# Data poisoning on ALE and PDP

Julia Chylak, Aleksandra Mysiak, Krzysztof Tomala

# Goals for this checkpoint

- ➔ Quantify the differences between PD and ALE using aggregated plots
- ➔ Is there a difference between different datasets?
- ➔ Can we limit how much we change the data distributions but still disturb the explanations?

# Data distribution invariance

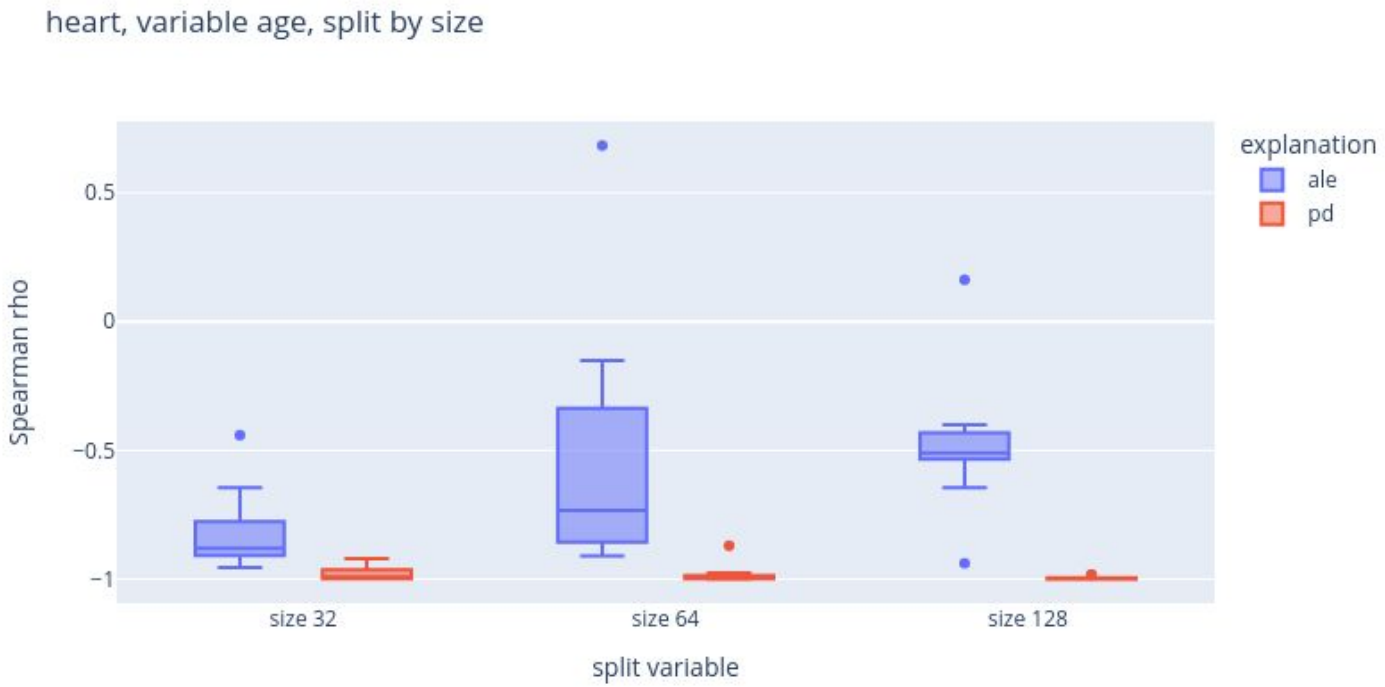
Average of squared differences of sorted values
as an additional loss term

- Not exactly mathematically sound
- Simple and quick to calculate
- Can be plugged into autograd
- And it works!

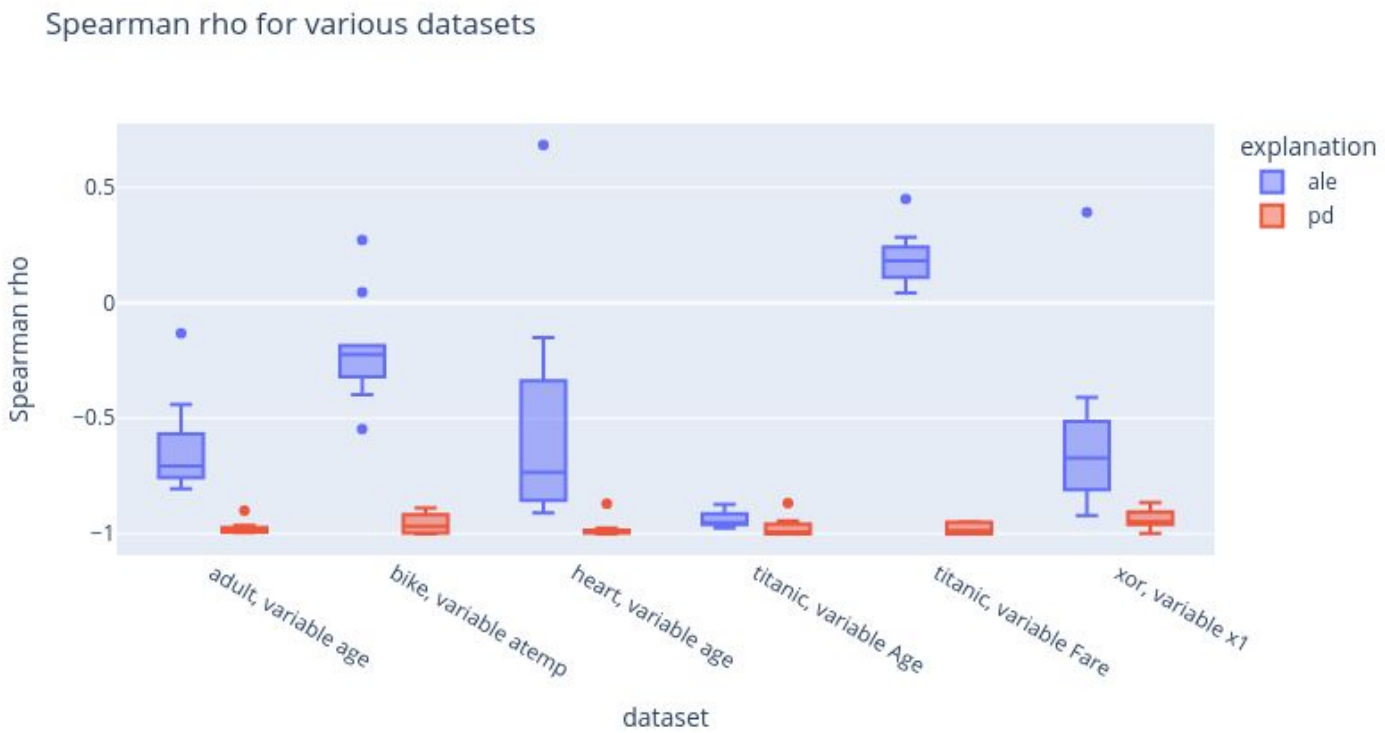
```
1   def loss_dist(X_original, X_changed):
2       x1 = tf.sort(X_original, axis=0)
3       x2 = tf.sort(X_changed, axis=0)
4       ret = tf.reduce_mean((x1 - x2) ** 2)
5       return ret
```

# Results

# ALE and PD difference

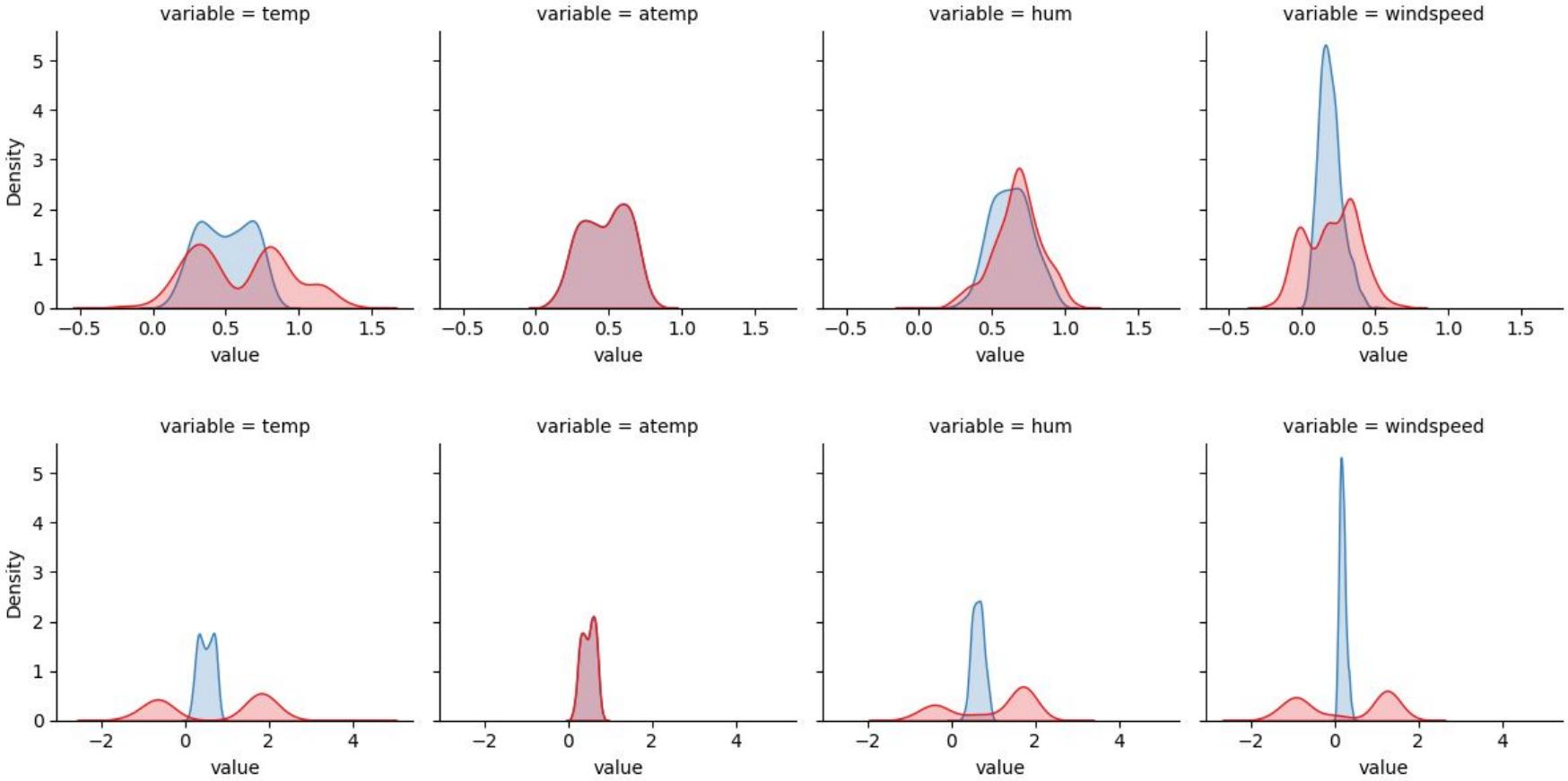


# Dataset impact

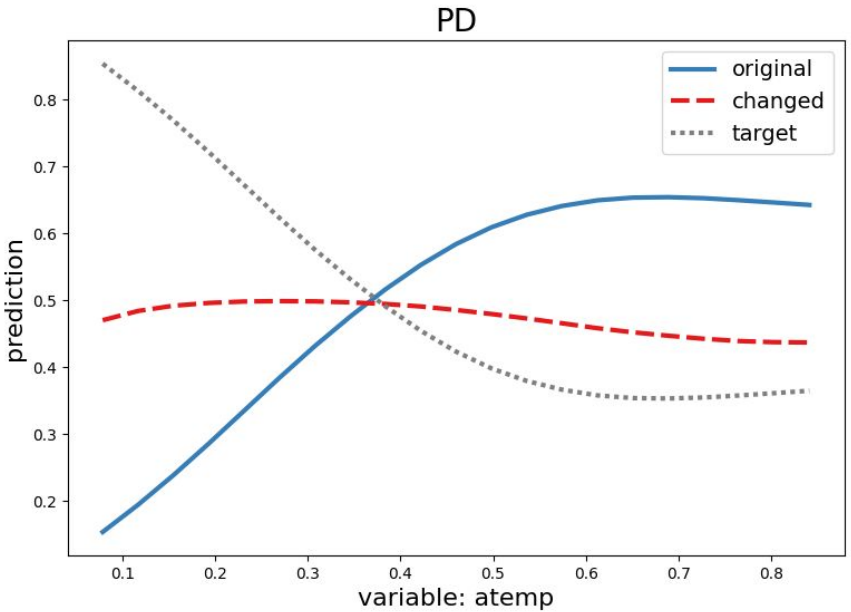


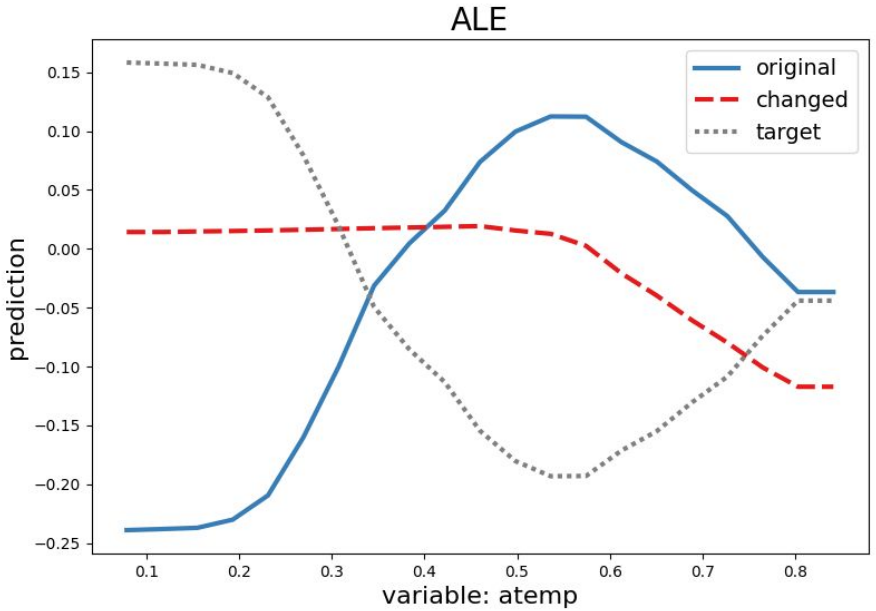
# Data distribution restriction

# Data distribution restriction

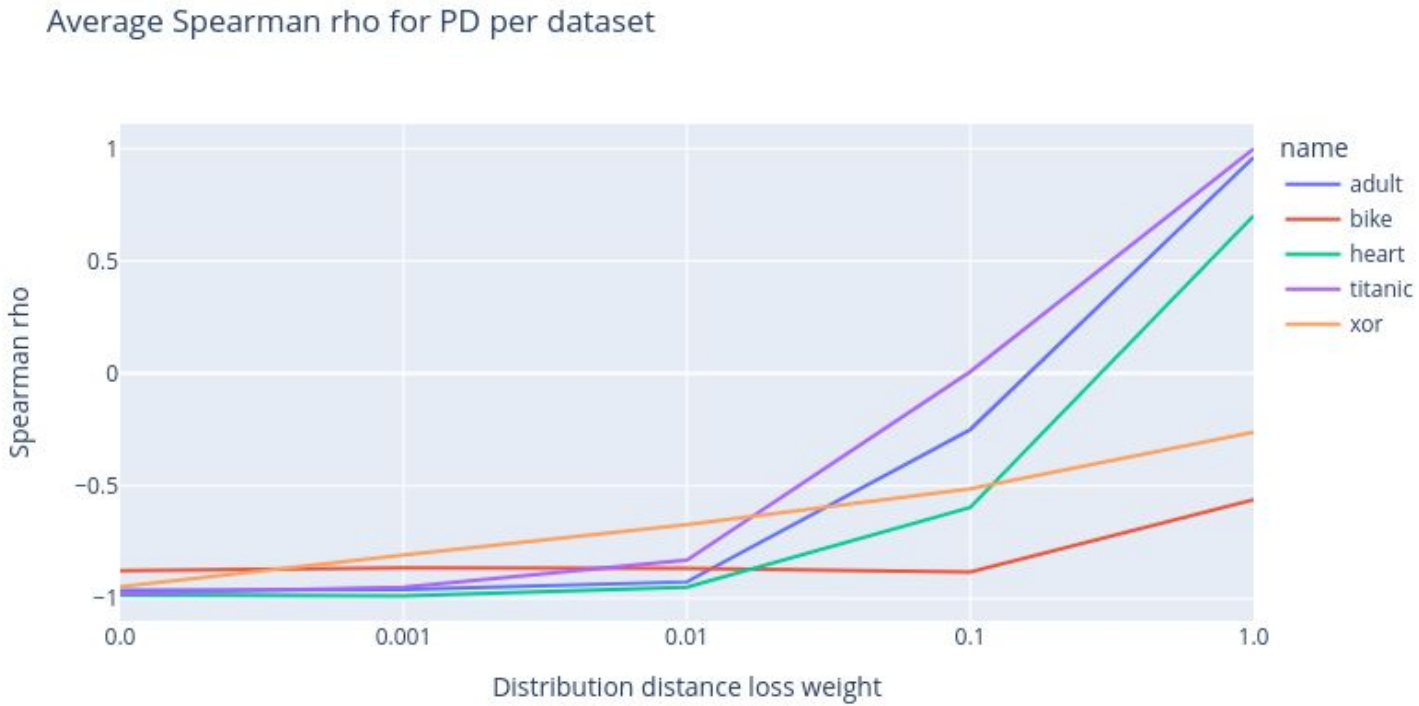


# Data distribution restriction



# Data distribution restriction



# Summary

- ➔ ALE is susceptible to PDP-directed attacks, although to a lesser degree
- ➔ Usually no qualitative difference between datasets
- ➔ A simple quadratic loss is enough to limit marginal distribution change
- ➔ PDP can be poisoned with realistic marginal distributions

Thank you for you attention!