University of Amsterdam
Korteweg de Vries Institute for Mathematics

Theoretical limitations of Explainability methods

Xomnia Data & Drinks

2023-09-14

Programme for today

- Introduce myself and my research
- Brief XAI introduction

- Attribution methods that provide recourse and are robust cannot exist
- The risk of counterfactual explanations

- Conclusion

Short introduction

Short Introduction

- PhD student at the UvA: *Formalising Explainable AI*
- Previously, MSc in Mathematics & worked at Amsterdam Data Collective

- All work presented was created in collaboration with:

Dr. Tim van Erven

Dr. Rianne de Heide

Dr. Damien Garreau

Explainable Artificial Intelligence

Setting

Leading example

2 parties:

- Credit Loan Applicant (A)

- Credit Loan Provider (P)

Loan application process:

- (A) provides (P) with a set of features:

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

- (P) has an automated decision system f

$$f(x) = 1 \quad \text{if accepted}$$
$$f(x) = -1 \quad \text{if not}$$

- (A) get decision and an explanation from (P)

Call for XAI

Some Reasons

- Fairness: Biases can be detected earlier
- Trustworthiness
- Increases reliability
- Regulation

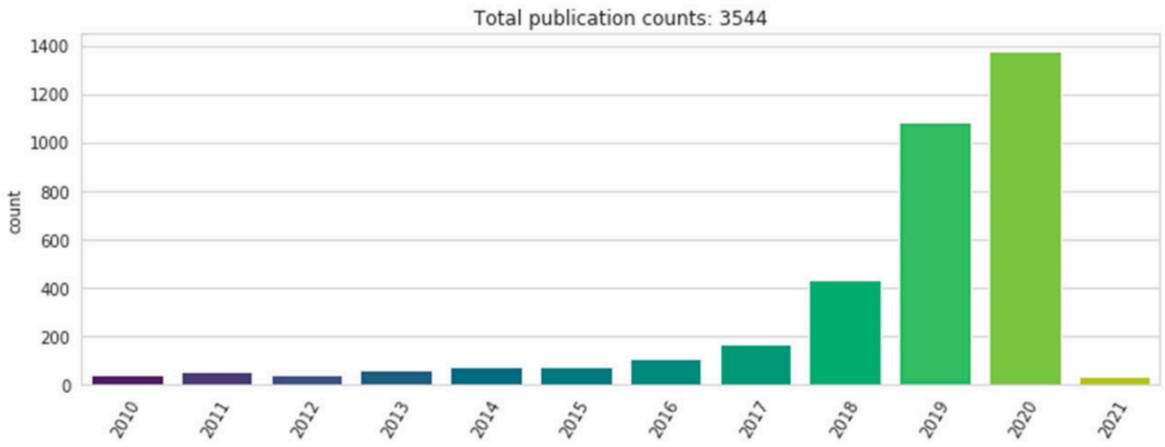
Explanations

Explosion

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]

(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]

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]	



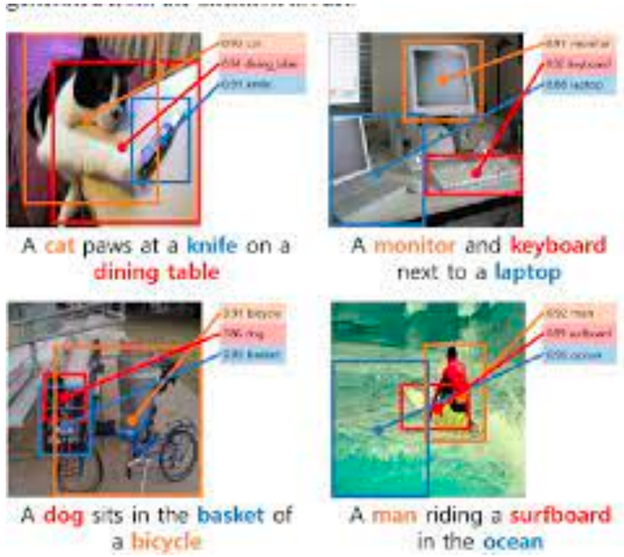
Explanations

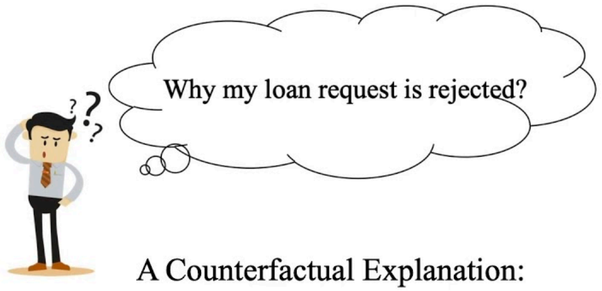
Examples

- Highlight important features
- Counterfactual explanation
- Caption generation
- Example based
- Activation Probing

Text with highlighted words

Why does the older generation think that just because they don't understand video games and technology, they feel like they have to hate them and blame every bad thing on them?





A Counterfactual Explanation:

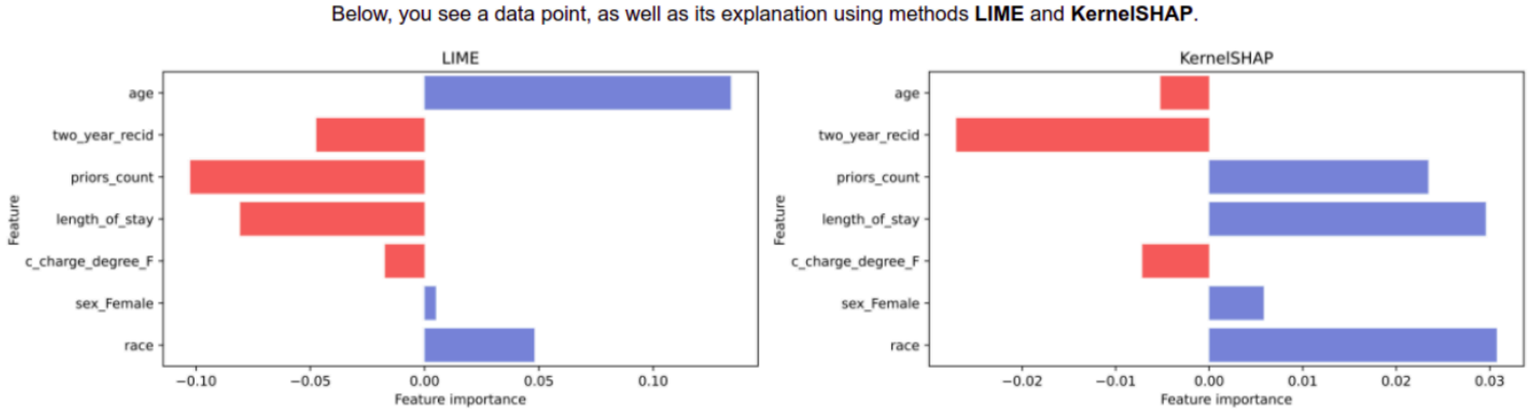
If you had an income of $40,000 rather than $30,000, your loan request would have been approved.

the minimal change made to alter the decision

Explanations

Some issues

- Easily manipulated
- Disagreement problem
- Post-Hoc can be unfaithful



Explanations

Call for Rigor

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

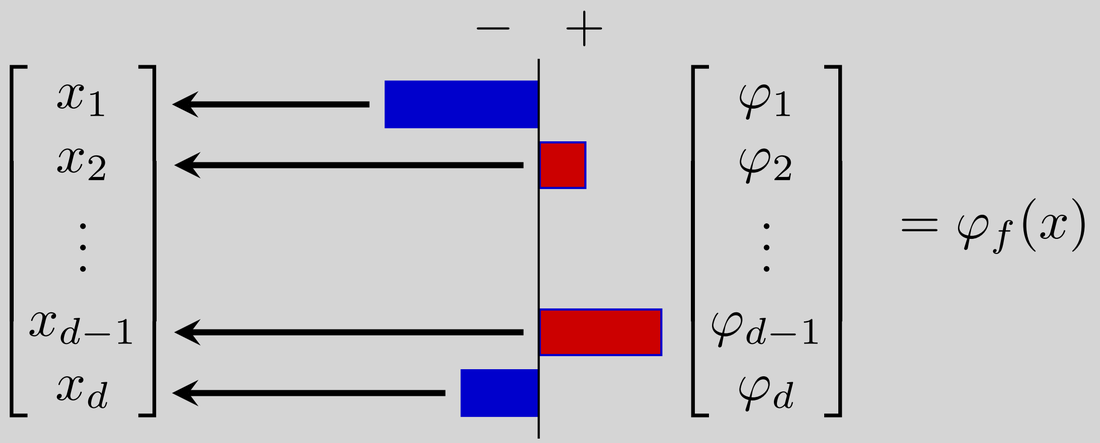
"Interpretability research suffers from an over-reliance on intuition-based approaches that risk—and in some cases have caused—illusory progress and misleading conclusions",

[Leavitt, Morcos, 2020]

Attribution methods &

Counterfactual methods

Attribution method



Attribute how important each feature was

- Positive value (typically) means that feature correlates positively with the outcome
- Negative value (typically) means that feature correlates negatively with the outcome

Not always the correct interpretation!

"Your income of € 40 000 contributed positively.

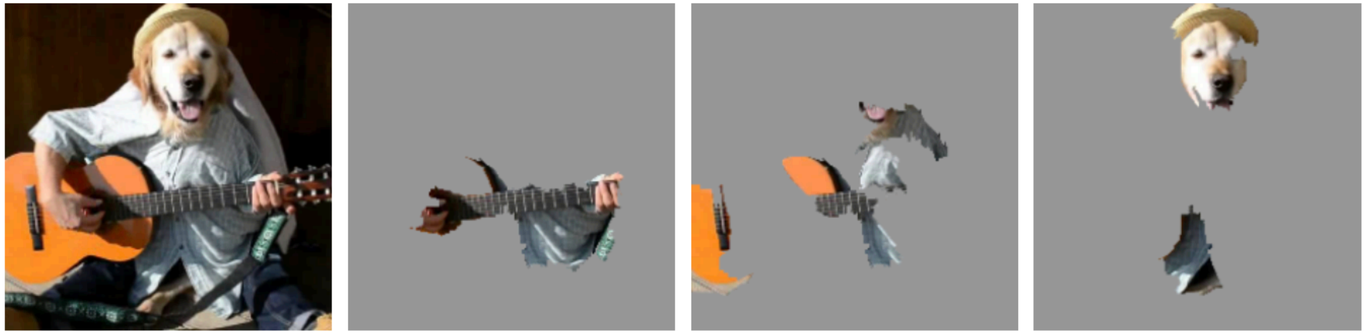
However, your 5 different credit cards contributed negatively"

Attribution methods

Examples

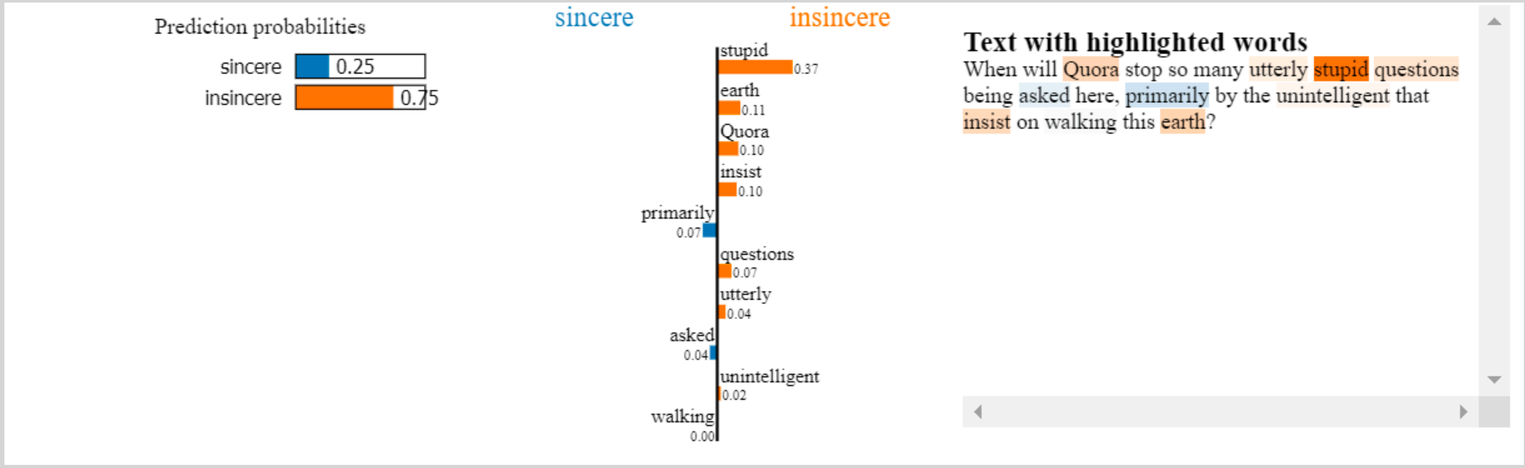
- LIME
- SHAP
- …

Pictures



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

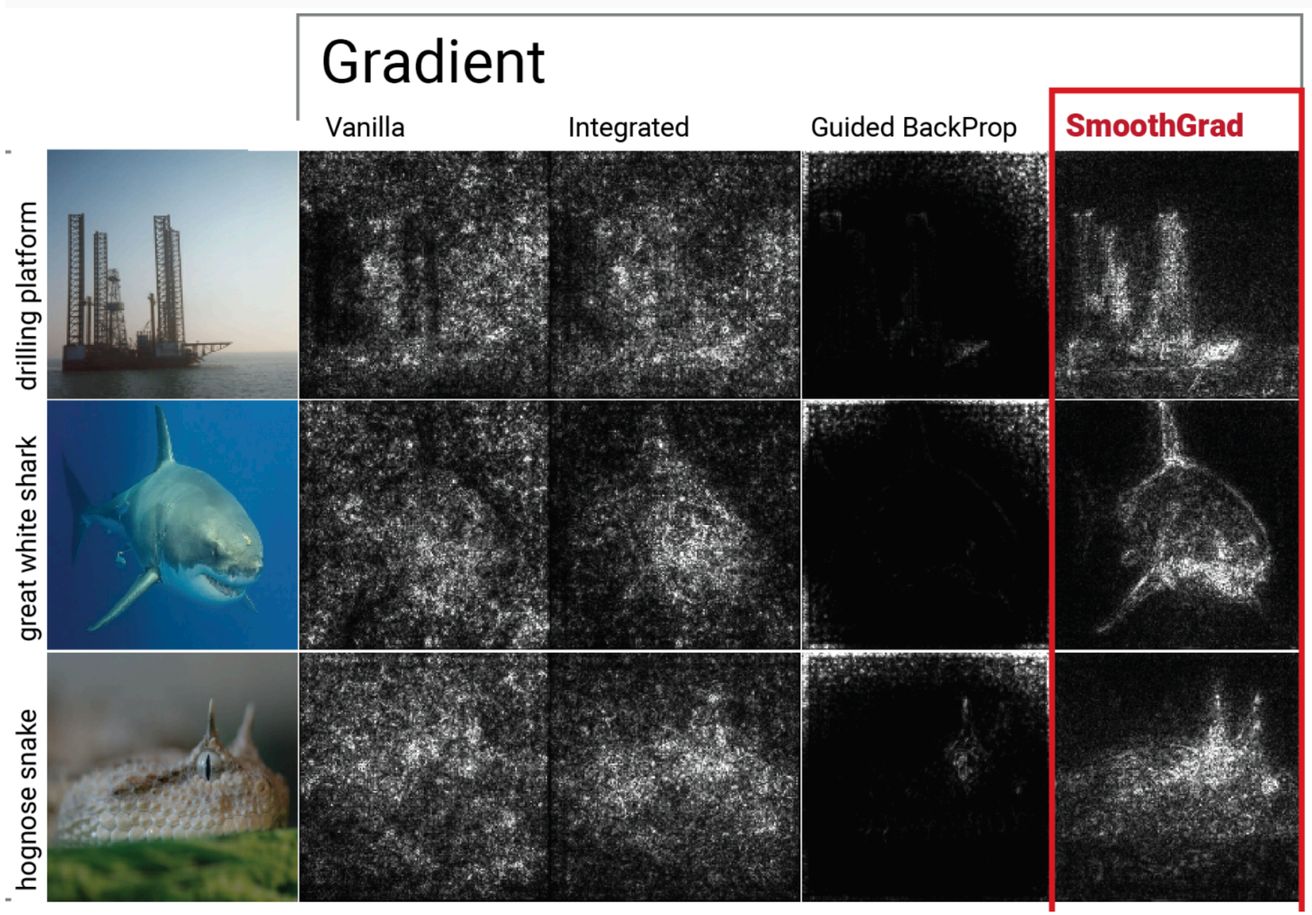
Text



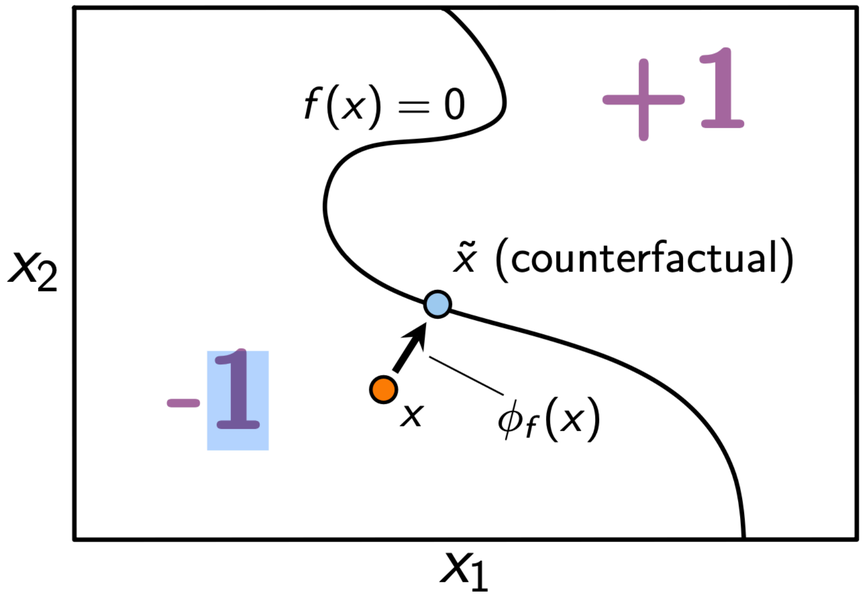
Attribution methods

Examples

- Grad-Cam
- SmoothGrad
- Integrated Gradients
- …



Counterfactual method



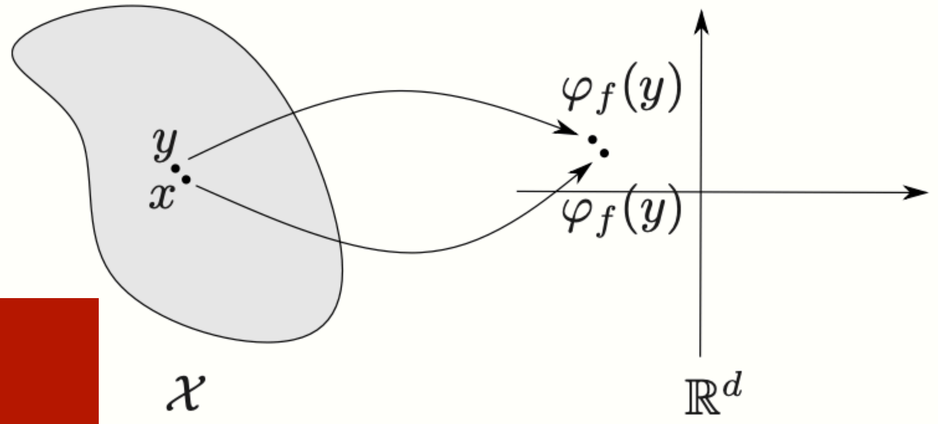
- Tell (A) how to change the decision from -1 to +1
- Minimal cost for (A)
- Provide ***Recourse***

"If you would have had an income of € 40 000 instead of €35 000, your loan request would have been approved."

Attribution methods that provide recourse cannot be robust

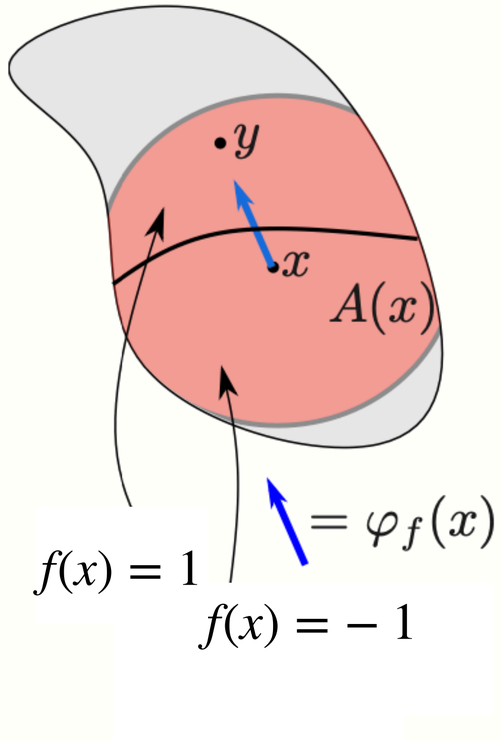
Desirable properties

Attribution methods



There are 2 desirable properties:

- Explanations need to be ***robust:***
  - If features are similar, explanations should be similar
- Explanations should act as counterfactuals **(Recourse sensitive):**
  - Positive attribution -> increase that value
  - Negative attribution -> decrease that value
  - Goal: change the decision from -1 to 1



Loan Applicant Example

Robustness

Features:

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

Features:

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

“Your income of € 40 000 contributed positively.

However, your 5 different credit cards contributed negatively”

Loan Applicant Example

Recourse sensitivity

“Your income of € 40 000 contributed positively.

However, your 5 different credit cards contributed negatively”

“If I decrease my number of credit cards, I will get the loan!

Impossibility result

Attribution methods cannot always

- Be Robust
- Provide Recourse

Impossibility result

Specifically,

- Someone develops an attribution method
- I can construct a classifier, for which
  - Either **robustness** fails,
  - Or **recourse sensitivity**,

Attribution methods cannot always

Provide recourse

Be robust

A taste of hope

There are models f, that

- Allow attributions
- Which are
  - **robustness**,
  - **recourse sensitivity**,

- Linear models
- Monotone models

In general,

- Very difficult to check,
- Pretty easy to break

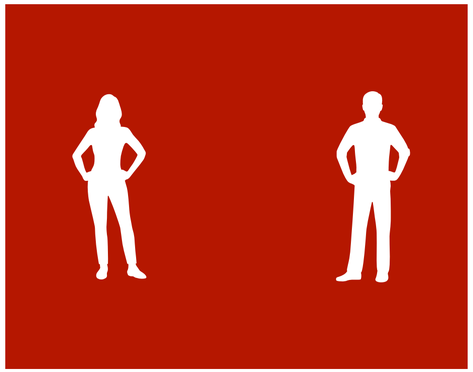
Risk of Recourse

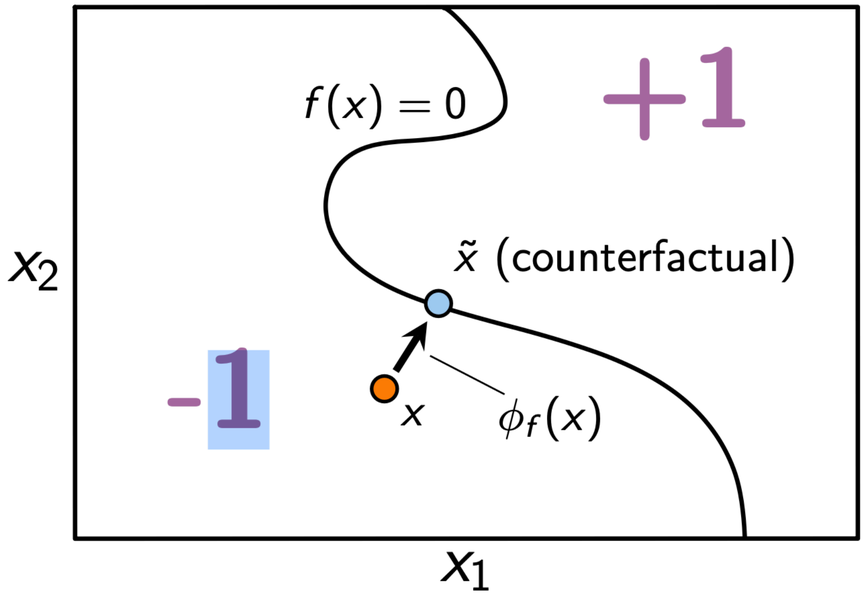
Counterfactual methods

No Attribution methods

- Forget about attribution methods for now
- Consider Counterfactual Explanations

- What happens when we look at the population?

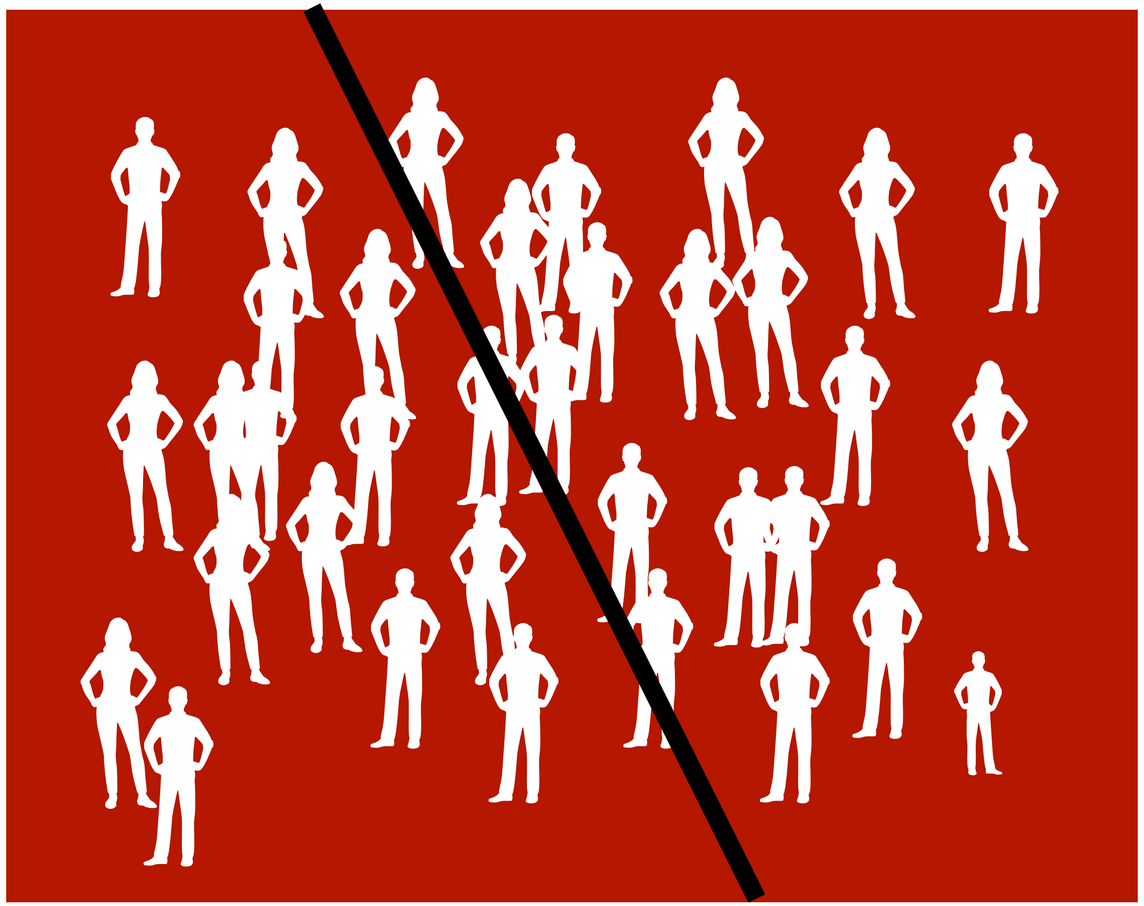




Counterfactual methods

No Attribution methods

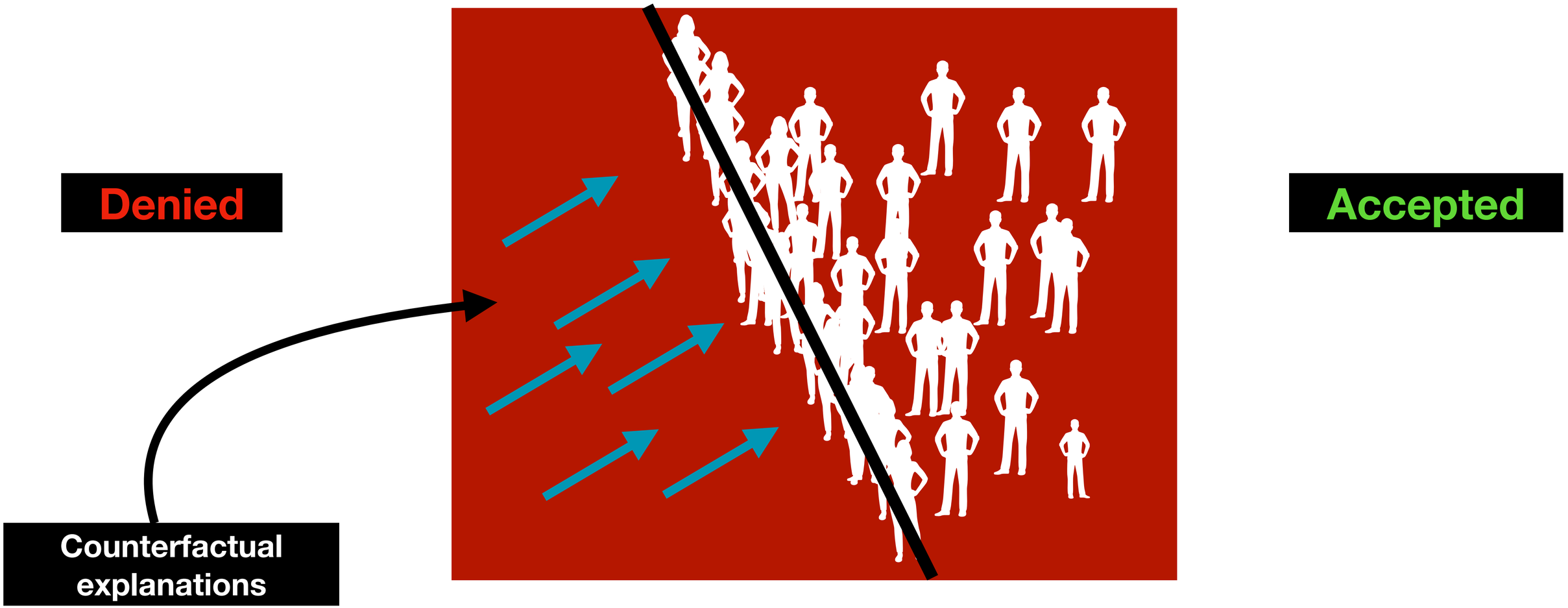
Denied



Accepted

Counterfactual methods

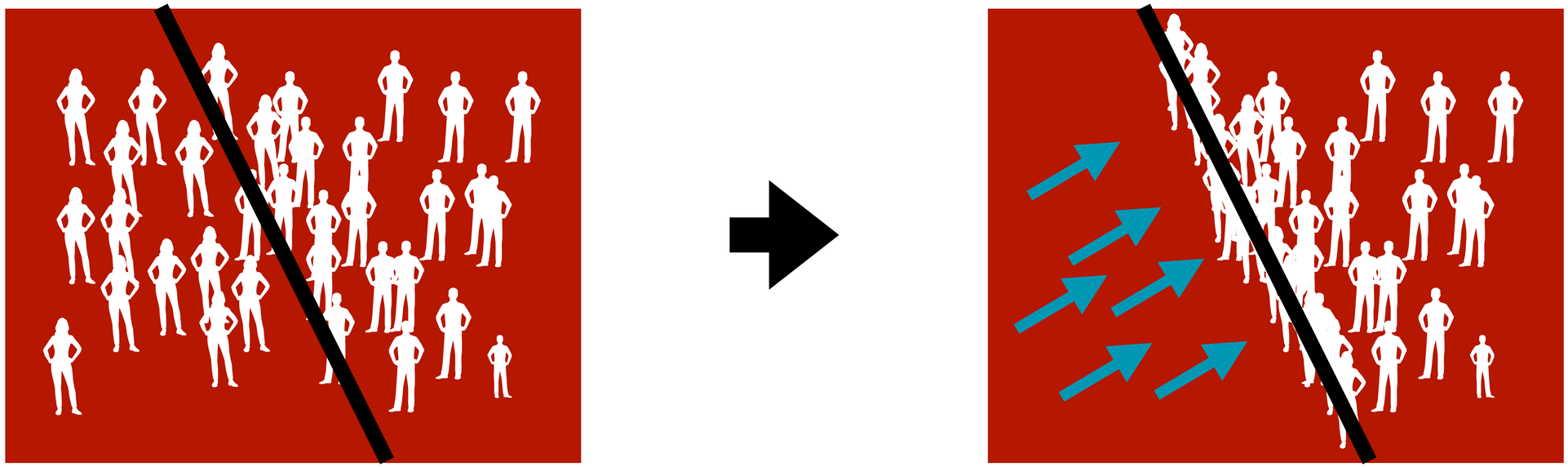
No Attribution methods



The risk of recourse

- What happens when we look at the population?
  - Underlying data distribution changes!
  - Modelling assumptions are violated!

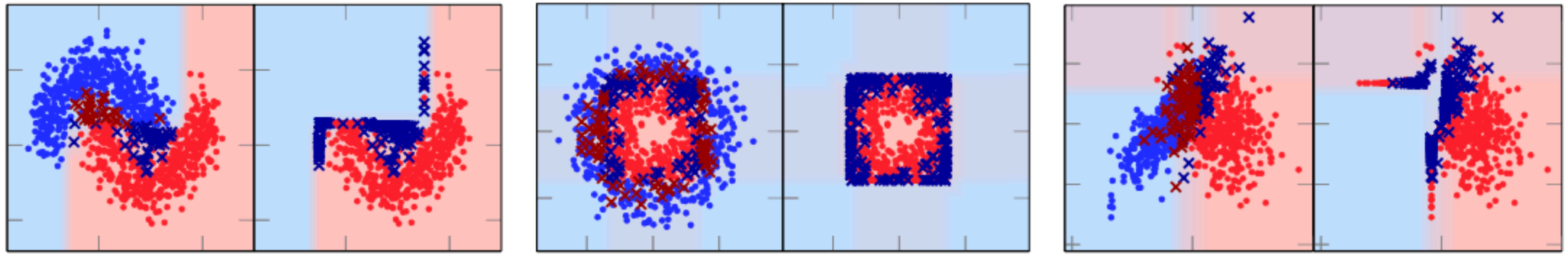
Accuracy will drop in most cases!



The risk of recourse

- What happens when we look at the population?
  - Underlying data distribution changes!
  - Modelling assumptions are violated!

Accuracy will drop in most cases!



The risk of recourse

Strategising against

Can (P) strategise against this accuracy drop?

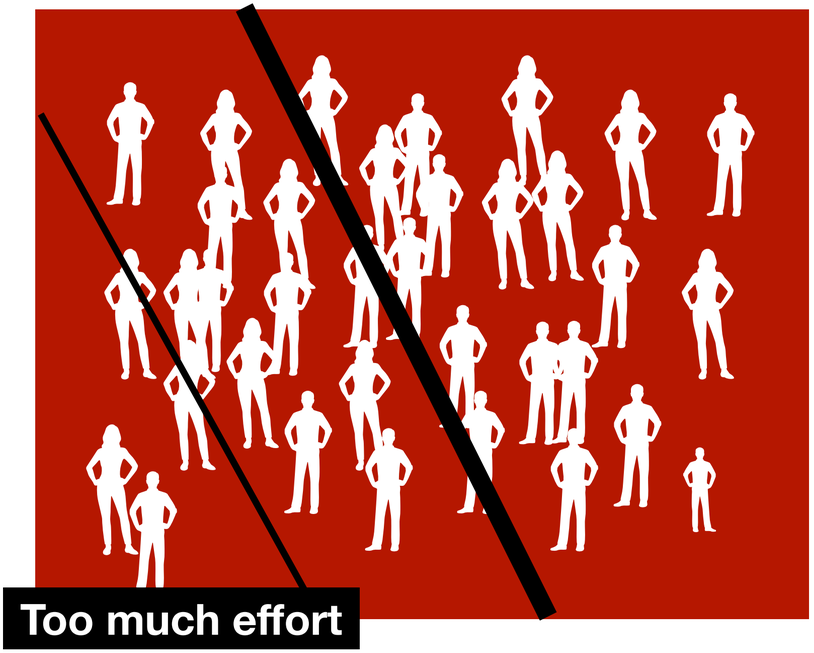
- Need to assume that not everyone gets an explanation

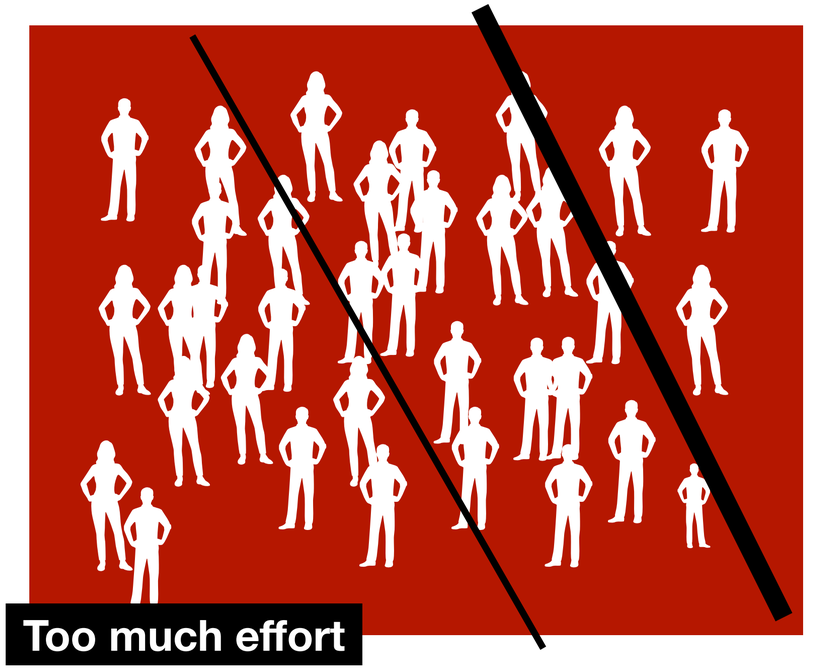
Yes and No

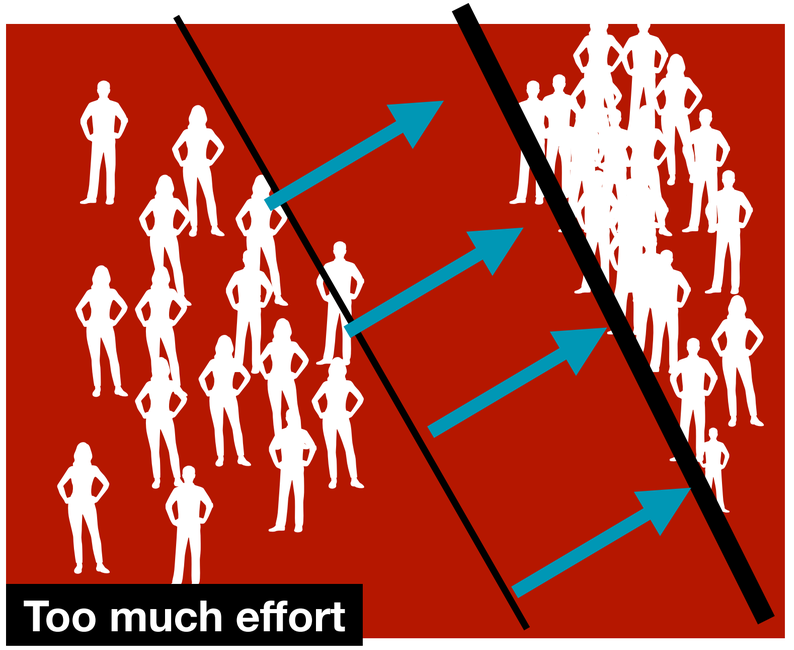
Too much effort

The risk of recourse

Strategising against







Exactly the same people get a loan

More effort

Practical implications

Practical implications

- Define the goal of your explanation, then pick a method
- When using attribution methods, be careful interpreting the scores
- When using counterfactual methods, ask the question
  - “How harmful is misclassification?”
- Whenever possible build simple models, increase complexity when necessary!

Some references

- *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]

Thank you for your attention!