Ethics and Accountability of Algorithmic Decision Making Systems

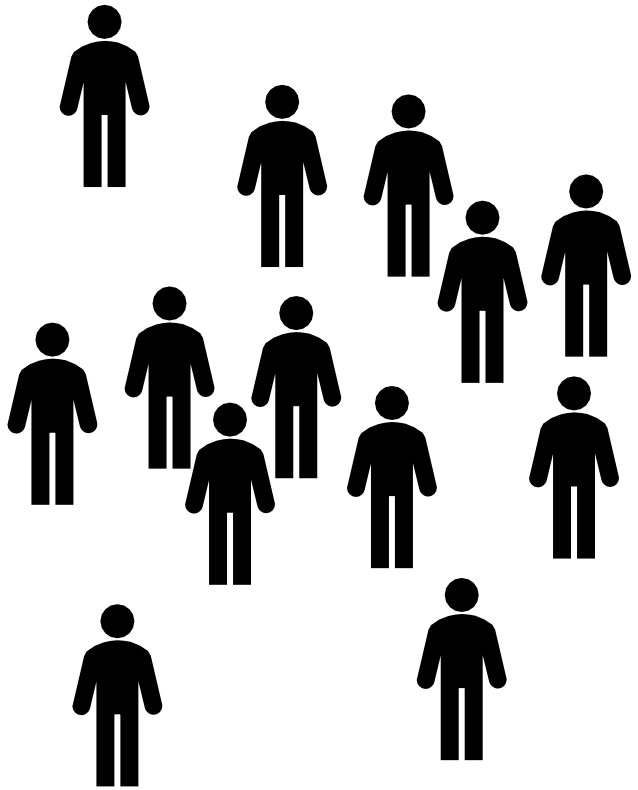
Prof. Dr. K.A. Zweig
TU Kaiserslautern
Algorithm Accountability Lab
@nettwerkerin

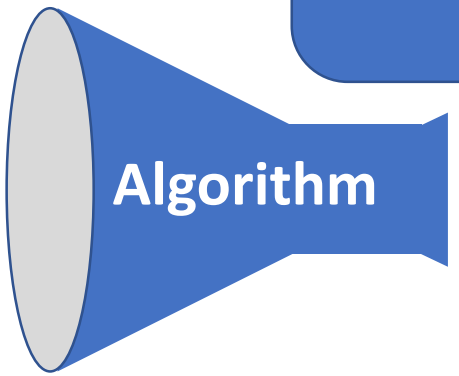
AI raises fears

AI will create poetry…

… and penalty

Algorithmic Decision Making Systems (ADM)



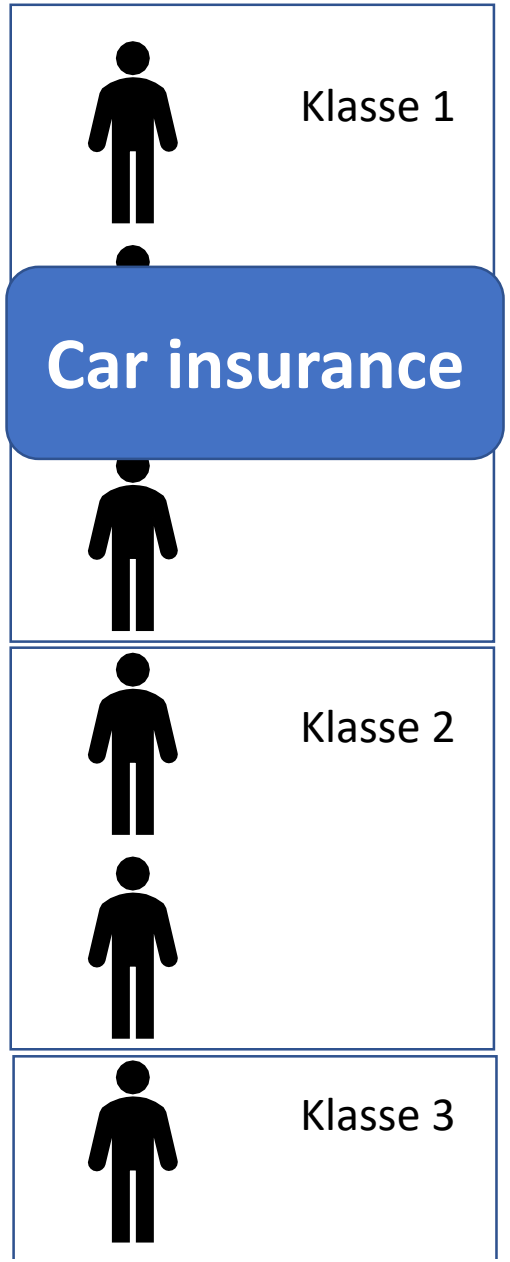


SCHUFA (credit worthiness)

Scoring

or

Car insurance



Classification

Who should judge humans?

Menschenbild

conception of man

Are humans the best option to judge other humans?

Mankind – so irrational!

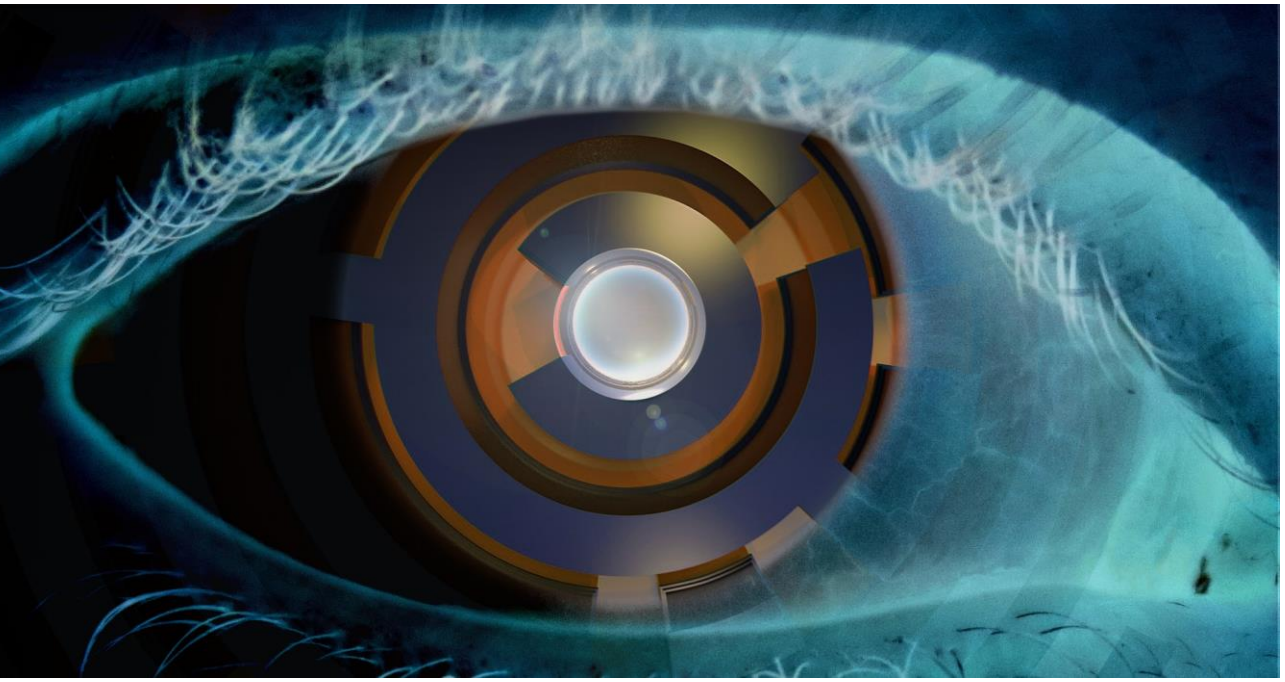
- Study: Judges have to review prison release proposals regularly.
- Shown: Time from last break reduced likelihood for a positive decision[1].
- Many more studies seem to show:
  - Humans are irrational and biased.

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

Problematic situation in the USA

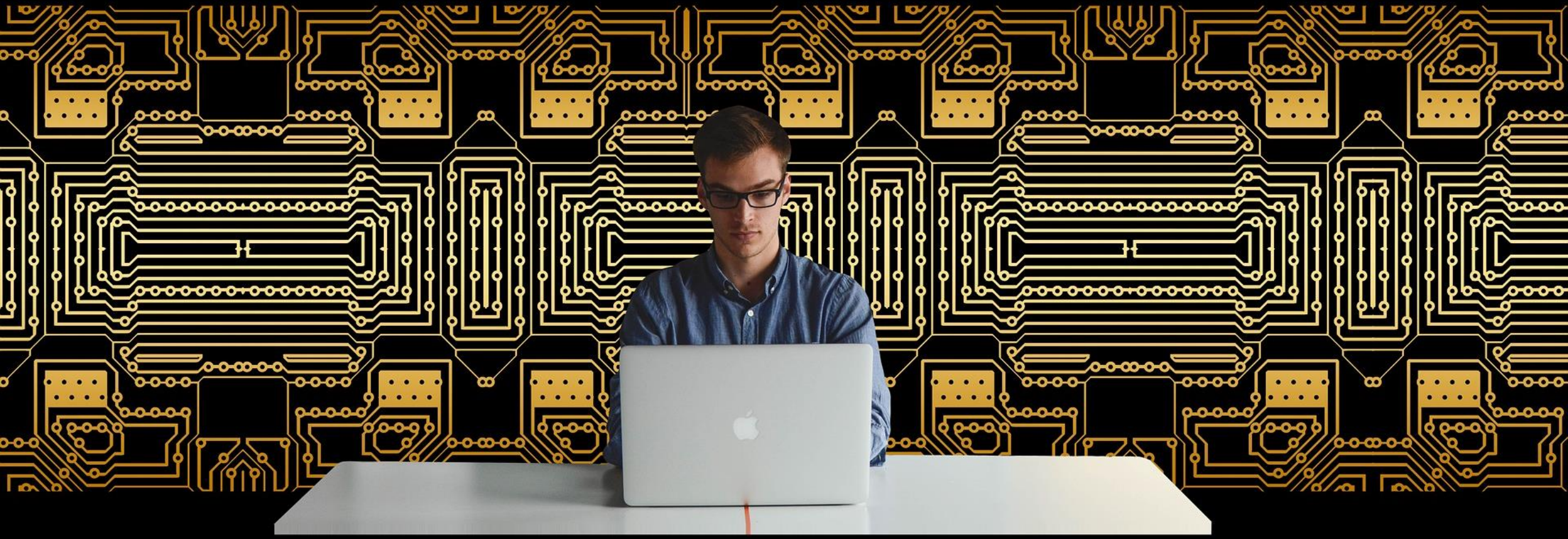
- Second highest incarceration rate worldwide.
- 6x higher rate of Afroamericans und 2x more of Latinos.
- Dramatic prognosis: every third boy at the age of 10 now will be in prison at least once in his life.

American Civil Liberties Union

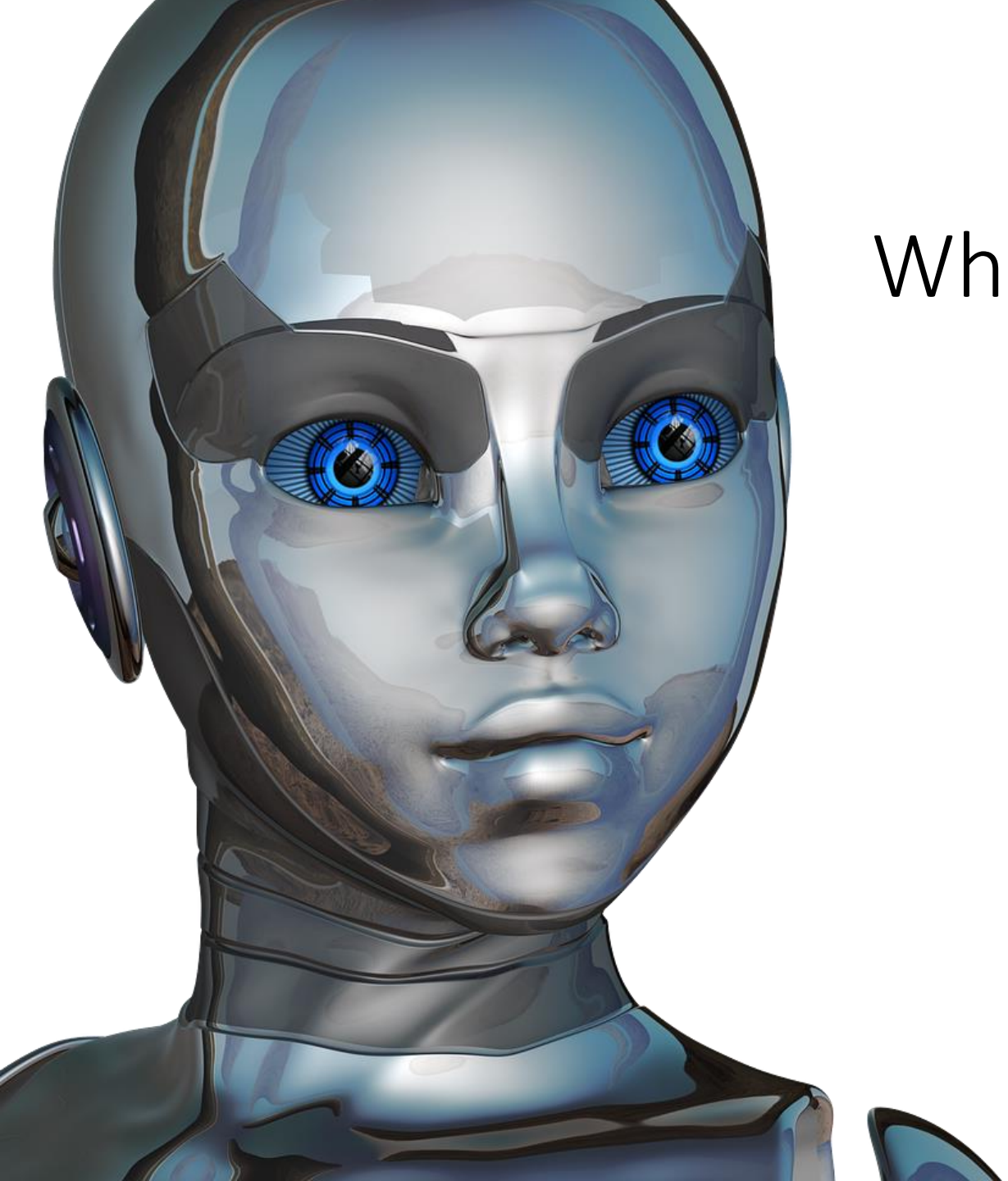


- American civil liberties states:
- ADM systems need to be used in all stages of the leagl process, …
- … to ensure fairness and objectivity.
- They propose that computer should learn the necessary decision rules from data.

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



Can computers learn?



What is learning?

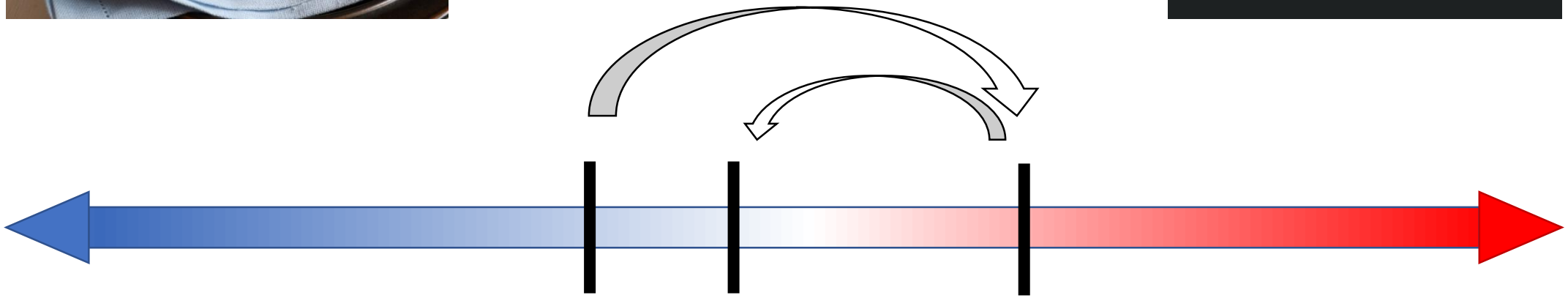
Simple:

Repeat some learned behavior in some defined situation.

Generalized:

Choose the correct behavior from a range of possibilities in the same kind of situation.

Sebastian learns „hot“ and „warm“



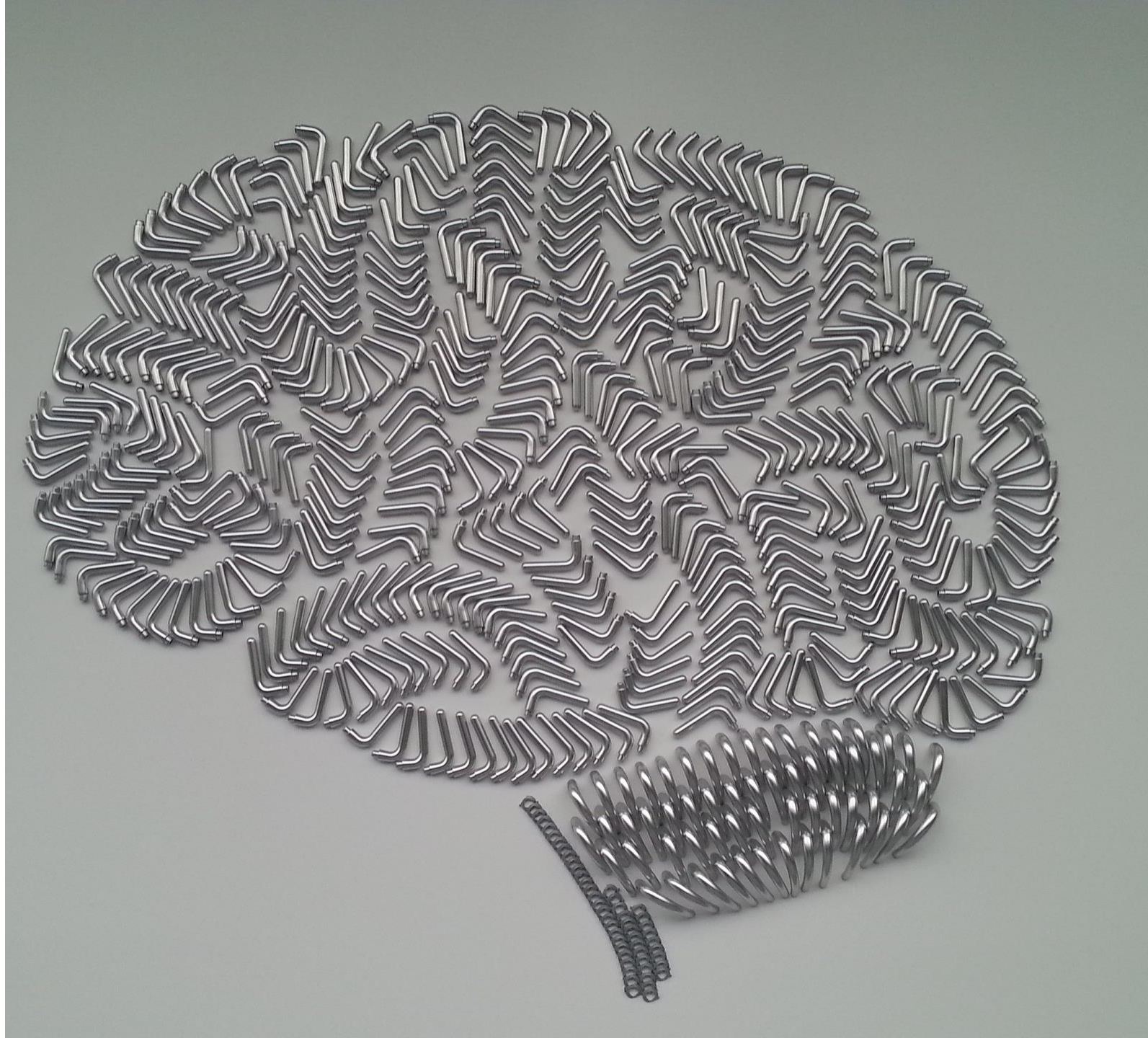
Too cautious, eats only cold meals

No steam, please

Too daring

Sebastian learns…

- By **feedback**: unexpectedly hot, unexpectedly cold
- By **saving rules in some structure**: in neurons and their connections.
- By **many data points** (experiences)**.**
- By **generalizing the learned rules**.

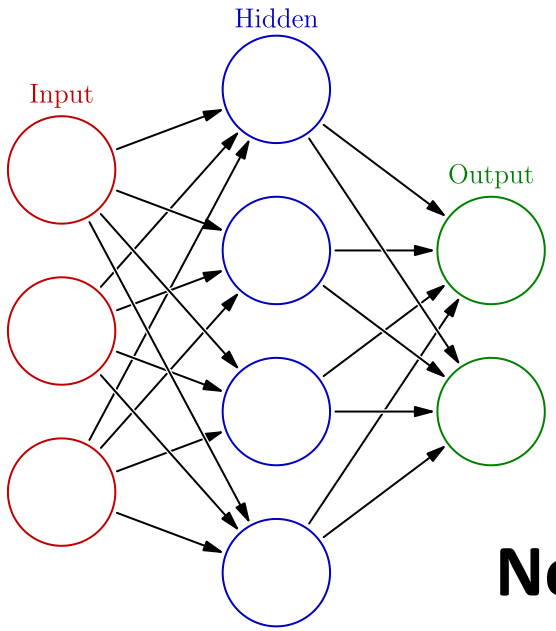


Computers learn..

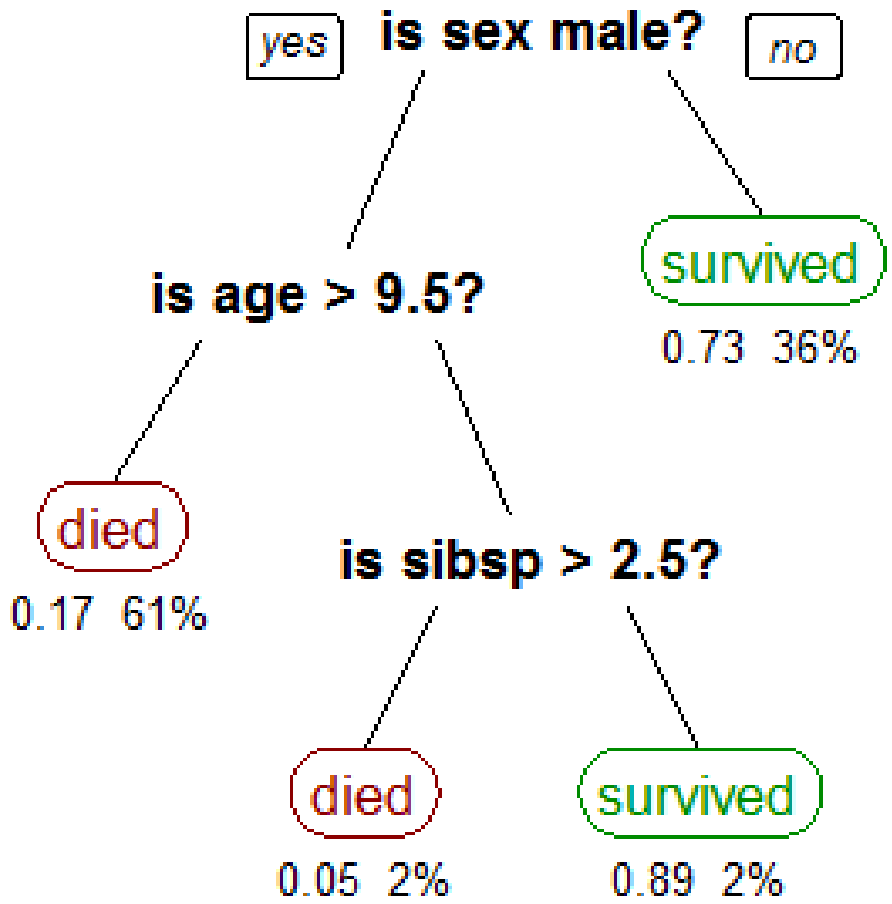
By giving them **a structure** for saving learned rules.

By giving them **feedback**.

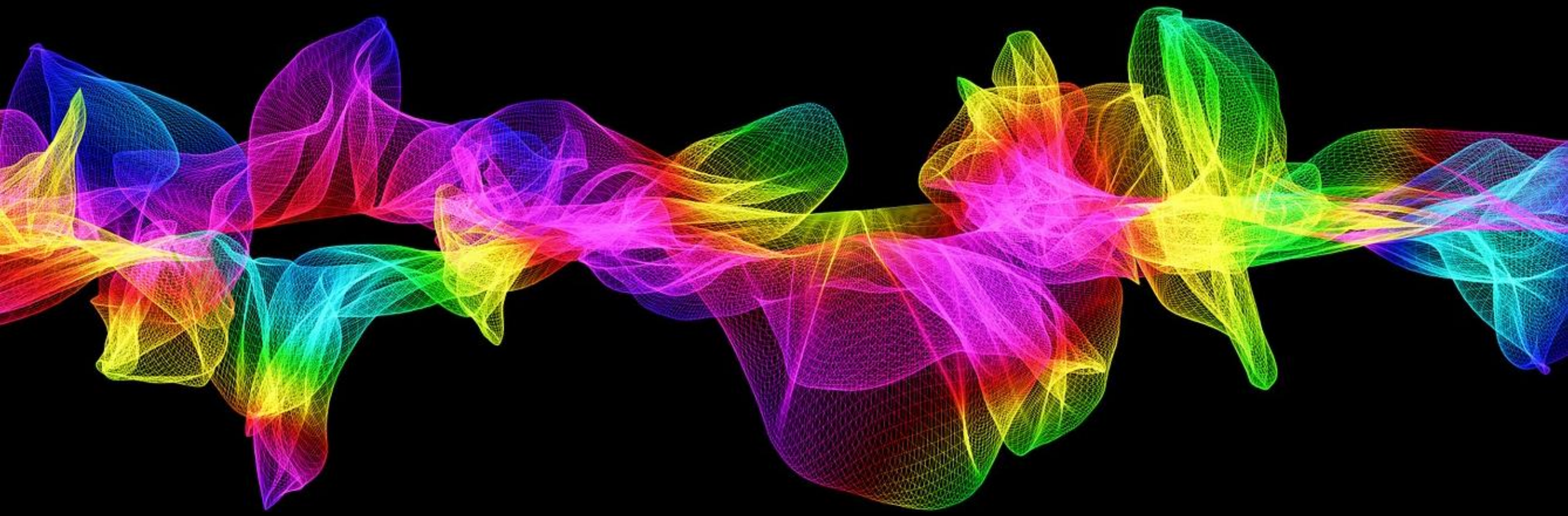
By learning .



Neural Net



Decision trees

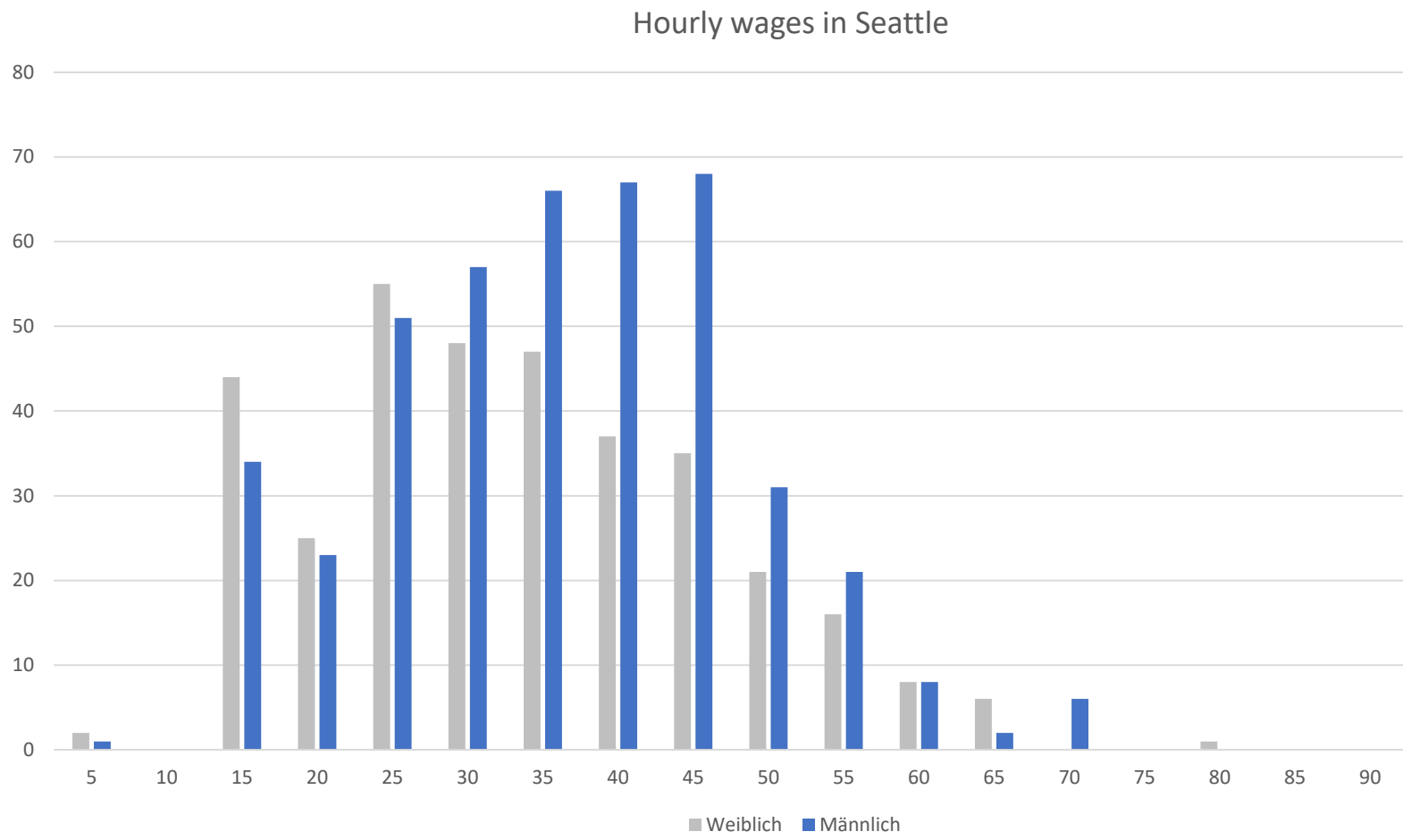


“Learn” from correlations

Wages in Seattle

You have to welcome a new employee. Is it Mr. or Ms Miller?

You know that the person gets less than $25/h. Is it rather Mr. or Ms Miller?

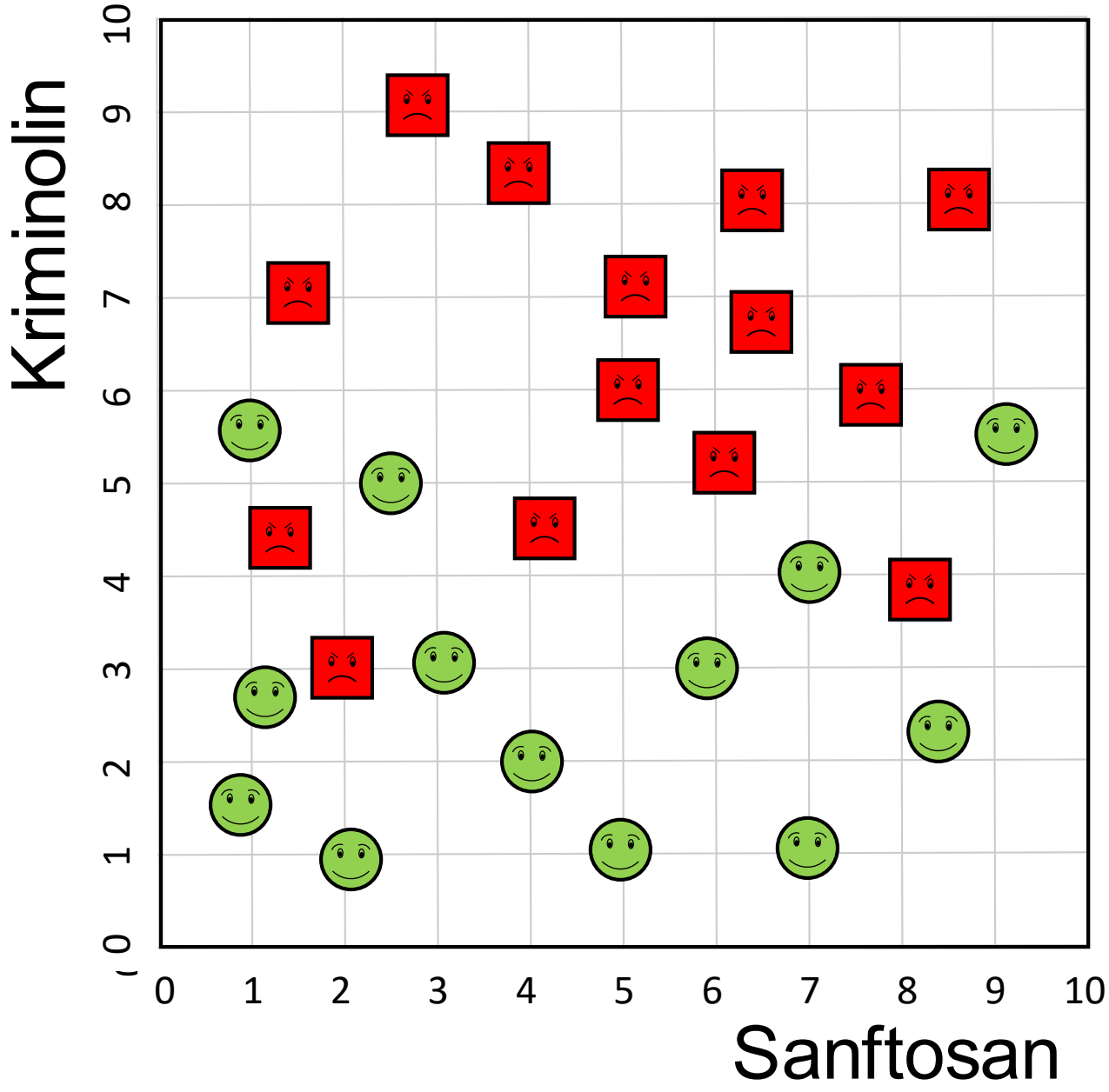


@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

“Learning” with SVMs

Aggressive criminals

Innocent citizens

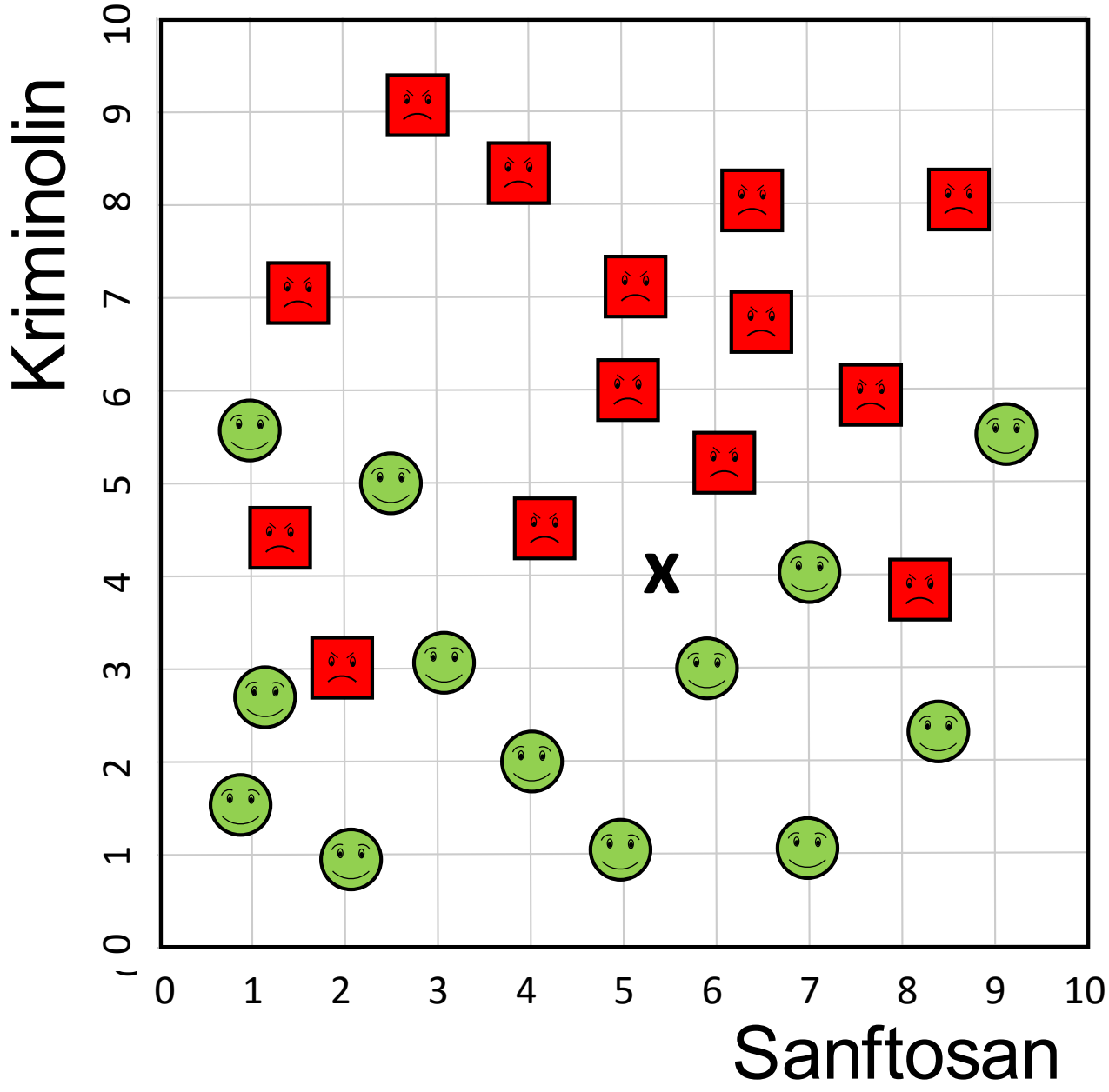


Aggressive criminals

Innocent citizens

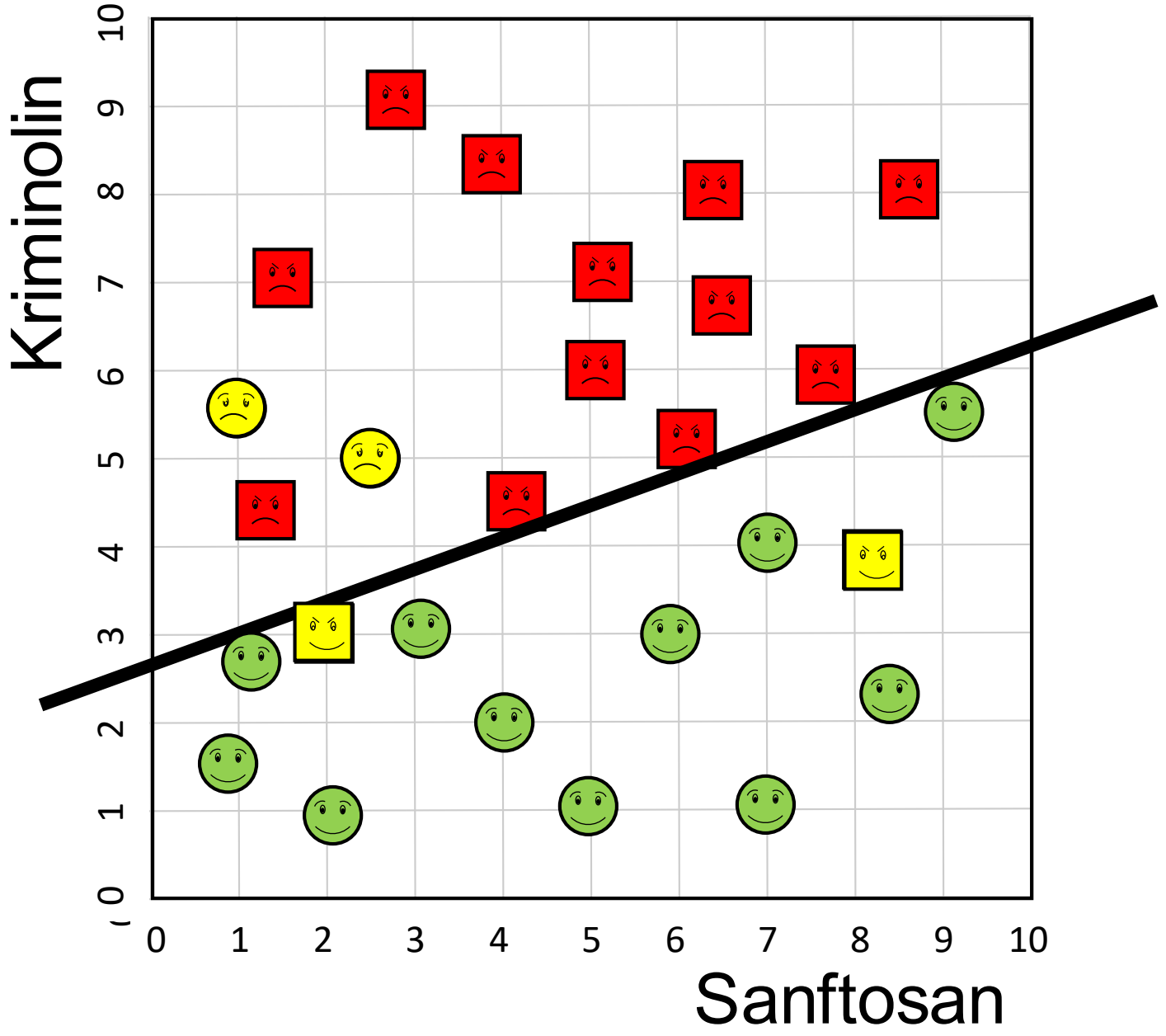
What do you think of Ms Miller?

5.5 Sanftosan
4.0 Kriminolin



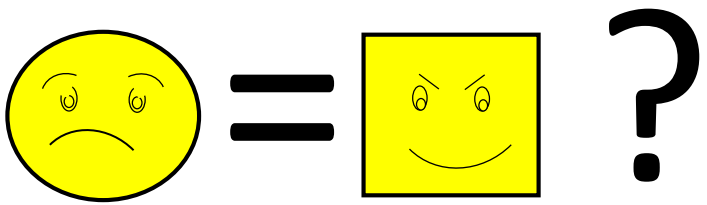
Aggressive criminals

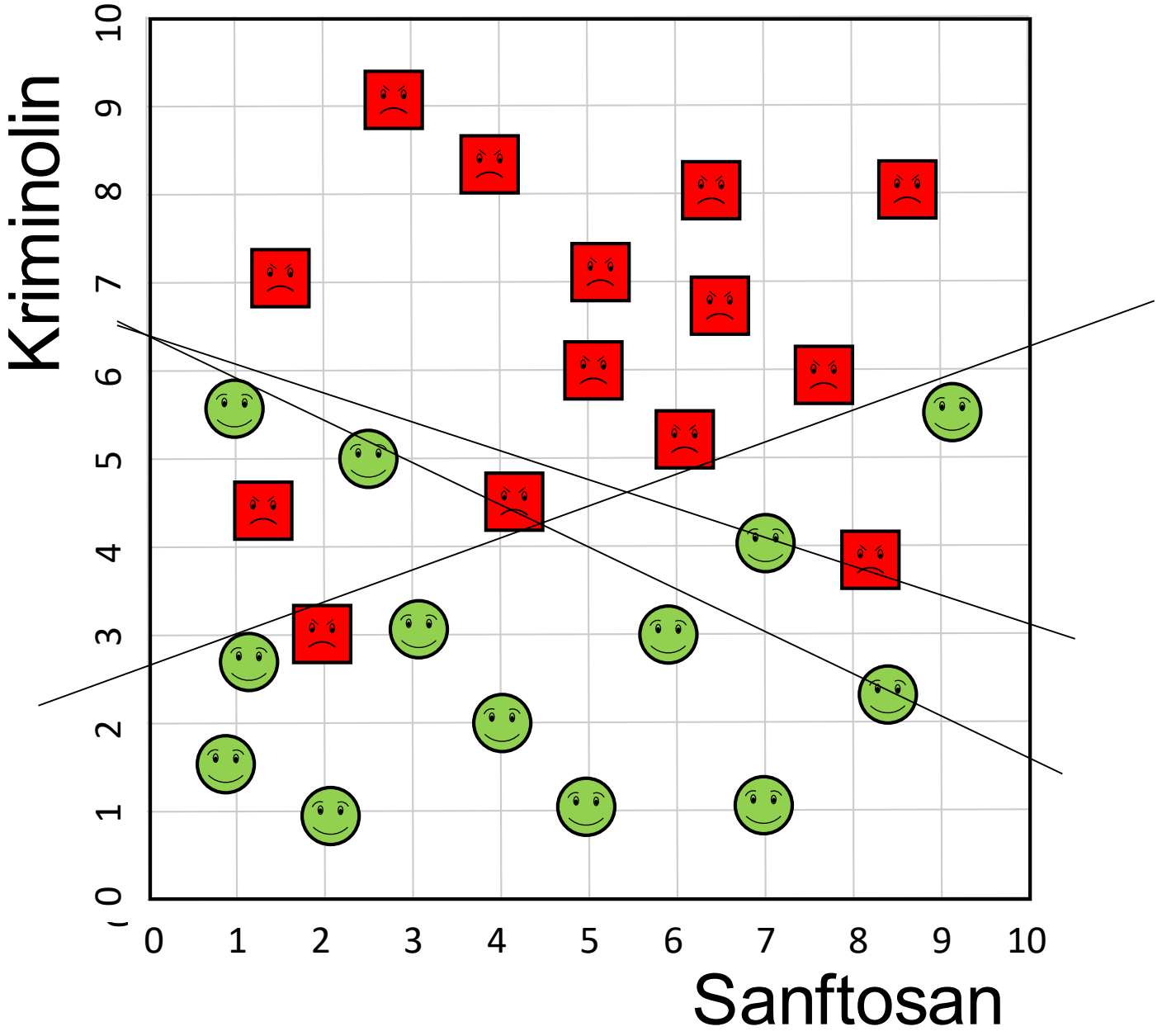
Innocent citizens



Aggressive criminals

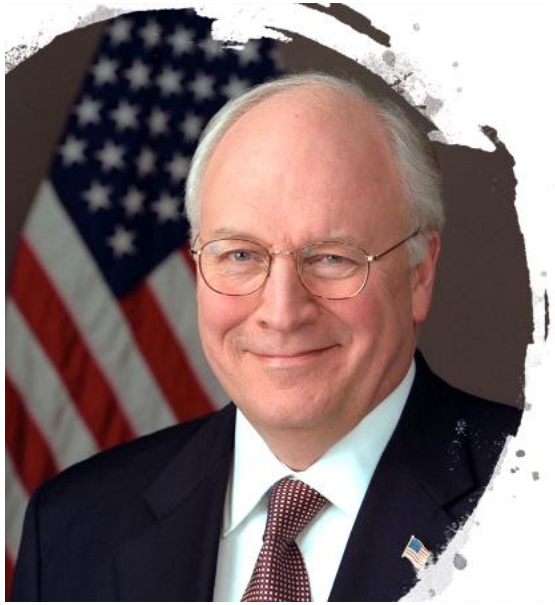
Innocent citizens





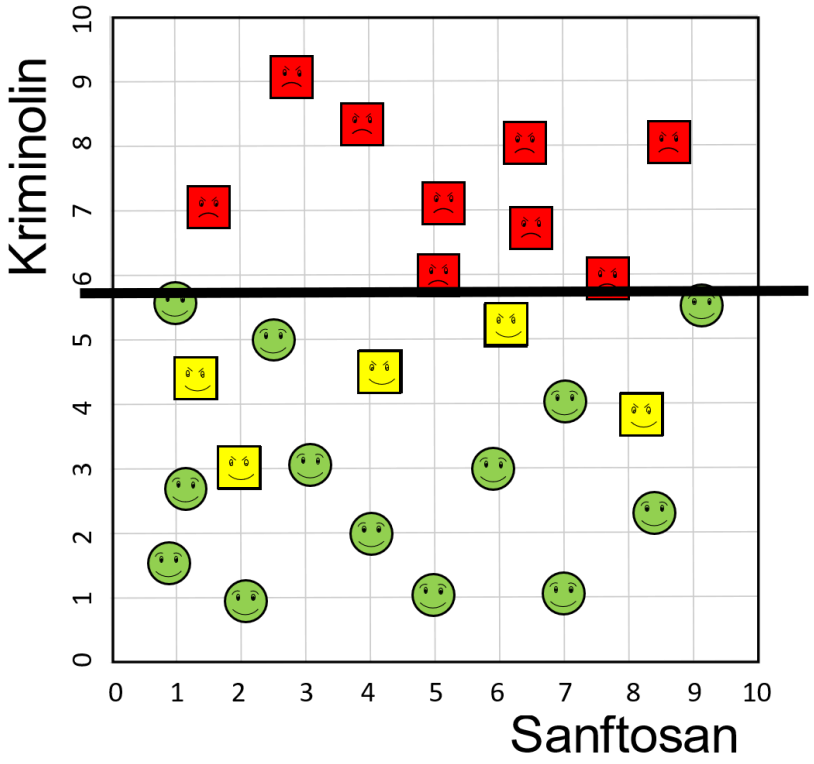
„It is better that ten guilty persons escape than that **one** innocent suffer.“

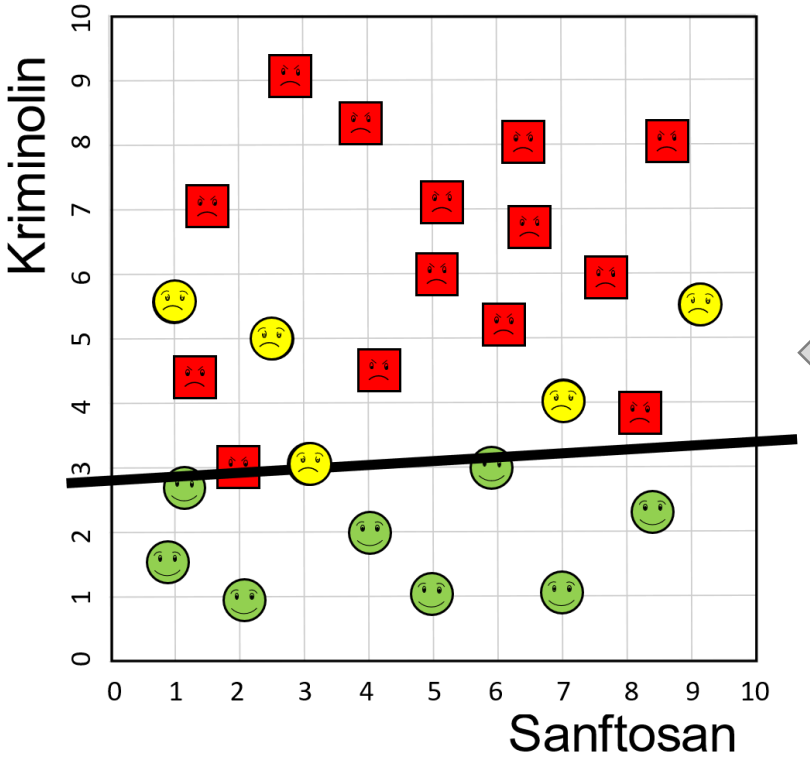
William Blackstone,
Rechtsphilosoph, 1760

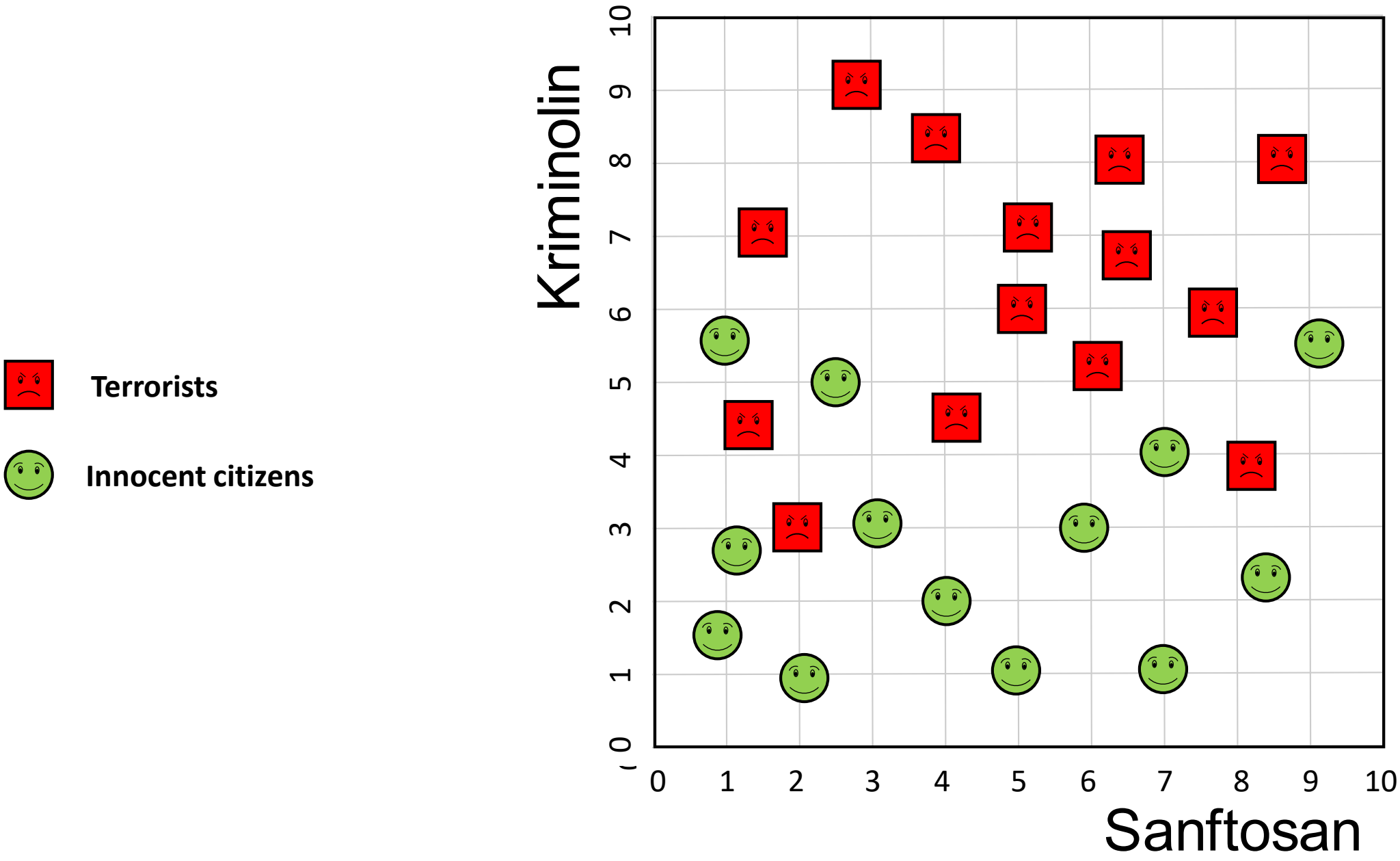


"I am more concerned with bad guys who got out and released than I am with a few that, in fact, were innocent."

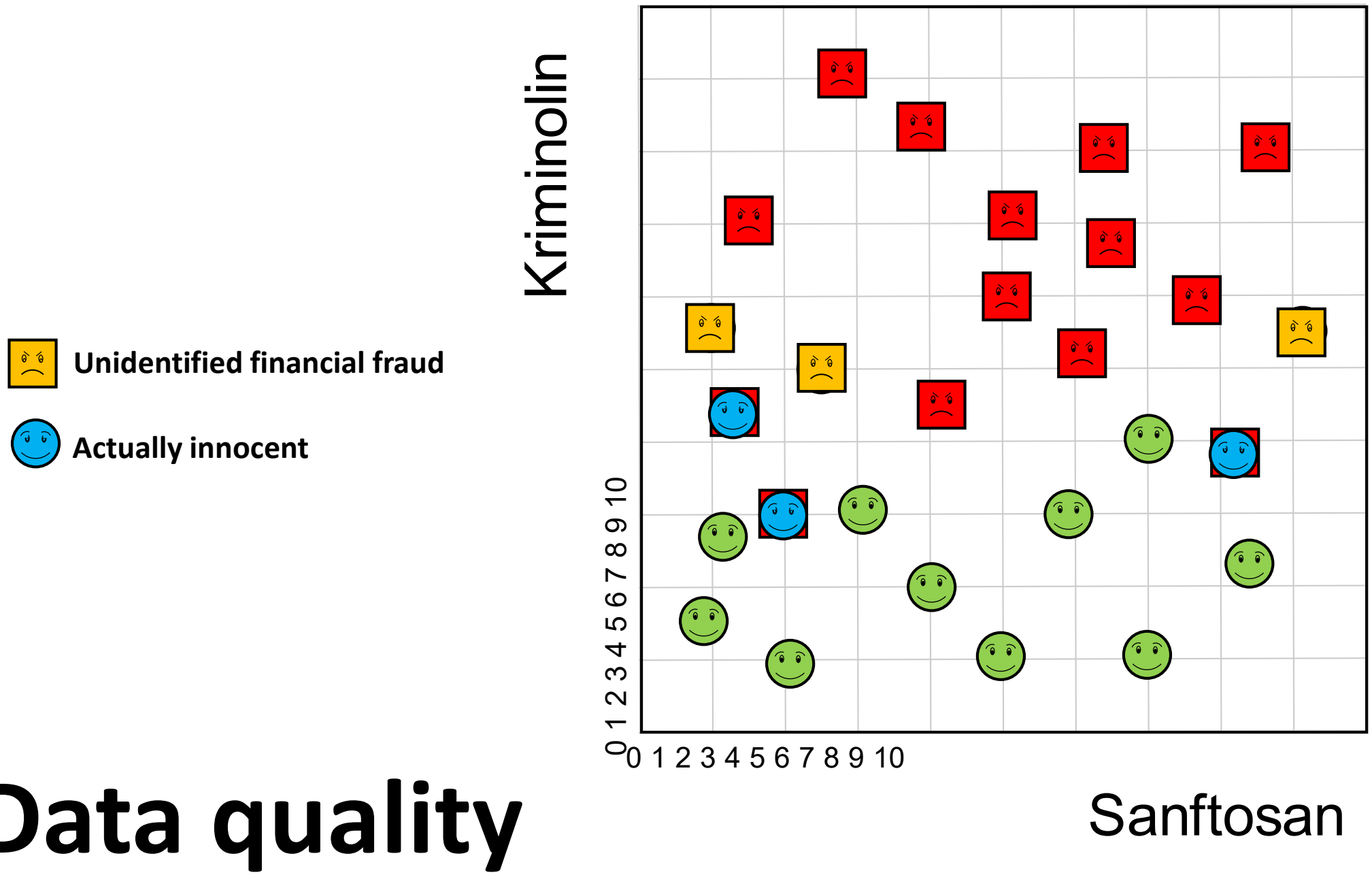
Dick Cheney, ehemaliger
Vizepräsident der USA,






Terrorists
Innocent citizens
Kriminolin
Sanftosan
0
1
2
3
4
5
6
7
8
9
10

Data quality



Learning with formulae

Recidivism risk assessment of criminals

Data

- Data Mining methods use, e.g.:
  - Age at first arrest
  - Age now
  - Financial situation
  - Criminal relatives (!)
  - Gender
  - Number and kind of previous convictions
  - Time point of last criminal action
  - A survey
  - But (of course) not the race of a person.
- To learn something, we need this data plus the information whether the person has recidivated or not.

Approaches: Regressions

- In practice, algorithm designer very often decide which data most likely correlate with „recidivism“.
- The result of the algorithm should be a single number.
- The higher the number, the higher the risk.
- Example formula:

$$\begin{aligned} & 3 * \ \# \text{ previous convictions} \\ & - 2 * \ \# \text{ days since last arrest} \\ & + 3 * (1 \text{ if man, } 0 \text{ else}) \\ & + 2,5 * (\ 1 \text{ if violent act involved, } 0 \text{ else}) + \ldots \end{aligned}$$

In general

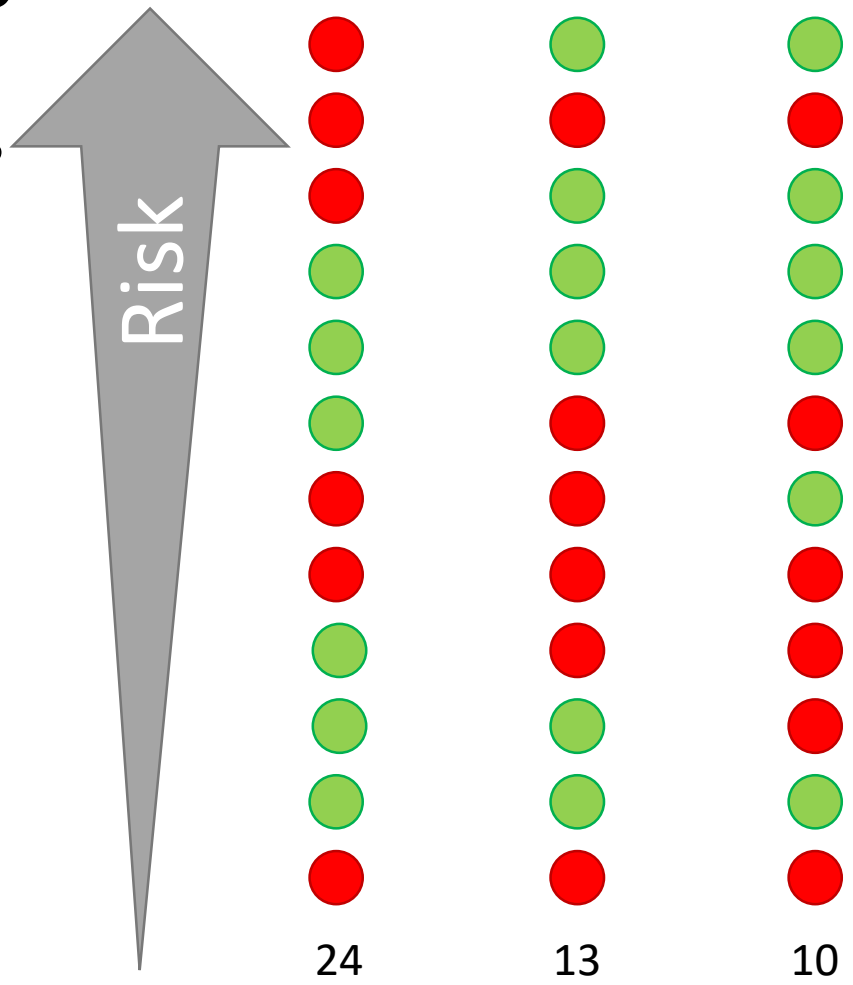
$$
\begin{aligned}
& w_1 * \ \# \text{ previous convictions} \\
& - w_2 * \ \# \text{ days since last arrest} \\
& + w_3 * (1 \text{ if man, } 0 \text{ else}) \\
& + w_4 * (\ 1 \text{ if violent act involved, } 0 \text{ else}) + \ldots
\end{aligned}
$$

The computer determines the weights and gets a feedback on its predictions and the actual observation of recidivism in that individual.

Quality of an algorithm

„Learning“ of weights

- Algorithm ‚tries out‘ weights and computes resulting risk for all persons in a test data set.
- Evalutes how many of the real recidivists get high risk scores.
- The weighting that maximizes this will be used for all further predictions.



Green balls: non-recidivating criminals;
Red balls: recidivating criminals.

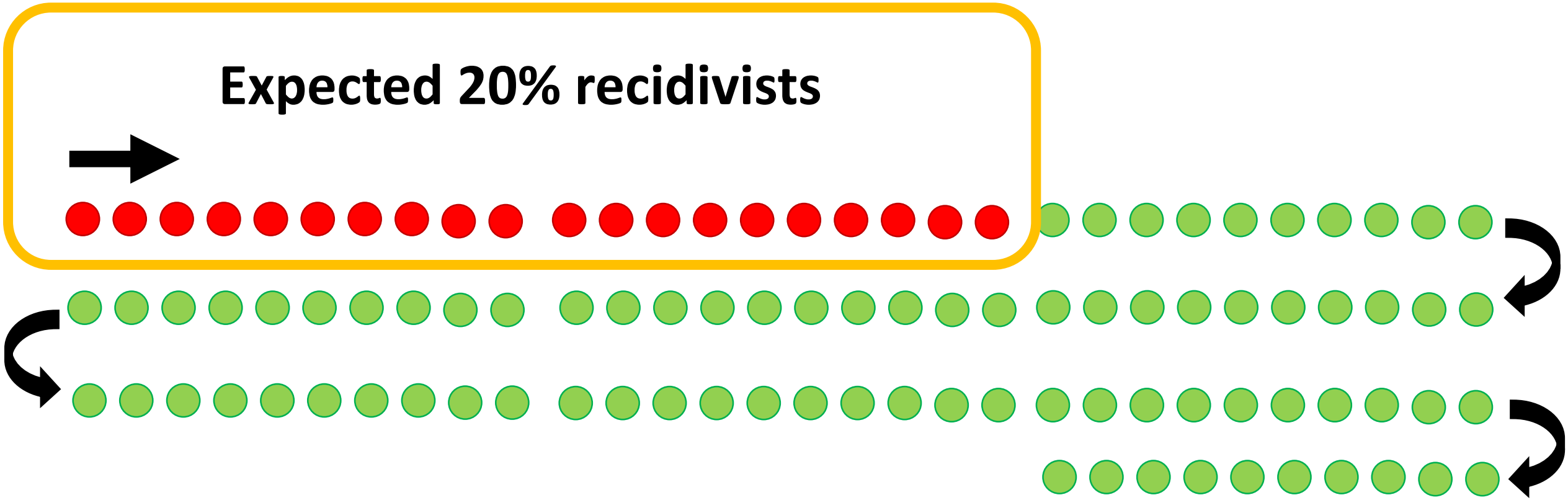
Optimal sorting: all reds on top, all greens on the bottom.

Quality measure: pairs of red/green balls where red is on top of the green.

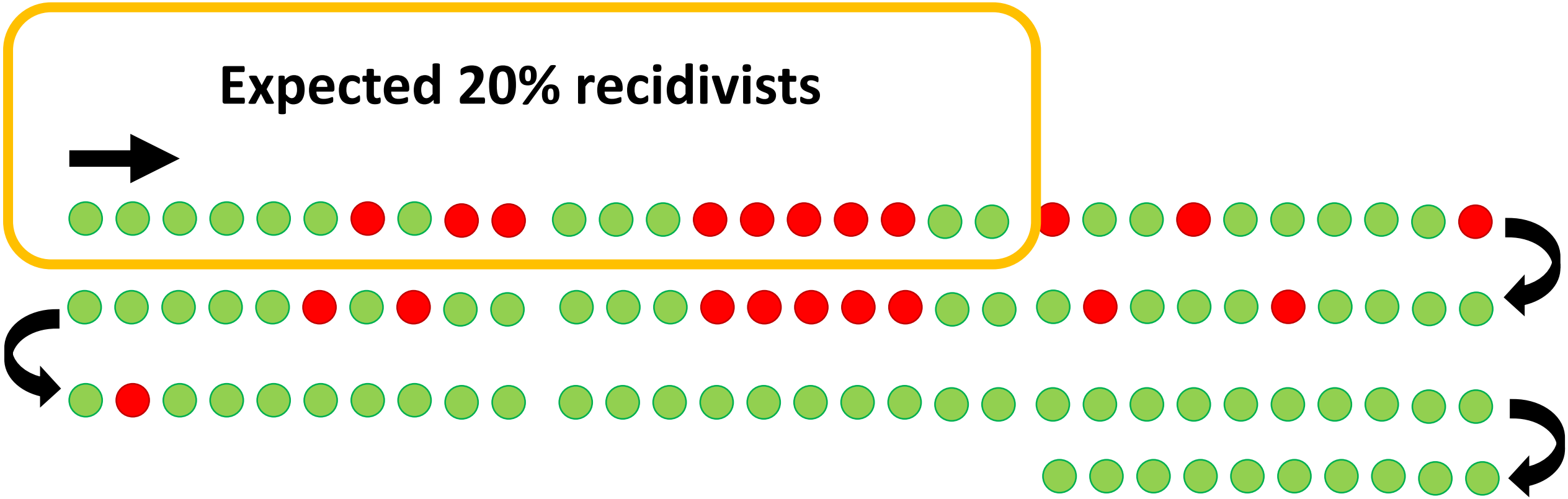
Oregon Recidivism Rate Algorithm

- 72 of 100 pairs are correctly sorted (72% success rate! Yeah!)
- Does this resemble they way judges make a decision?
- No, instead of judging pairs, they see a sequence of defendants, of which they are most interested in the ones with highest risk.
- Experience guides where to cut the risk score:
  - E.g., recidivism rate of joung criminals is about 20%.

Optimal Sorting



Possible sorting of an algorithm with 75% correctly sorted pairs



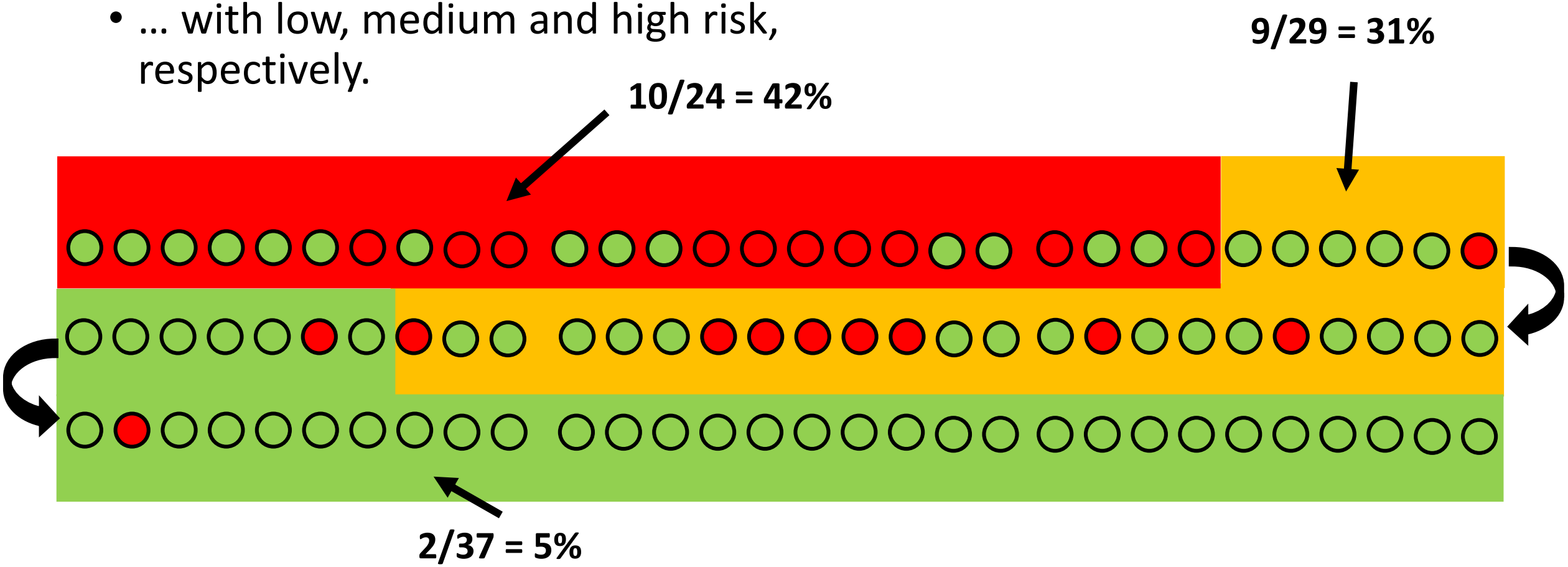
Buying a hunting dog ，

to shepherd!

Buying that software is like…

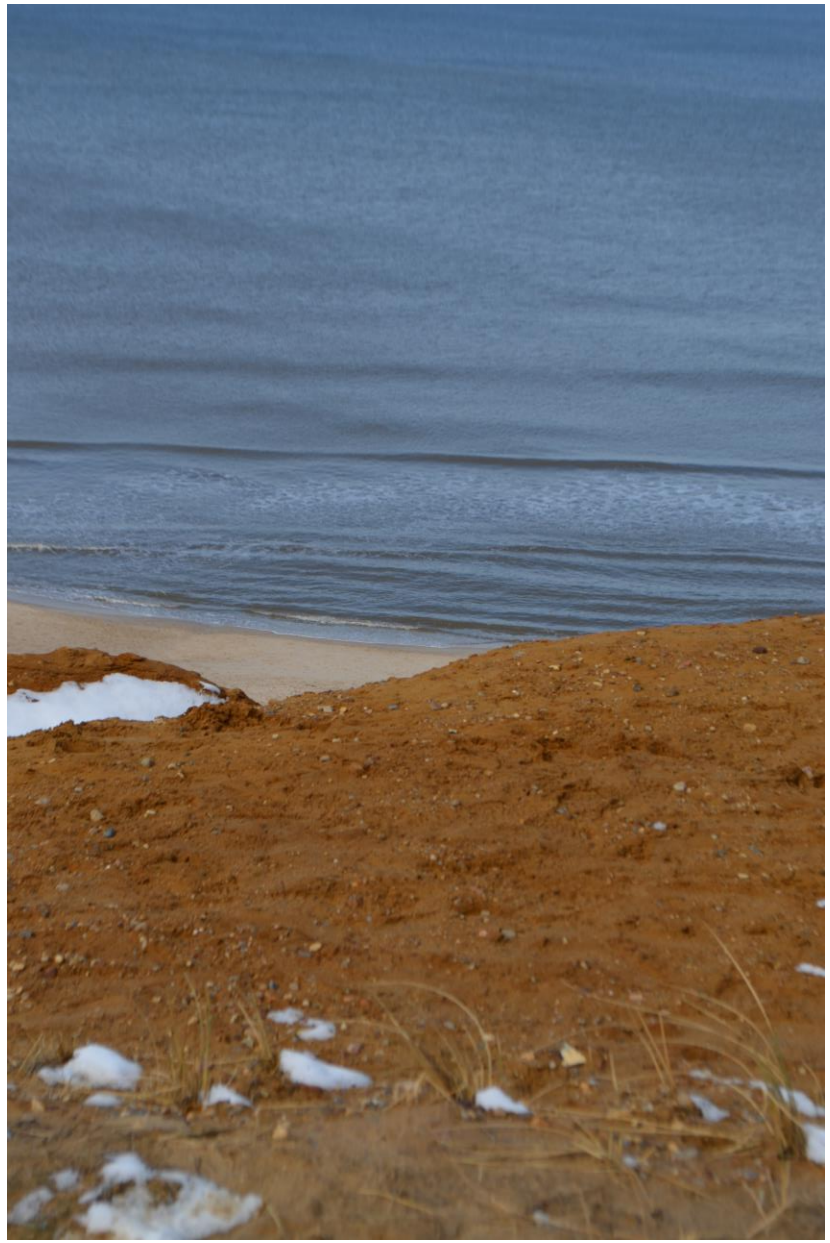
From scoring to classification

- ACLU states: criminals should be sorted into three categories…
- … with low, medium and high risk, respectively.



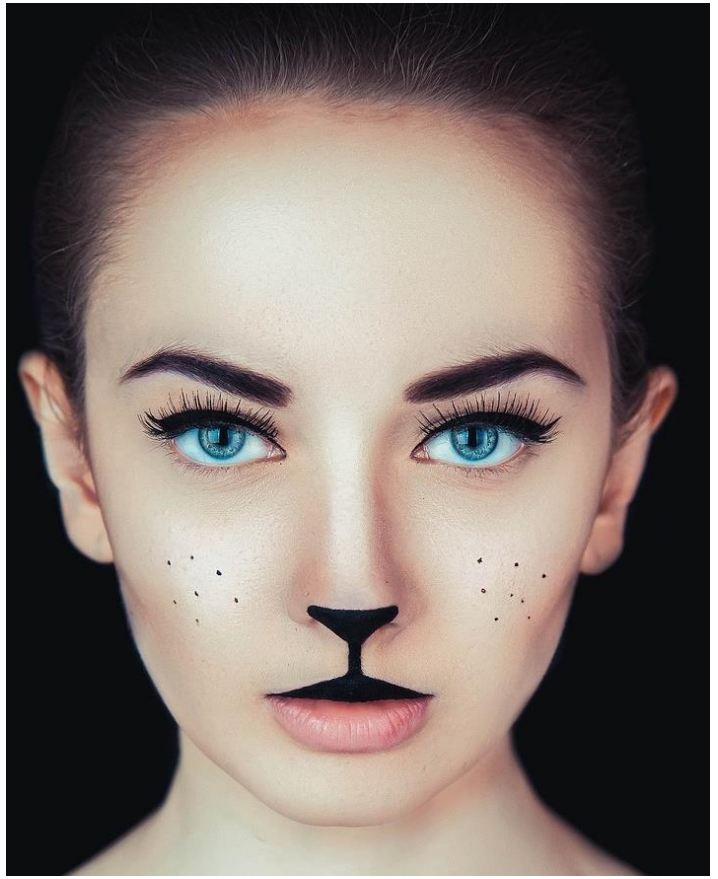
Statistical predictions of human behavior

Weather forecasts



40% a criminal….

- If humans were cats with 7 lives, to have a 40% risk of recidivism it would mean that they expectedly recidivate in three of them.
- But no! Humans are no cats.
- **Algorithmically legitimated prejudices**
  - Of 100 persons that are „like this person“ 40% are recidivating.



https://pixabay.com/de/m%C3%A4dchen-katze-k%C3%A4tzchen-mode-1048728/

Rule of thumb

AI is used primarily where there **are no simple rules**!

They often search for patterns in **highly noisy** data.

The patterns are of a **statistical nature.**

Often try to identify a small group of people (Problem of imbalance)

Can algorithms discriminate people?

Equality

This is what happens when you search for „CEO“ on Google

And this, if I search for ‘boss’ on Pixabay….

Discrimination

- Google shows job ads with a lower average salary to female surfers.
- Based on one perspective, recidivism risk assessment is rassistic.
- Discrimination in training data will be learned.
- If training data contains to little data about minorities, their properties will not be learned.

Algorithms in a democracy

In general

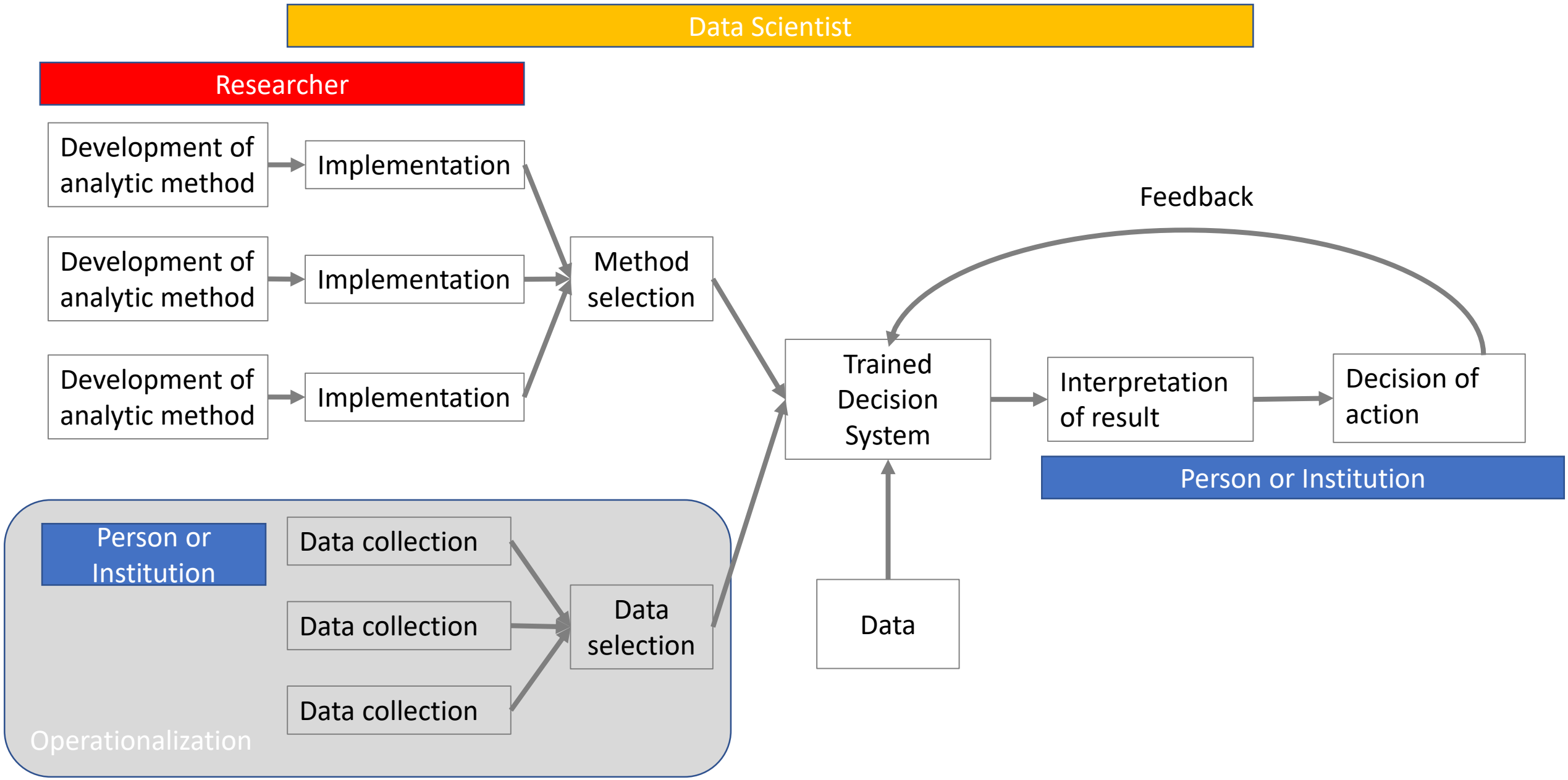
In principle, ADM systems can be used for many different, difficult questions:

- Automatic performance evaluation
- Credit approval
- Job application evaluation.
- Performance evaluation of employees.
- Algorithms that predict the time point of death(real!)
- Terrorist identification
- …

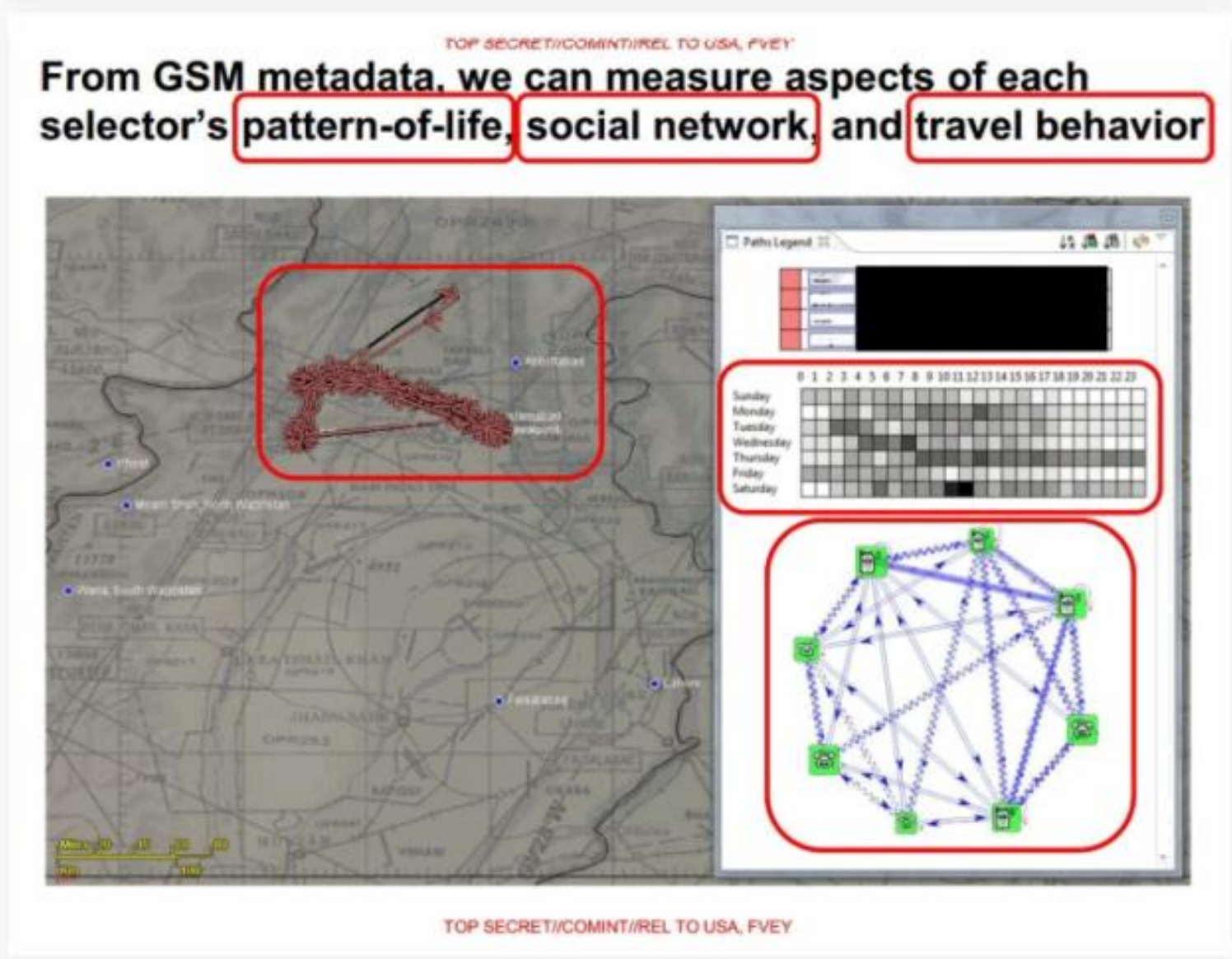
Your task today….

Develop an ADM system that identifies terroristic couriers!

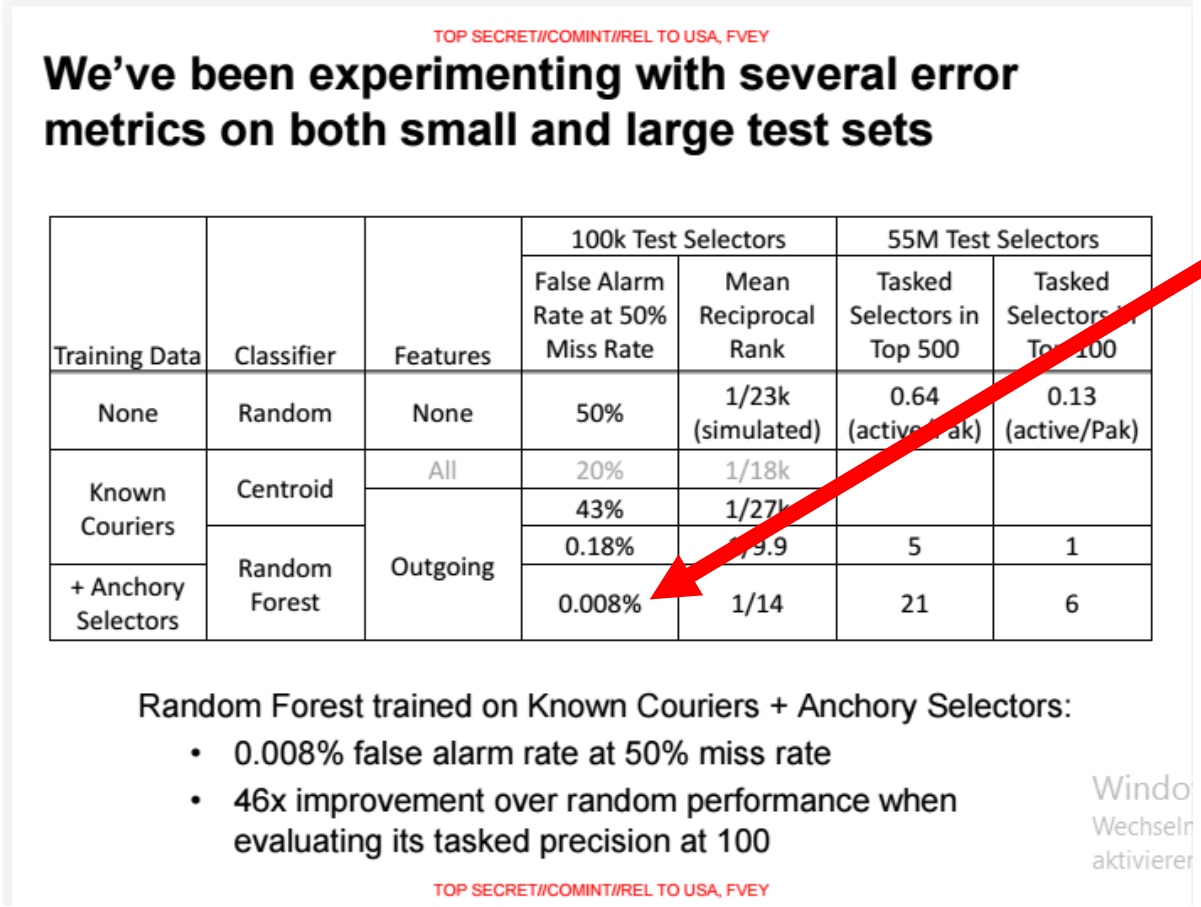
Design process



Capturing terrorists with network analysis



Terrorist identification 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:

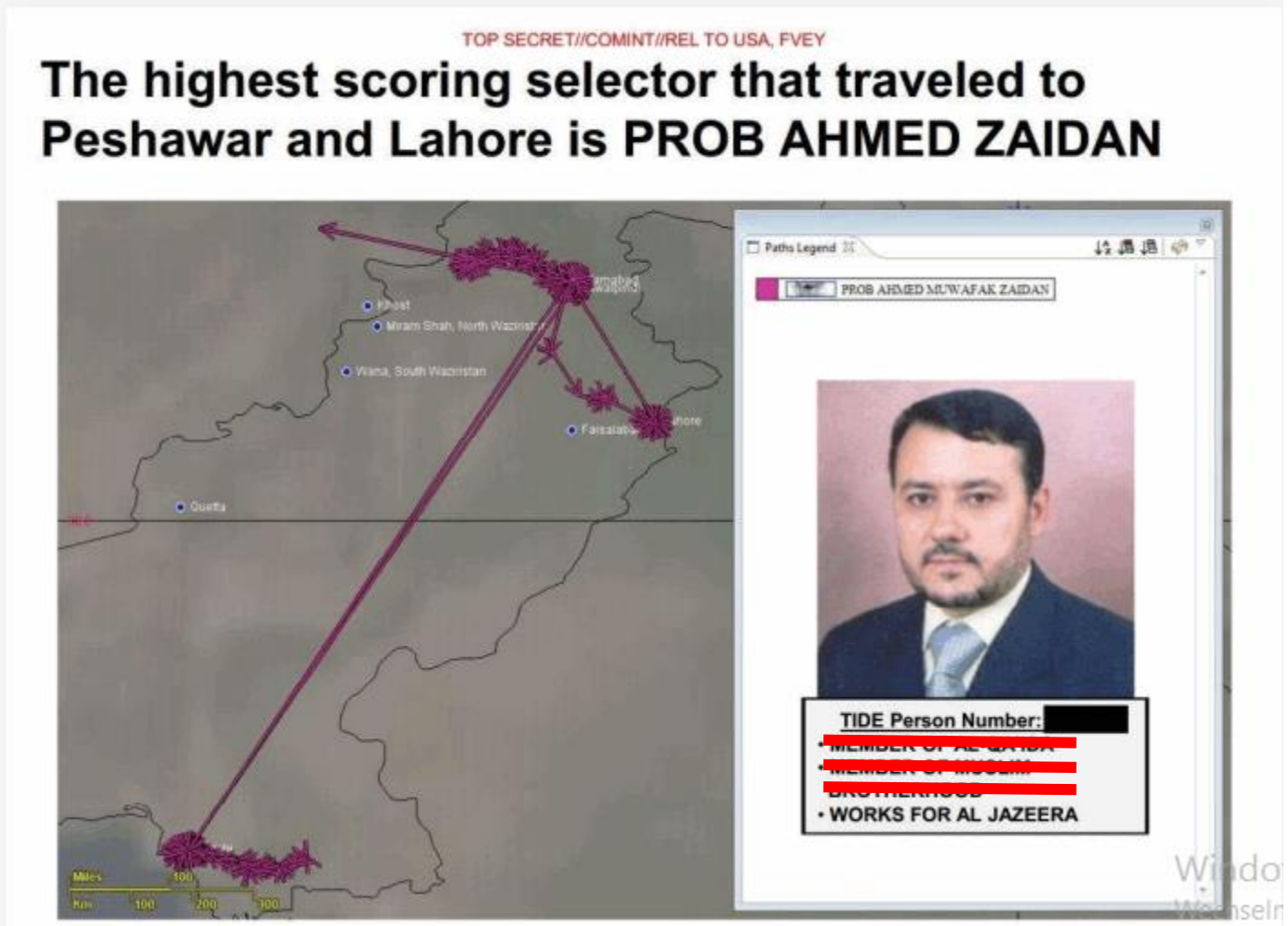
- 0.008% false alarm rate at 50% miss rate
- 46x improvement over random performance when evaluating its tasked precision at 100

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

These are 4,400 innocent persons to catch 50% of the (suspected) terrorists.

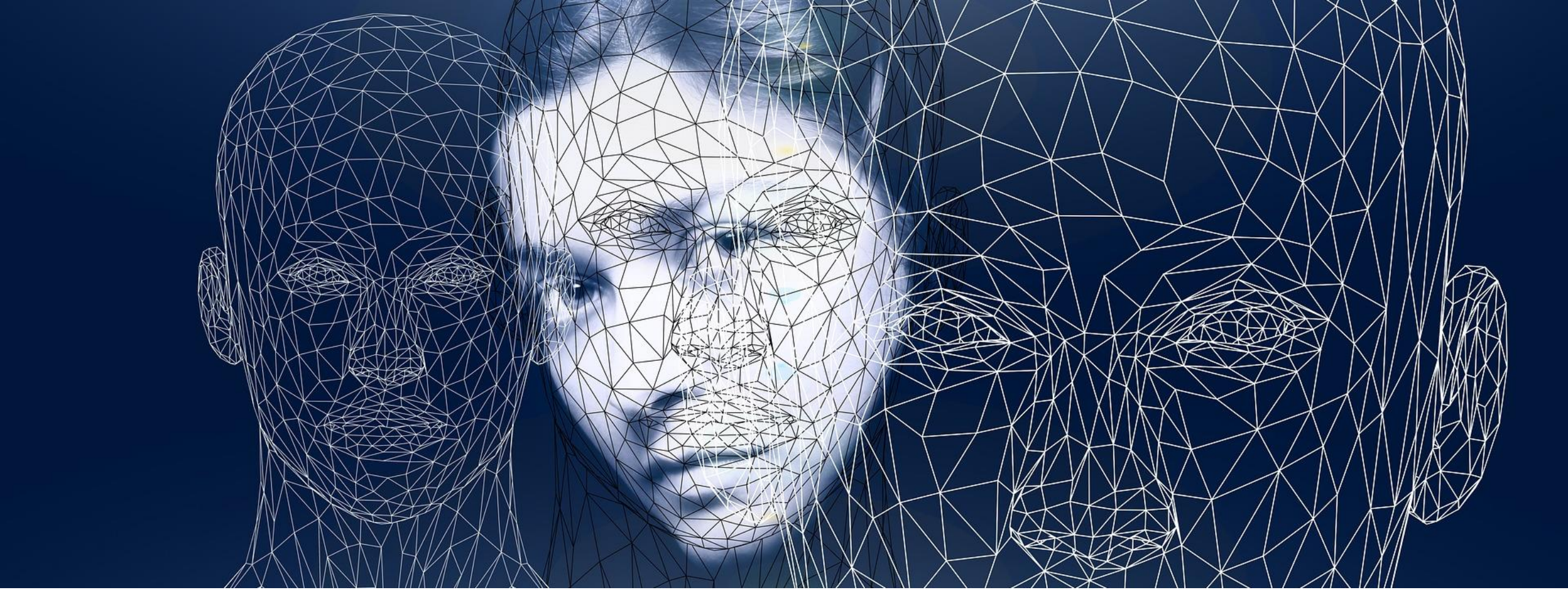
https://theintercept.com/document/2015/05/08/skynet-courier/
https://theintercept.com/2015/05/08/u-s-government-designated-prominent-al-jazeera-journalist-al-qaeda-member-put-watch-list/

Most suspicious person according to algorithm is…



How good are these robo-judges?

- Very bad: COMPAS
  - High risk category:
    - General recidivism: correct in 50% of all individuals!
    - Serious crimes: only 20% correct!
- An American terrorist identification system boasts:
  - „Only 0.008% false positives!“
  - With 55 million inhabiltants these are about 4,400 innocents to identify a few hundred.
- However, in detecting cancer they are sometimes better than physicians.



Socio-informatic system analysis

Problems with the embedding of an ADM into a social process

- Deciders follow an ‚attention economy‘.
- **„Best practice“** might require the usage of the software.
- **Delegation of responsibility!**
- Sometimes a false-negatively categorized person cannot prove the decision wrong!
  - E.g., rejected applicants for a job,
  - Rejected credits,
  - Suspects kidnapped andkept in camps.

My stance

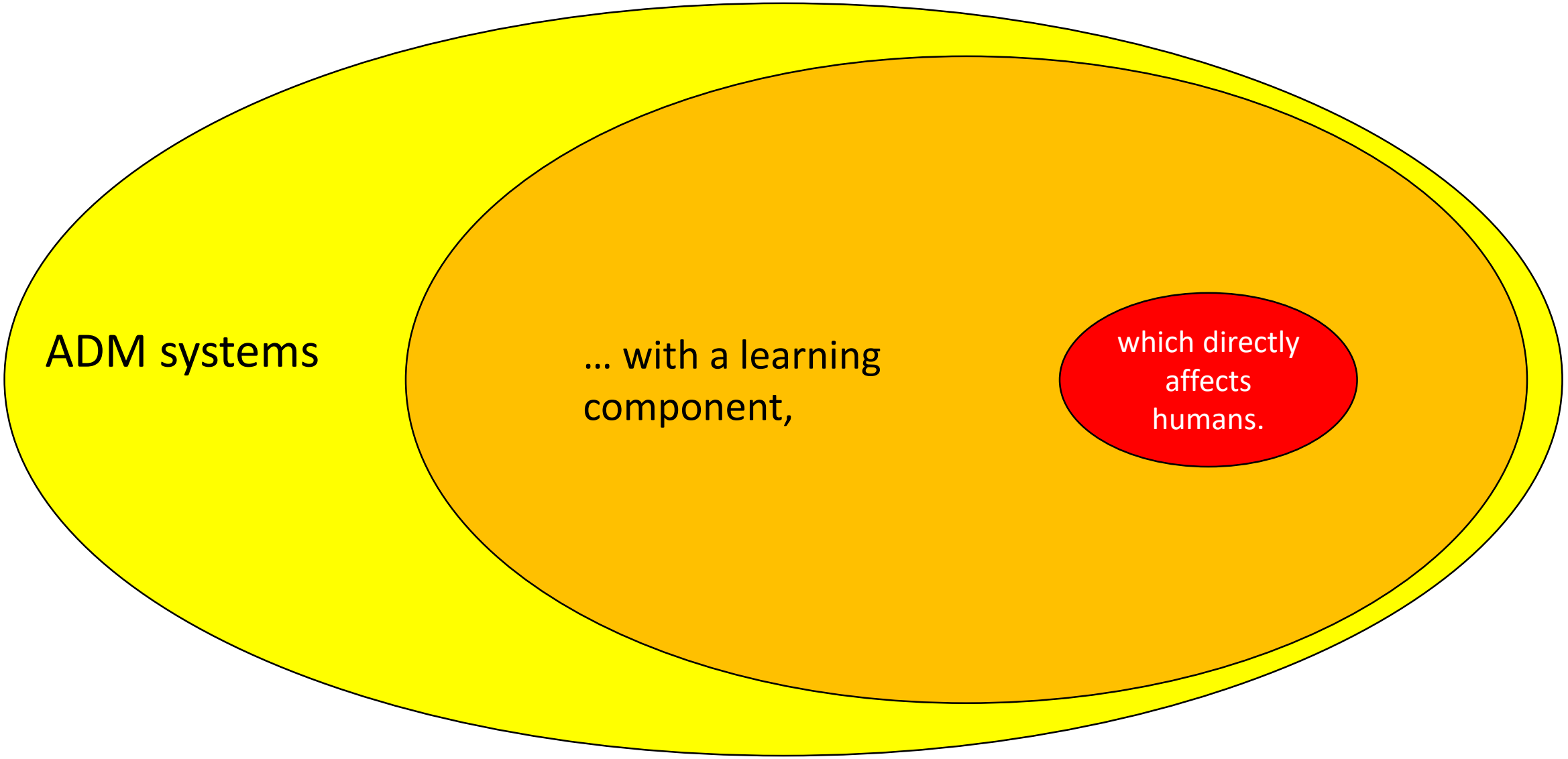
- ADM systems **could** help to make better decisions.
  - They can search through huge data sets.
  - They could identify new “patterns”.
  - Could avoid discriminations.
- **However**, today, they’re not yet there. And they might not be able to, especially when very few persons have to be identified in a crowd.

Problems of ADM systems in people and risk assessment

1. **Who decides what a „good“ decision is?**
2. **ADM systems result in probabilities, not truths.**
3. **ADM systems can discriminate**
4. **The help to identify small groups but with many false positives.**
5. **ADM systems can change social processes.**
6. **The reaction of the social system can increase the problem.**

@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern

ADM systems to be regulated

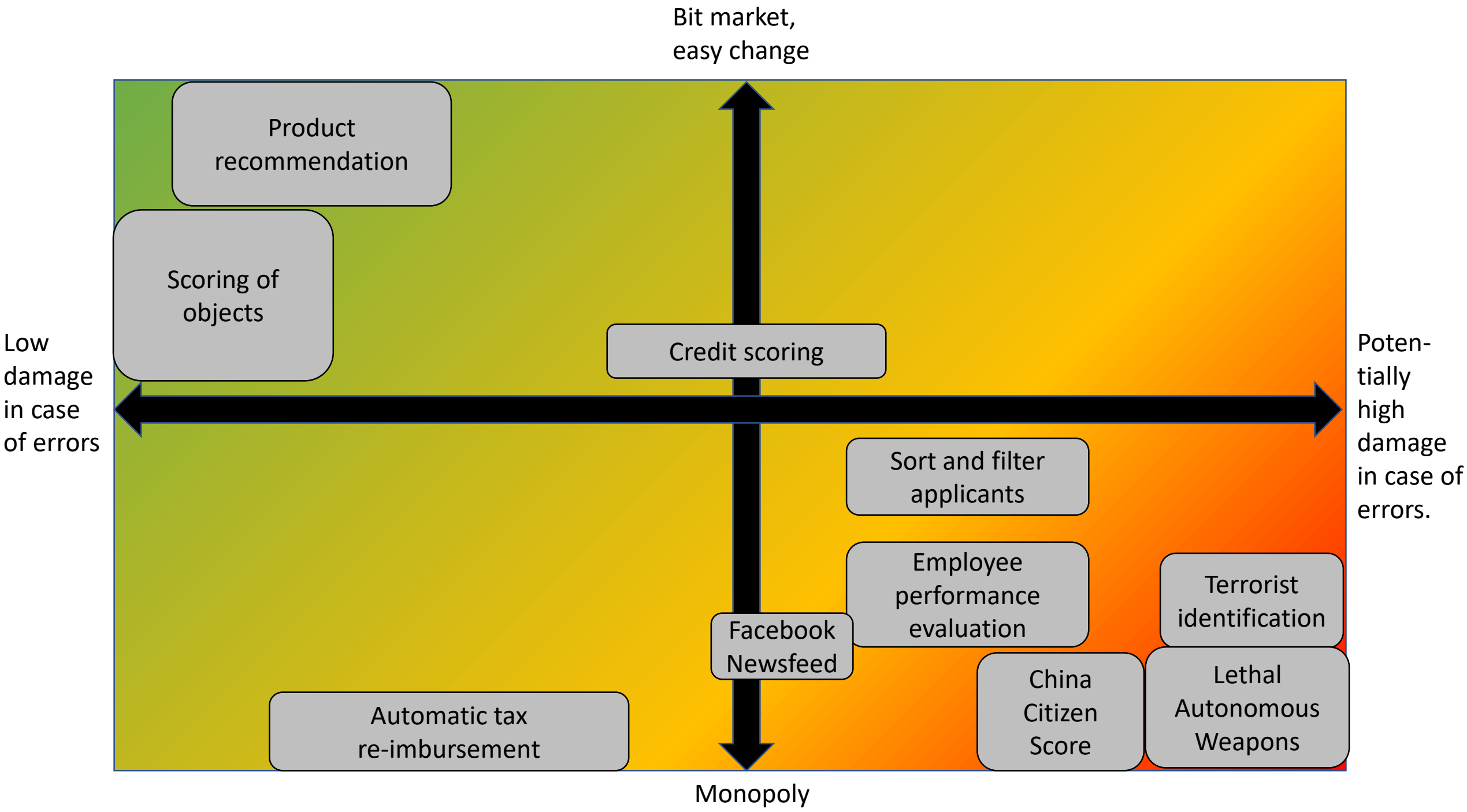


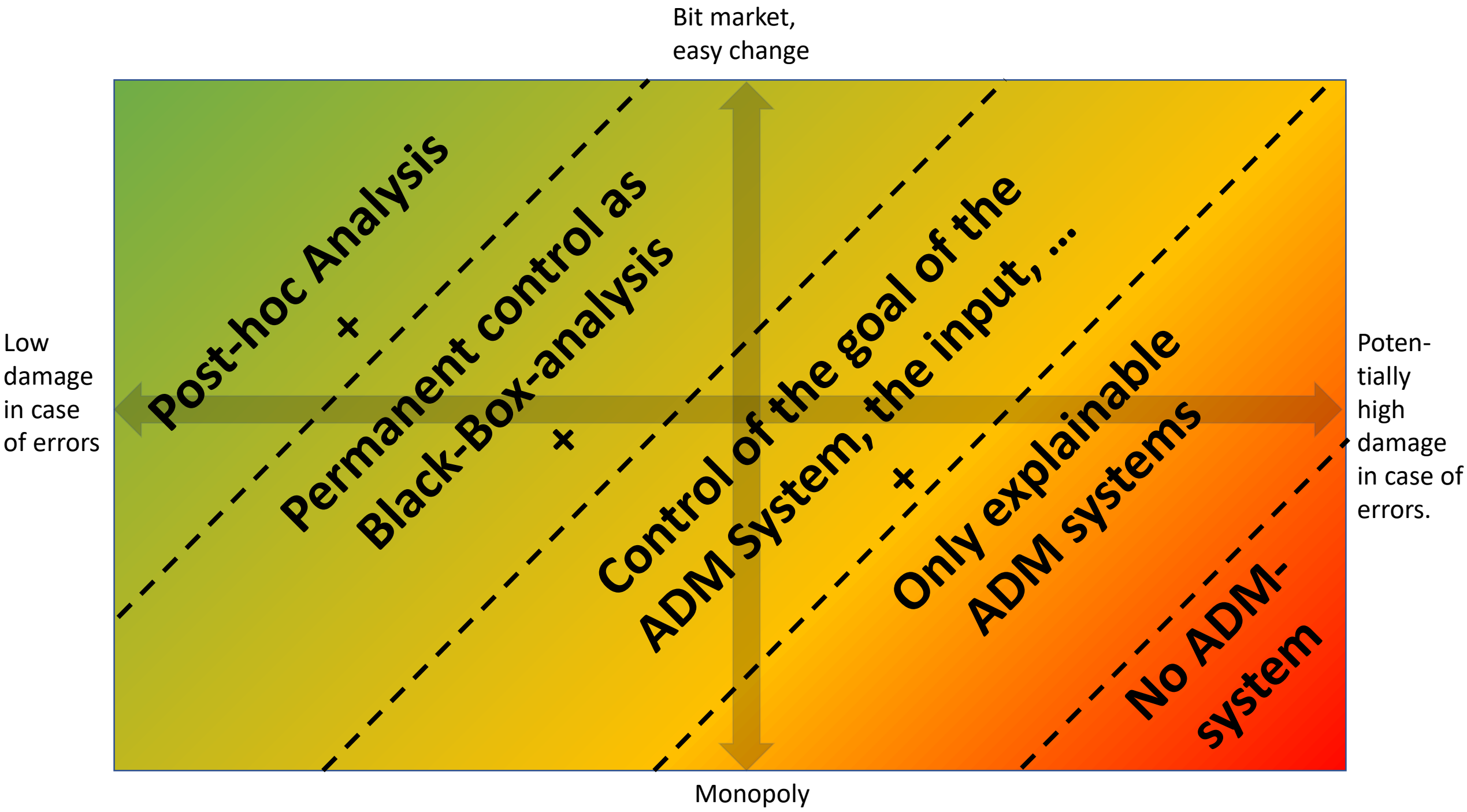
Do all of them need to be regulated the same?

1. Potential for damage in case of errors

$$\sum \begin{array}{l} \text{Potential for harm for individual (in case of error)} \\ +\text{Potential for harm for society (in case of errors)} \end{array}$$

2. Number of competitors and ease of re-evaluation by other ADM systems


Bit market,
easy change
Product recommendation
Scoring of objects
Credit scoring
Low damage in case of errors
Poten-tially high damage in case of errors.
Sort and filter applicants
Employee performance evaluation
Terrorist identification
Facebook Newsfeed
China Citizen Score
Lethal Autonomous Weapons
Automatic tax re-imbursement
Monopoly


Bit market,
easy change
Low
damage
in case
of errors
Poten-
tially
high
damage
in case of
errors.
Monopoly
Post-hoc Analysis
+
Permanent control as
Black-Box-analysis
+
Control of the goal of the
ADM System, the input, ...
+
Only explainable
ADM systems
No ADM-
system

References (sorry, German only!)

1. Brochure of the Bayerische Landesmedienanstalt
Google for „BLM Dein Algorithmus - meine Meinung“

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

2. Study for the Bertelsmann foundation (2018)