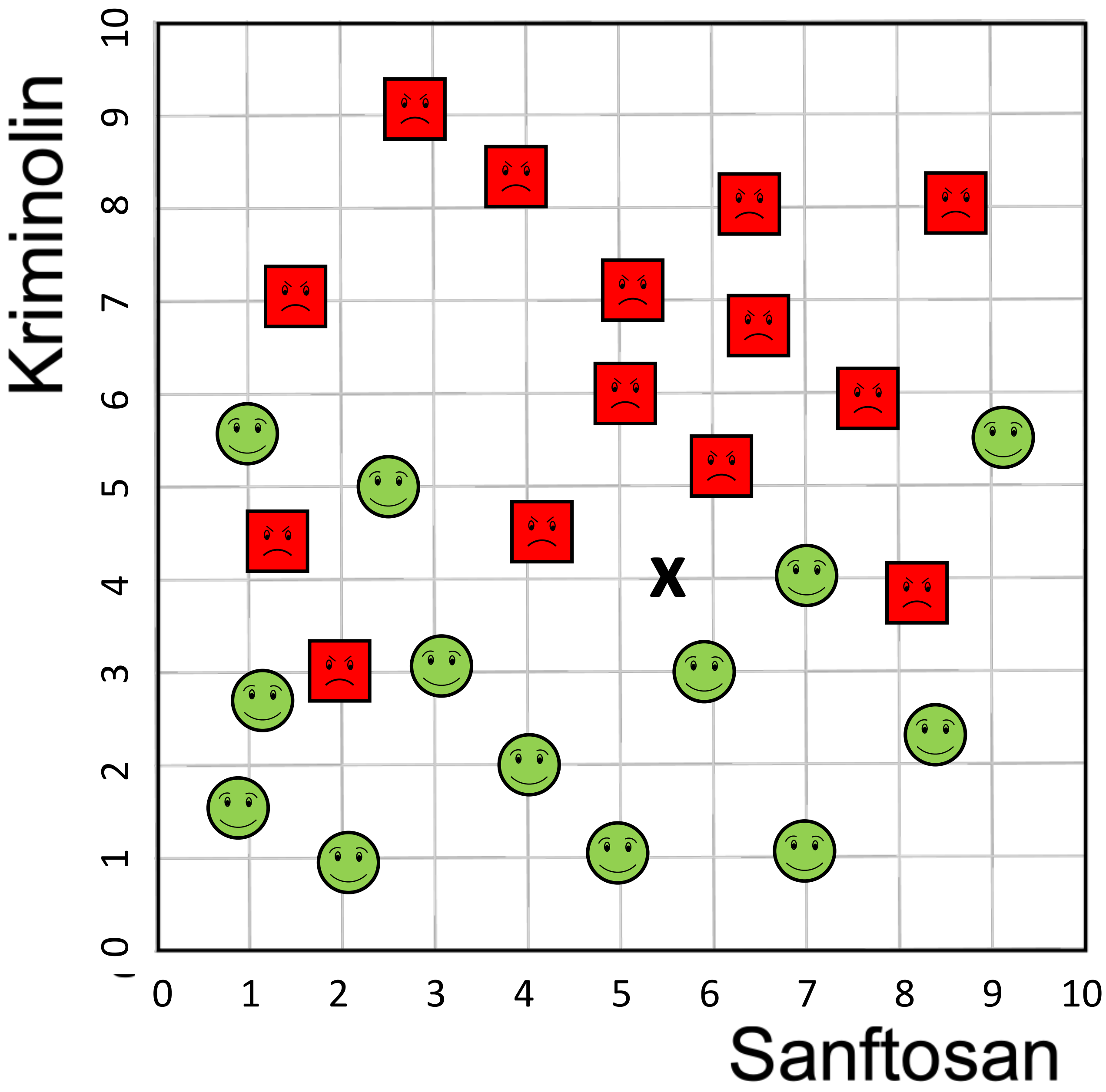
The dividing line now serves as a rule for deciding whether a person is considered a criminal or appears to be innocent.



TRUSTED AI

🟥 Malicious criminal

🟢 Innocent citizen



Judge Mrs. Miller:

5.5 Sanftosan
4.0 Kriminolin

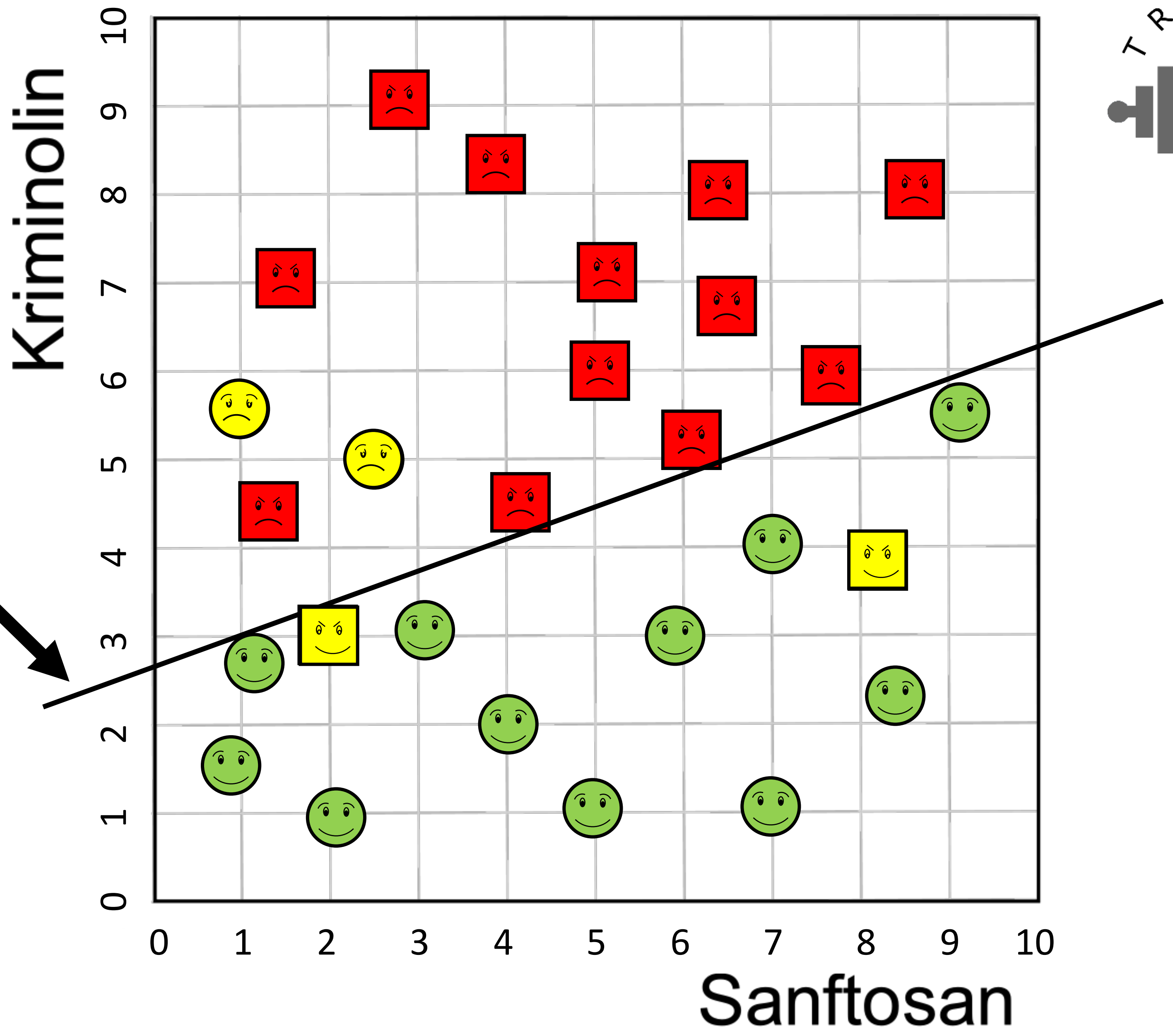
TRUSTED AI

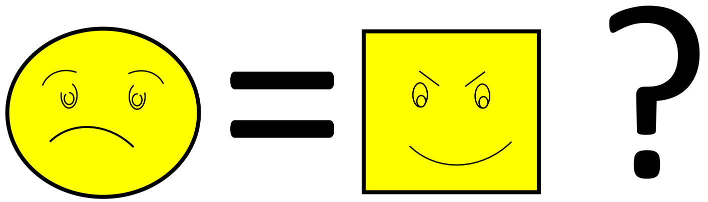
**One of the possible dividing lines**

All possible dividing lines generate errors:

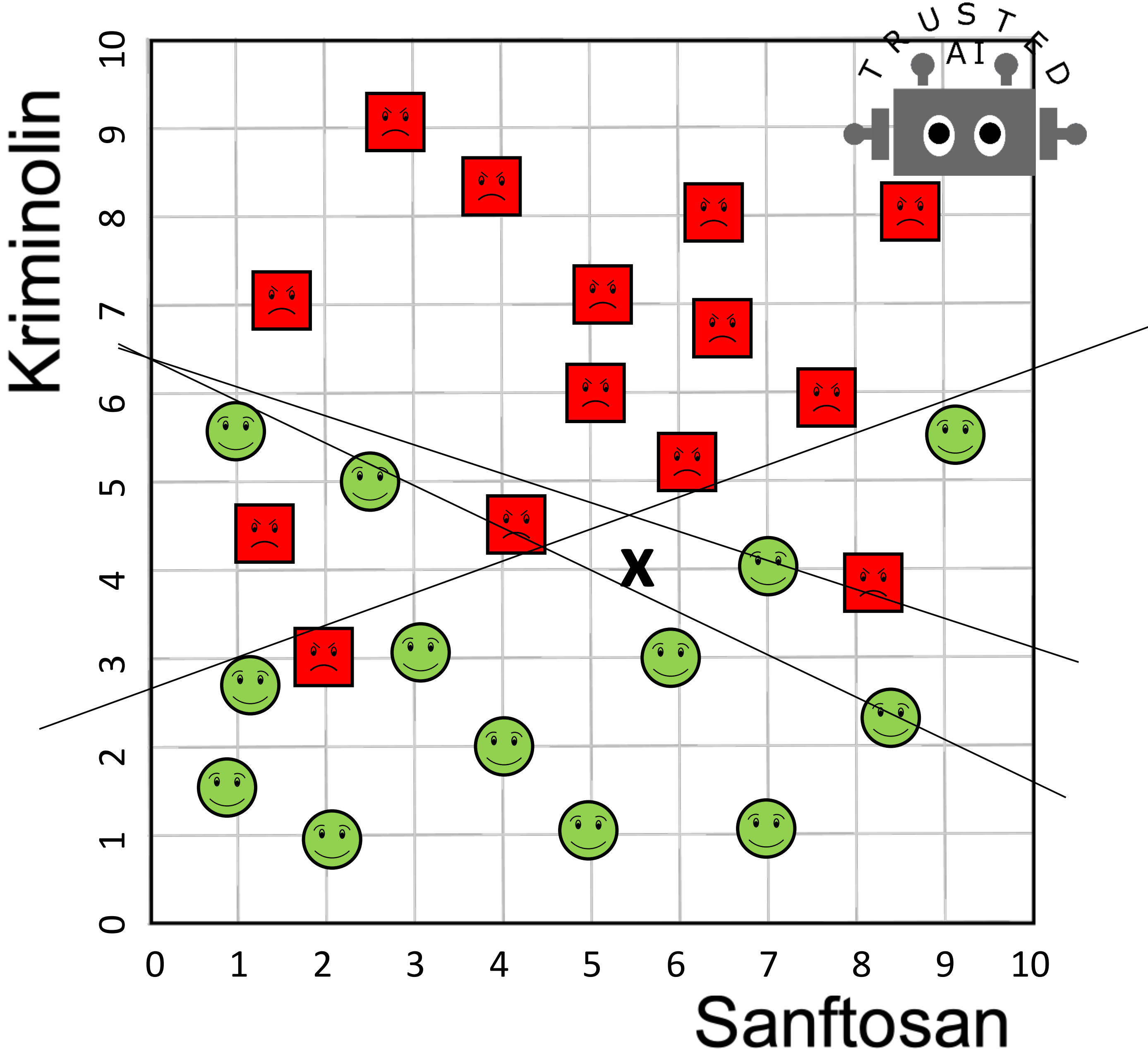
🟨 Malicious criminals who remain undetected

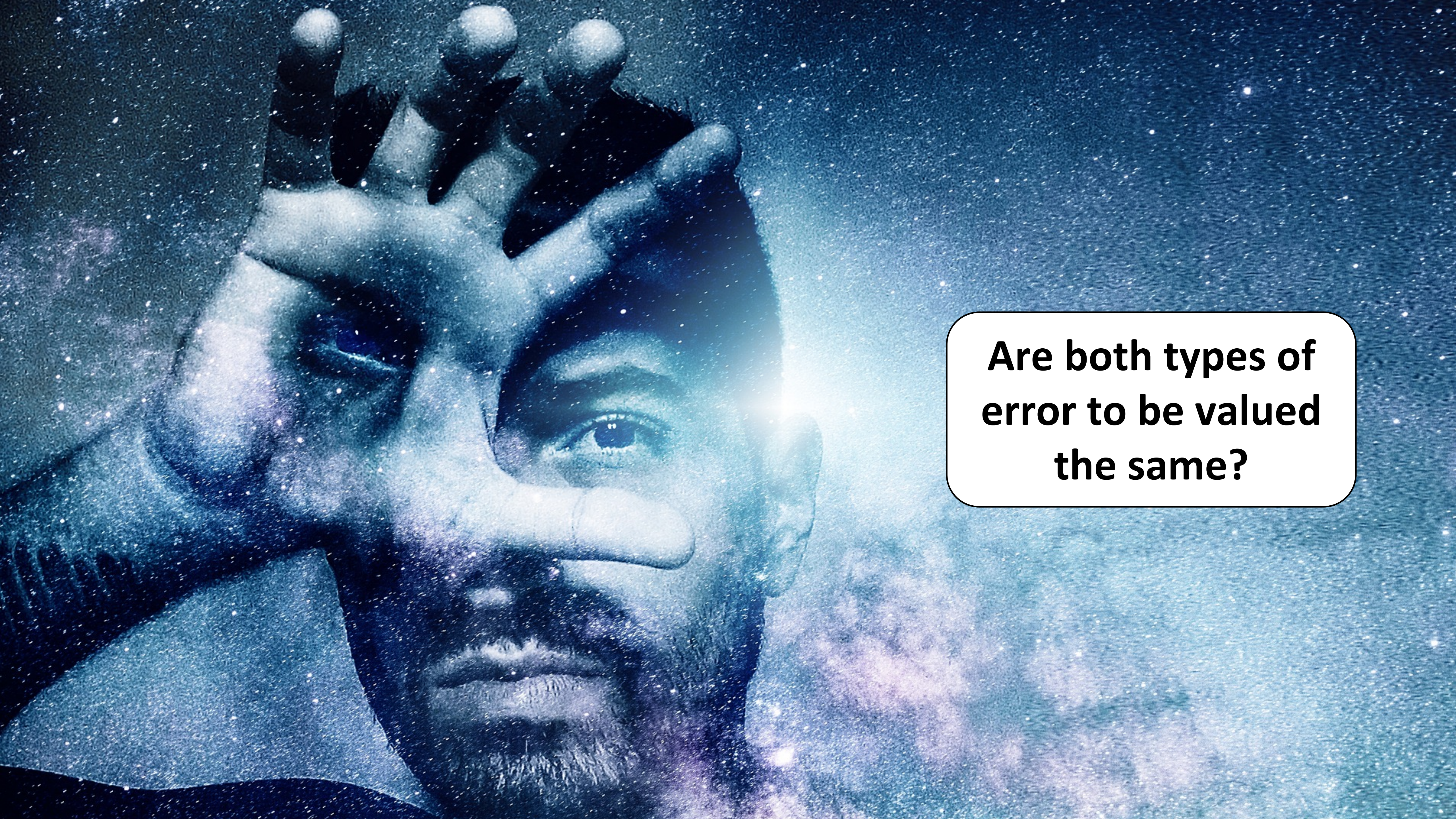
🟡 Innocent citizens, mistaken to be criminals





If both errors are considered equally bad, there are several optimal dividing lines with as few errors as possible.

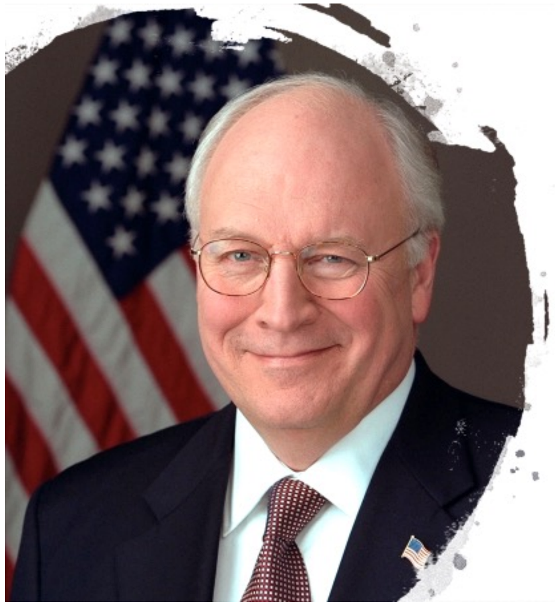




Are both types of error to be valued the same?

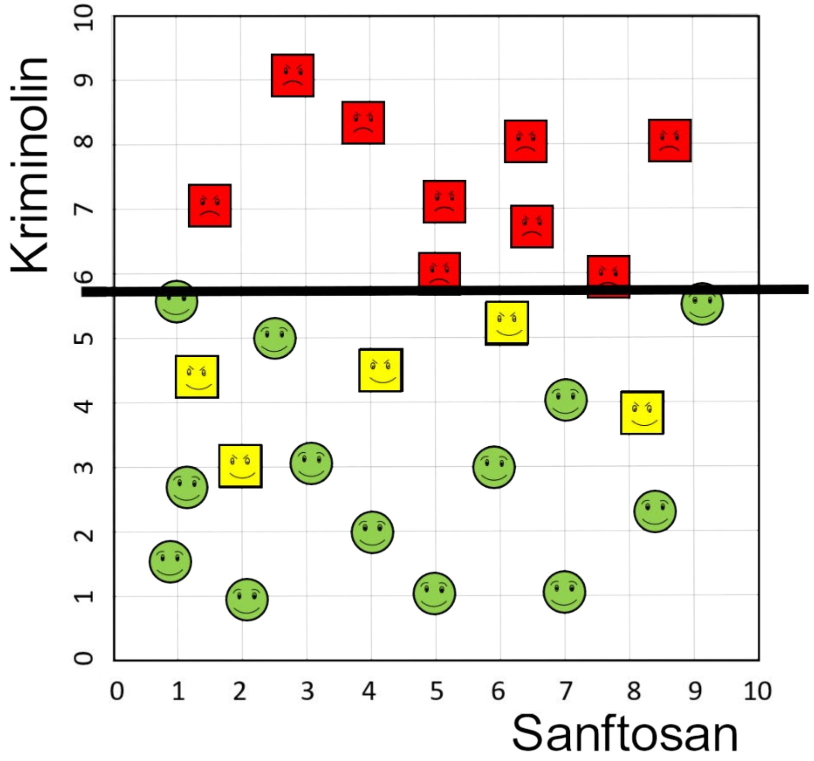
„It is better that ten guilty persons escape than that **one** innocent suffer.“

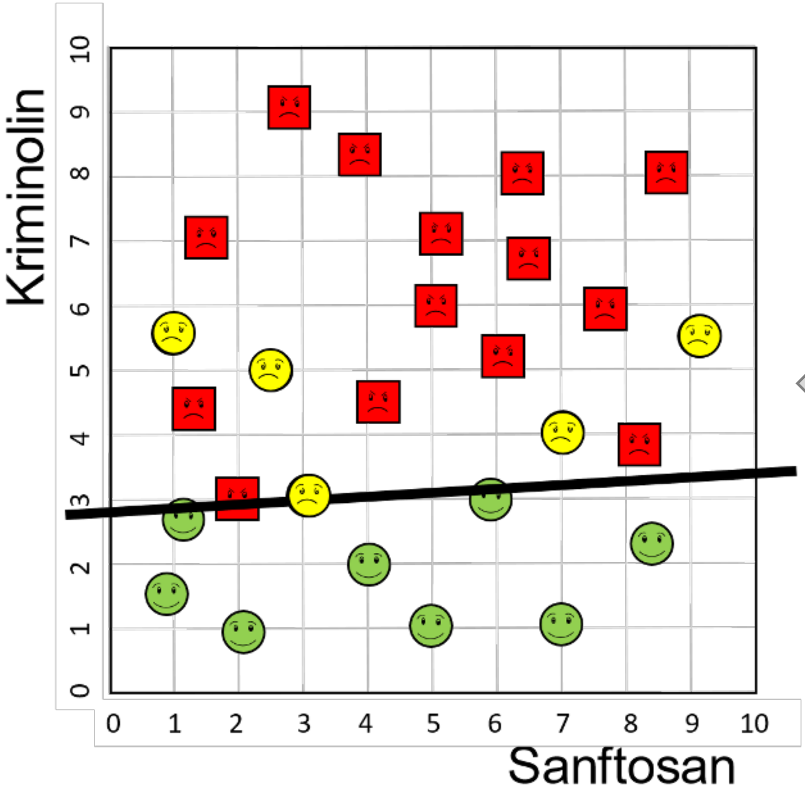
William Blackstone,
Rechtsphilosoph, 1760

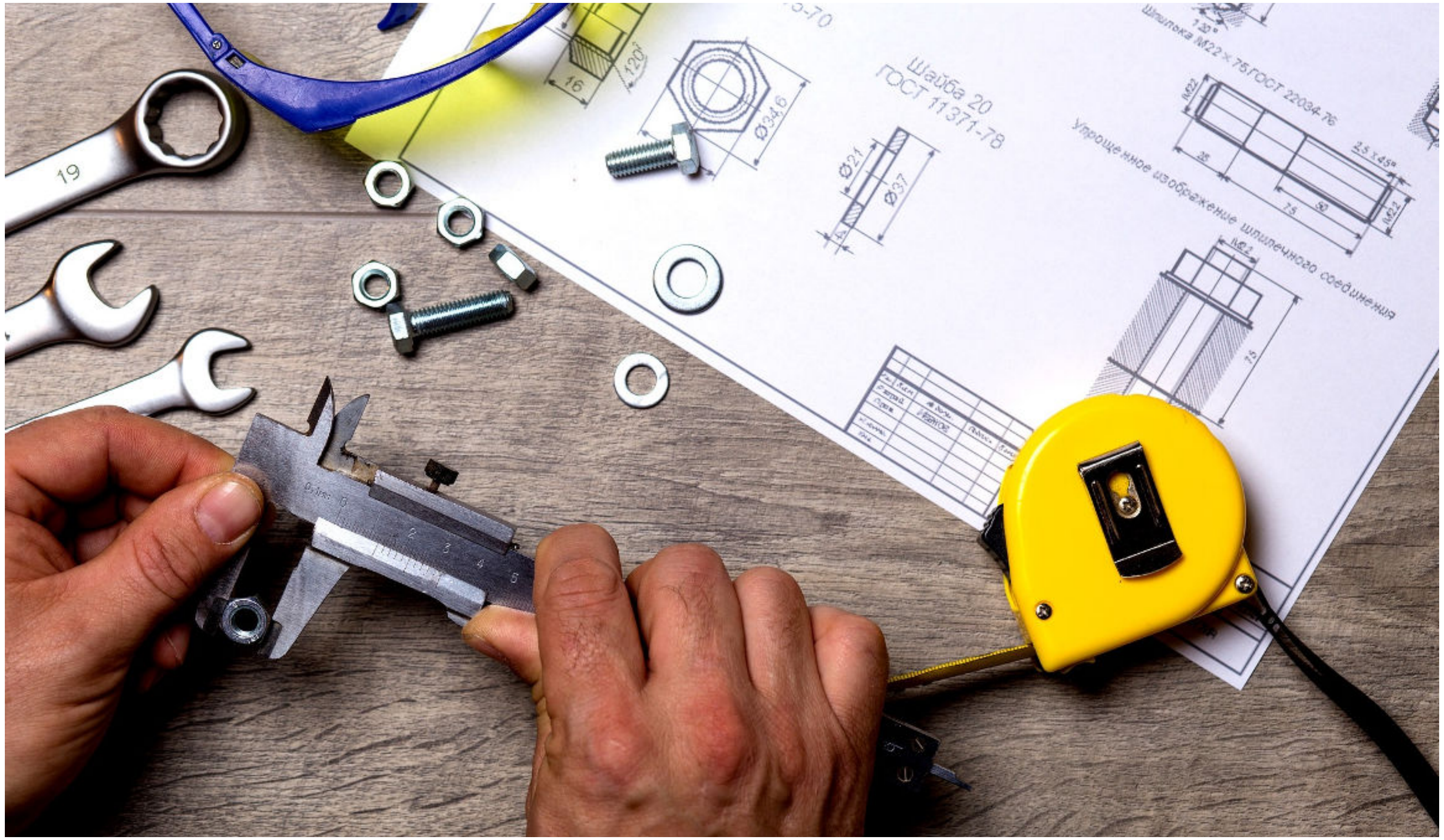


"I am more concerned with bad guys who got out and released than I am with a few that, in fact, were innocent."

Dick Cheney, ehemaliger
Vizepräsident der USA,






19
Шайба 20
ГОСТ 11371-78
Ø34,6
Ø21
Ø37
Шпилька М22×75 ГОСТ 22034-76
Упрощенное изображение шпилечного соединения

# Quality measures

- Sensitivity
- Specificity
- Accuracy
- More than 25 additional measures

## 1. Observation

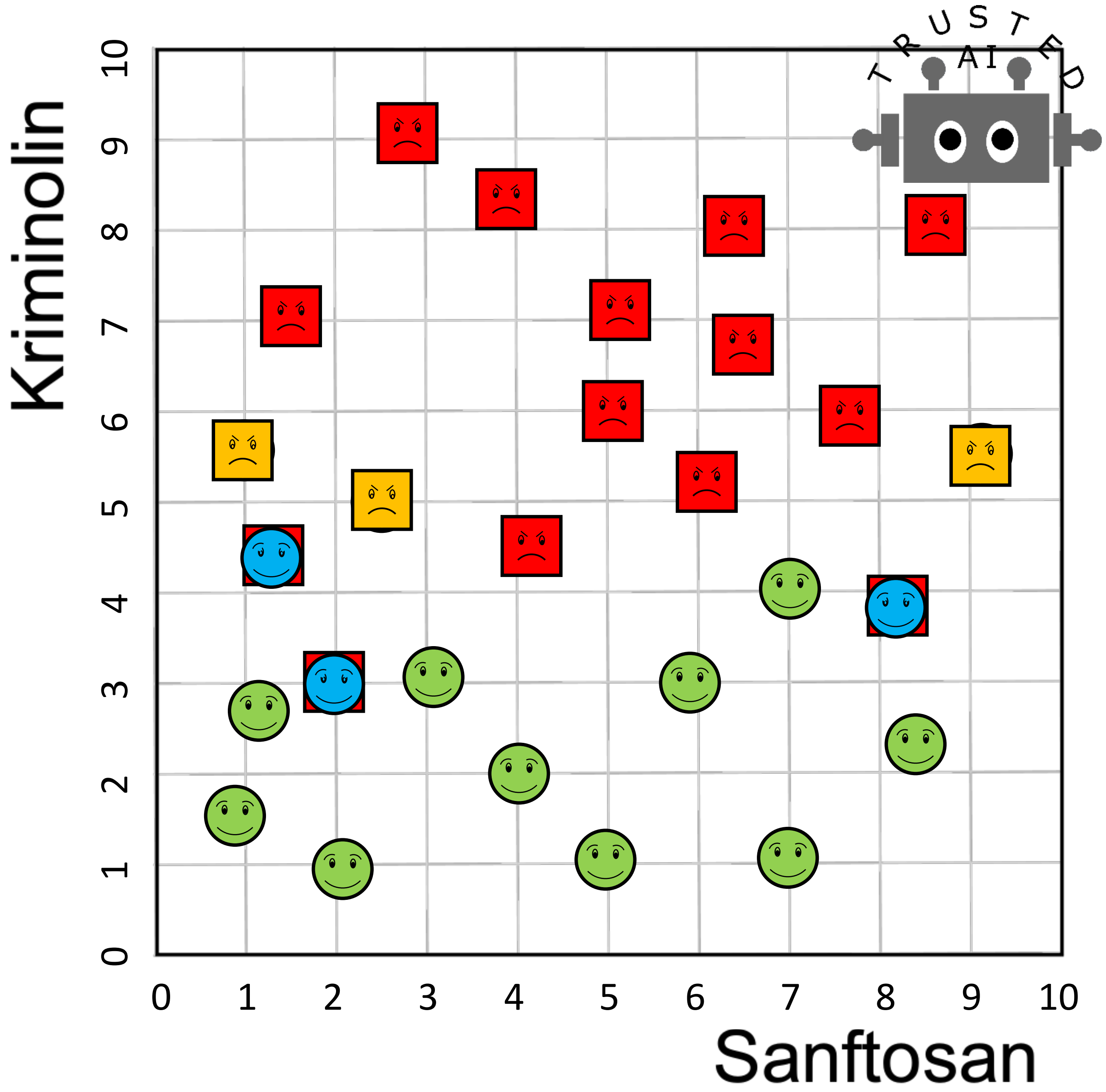
What should be optimized by an
artificial intelligence,
is a societal decision!

# Data quality

🟨 Tax fraudsters not yet detected

🔵 Innocent in prison

Incorrect data point assignments affect the training of the Support Vector Machine and thus subsequent decisions



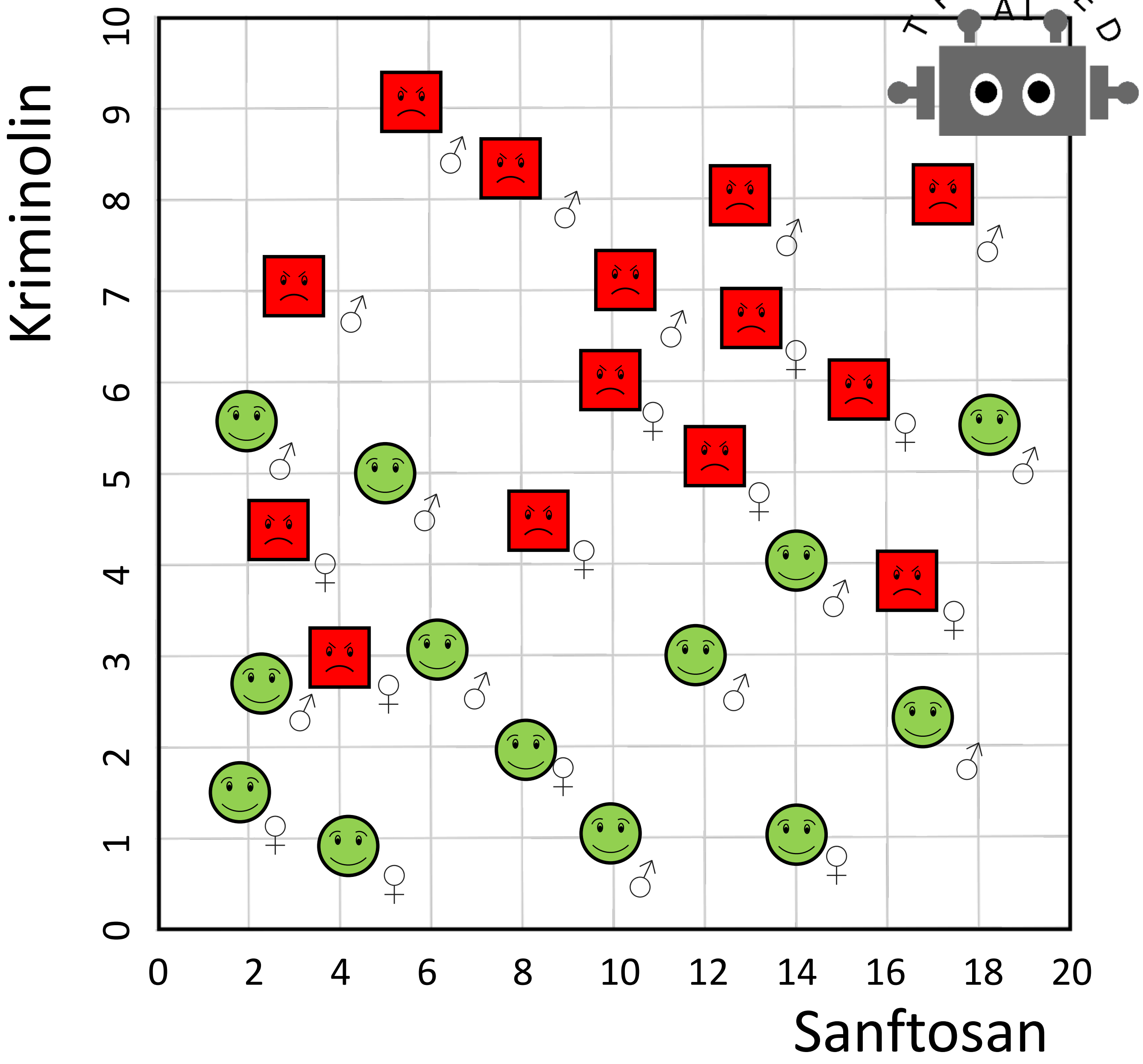
## 2. Observation

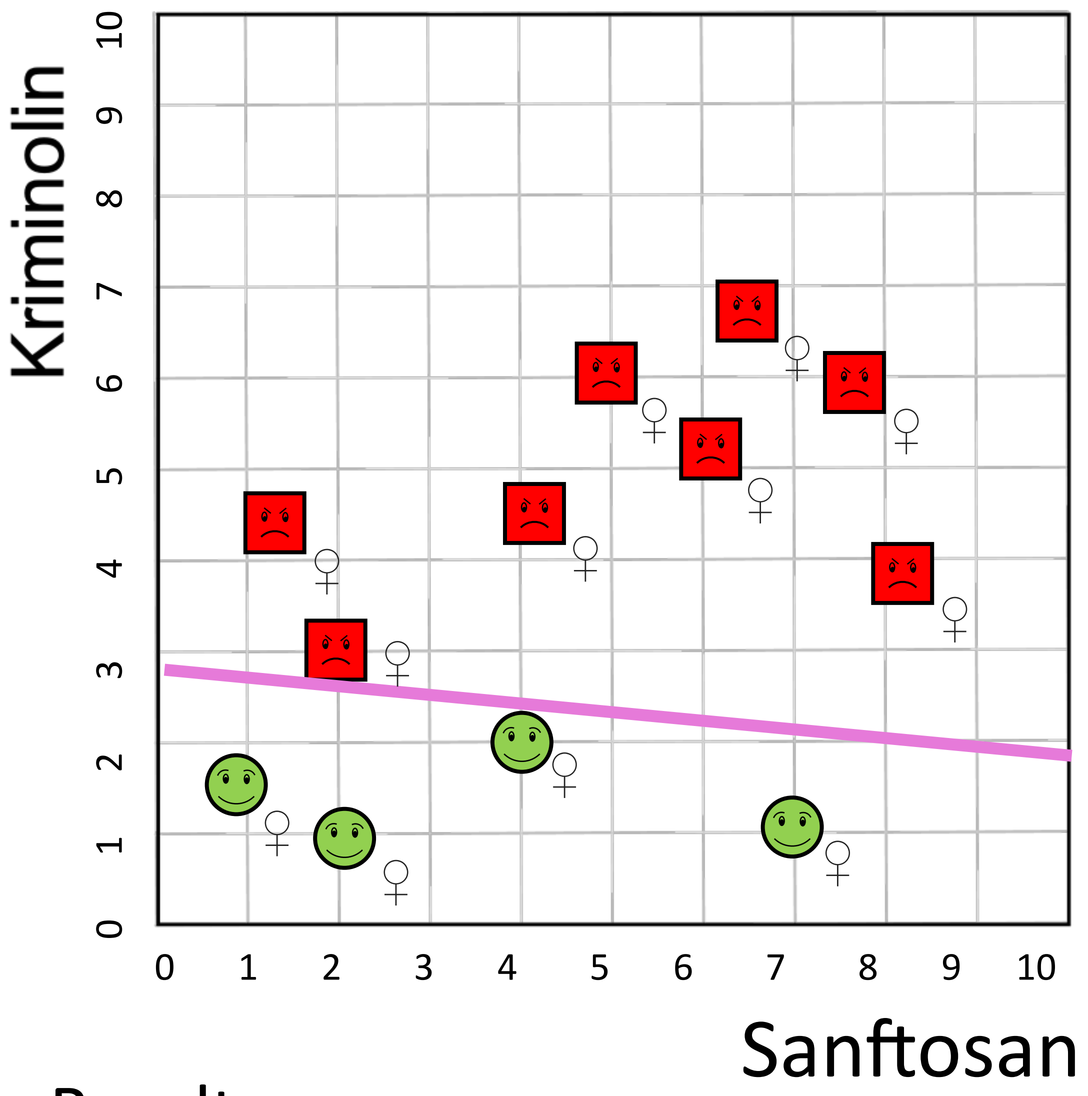
How well the machine learns is directly dependent on the quality of the data.

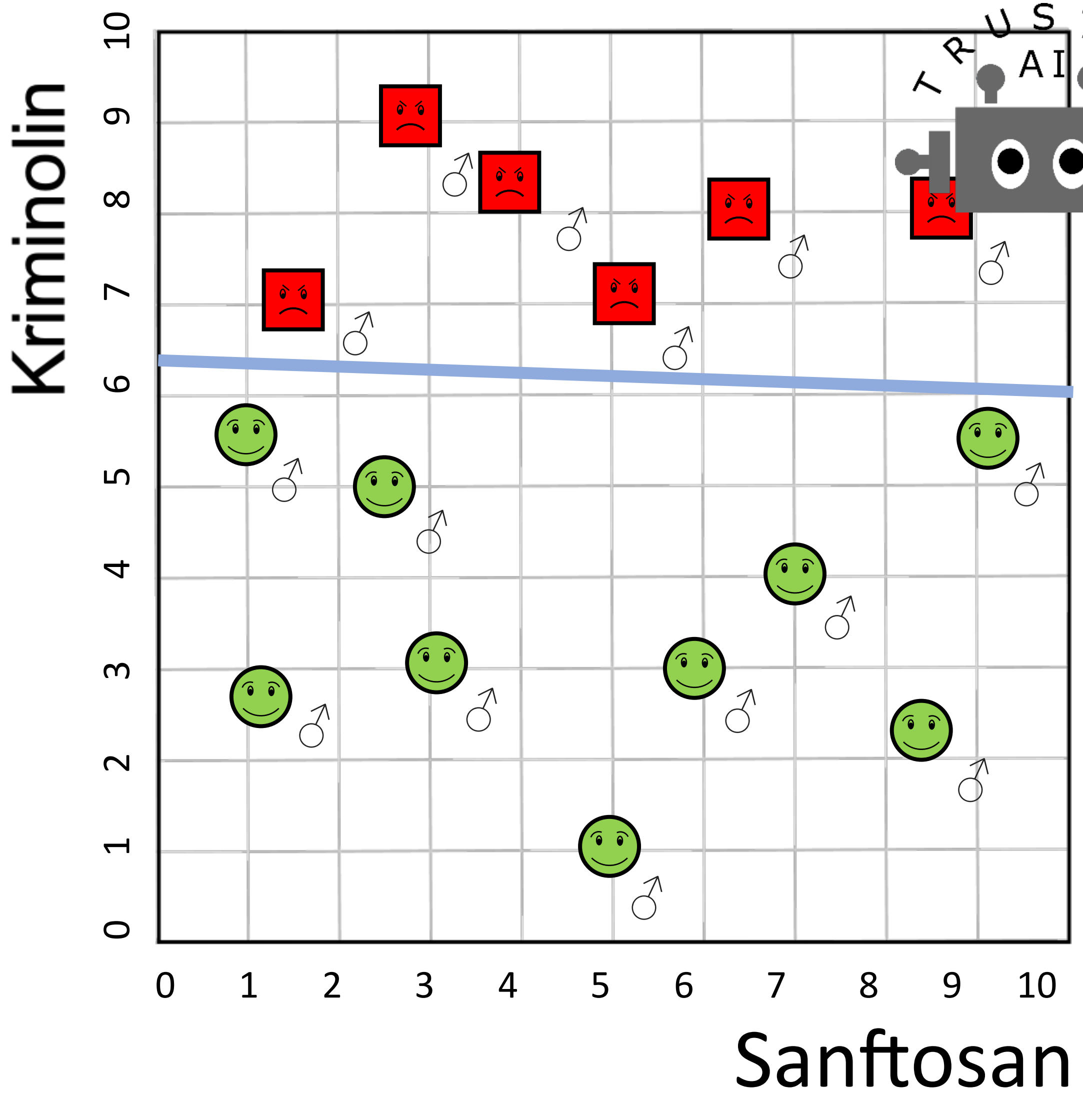
# Discrimination

🟥 Criminal

🟢 Innocent





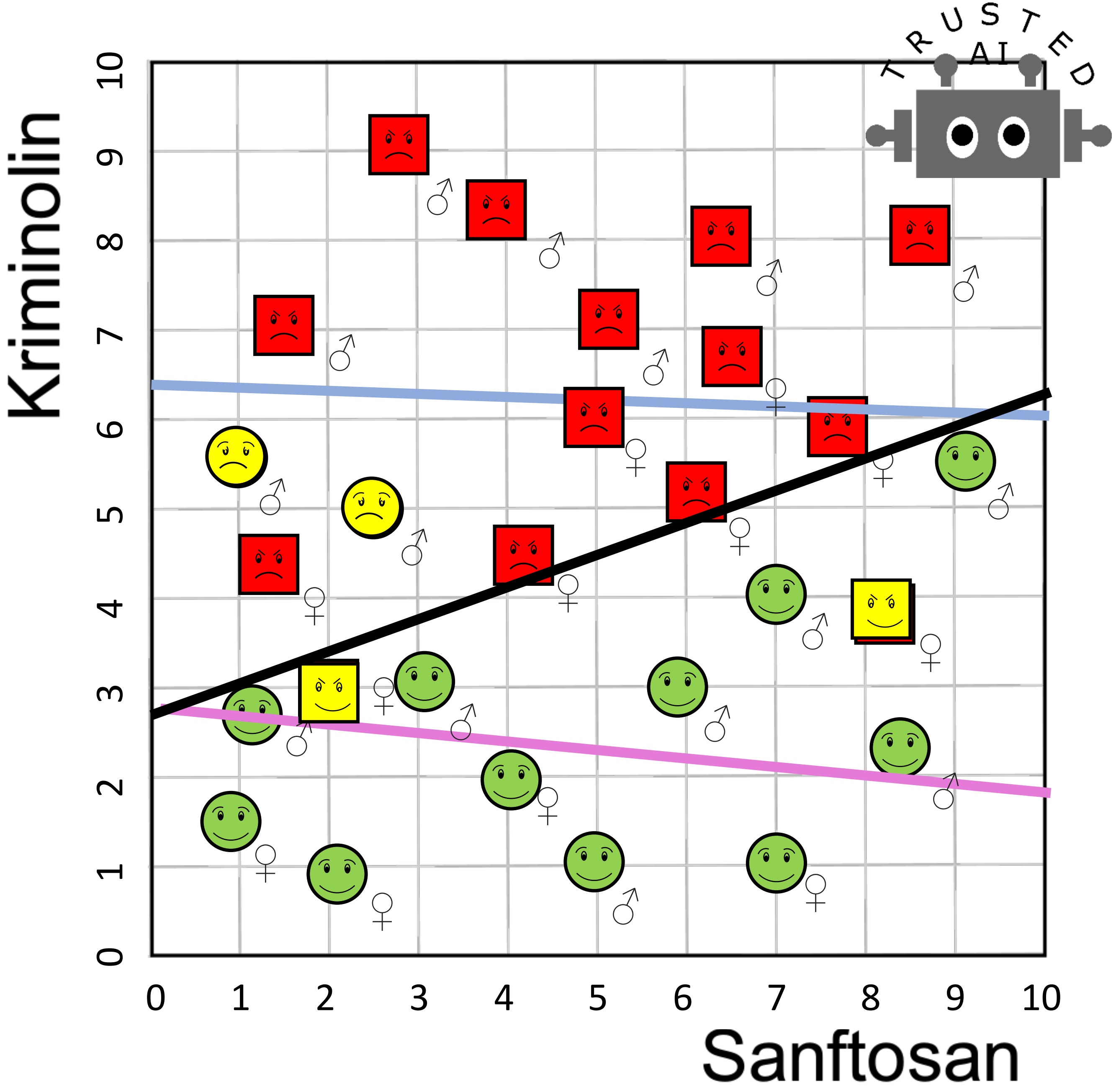


Result:

In this fictional example, an optimal decision rule without error is found for each subset.

On the other hand, if we put both groups together, the trained Support Vector Machine discriminates males :

Two female criminals are considered innocent, and two innocent male citizens are considered criminals.



## 3. Observation

Protected information can be important
in making better decisions.
Discrimination is not per se avoided
by withholding the information.

## 3. Observation (cont.)

The legally protected property may be necessary in order to make optimal decisions.

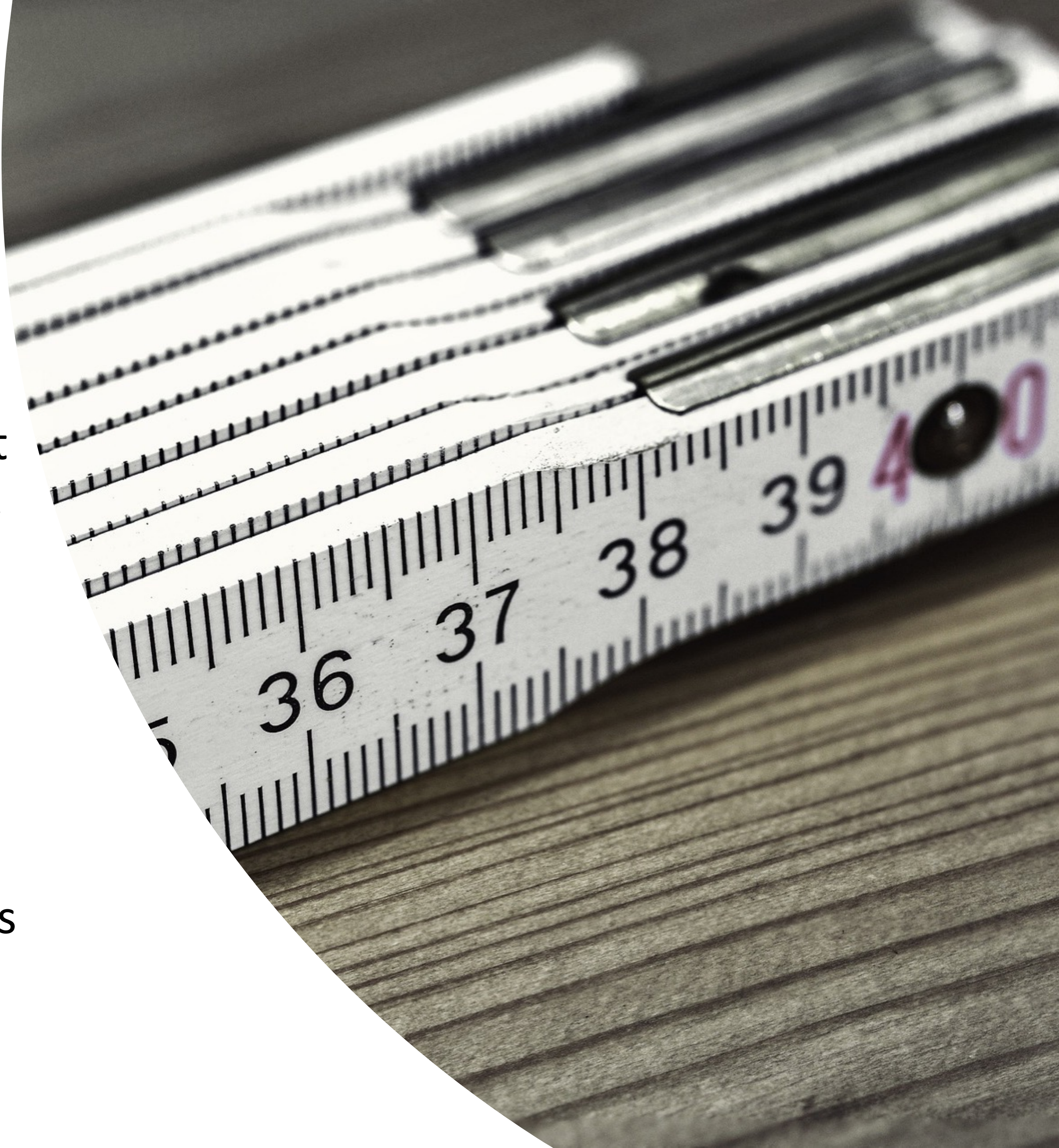
(Haeri & Zweig,2020; Hoffmann et al. 2022)

# Discrimination

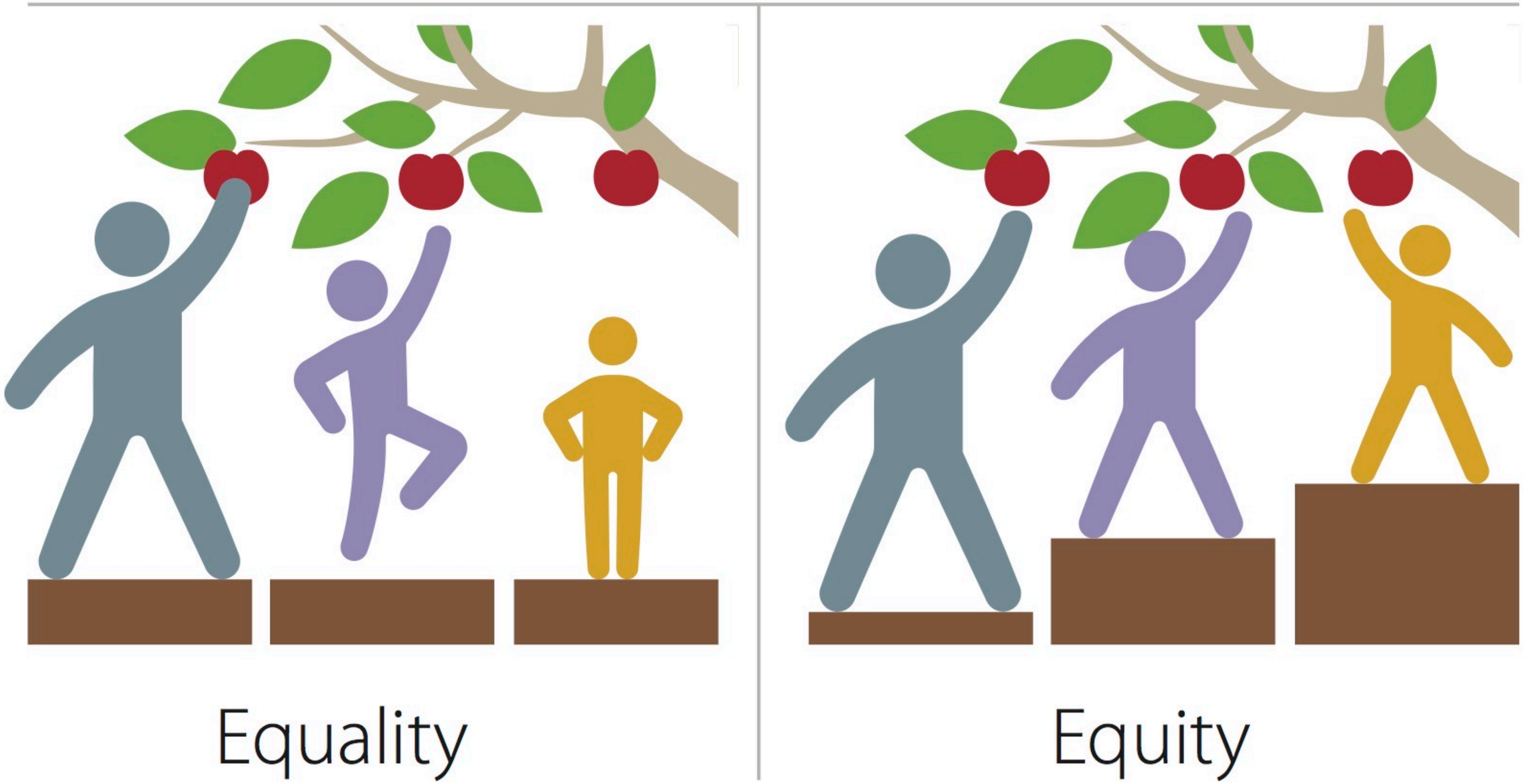
- Discriminations in training data are "learned along".
- If training data contains too little data about minorities, their properties will not be "learned along".

# Measuring discrimination

- Using fairness measure(s)
- Require (statistical) equality quality for subgroups.
  - Buolamwini: Subgroups should at least have 80% of maximum values (Buolamwini, 2017, S.49).
- Sensitive information is required for testing ↔ Data protection!
- Attention: Most fairness measures contradict each other (Zweig & Krafft, 2018).
  - There is no simple solution → societal decision (selection might even requires democratic legitimacy in important cases.



# Anyone always loses



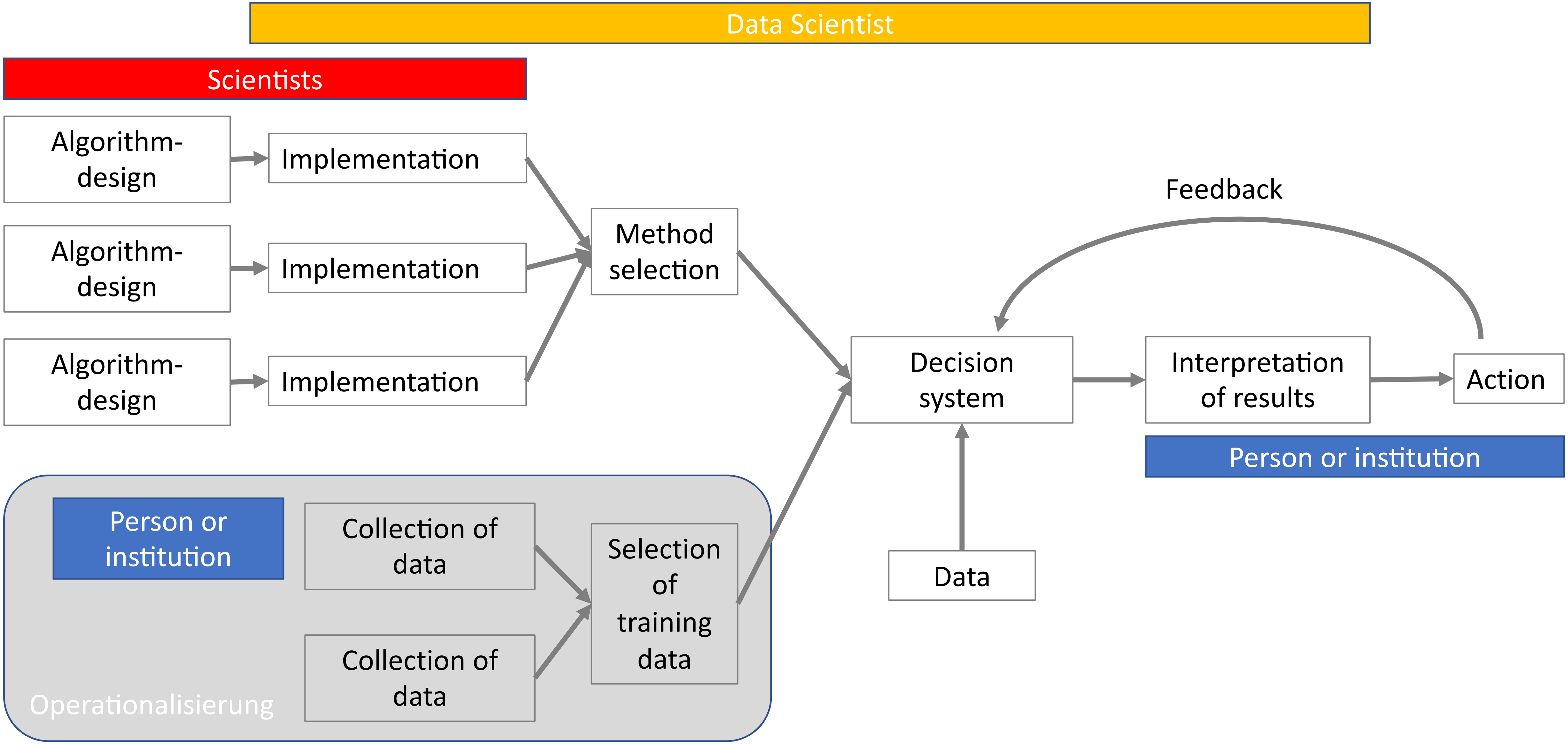
By: MPCA Photos
https://www.flickr.com/photos/mpcaphotos/31655988501
https://creativecommons.org/licenses/by-nc/2.0/

## 4. Observation

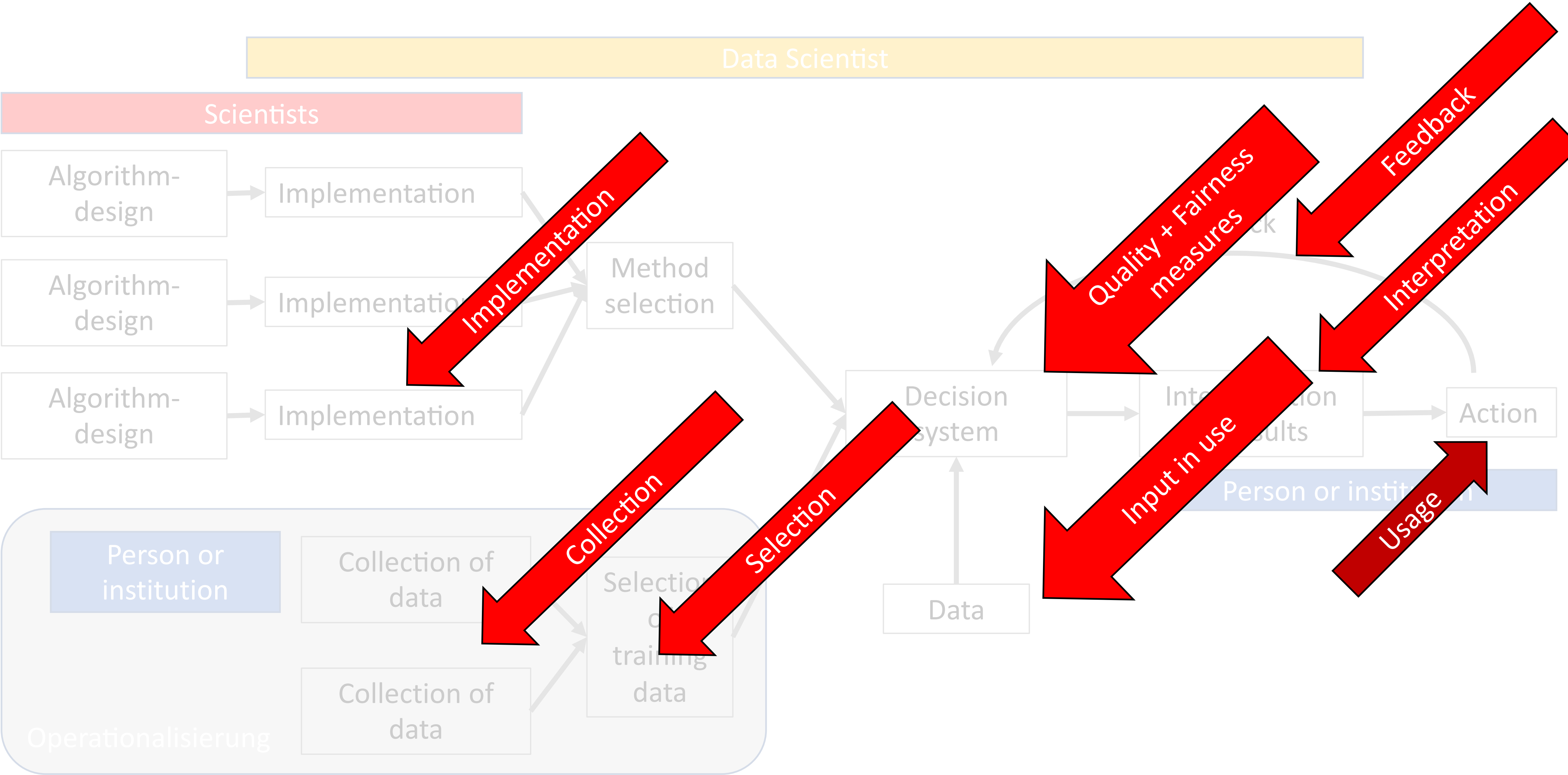
What "fair" means is
a societal decision,
but can also be shaped
by corporate philosophy.

# Who is responsible?

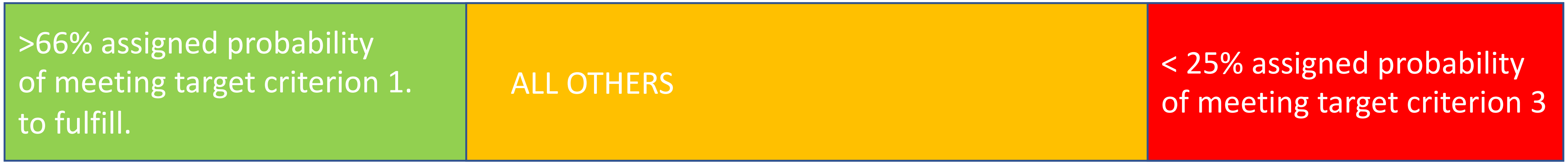
# Long chain of responsibilities



# Where can discrimination be introduced?



# Discrimination depends on the exact usage



>66% assigned probability of meeting target criterion 1. to fulfill.

ALL OTHERS

< 25% assigned probability of meeting target criterion 3

- Divides unemployed into 3 classes:
  - High chances of integration - no further measures needed.
  - Medium chances of integration - with measures.
  - Low chances of integration - measures not useful.

# Result:

- Assigns higher risk to the elderly (>50), women, caregivers.
- **Discrimination?**
- **Depends on the usage!**

# Fair usage?

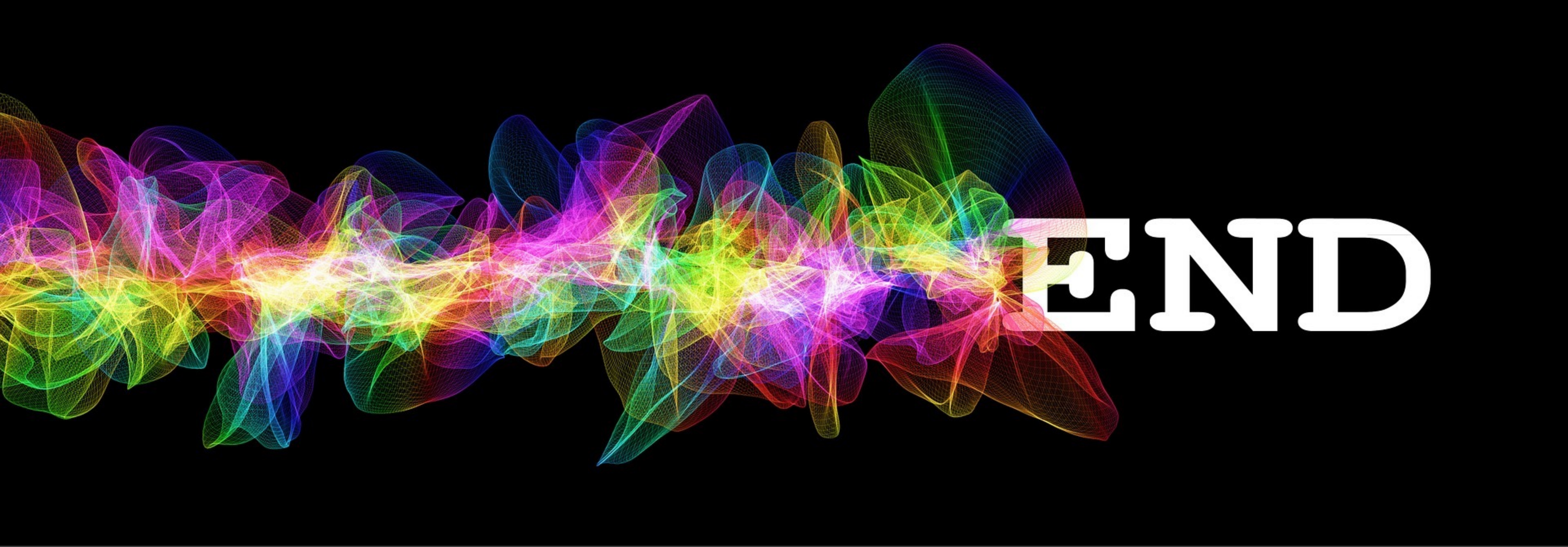
- The system is used to balance against societal discriminination
- The overall system can only have a balancing effect if the ADM system reflects actual discriminates.
- According to the AMAS director, people disadvantaged by the labor market are more often supported now [1].

[1] https://www.johanneskopf.at/2019/09/24/offener-brief-fr-prof/

# Important: Social compatibility rules ("Sozialverträglichkeitsr egeln")

- Classification must be discussed with citizen in dialogue.
- Only supportive use.
- Recalculated every year.
- Only data from the last 4 years.

Gamper, Kernbeiß & Wagner-Pinter: „Das Assistenzsystem AMAS – Zweck, Grundlagen, Anwendung (Dokumentation), Mai 2020, http://www.forschungsnetzwerk.at/downloadpub/2020_Assistenzsystem_AMAS-dokumentation.pdf



# Diskussion

# Literaturverzeichnis

Buolamwini, J. A. (2017). Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers (Doctoral dissertation) Massachusetts Institute of Technology.

Haeri, M. A., & Zweig, K. A. (2020, December). The Crucial Role of Sensitive Attributes in Fair Classification. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 2993-3002). IEEE.

Hanna Hoffmann, Verena Vogt, Marc P. Hauer, Katharina Zweig (2022). Fairness by awareness? On the inclusion of protected features in algorithmic decisions. In: Computer Law & Security Review. Elsevir.

Kurzweil R. (1990) The Age of Intelligent Machines. Cambridge, Mass: MIT Press.

Zweig, K. A., & Krafft, T. D. (2018). Fairness und Qualität algorithmischer Entscheidungen. *57518*, 204-227 in *(Un)Berechenbar?,* Fraunhofer FOKUS, Kompetenzzentrum ÖFIT.