University of Amsterdam
Korteweg de Vries Institute for Mathematics

Theoretische beperkingen van Explainability methodes

Ministerie van Justitie en Veiligheid

2024-03-28

Programma

- Kleine introductie van mijzelf en mijn onderzoek
- Introductie van Explainable AI (XAI)

- Attribution methods en een onmogelijkheidheidsresultaat
- Het risico van counterfactual methods

- Conclusie met wat praktische implicaties

Korte introductie

Korte introductie

- Promovendus aan de UvA: *Het formaliseren van Explainable AI*
- Hiervoor, MSc in Wiskunde & gewerkt als werkstudent bij een Data Science Consultancy
- Al het werk in deze presentatie is gecreëerd in samenwerking met

Dr. Tim van Erven
Universiteit van Amsterdam

Dr. Rianne de Heide
Vrije Universiteit

Dr. Damien Garreau
Universiteit van Würzberg

Explainable Artificial Intellegence (XAI)

Roep voor XAI

Een paar redenen

- Fairness: Biases in je modellen zouden eerder gedetecteerd kunnen worden
- Kan Vertrouwen verhogen
- Kan Betrouwbaarheid verbeteren
- Regulering

Uitleg methodes

Explosie

Methods
CAM with global average pooling [42], [91]
+ Grad-CAM [43] generalizes CAM, utilizing gradient
+ Guided Grad-CAM and Feature Occlusion [68]
+ Respond CAM [44]
+ Multi-layer CAM [92]
LRP (Layer-wise Relevance Propagation) [13], [53]
+ Image classifications. PASCAL VOC 2009 etc [45]
+ Audio classification. AudioMNIST [47]
+ LRP on DeepLight. fMRI data from Human Connectome Project [48]
+ LRP on CNN and on BoW(bag of words)/SVM [49]
+ LRP on compressed domain action recognition algorithm [50]
+ LRP on video deep learning, *selective relevance method* [52]
+ BiLRP [51]
DeepLIFT [57]
Prediction Difference Analysis [58]
Slot Activation Vectors [41]
PRM (Peak Response Mapping) [59]
LIME (Local Interpretable Model-agnostic Explanations) [14]
+ MUSE with LIME [85]
+ Guidelinebased Additive eXplanation optimizes complexity, similar to LIME [93]
Also listed elsewhere: [56], [69], [71], [94]
Others. Also listed elsewhere: [95]
+ Direct output labels. Training NN via multiple instance learning [65]
+ Image corruption and testing Region of Interest statistically [66]
+ Attention map with autofocus convolutional layer [67]
DeconvNet [72]
Inverting representation with natural image prior [73]
Inversion using CNN [74]
Guided backpropagation [75], [91]
Activation maximization/optimization [38]
+ Activation maximization on DBN (Deep Belief Network) [76]
+ Activation maximization, multifaceted feature visualization [77]
Visualization via regularized optimization [78]
Semantic dictionary [39]
Network dissection [36], [37]
Decision trees
Propositional logic, rule-based [82]
Sparse decision list [83]
Decision sets, rule sets [84], [85]
Encoder-generator framework [86]
Filter Attribute Probability Density Function [87]
MUSE (Model Understanding through Subspace Explanations) [85]

(2019) A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI

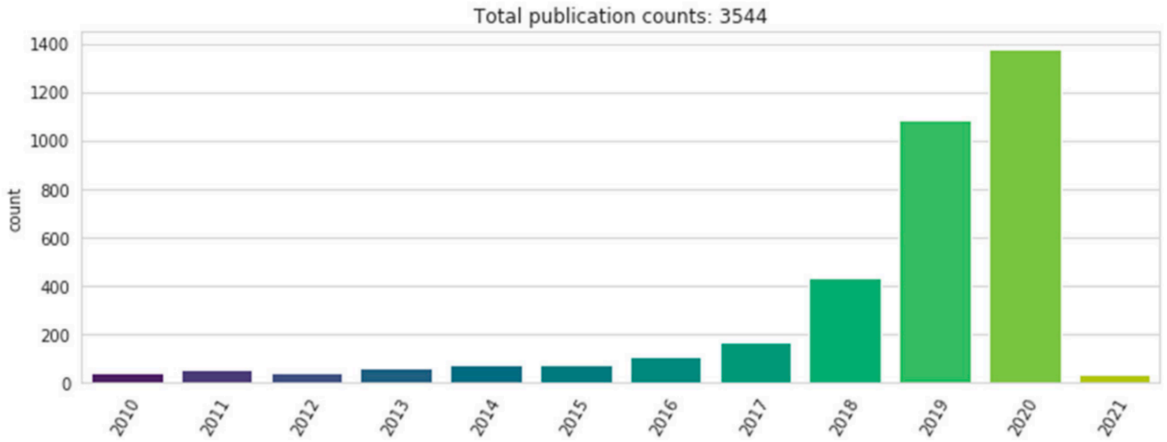
(2014.03) SEDC [129]
(2015.08) OAE [51]
(2016.05) HCLS [110, 112]
(2017.06) Feature Tweaking [186]
(2017.11) CF Expl. [196]
(2017.12) Growing Spheres [114]
(2018.02) CEM [55]
(2018.02) POLARIS [209]
(2018.05) LORE [80]
(2018.06) Local Foil Trees [190]
(2018.09) Actionable Recourse [189]
(2018.11) Weighted CFs [77]
(2019.01) Efficient Search [175]
(2019.04) CF Visual Expl. [76]
(2019.05) MACE [99]
(2019.05) DiCE [145]
(2019.05) CERTIFAI [179]
(2019.06) MACEM [56]
(2019.06) Expl. using SHAP [165]
(2019.07) Nearest Observable [201]
(2019.07) Guided Prototypes [191]
(2019.07) REVISE [95]
(2019.08) CLEAR [202]
(2019.08) MC-BRP [123]
(2019.09) FACE [162]
(2019.09) Equalizing Recourse [83]
(2019.10) Action Sequences [163]
(2019.10) C-CHVAE [156]
(2019.11) FOCUS [124]
(2019.12) Model-based CFs [127]
(2019.12) LIME-C/SHAP-C [164]
(2019.12) EMAP [41]
(2019.12) PRINCE [71]
(2019.12) LowProFool [18]
(2020.01) ABELE [79]
(2020.01) SHAP-based CFs [66]
(2020.02) CEML [11–13]
(2020.02) MINT [100]
(2020.03) ViCE [74]
(2020.03) Plausible CFs [22]
(2020.04) SEDC-T [193]
(2020.04) MOC [52]
(2020.04) SCOUT [199]
(2020.04) ASP-based CFs [28]
(2020.05) CBR-based CFs [103]
(2020.06) Survival Model CFs [106]
(2020.06) Probabilistic Recourse [101]
(2020.06) C-CHVAE [155]
(2020.07) FRACE [210]
(2020.07) DACE [96]
(2020.07) CRUDS [60]
(2020.07) Gradient Boosted CFs [5]
(2020.08) Gradual Construction [97]
(2020.08) DECE [44]
(2020.08) Time Series CFs [16]
(2020.08) PermuteAttack [87]
(2020.10) Fair Causal Recourse [195]
(2020.10) Recourse Summaries [167]
(2020.10) Strategic Recourse [43]
(2020.11) PARE [172]

(2020) A survey of algorithmic recourse: definitions, formulations, solutions, and prospects

Methods	H
Linear probe [101]	
Regression based on CNN [106]	
Backwards model for interpretability of linear models [107]	
GDM (Generative Discriminative Models): ridge regression + least square [100]	
GAM, GA^{2}M (Generative Additive Model) [82], [102], [103]	
ProtoAttend [105]	
Other content-subject-specific models:	N
+ Kinetic model for CBF (cerebral blood flow) [131]	N
+ CNN for PK (Pharmacokinetic) modelling [132]	N
+ CNN for brain midline shift detection [133]	N
+ Group-driven RL (reinforcement learning) on personalized healthcare [134]	N
+ Also see [108]–[112]	N
PCA (Principal Components Analysis), SVD (Singular Value Decomposition)	N
CCA (Canonical Correlation Analysis) [113]	
SVCCA (Singular Vector Canonical Correlation Analysis) [97] = CCA+SVD	
F-SVD (Frame Singular Value Decomposition) [114] on electromyography data	
DWT (Discrete Wavelet Transform) + Neural Network [135]	
MODWPT (Maximal Overlap Discrete Wavelet Package Transform) [136]	
GAN-based Multi-stage PCA [118]	
Estimating probability density with deep feature embedding [119]	
t-SNE (t-Distributed Stochastic Neighbour Embedding) [77]	
+ t-SNE on CNN [120]	
+ t-SNE, activation atlas on GoogleNet [121]	
+ t-SNE on latent space in meta-material design [122]	
+ t-SNE on genetic data [137]	
+ mm-t-SNE on phenotype grouping [138]	
Laplacian Eigenmaps visualization for Deep Generative Model [124]	
KNN (k-nearest neighbour) on multi-center low-rank rep. learning (MCLRR) [125]	
KNN with triplet loss and *query-result activation map pair* [139]	
Group-based Interpretable NN with RW-based Graph Convolutional Layer [123]	
TCAV (Testing with Concept Activation Vectors) [96]	
+ RCV (Regression Concept Vectors) uses TCAV with Br score [140]	
+ Concept Vectors with UBS [141]	
+ ACE (Automatic Concept-based Explanations) [56] uses TCAV	
Influence function [129] helps understand adversarial training points	
Representer theorem [130]	
SocRat (Structured-output Causual Rationalizer) [127]	
Meta-predictors [126]	
Explanation vector [128]	
# Also listed elsewhere: [14], [43], [85], [94]	N
# Also listed elsewhere: [14], [60], [85] etc	N
CNN with separable model [142]	
Information theoretic: Information Bottleneck [98], [99]	
Database of methods v.s. interpretability [10]	N
Case-Based Reasoning [143]	
Integrated Gradients [69], [94]	

(2019) A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI

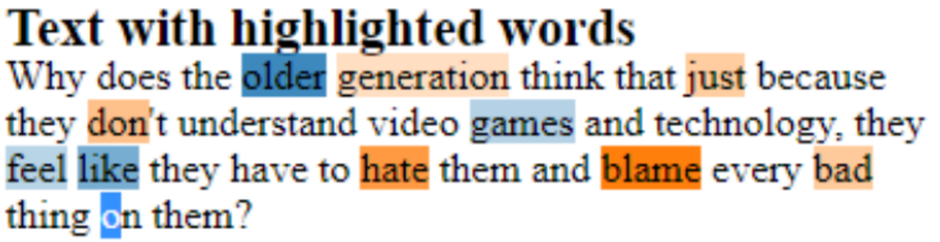
(2021) Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics



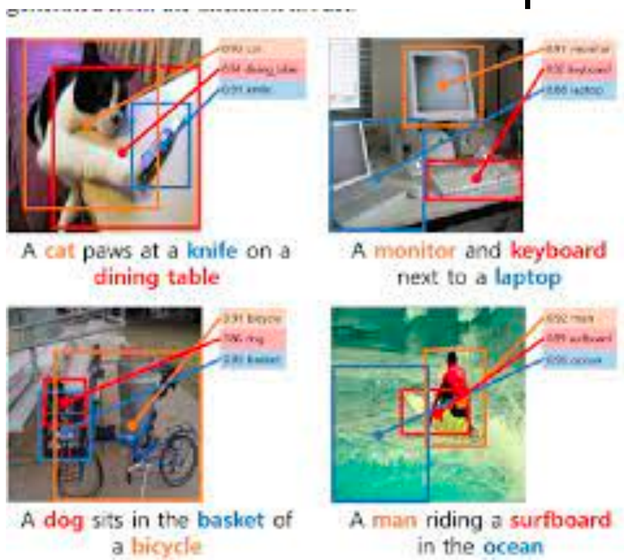
Uitleg methodes

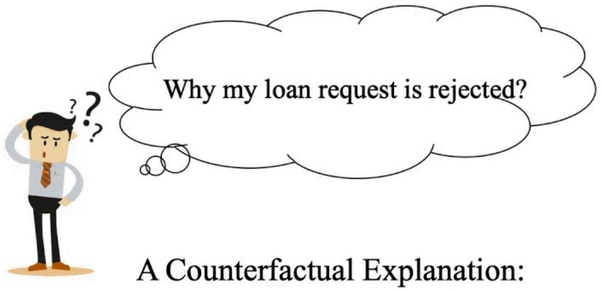
Voorbeelden

- Licht belangrijke features uit
- Counterfactuals



- Onderschrift generatie
- Prototype voorbeelden
- Netwerk probing

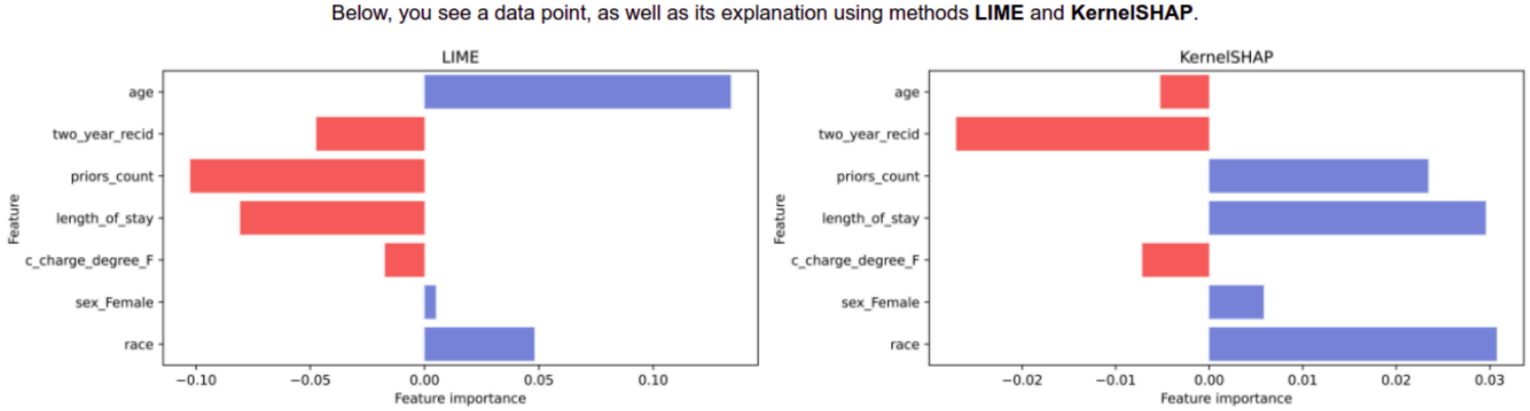




Uitleg methodes

Een paar problemen

- Makkelijk te manipuleren
- Methodes zijn het oneens met elkaar
- Plausibele uitleggen kunnen oneerlijk zijn over wat het model daadwerkelijk doet



Uitleg methodes

Roep om formele analyses

- Mythos of model interpretability, [Lipton, 2017]
- Towards a rigorous science of interpretable machine learning, [Doshi-Velez, Kim, 2017]

“Interpretability research suffers from an over-reliance on intuition-based approaches that risk—and in some cases have caused—illusory progress and misleading conclusions”,

[Leavitt, Morcos, 2020]

Attribution methods &

Counterfactual methods

Setting

Lopend voorbeeld

2 partijen:

- Krediet lening aanvrager (A)

- Krediet lening verstrekker (V)

Geautomatiseerde aanmeldingsprocedure:

- (A) voorziet (V) van een lijst met features:

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

- (V) heeft een model voor kredietwaardigheid f

$$f(x) = +1 \quad \text{als aangenomen}$$
$$f(x) = -1 \quad \text{als afgewezen}$$

- (A) krijgt de beslissing en een uitleg van (P)

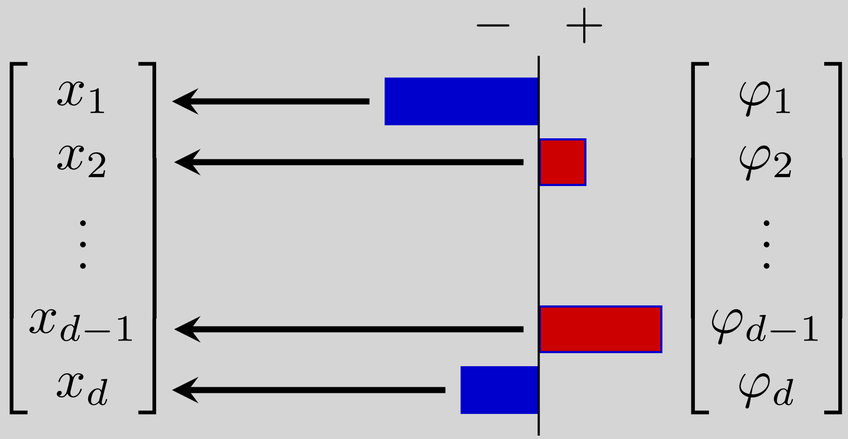
Attribution method

Licht uit hoe belangrijk bepaalde features zijn:

- Positieve waardes betekenen een positieve correlatie tussen de feature en de uitkomst

- Negatieve waardes betekenen een negatieve correlatie tussen de feature en de uitkomst

Niet altijd de correcte interpretatie!



"Uw inkomen van € 40 000 droeg positief bij.

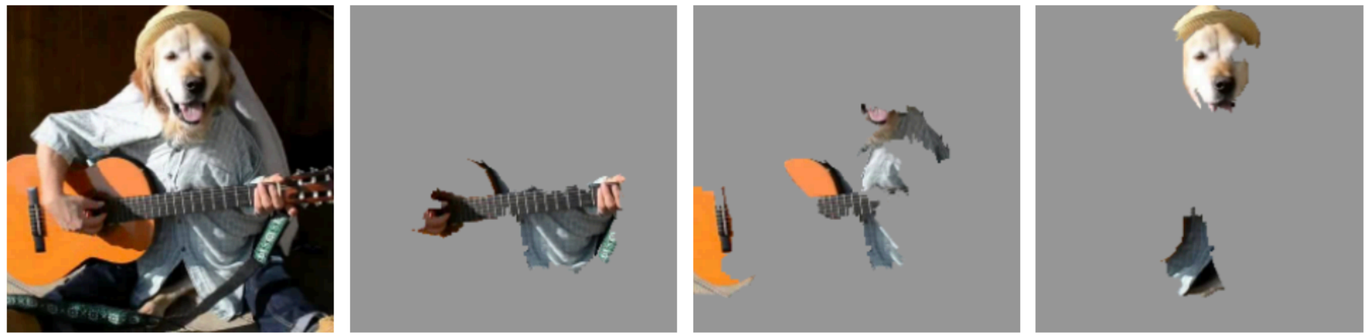
Echter, Uw 5 verschillende credit cards droegen negatief bij aan uw aanmelding."

Attribution methods

Voorbeelden

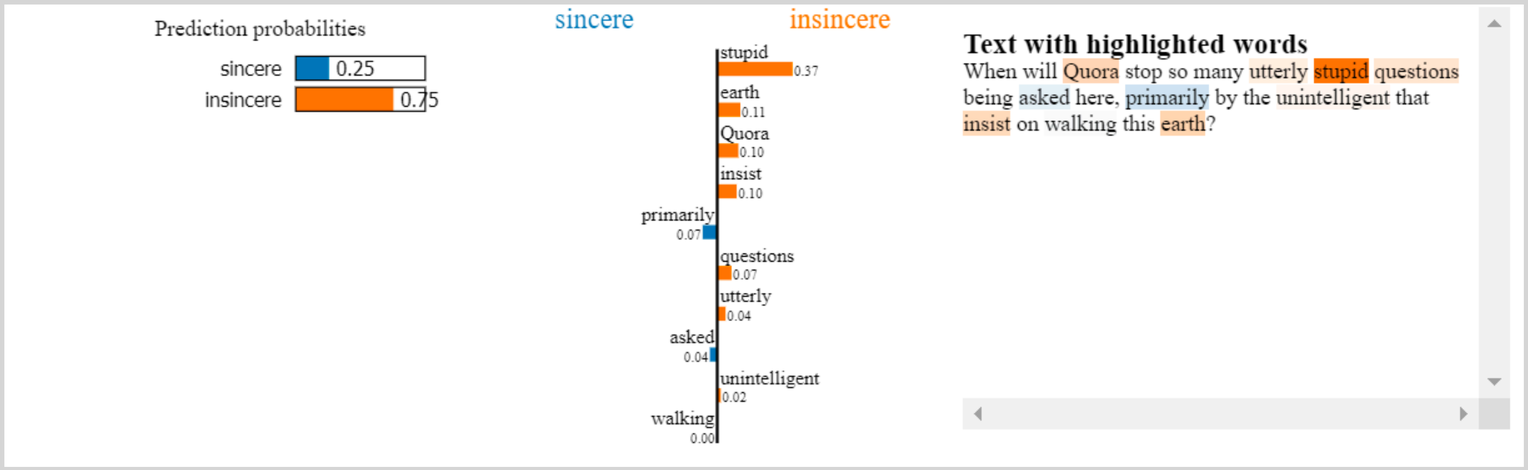
- LIME
- SHAP
- …

Pictures



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

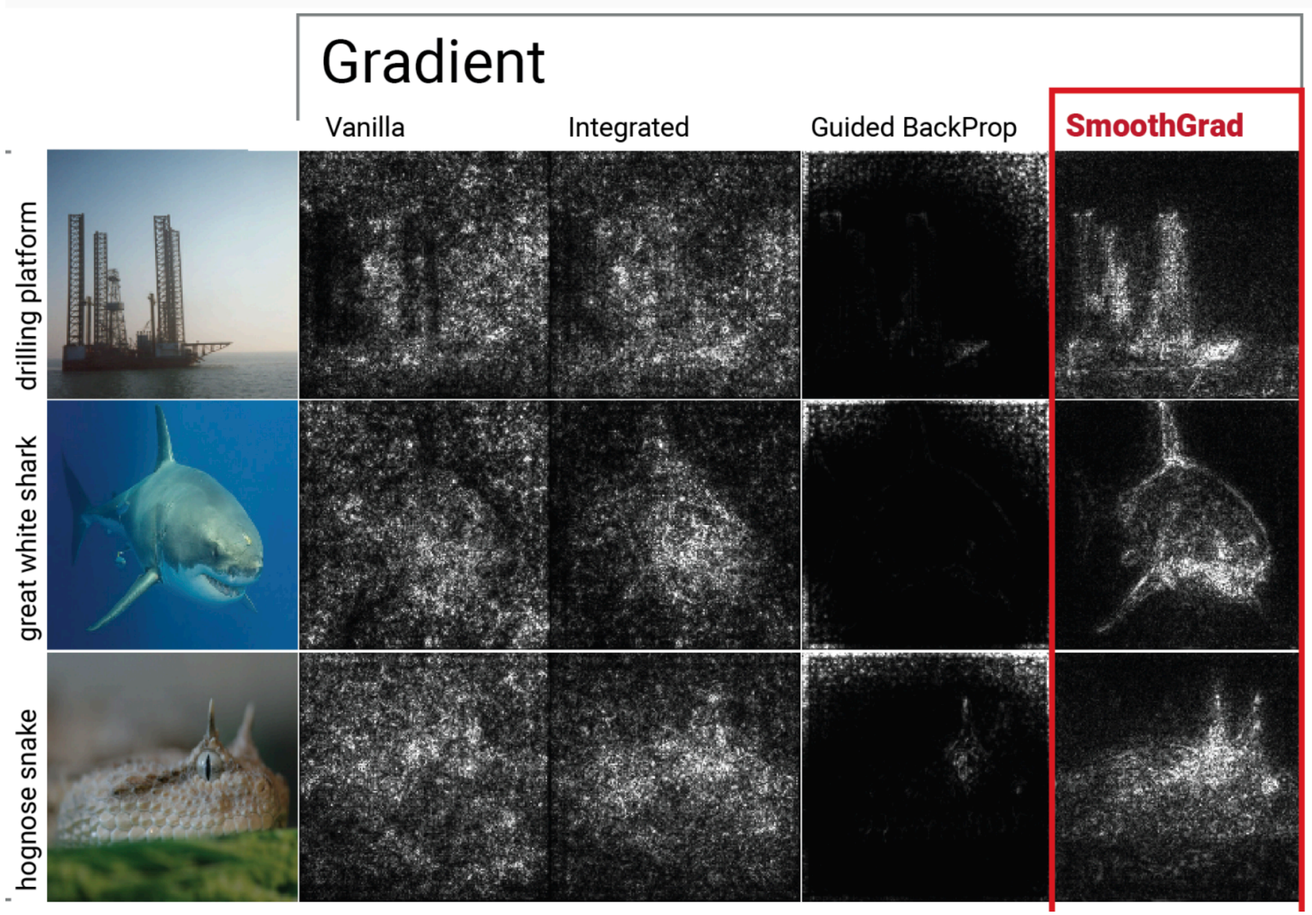
Text



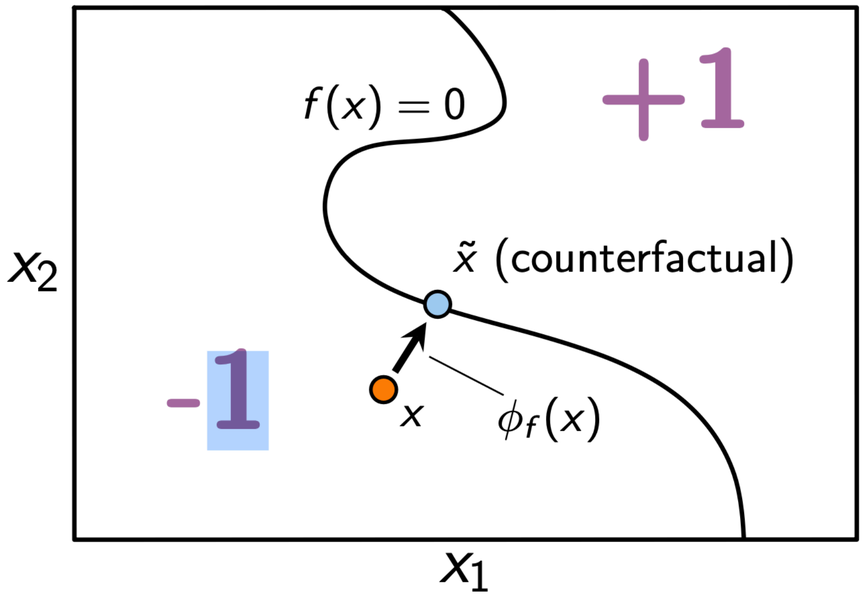
Attribution methods

Voorbeelden

- Grad-Cam
- SmoothGrad
- Integrated Gradients
- …



Counterfactual method



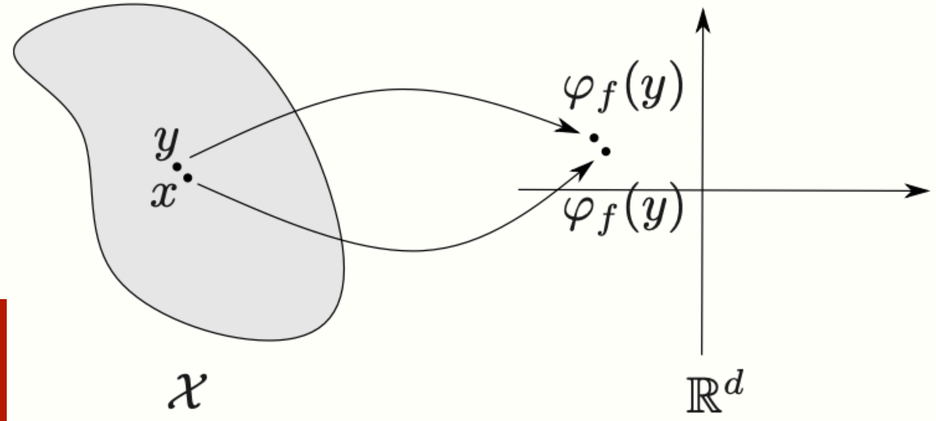
- Vertel (A) hoe de beslissing veranderd kan worden van -1 naar +1
- Minimale moeite voor (A)
- Verstrekt "***Recourse***"

"Als u een inkomen zou hebben van € 40 000 in plaats van €35 000,
dan zou je leningsaanvraag geaccepteerd zijn."

Attribution methods de voorzien in recourse kunnen niet robust zijn

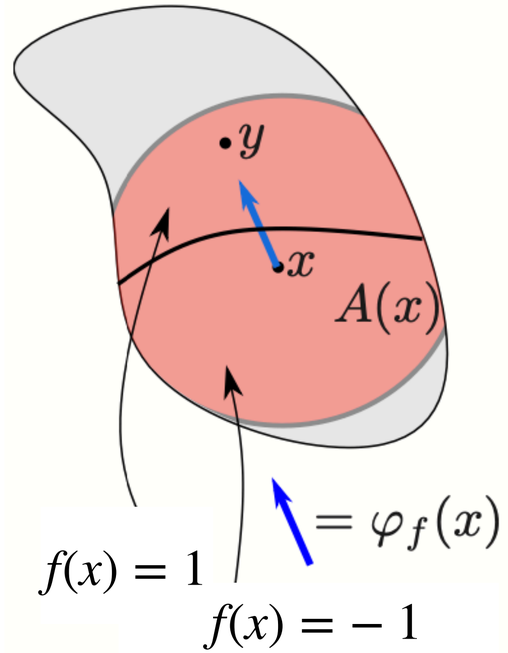
Wenselijke Eigenschappen

Attribution methods



Er zijn ten minste 2 wenselijke eigenschappen:

- Een uitleg moet ***robuust*** zijn*:*
  - Als 2 inputs vergelijkbaar zijn, moet de uitleg ook vergelijkbaar zijn
- Een uitleg zou kunnen dienen als counterfactual **(Recourse gevoelig):**
  - Positieve attribution -> verhoog die waarde
  - Negatieve attribution -> verlaag die waarde
  - Doel: verander de beslissing van -1 to 1



Lening Aanvrager Voorbeeld

Robustheid

Features:

- Inkomen: 40.000
- Creditcards: 5
- …
- Gender: Female

Features:

- Inkomen: 40.000
- Creditcards: 5
- …
- Gender: Male

“Uw inkomen van € 40 000 droeg positief bij.

Echter, uw 5 verschillende creditcards droegen negatief bij.”

Loan Applicant Example

Recourse sensitivity

“Uw inkomen van € 40 000 droeg positief bij.

Echter, uw 5 verschillende creditcards droegen negatief bij.”

“Als ik het aantal creditcards verlaag, krijg ik waarschijnlijk een lening!”

Onmogelijkheidsresultaat

Attribution methods kunnen niet altijd

- Robuust zijn
- Recourse verschaffen

Onmogelijkheidsresultaat

Specifiek:

- Stel iemand ontwikkelt een attribution method
- Dan kan ik een model construeren waar,
  - **Robuustheid** faalt voor die methode,
  - Of **Recourse sensitivity** faalt,

Attribution methods kunnen niet altijd

Robuust zijn

Recourse verschaffen

Er is wat hoop

Er zijn modellen f, die

- Wel attributions toelaten
- Met de eigenschappen
  - **Robuustheid**,
  - **Recourse Sensitivity**,

- Lineaire modellen
- Monotone modellen

Over het algemeen,

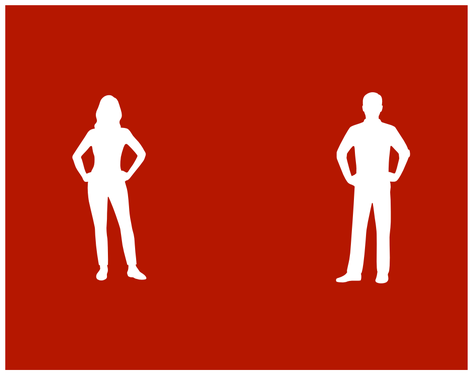
- Moeilijk van te voren te controleren,
- Erg makkelijk te breken

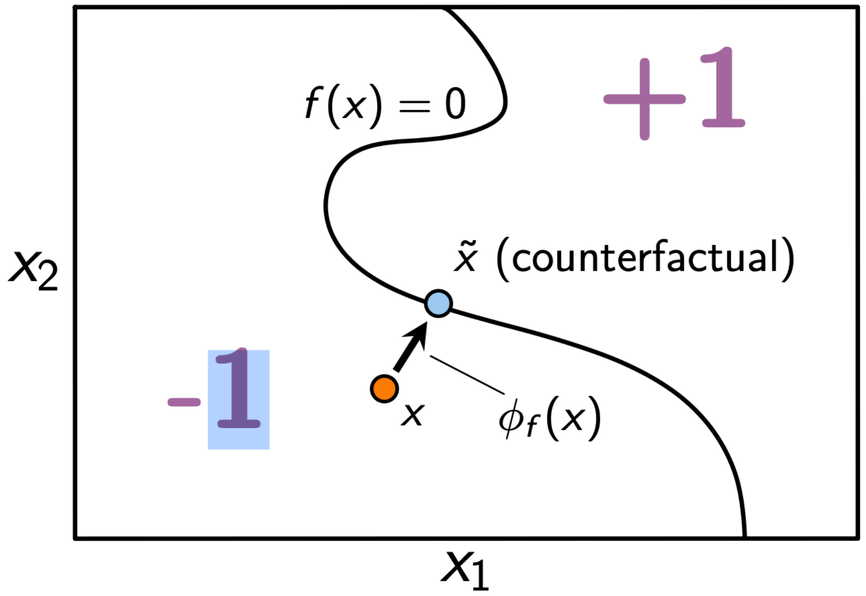
Het risico van Recourse

Counterfactual methods

- We vergeten attribution methods even
- Bekijk Counterfactual Explanations

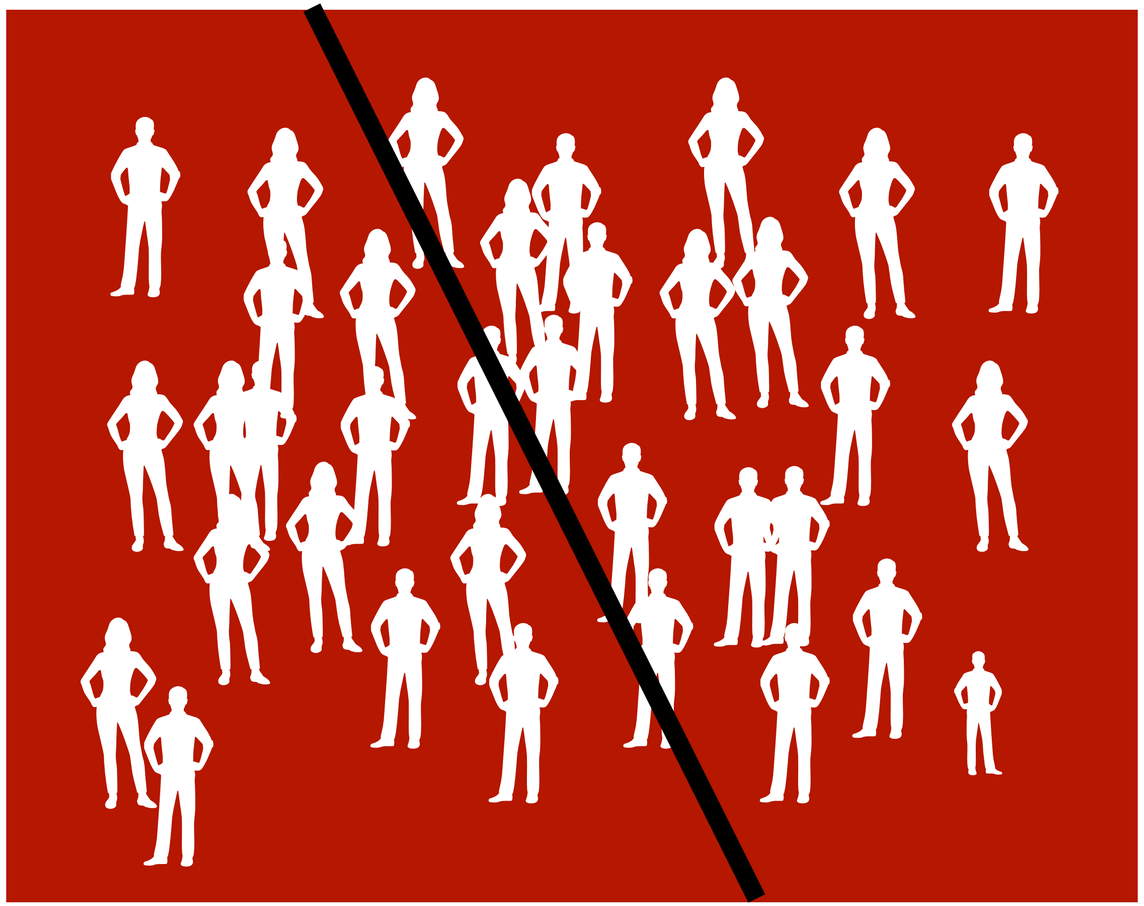
- Wat gebeurt er als we naar de gehele populatie kijken





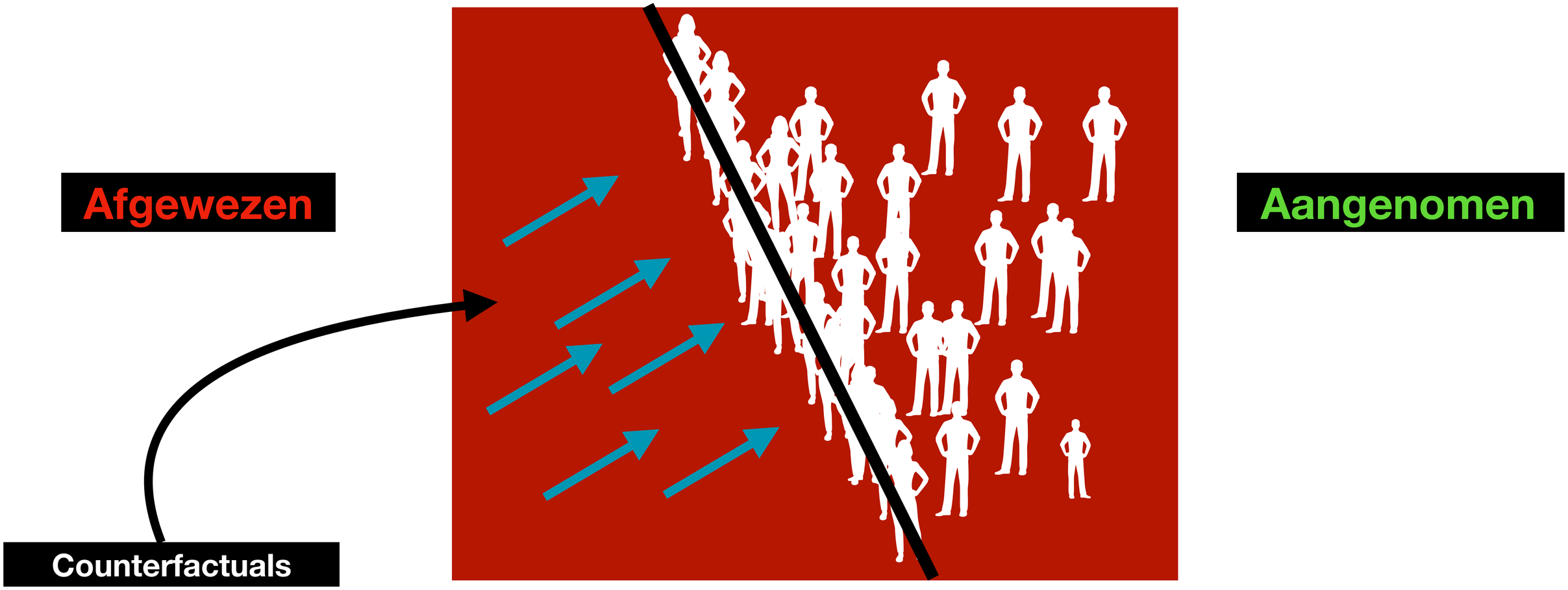
Counterfactual methods

Afgewezen



Aangenomen

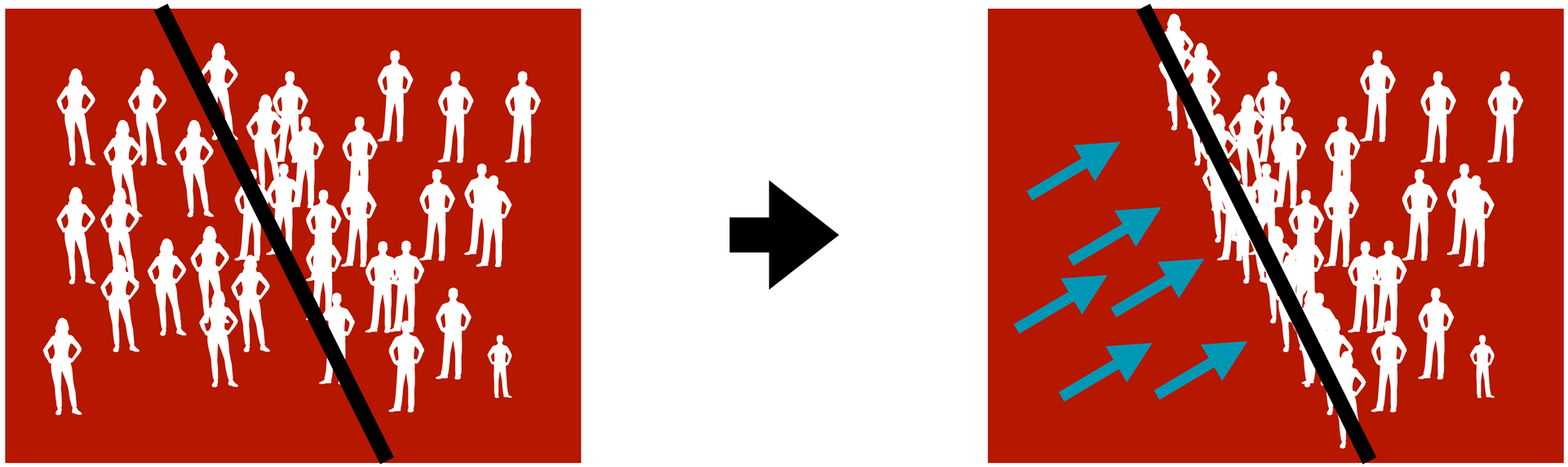
Counterfactual methods



Het risico van Recourse

- Wat gebeurt er als we naar de gehele populatie kijken?
  - Onderliggende data verdeling verandert!
  - Modeleer aannames gelden niet meer!

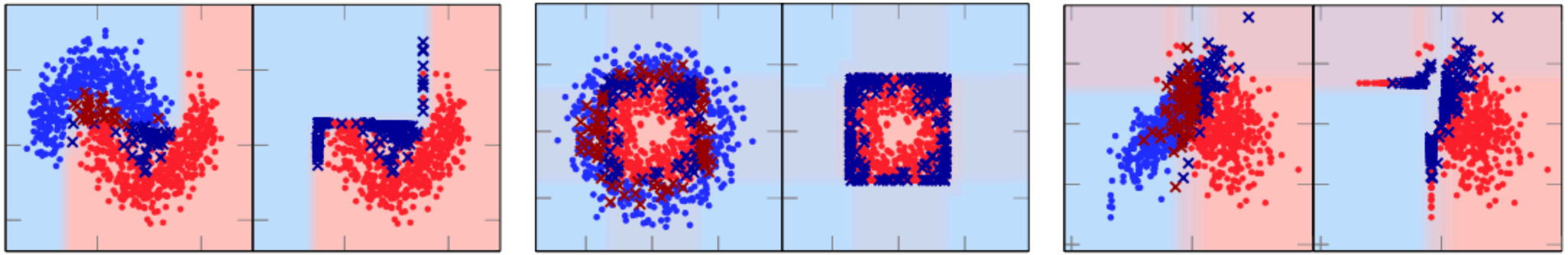
Nauwkeurigheid zal dalen!



The risk of recourse

- Wat gebeurt er als we naar de gehele populatie kijken?
  - Onderliggende data verdeling verandert!
  - Modeleer aannames gelden niet meer!

Nauwkeurigheid zal dalen!



Het risico van Recourse

Strategieën tegen dit fenomeen

Kan (V) zich beschermen tegen deze daling in nauwkeurigheid?

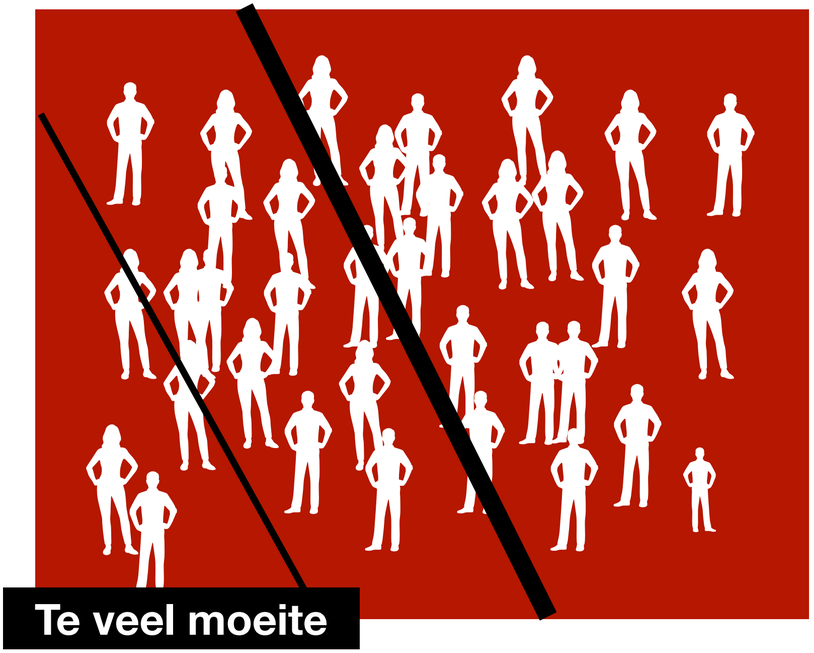
- We moeten ten minste aannemen dat niet iedereen een uitleg krijgt.

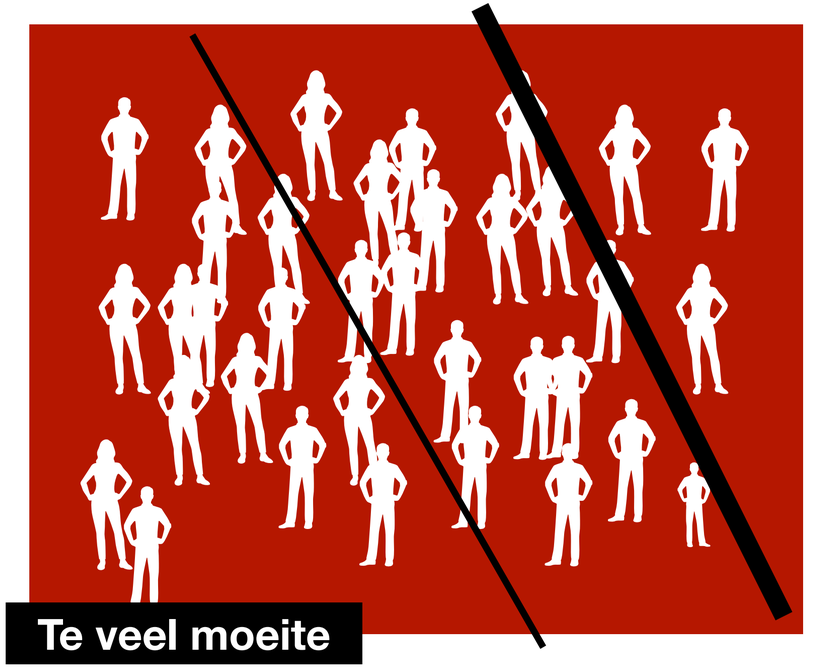
Ja en Nee

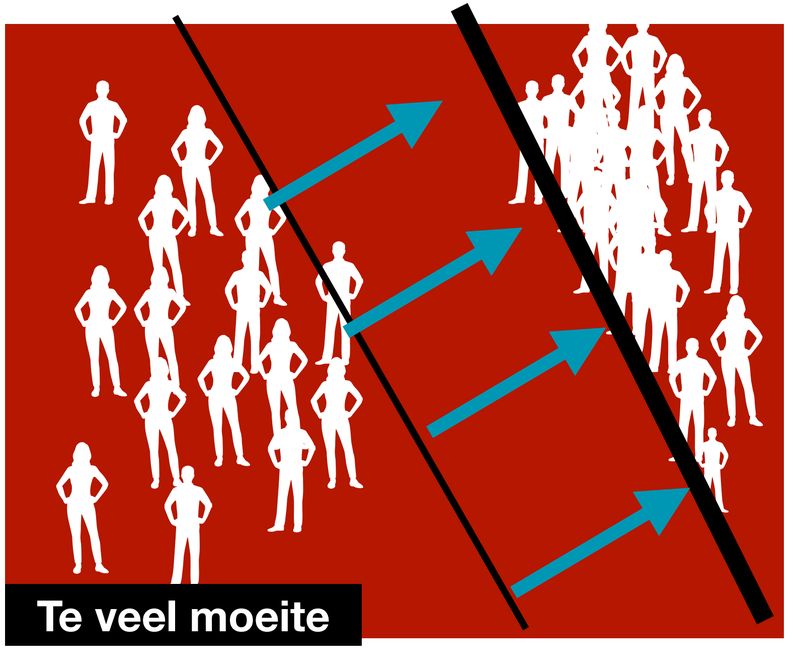
Te veel moeite

Het risico van Recourse

Strategieën tegen dit fenomeen







Exact dezelfde mensen krijgen een lening

Veel meer moeite

Praktische implicaties

Praktische implicaties

- Definieer een duidelijk doel van je uitleg, kies daarna een methode.
- Bij het gebruik van attribution methods, wees voorzichtig met het interpreteren van de score.
- Bij het gebruik van counterfactual methodes, stel de vraag
  - “Hoe pijnlijk/gevaarlijk is een misclassificatie?”
- Als het mogelijk is gebruik simpele/interpreteerbare modellen, verhoog de complexiteit alleen als dat nodig is!

Referenties

- *Mythos of model interpretability*, [Lipton, 2017]
- *Towards a rigorous science of interpretable machine learning*, [Doshi-Velez, Kim, 2017]
- *Towards falsifiable interpretability research* [Leavitt, Morcos, 2020]
- *Attribution-based Explanations that Provide Recourse Cannot be Robust* [F, De Heide, van Erven, 2022]
- *The Risks of Recourse in Binary Classification* [F, Garreau, van Erven, 2023]

Bedankt voor jullie aandacht