# Fairness-aware Classifier with Prejudice Remover Regularizer

**Toshihiro Kamishima***, Shotaro Akaho*, Hideki Asoh*, and Jun Sakuma†

**National Institute of Advanced Industrial Science and Technology (AIST), Japan*

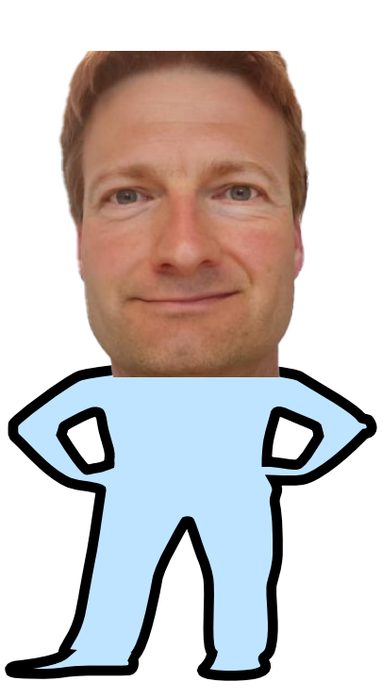
*†University of Tsukuba, Japan; and Japan Science and Technology Agency*

Test of Time Awards Talk @ ECMLPKDD 2022, 2022-09-21

BARCELONA, SPAIN

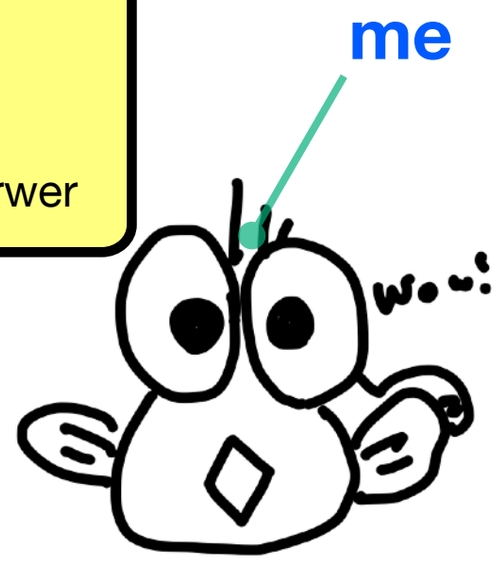
# ECMLPKDD2010

European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases

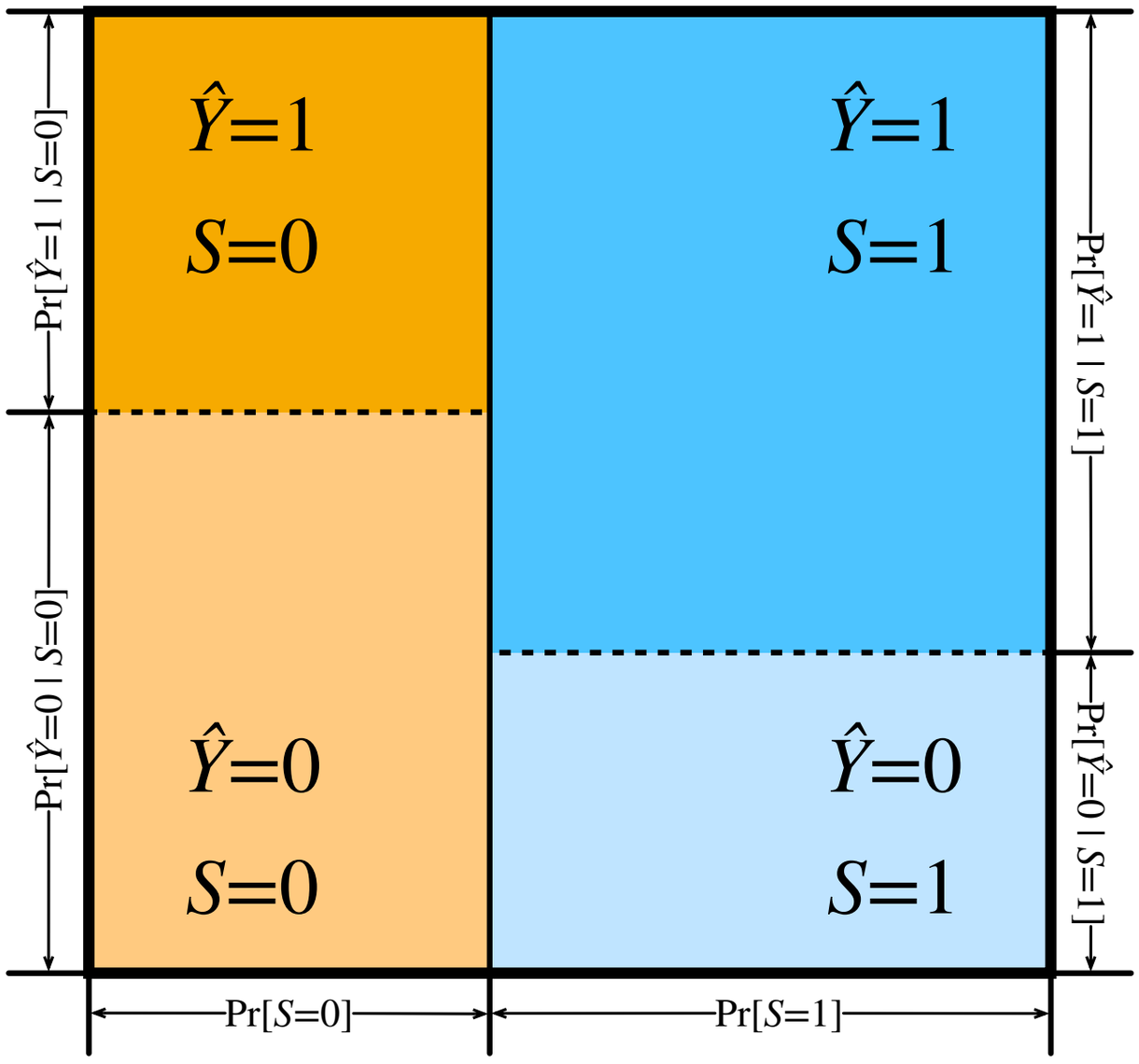


## Three Approaches for Making the Naïve Bayes Classifier Discrimination-Free

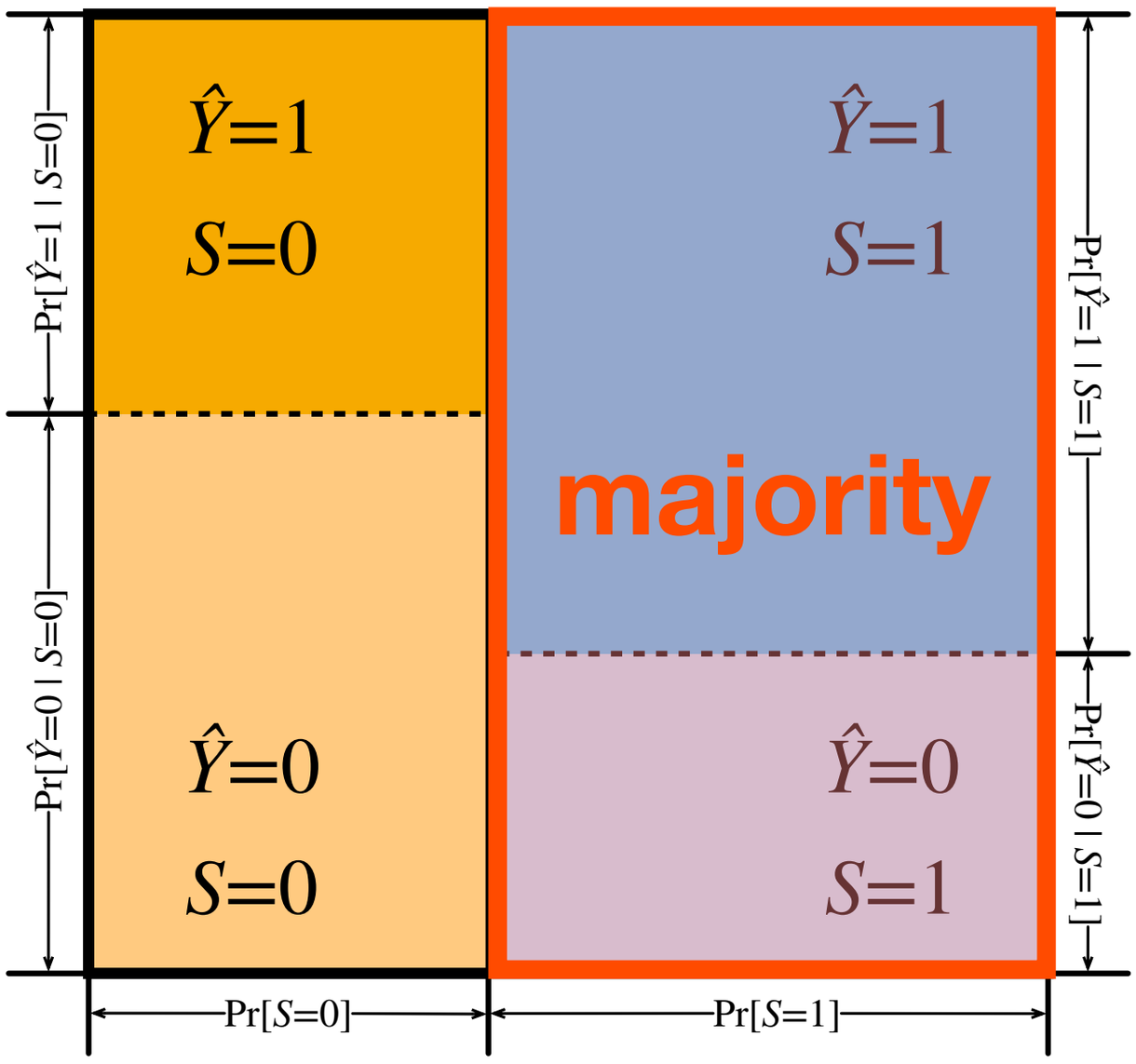
Toon Calders & Sicco Verwer



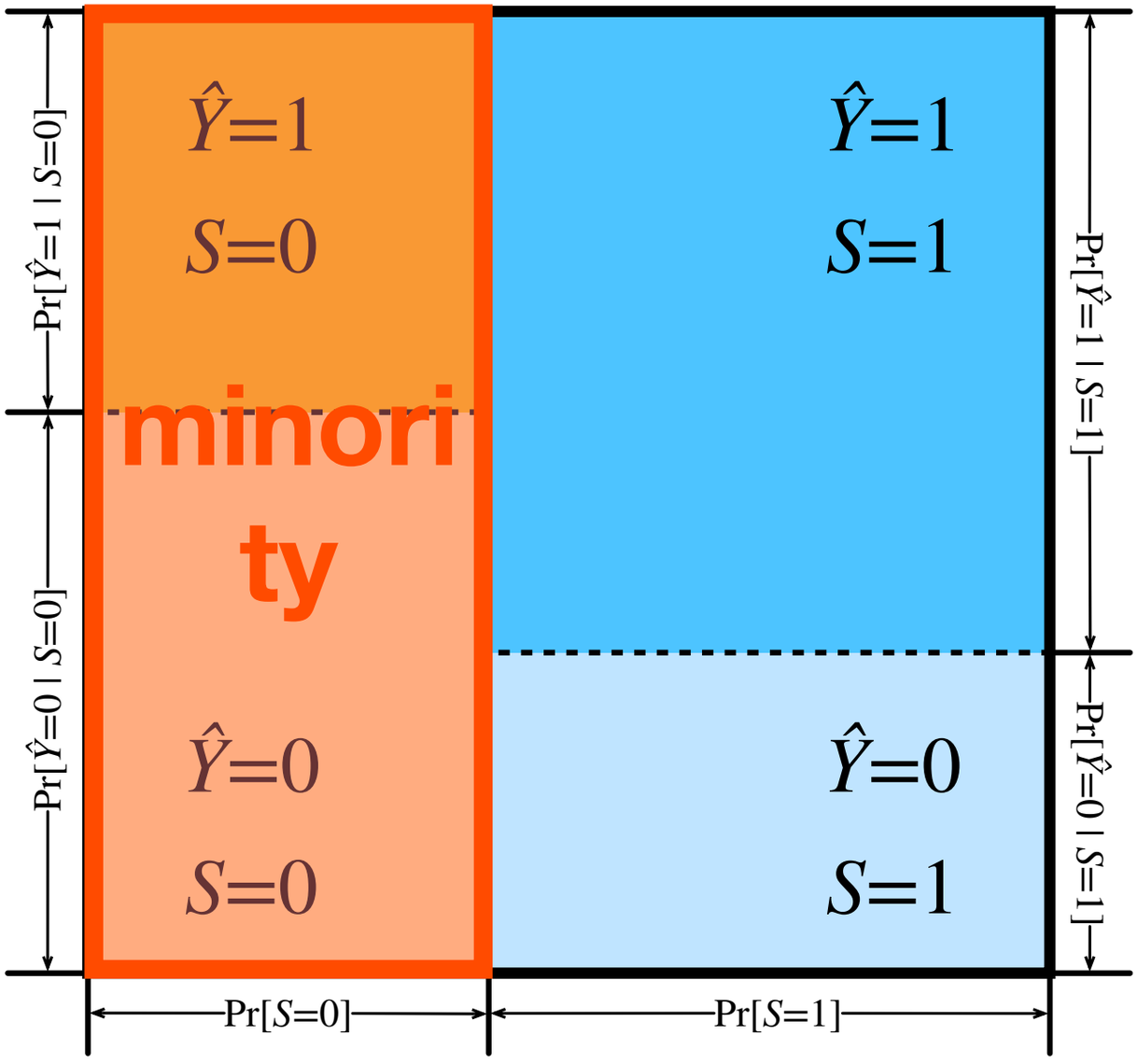
Fairness between two sensitive groups: **Orange** and **Blue**



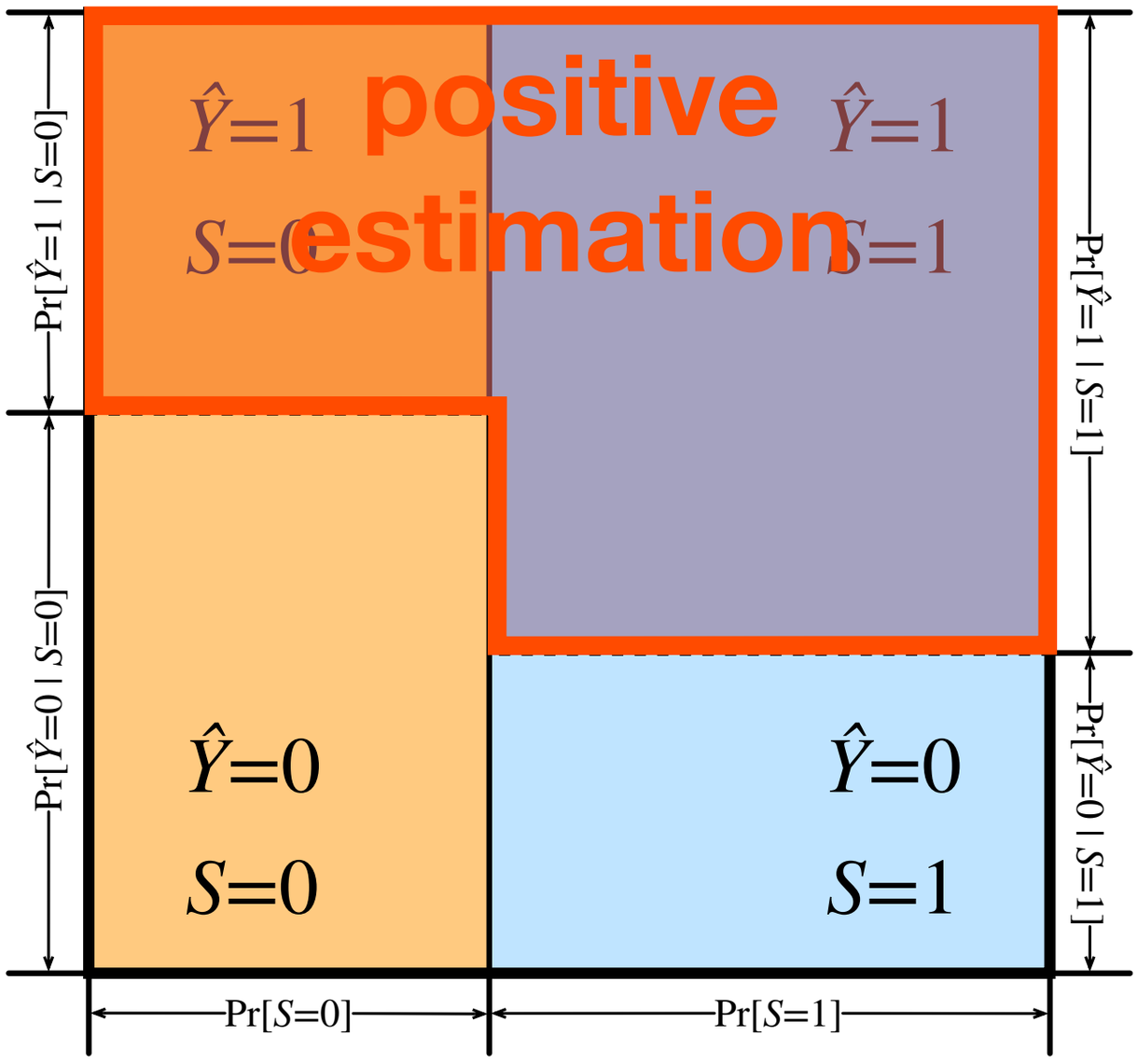
Fairness between two sensitive groups: **Orange** and **Blue**



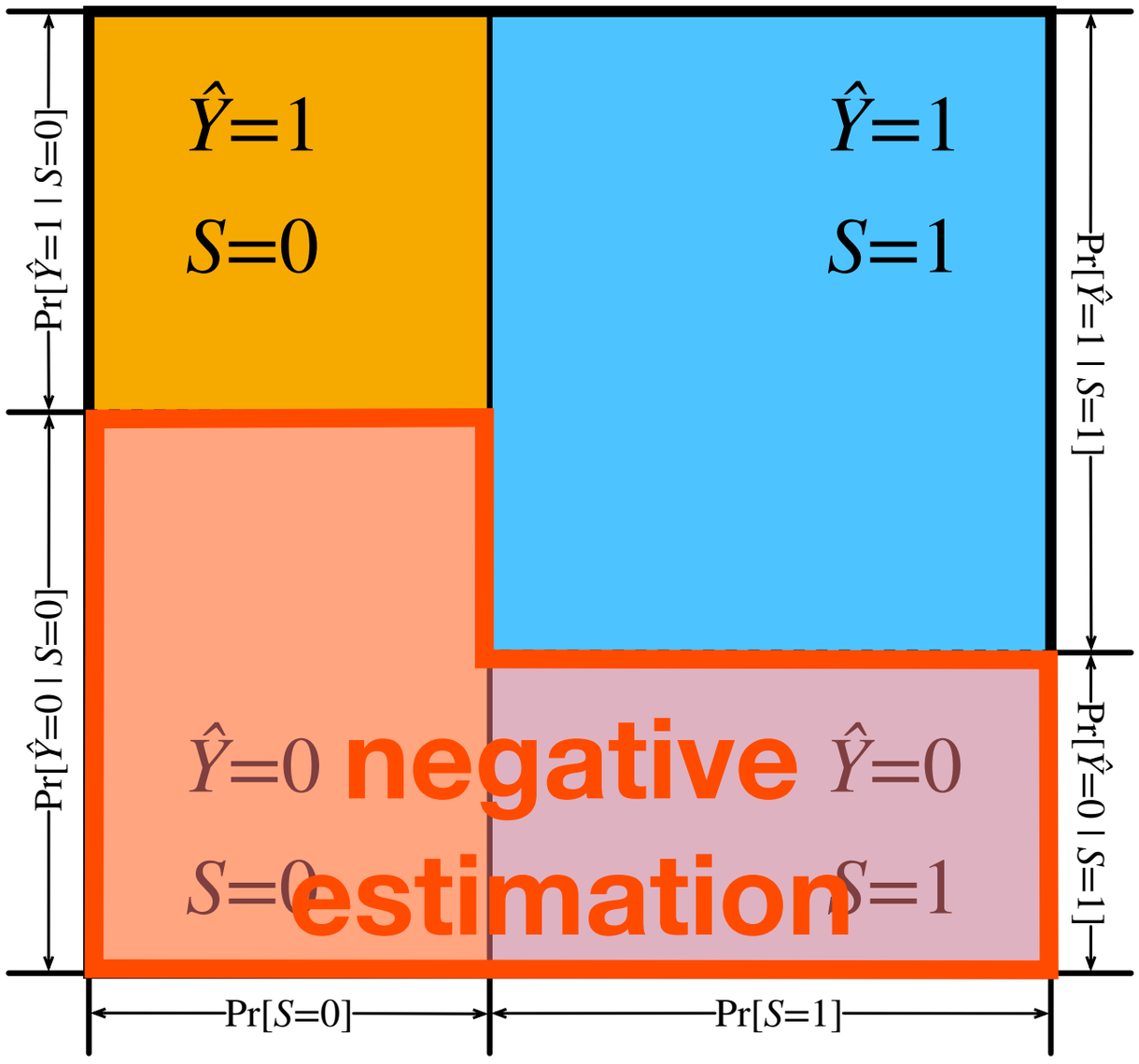
Fairness between two sensitive groups: **Orange** and **Blue**



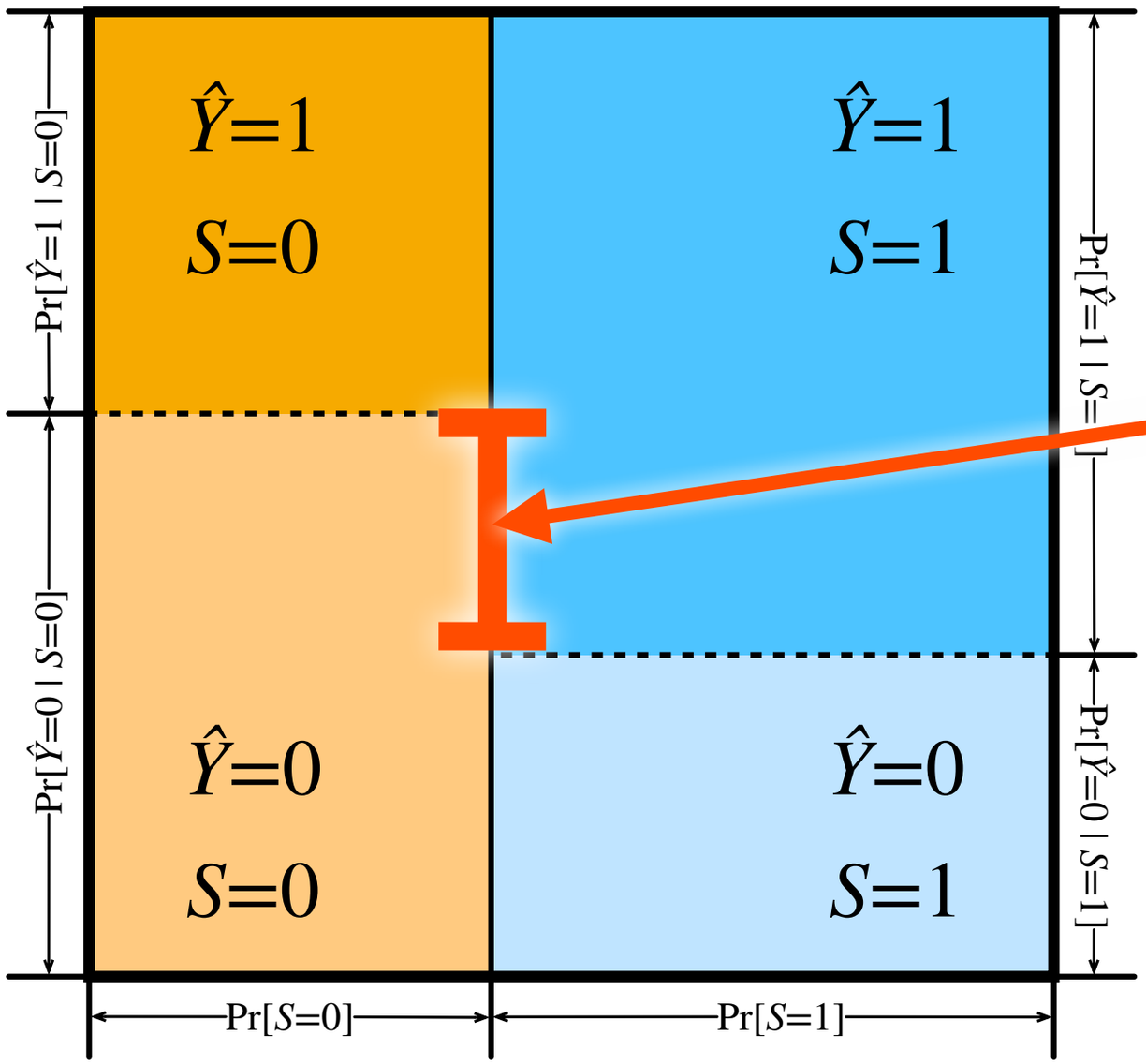
Fairness between two sensitive groups: **Orange** and **Blue**



Fairness between two sensitive groups: **Orange** and **Blue**



Fairness between two sensitive groups: **Orange** and **Blue**

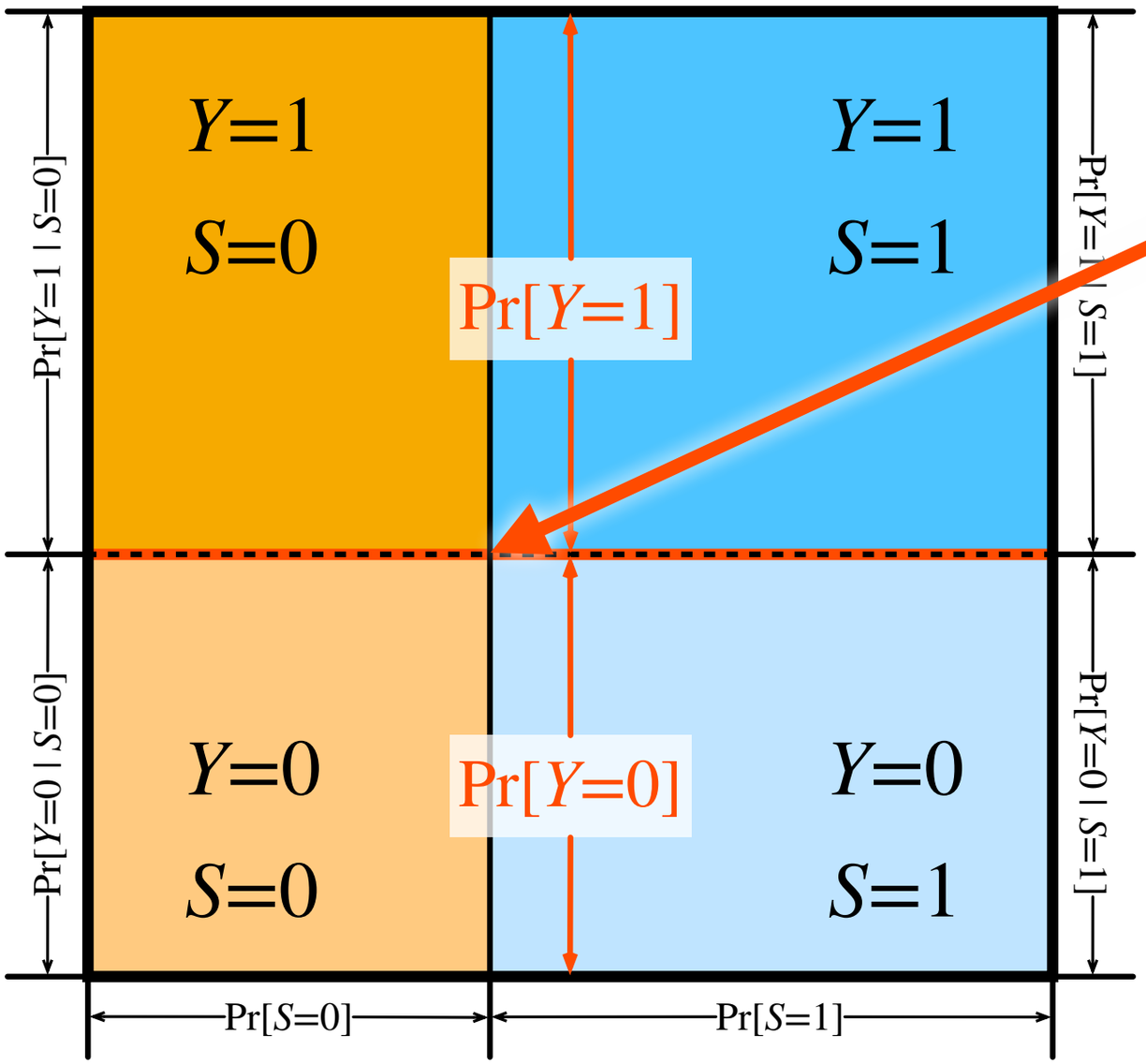



**discrimination score**

$$\left| \Pr[\hat{Y}=1 \mid S=0] - \Pr[\hat{Y}=1 \mid S=1] \right|$$

Distributive fairness means the equality of determination ratios between sensitive groups

# Two Naïve Bayes method

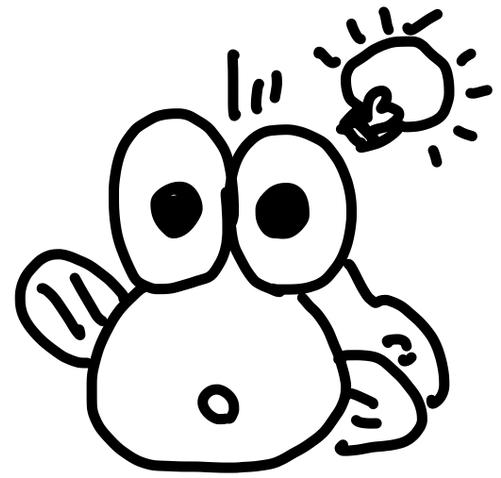



**Two Naïve Bayes method**

(1) Learn a pair of naïve Bayes classifiers for each sensitive group

(2) A joint distribution of $Y$ and $S$ is iteratively modified so as to satisfy the fairness condition

# Formulation as a Constrained Optimization Problem



# Formulation as a Constrained Optimization Problem

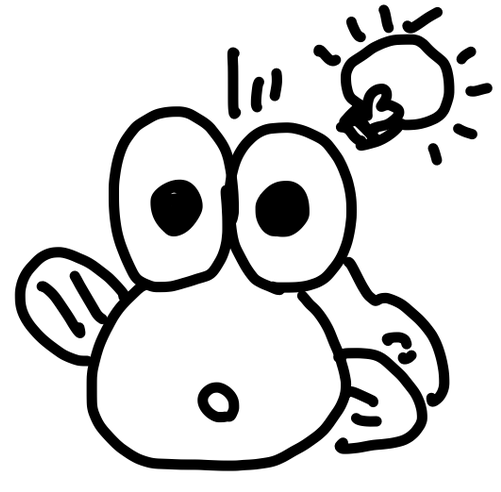
**Minimize Objective Function**

classification loss
$\text{loss}(\mathscr{D} \mid \boldsymbol{\Theta})$

$- \lambda$

fairness constraint
$-\text{I}(Y, S)$

balancing fairness and utility



**simple**

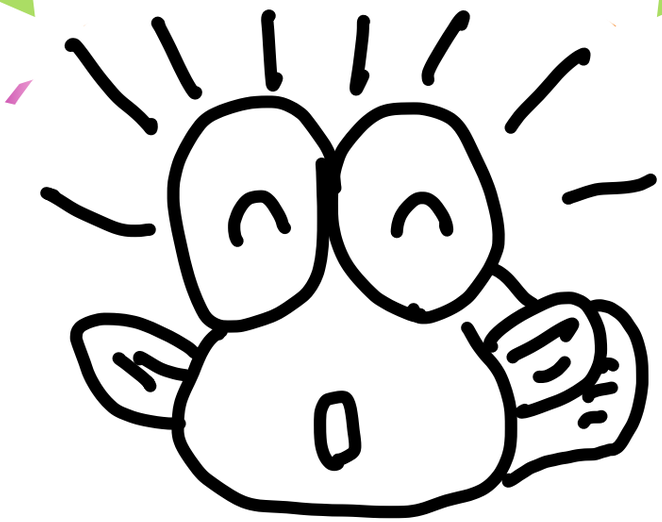
**simple**

**flexible**

# Good Idea!

simple

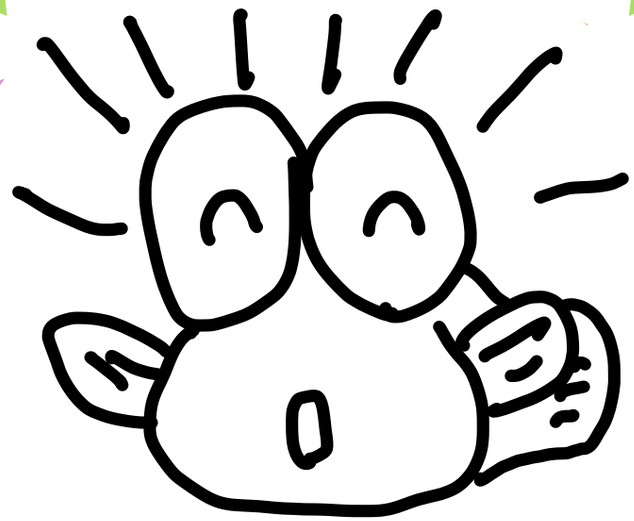
flexible



# Good Idea!

**simple**

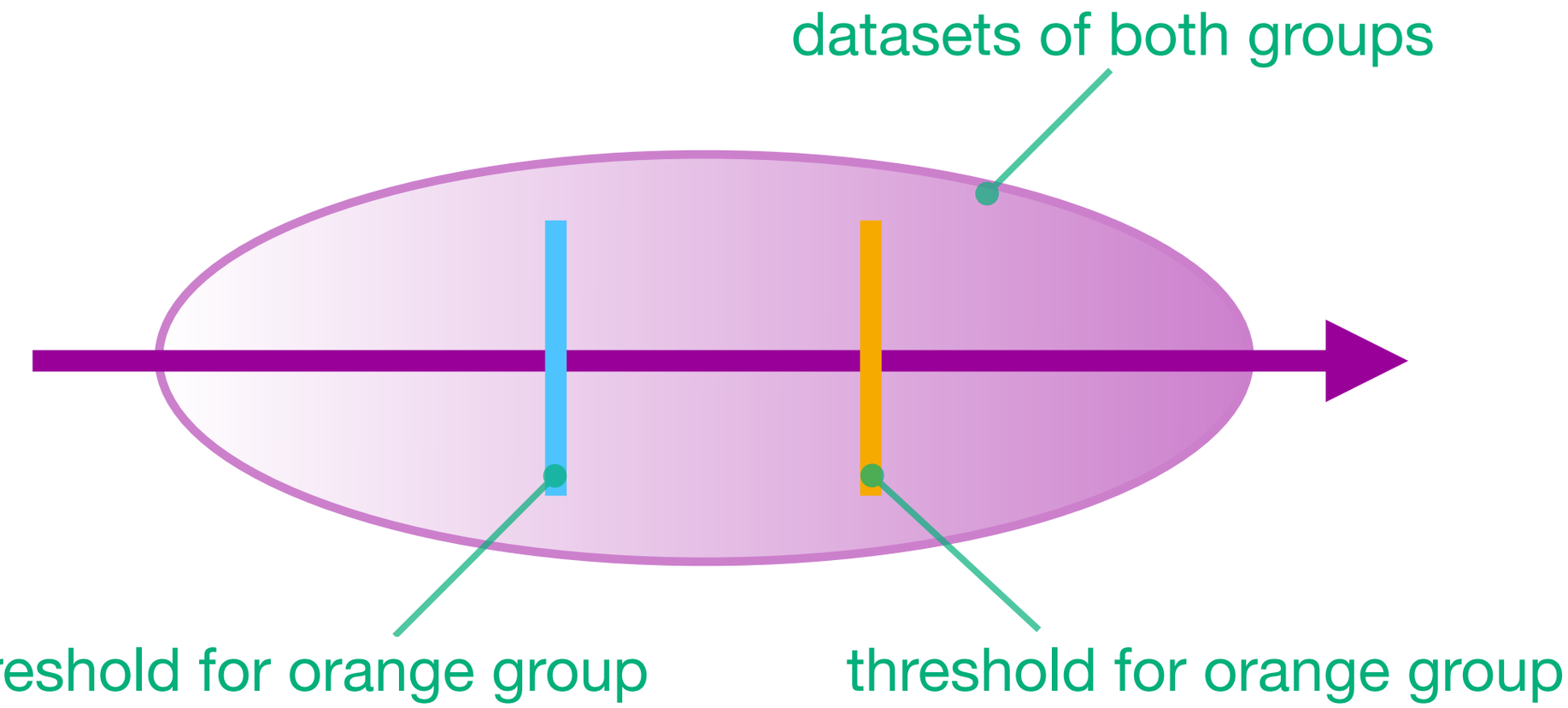
**flexible**



**It would be the happiest moment for researchers**

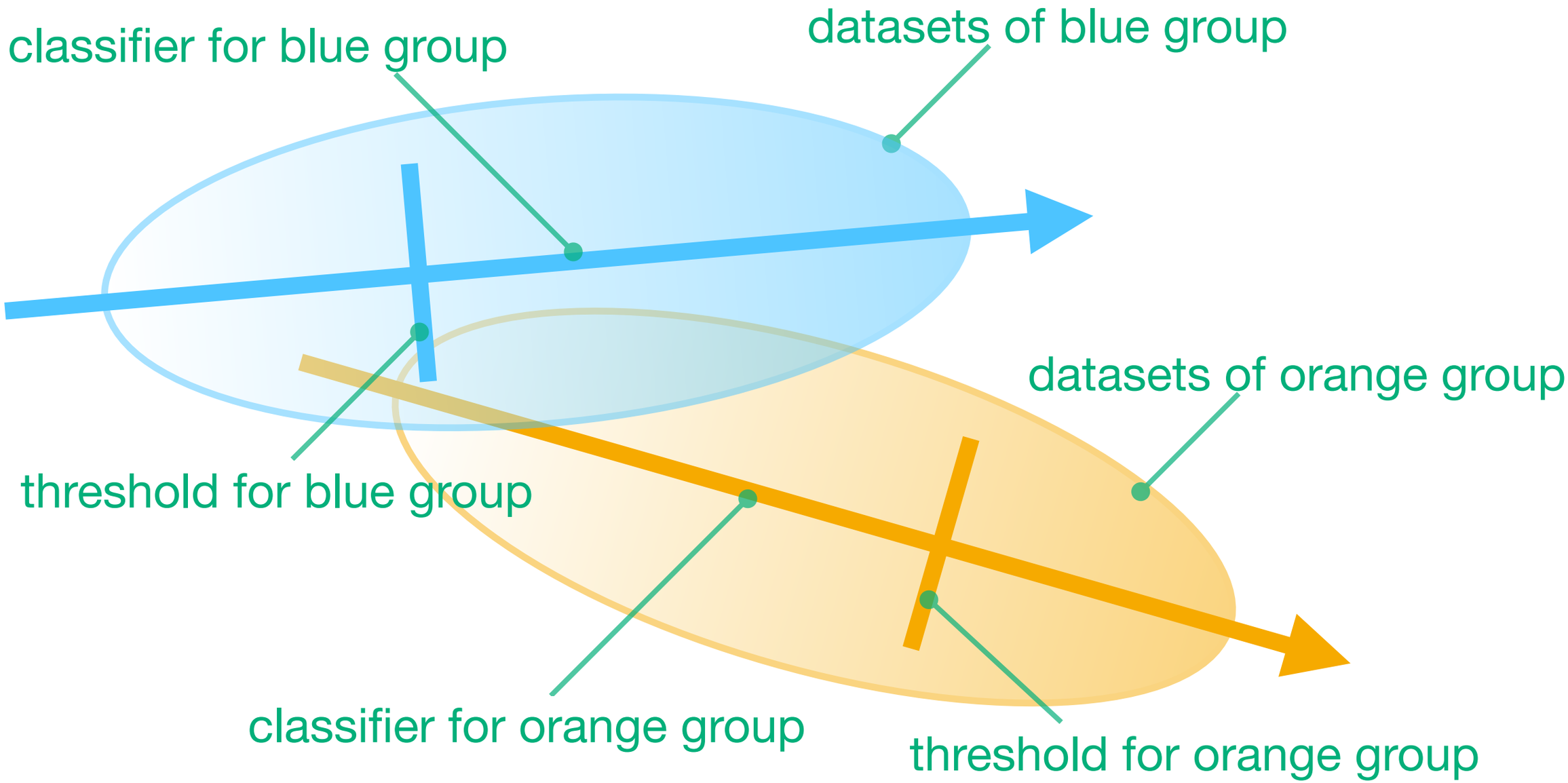
However, it was the beginning of the long winding road

# β1 Model




**This model was not converged**

# β2 Model




**This model was converged,**

**but it was unfair than an original classifier**

“Fairness-aware Learning through Regularization Approach” , IEEE International Workshop on Privacy Aspects of Data Mining, 2011

# ECMLPKDD2012 model

The β2 model was unfair than an original model

↓

The objective function is non-convex, and its local minimas were bad

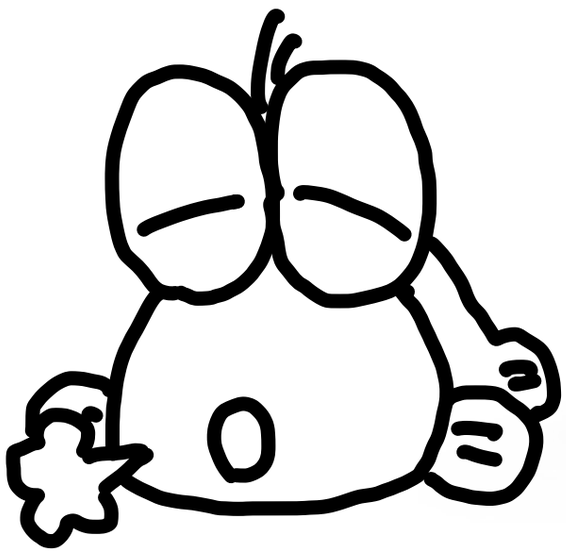
↓

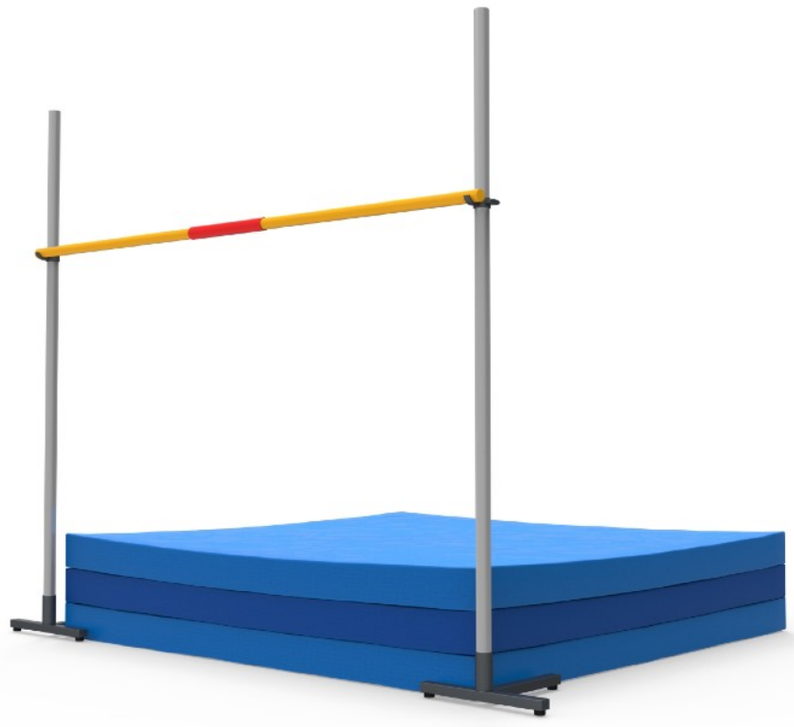
**Initializing classifiers without constraint terms**

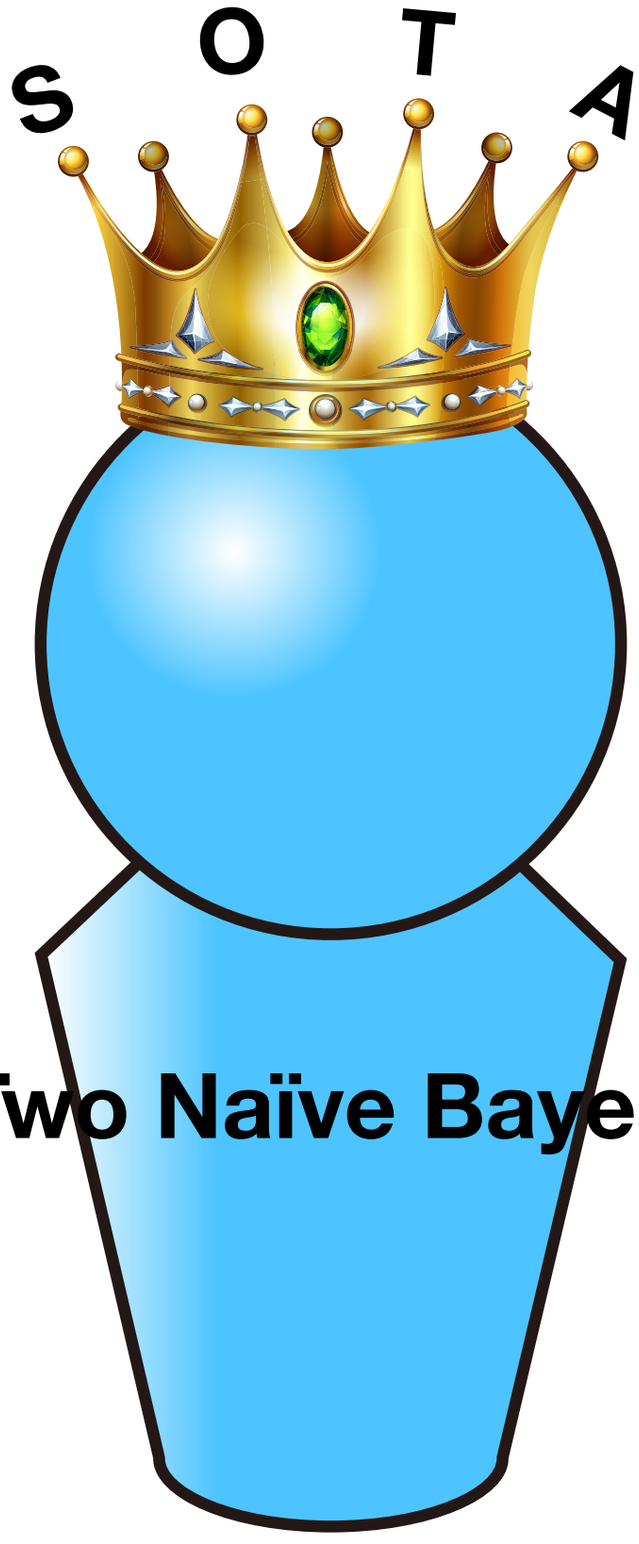
↓

This model was fairer than an original classifier

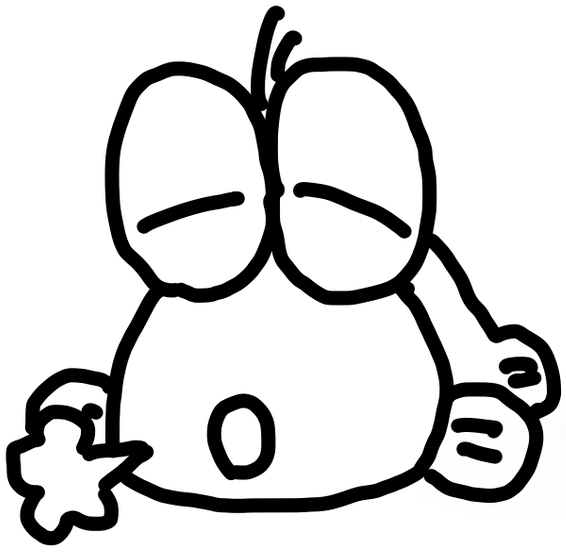
“Fairness-aware Classifier with Prejudice Remover Regularizer” , ECMLPKDD, 2012

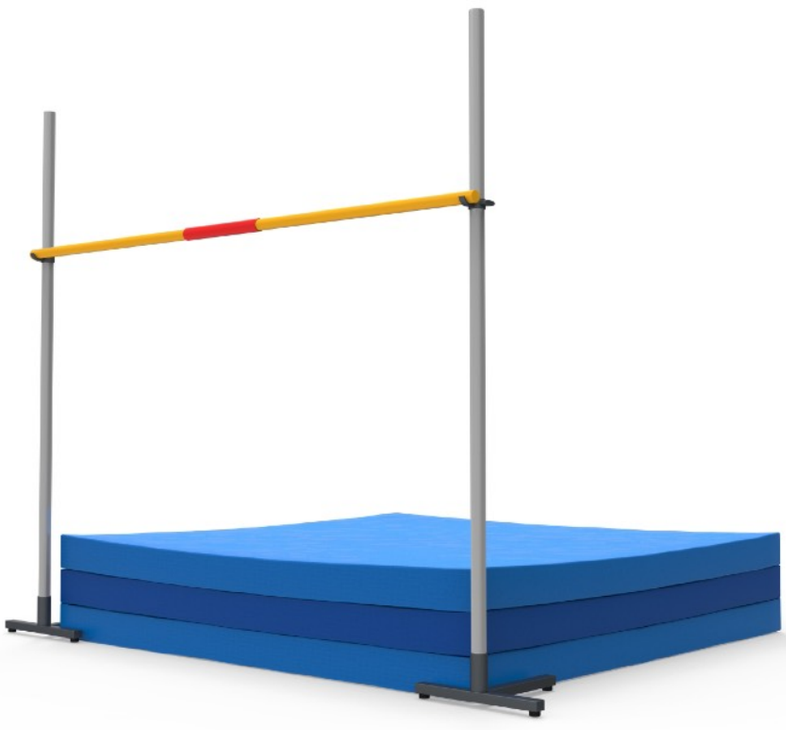






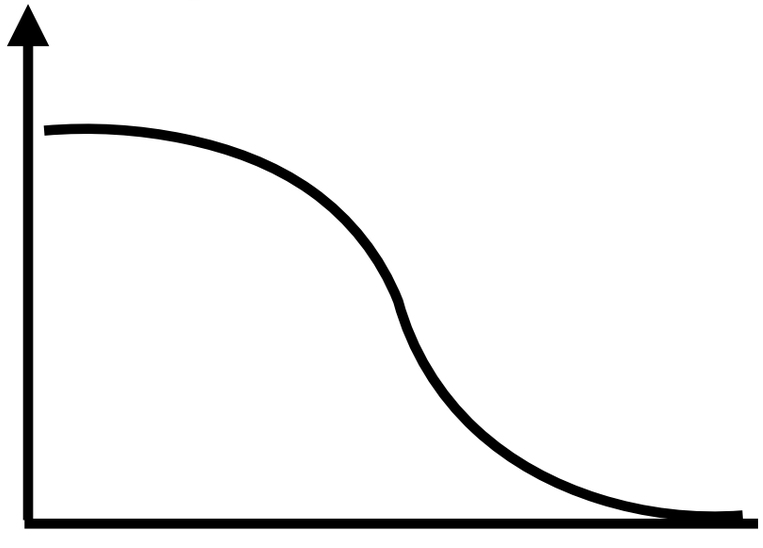

**Two Naïve Bayes**





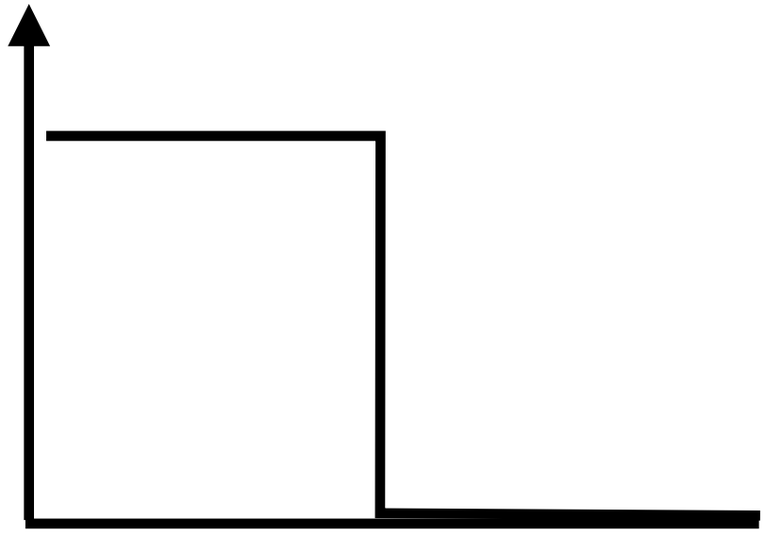
# Actual Independence

**probabilistic**



$$\hat{Y} \sim \Pr[\hat{Y} = 1 \mid f(\mathbf{X})]$$

**deterministic**



$$\hat{Y} = 1 \text{ if } \Pr[\hat{Y} = 1 \mid f(\mathbf{X})] \geq 0.5$$

**Actual Independence:** Class labels are deterministically generated by applying a decision rule

**effectiveness improvement**

"Model-based and Actual Independence for Fairness-aware Classification" , Data Mining and Knowledge Discovery, 2018

# Fair Recommendation

## Prediction Function

**a prediction function is selected according to a sensitive value**

$$\hat{r}(x, y, s) = \mu^{(s)} + b_x^{(s)} + c_y^{(s)} + \mathbf{p}_x^{(s)} \mathbf{q}_y^{(s)\top}$$

sensitive feature

## Objective Function

**independence parameter:** control the balance between the independence and accuracy

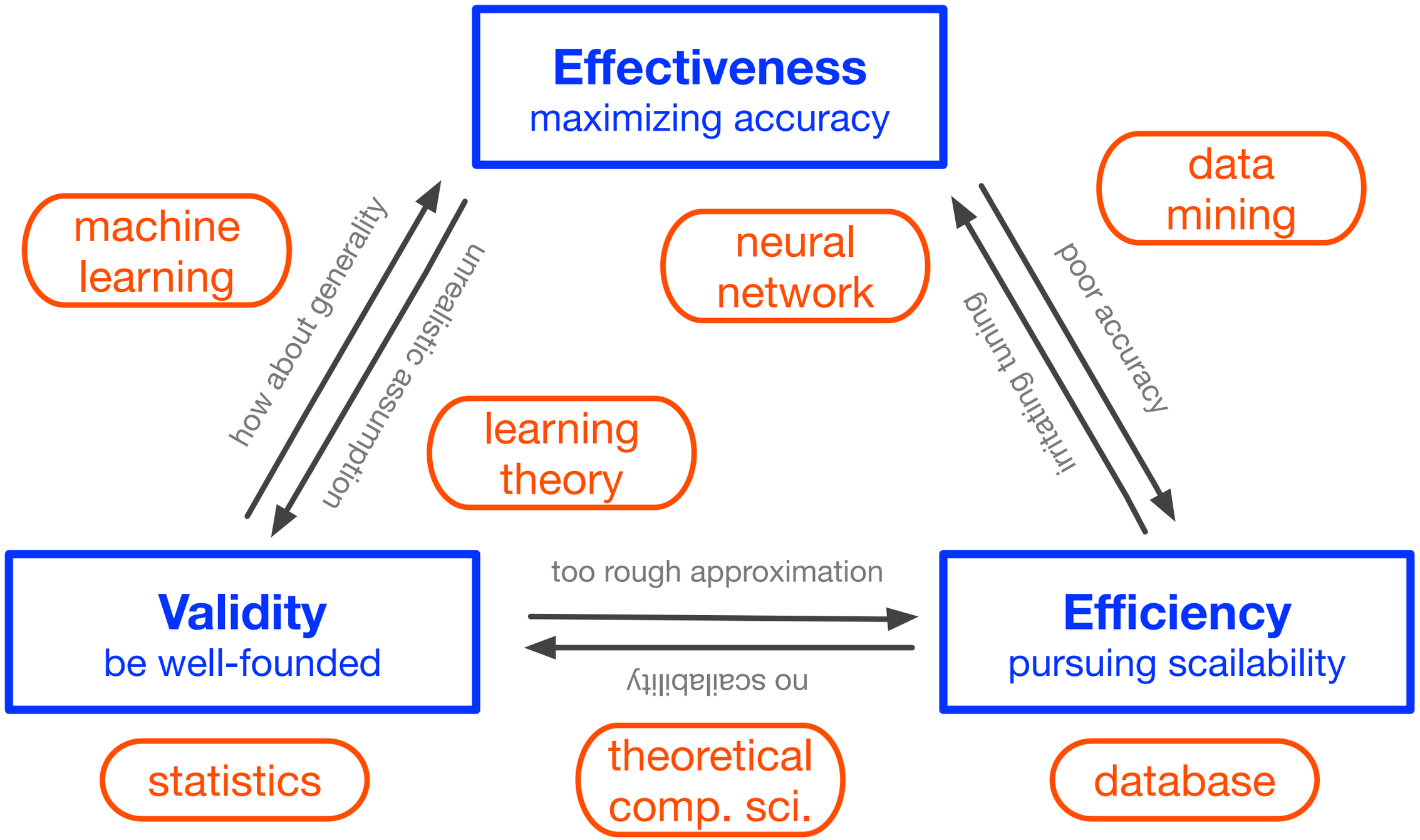
$$\sum_D (r_i - \hat{r}(x_i, y_i))^2 - \eta \, \mathrm{indep}(R, S) + \lambda \, \|\Theta\|^2$$


“Enhancement of the Neutrality in Recommendation” , The 2nd Workshop on Human Decision Making in Recommender Systems, 2012

“Efficiency Improvement of Neutrality-enhanced Recommendation” , The 3rd Workshop on Human Decision Making in Recommender Systems, 2013

“Recommendation Independence” , Proc of the Conf. on Fairness, Accountability and Transparency, 2018

# Interests of Data Science Communities



My survey slide about Fairness-Aware Machine Learning is available at:

**https://www.kamishima.net/faml/**


## Acknowledgements

- We'd like to be appreciate the colleagues in our fairness-aware ML research, Hiroshi Nakagawa, Hiromi Arai, Kazuto Fukuchi, Shoko Ema, and Junpei Komiyama.
- We greatly respect the pioneering works of the fairness-aware ML by the European community, Toon Calders, Dino Pedreschi, Indrė Žliobaitė, Salvatore Ruggieri, Faisal Kamiran, and Sara Hajian.
- We'd like to express our thanks for the cooperation of the fair RecSys community: Michael Ekstrand, Robin Burke, Pierre-Nicolas Schwab, Jean Garcia-Gathright, Nasim Sonboli, Luca Belli, and Amifa Raj.
- We also thank to the FAccT community, Suresh Venkatasubramanian, Sorelle Friedler, and Solon Barocas, Krishna Gummadi, and Md. Bilal Zafar.
- Thanks for the support of the JSPS KAKENHI 科研費 KAKENHI Grant Number 24500194, 15K00327, 18H03300, and 21H03504, and of our affiliations; National Institute of Advanced Industrial Science and Technology (AIST) and University of Tsukuba.