Fairness in Machine Learning models using Causality

Supervision:

Patrick Forré,	University of Amsterdam
Christos Louizos,	University of Amsterdam
Tamas Erkelens,	Municipality of Amsterdam
Barteld Braaksma,	Statistics Netherlands

Rik Helwegen

10-02-2021

Today

- Problem description
- Why Causality?
- FairTrade method
- Risk profiles in unlawful social welfare
- Limitations and Future work
- Conclusion

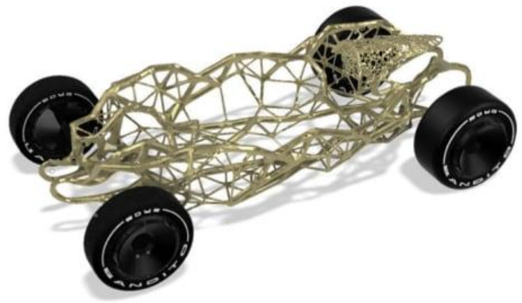
Problem description

Machine Learning (ML) more powerful

Expansion of applications

People's profiles as data points

Discrimination by sensitive attributes



Problem description

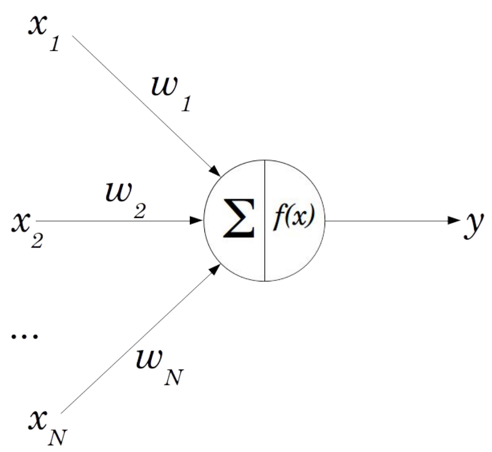
Sensitive attributes: Personal attributes deemed *unfair* to use for prediction models

What is fairness?

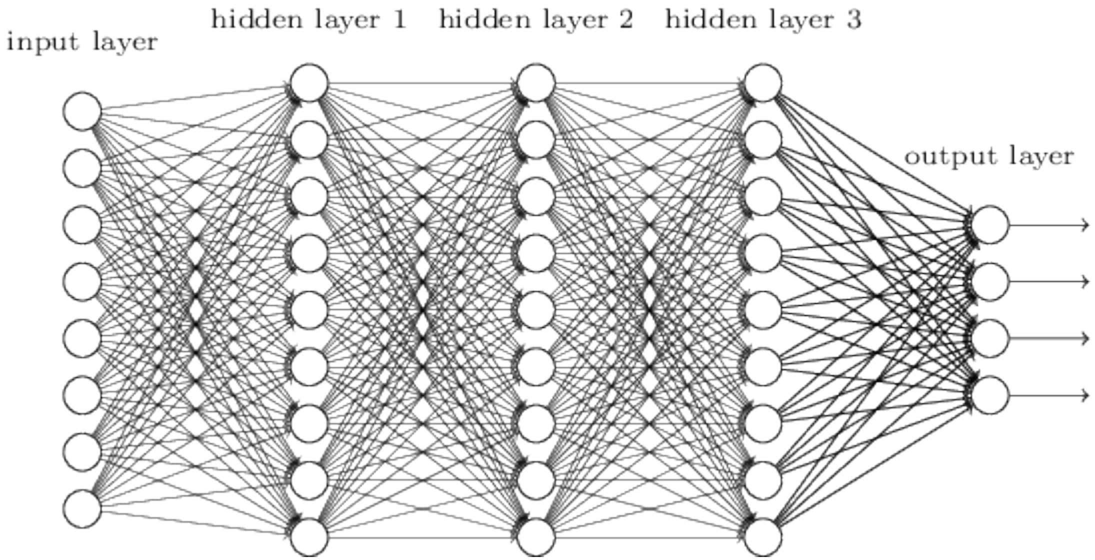
Discrimination by sensitive attributes

Problem description

No longer possible to use 'classical' correction due to increased model complexity



Regression model



Neural Network

Why Causality?

Gemeente
Amsterdam

Experiment 1

Fairness by Unawareness experiment

(leaving sensitive variable out of the model)

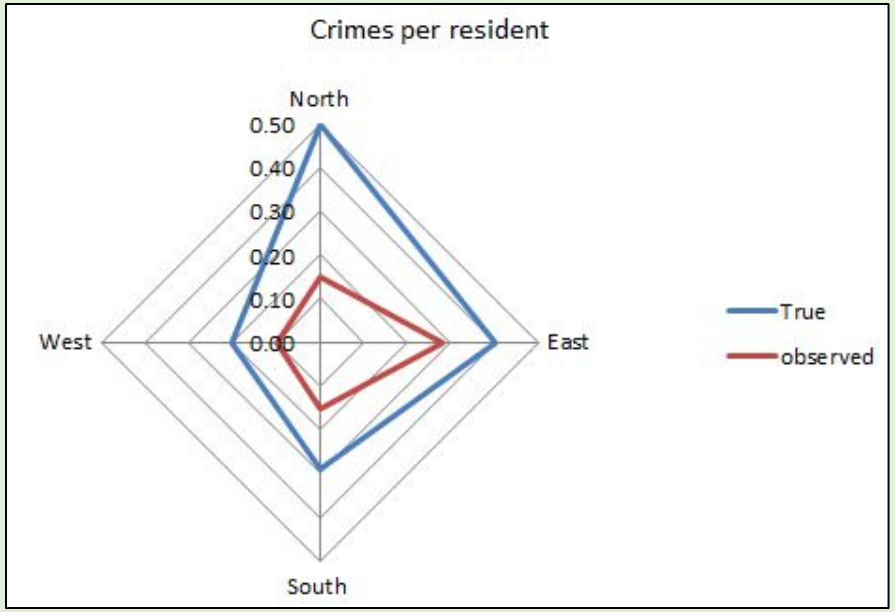
Simulation experiment:

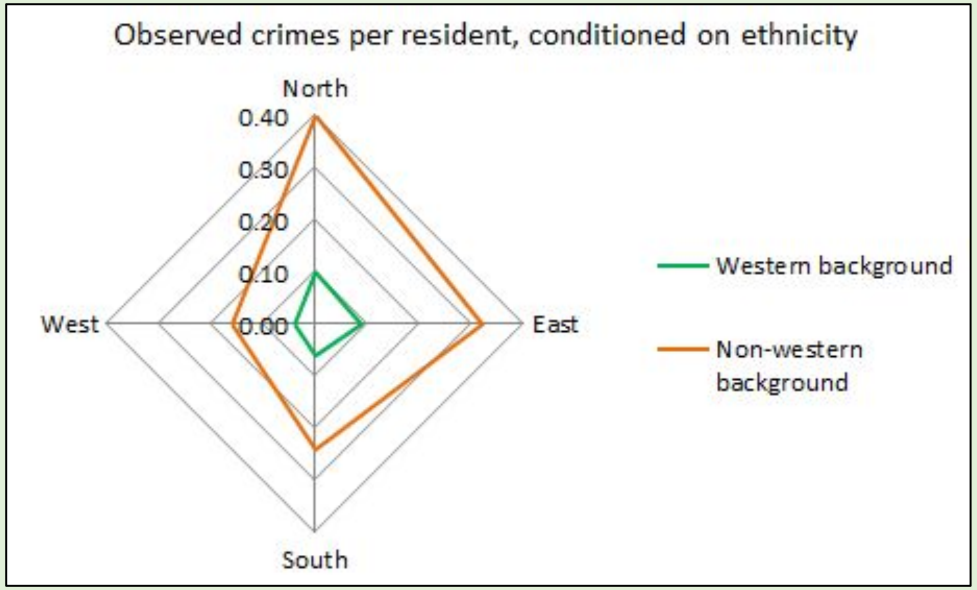
- Police wants to decide between checking neighbourhood North, East, South or West.
- Data is obtained with a **bias**, people with a non-western background have higher probability of being caught after a crime. Therefore the police wants to **exclude ethnicity** from the model.
- The most crimes occur in **North**.
- Most people with non-western background live in **East**.

Experiment 1

Fairness by Unawareness experiment

(leaving sensitive variable out of the model)



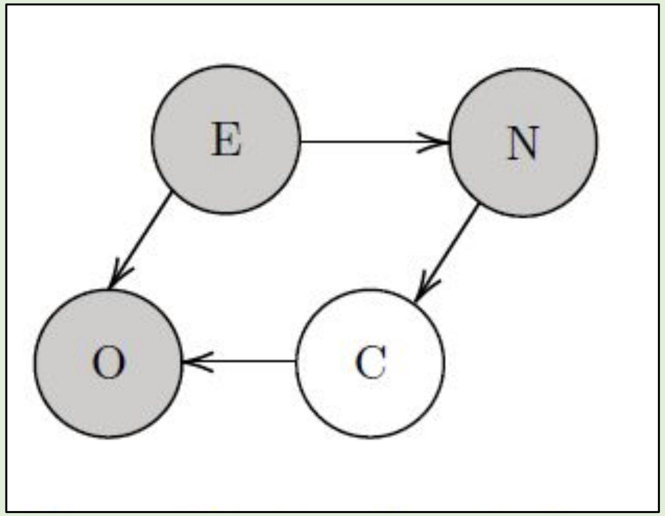


Simpsons Paradox

Experiment 1

Fairness by Unawareness experiment

(leaving sensitive variable out of the model)



Variable	Meaning	Values
E	Ethnicity	Western, Non-western
N	Neighbourhood	North, East, South, West
C	Crime committed	Yes, No
O	Observed Crime	Yes, No

Why Causality?

Solve fundamental problems of observational based fairness metrics by:

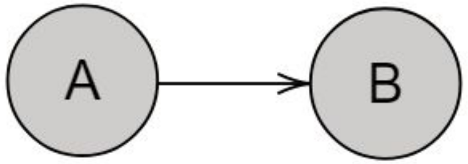
> Understanding data

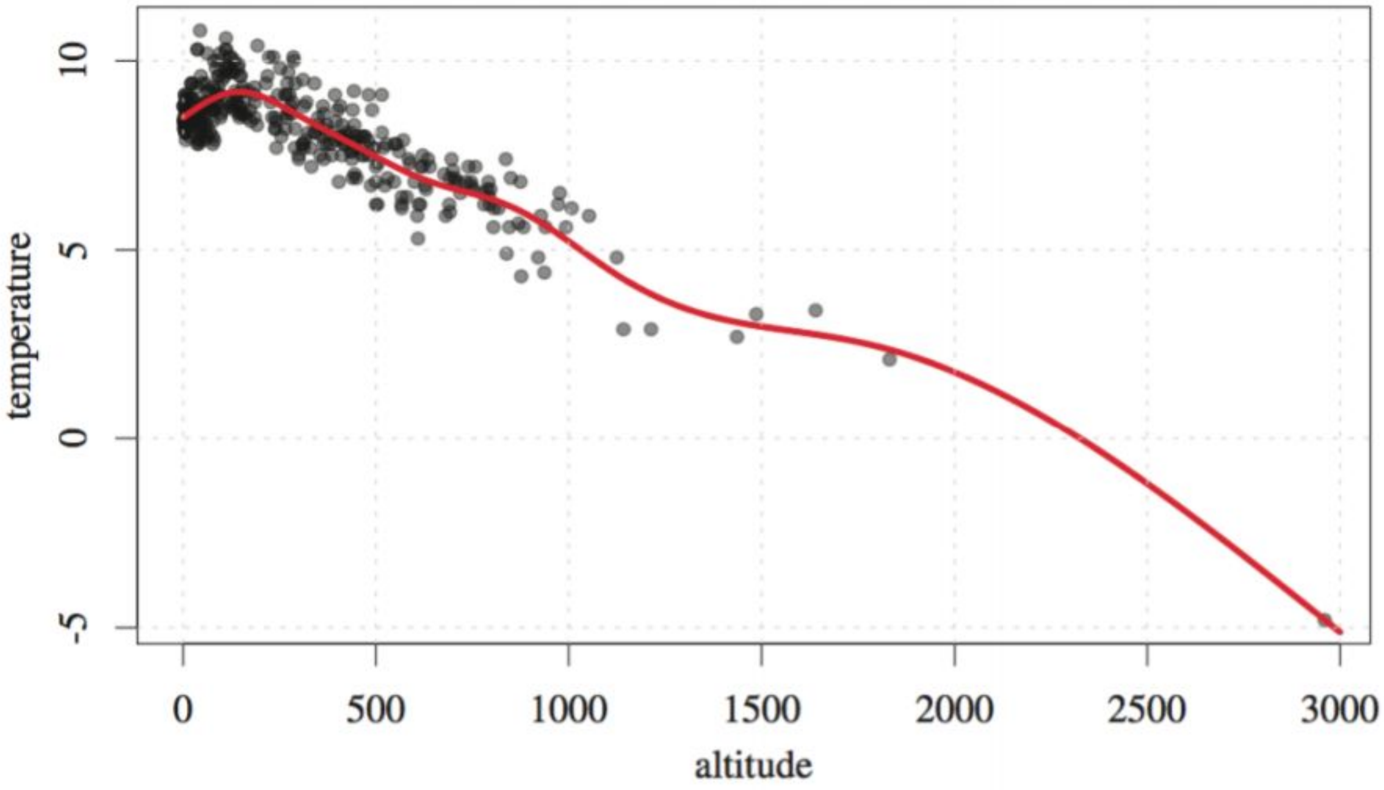
> Interpretation of model effects

> Control of explicit fairness demands

Why Causality?

Causality theory considers *causal effects:*





Why Causality?

- Causal solution to *what is fair?*:

Intuition: Intervening on the sensitive attribute should not influence the outcomes

For *Counterfactual Fair* models holds:

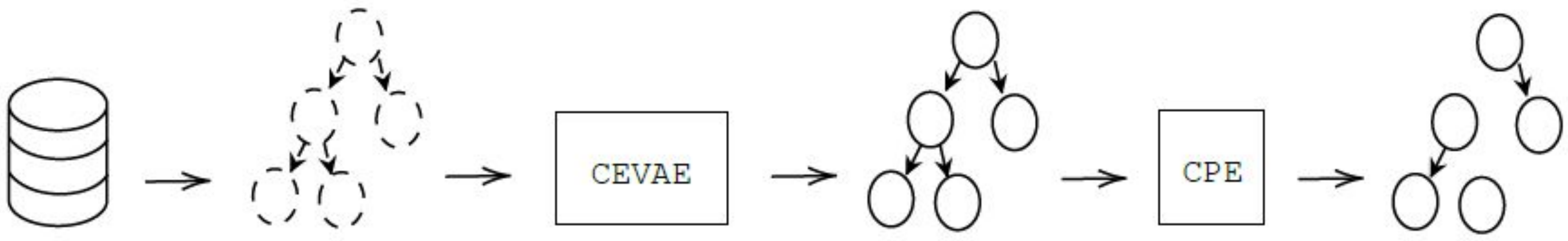
$$P\left(\hat{Y}(a,U) = y \middle| X = x, A = a\right) = P\left(\hat{Y}(a',U) = y \middle| X = x, A = a\right)$$

Symbol	Meaning
Y^	Predictor
A	Sensitive attribute
X	Covariates
U	Individual background

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems* (pp. 4066-4076).

FairTrade method

I. **Assume** Causal graph
II. **Infer** Causal relations
III. **Fair** prediction



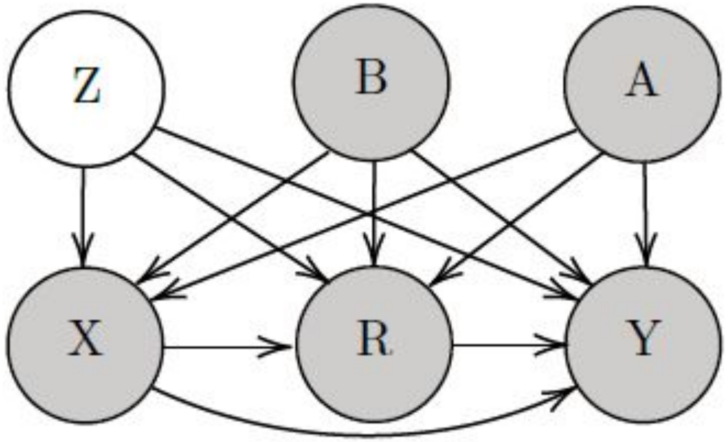
Gemeente Amsterdam

FairTrade method

I. **Assume** Causal graph

II. **Infer** Causal relations

III. **Fair** prediction



Symbol	Meaning
Y	Label
A	Sensitive attribute
Z	Unobserved confounder
B	Base variables
X	Other variables
R	Resolving variables

Gemeente Amsterdam

FairTrade method

I. **Assume** Causal graph

II. **Infer** Causal relations

III. **Fair** prediction

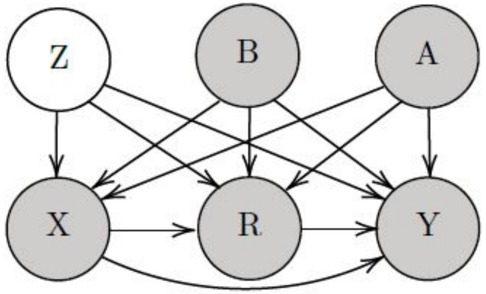
Causal Effect Variational Autoencoder (CEVAE):

1. Inference Step: Recover unobserved confounder

   q(z | a, b, x, r)

2. Generative Step: Reconstruct observed variables from parents

   p(x | z, a, b)
   p(r | z, a, b, x)
   p(y | z, a, b, x, r)



Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., & Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems* (pp. 6446-6456).

FairTrade method

I. **Assume** Causal graph

II. **Infer** Causal relations

III. **Fair** prediction

Causal Path Enabler (CPE):

Train *auxiliary* model which only has *fair* information as *input*, possible input:

- Non-descendants of the sensitive variable
- Background variables independent of the sensitive variable
- Resolving variables

Gemeente Amsterdam

FairTrade method

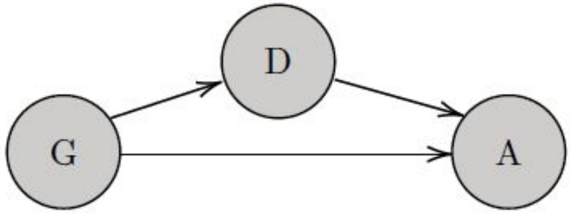
I. **Assume** Causal graph

II. **Infer** Causal relations

III. **Fair** prediction

Resolving variables

Variables deemed fair to use despite influence from sensitive variable



Variable	Meaning
G	Gender
D	Department choice
A	Admission rate

Berkley admission problem, is it fair if admission rate depends on gender via department choice?

FairTrade method – Possible improvements

I. **Assume** Causal graph
> Sensitivity analysis on assumption mistakes

II. **Infer** Causal relations
> Research on recovering of true effects

III. **Fair** prediction
> Formalisation step in input criteria CPE to enable PSE
> Evaluation counterfactual distributions

Gemeente Amsterdam

Risk profiles in unlawful social welfare

Situation Amsterdam:

- Around 40.000 people receive social welfare, including an unknown number of fraudulent cases
- Municipality wants to decrease unlawful social welfare, and also has committed to only fair algorithms in the city
- Set of fraud labels is suspected to be biased due to the passive 'signal' approach over the last years

Goal:

Create a classification model for risk profiles in social welfare, which is counterfactually fair with respect to *ethnicity.*

Gemeente Amsterdam

Risk profiles in unlawful social welfare

Gemeente Amsterdam

Experiment 3 - **Risk profiles in social welfare**

Goal:

Create a classification model for risk profiles in social welfare, which is counterfactually fair with respect to *ethnicity.*

Method:

FairTrade method

I. **Assume** Causal graph
II. **Infer** Causal relations
III. **Fair** prediction

Experiment 3 - **Risk profiles in social welfare**

Data

Proof of concept experiment on CBS data:

- Safe workspace with computing power
- Data from more municipalities to reduce biases
- More attributes to infer personal background

11.230 profiles with balanced fraud labels:

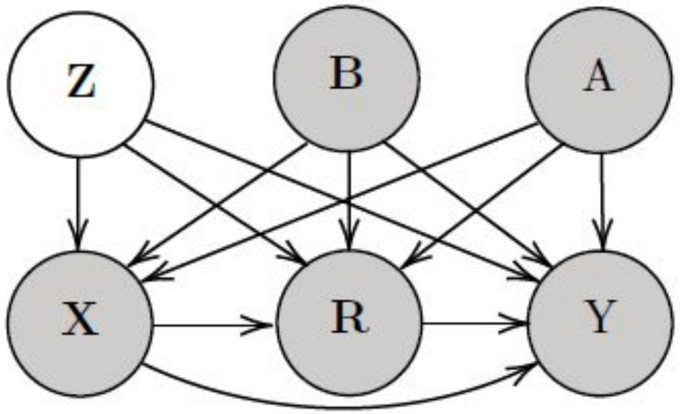
- Age
- Education
- Income
- Housing
- Jobs
- Property
- Crime history
- Debt
- Partner
- Household
- Other social benefits

Experiment 3 - **Risk profiles in social welfare**

I. **Assume** Causal graph

II. **Infer** Causal relations

III. **Fair** prediction



Symbol	Meaning
Z	Unobserved Confounders
A	Ethnicity
Y	Label unlawful social welfare
B	Base variables
R	Resolving variables
X	Other variables

B: {Income mutation, Gender}

R: {Involved in crime, Partner with debt, Recidivism}

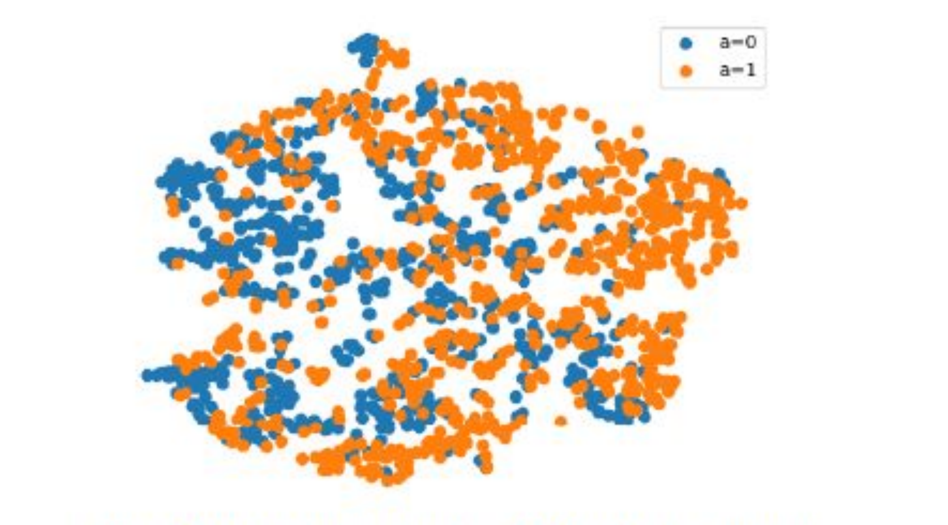
X: All other attributes

Experiment 3 - Risk profiles in social welfare

I. **Assume** Causal graph

II. **Infer** Causal relations

III. **Fair** prediction



(a) TSNE of latent space Z after 0 iterations

(d) TSNE of latent space Z after 16k iterations

Checking implied independencies: background independent of ethnicity?

Experiment 3 - **Risk profiles in social welfare**

I. **Assume** Causal graph

II. **Infer** Causal relations

III. **Fair** prediction

Statistical Parity: Group level fairness [0,1], equals 1 under counterfactual fairness.

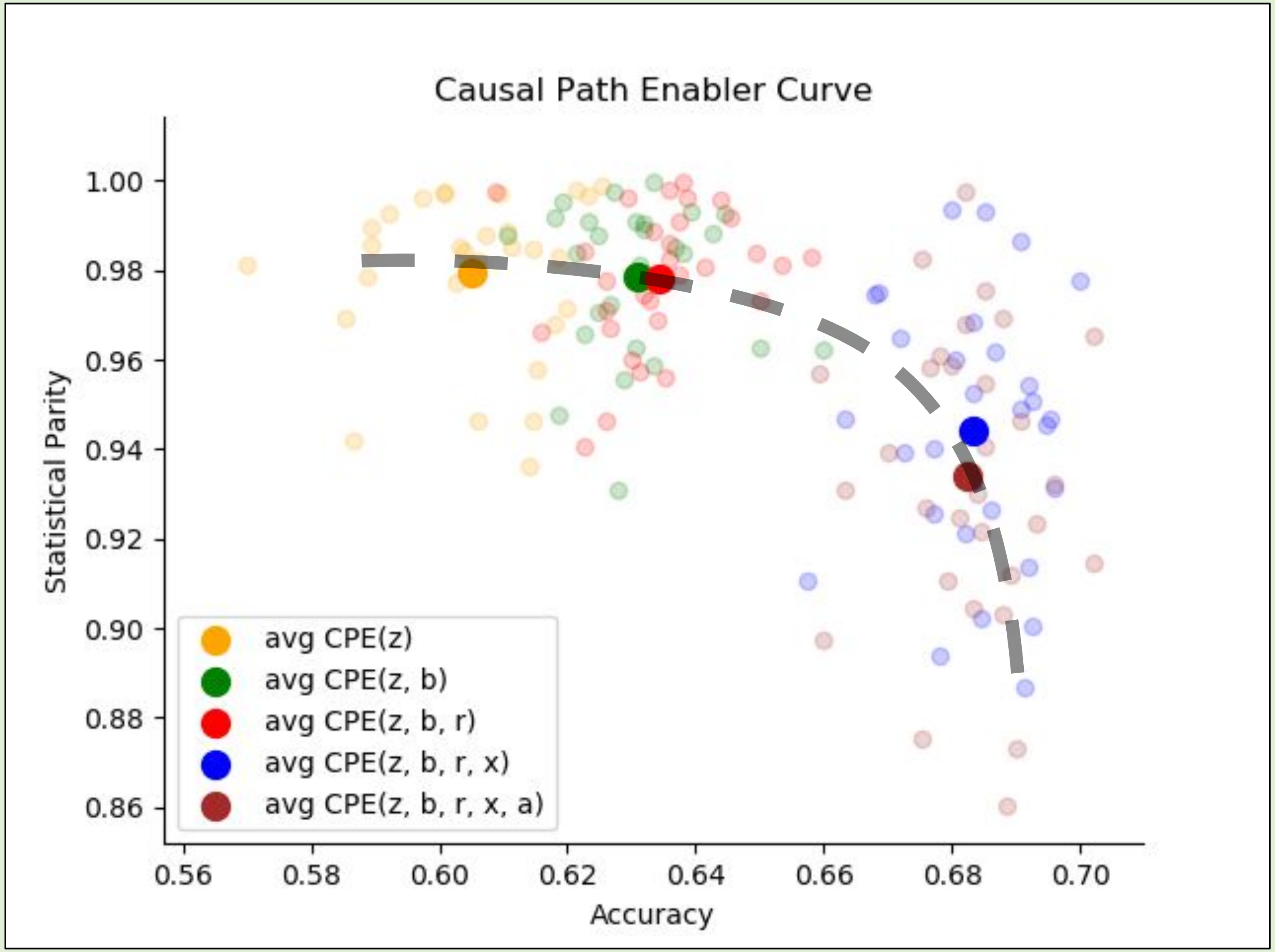
Baseline accuracies

Model	**Accuracy (std)**
Random Forest	0.6880 (0.01)
MLP	0.6785 (0.02)
Logistic Regression	0.6781 (0.02)

FairTrade model outcomes

	Accuracy (std)	**Statistical Parity (std)**
CPE(z)	0.605 (0.013)	**0.979** (0.017)
CPE(z,b)	0.631 (0.010)	0.978 (0.016)
CPE(z,b,r)	0.634 (0.011)	0.978 (0.015)
CPE(z,b,r,x)	0.683 (0.010)	0.944 (0.029)
CPE(z,b,r,x,a)	**0.682** (0.010)	0.934 (0.033)

Experiment 3 - **Risk profiles in social welfare**



Experiment 3 - **Risk profiles in social welfare**

Experiment outcomes:

- Fairness and accuracy show a trade-off curve for different levels of constraints
- Using counterfactual fair information, an accuracy of 63% can be obtained on a balanced data set, compared to a maximum of 68% for models without fairness constraints

Limitations and Future work

1. Evaluation
    1. Causal assumptions
    2. Counterfactual Fairness
    3. Approximate inference
    4. Path Specific Effects

2. Public debate on formal definitions of fairness

Gemeente Amsterdam

Conclusion

- Machine Learning has increasing impact on people's live
- Causality helps to formalise fairness
- The FairTrade method makes it possible to approximate fair models in practical applications
- A trade-off curve between fairness and accuracy is obtained for neural network based classification of exceptionally detailed real data profiles

Thank you for listening!