Holistic AI × The Alan Turing Institute

Trade-offs of Bias with other verticals in Trustworthy AI Part I

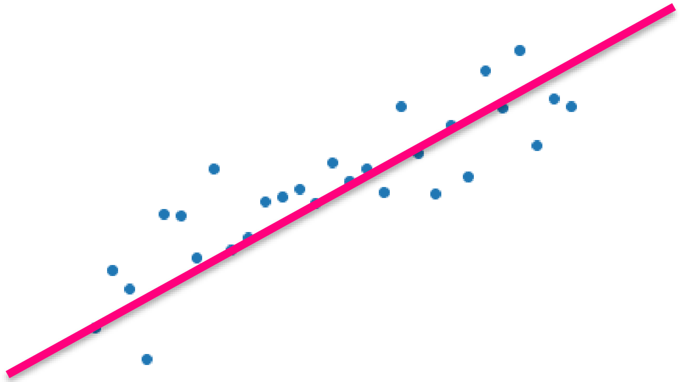
Turing Course

Contents

Regression

- modelling the relationship between a scalar response and one or more explanatory variables

- Many binary classification methods use linear regression as an intermediate step (e.g. Logistic regression)

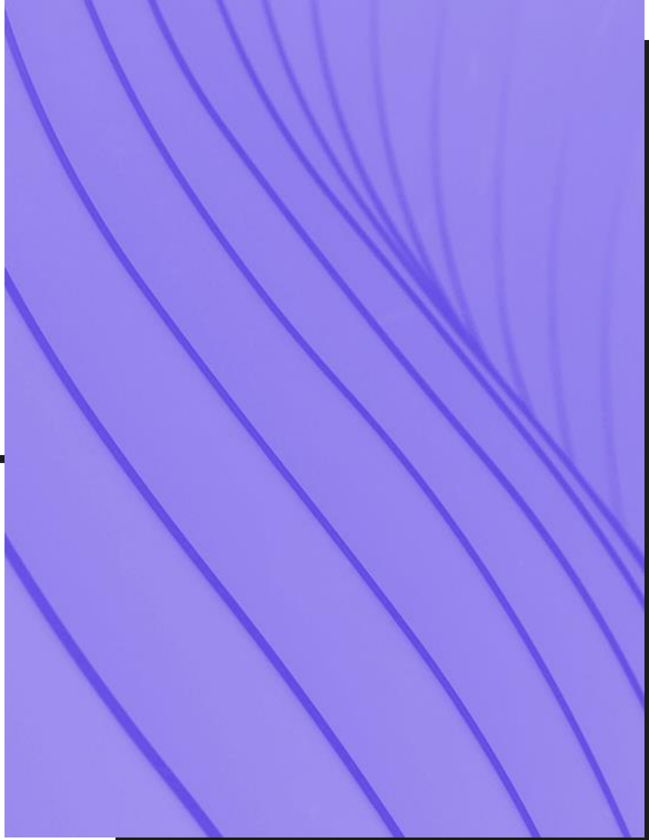


Multiclass classification

- Multiple binary classification problems. **one vs rest** and **one vs one**.
- Neural networks. probability of each class given (can be used for Regression)
- k-nearest neighbours.
- Naive Bayes.
- Decision trees.

I – Explainability

- Motivation
- Overview of methods
- SHAP example
- Interactions with Fairness



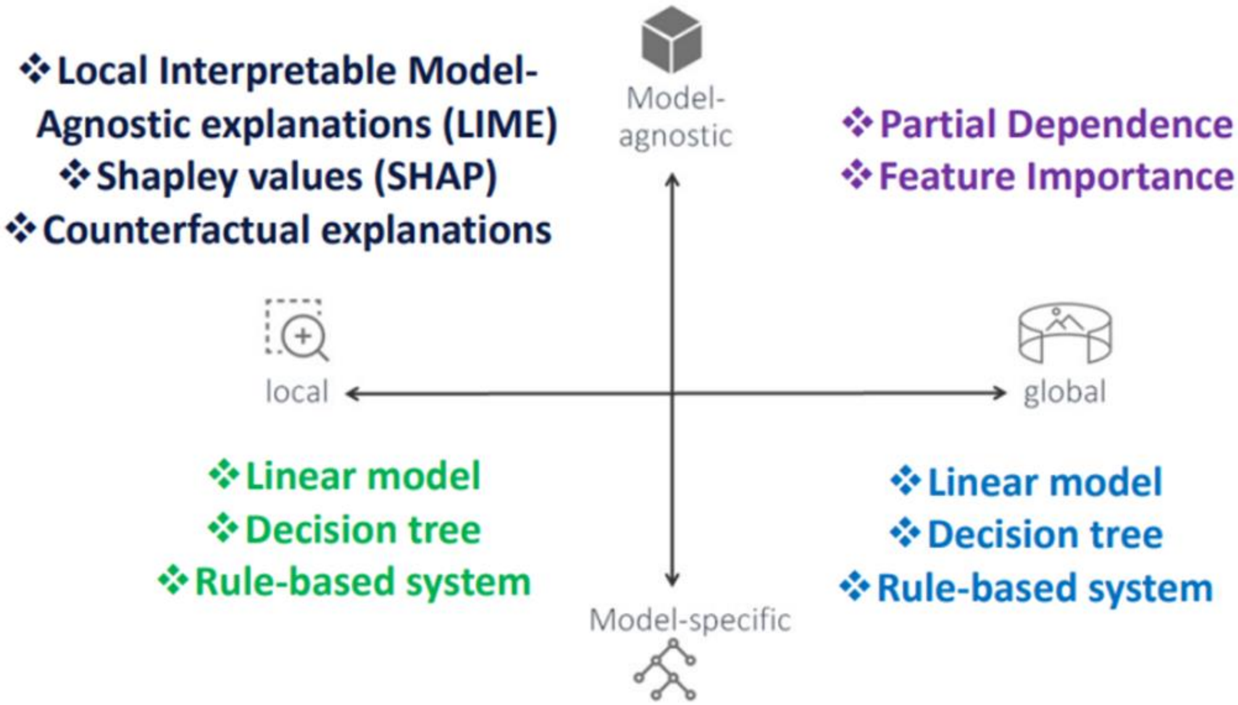
Motivation

– Ex1. Cancer diagnosis through classification of tumor in 3 classes.

– Ex2. Regression to predict health insurance premium.

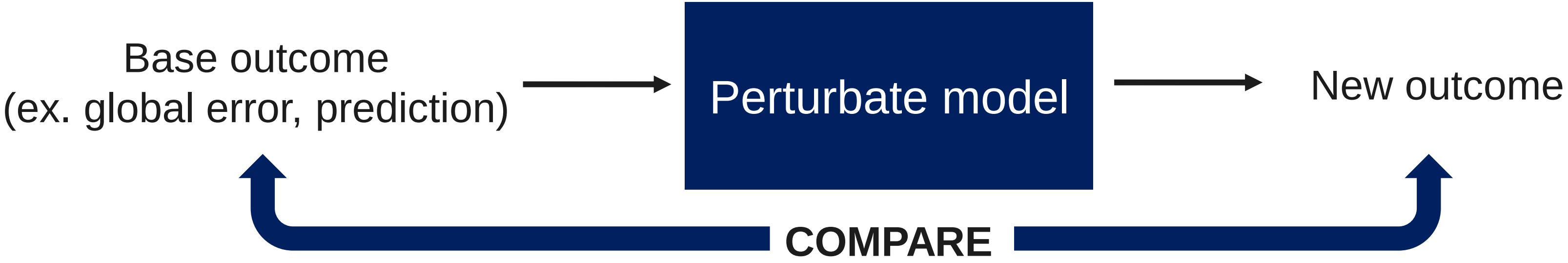
Overview

[Koshiyama et al.,2021]



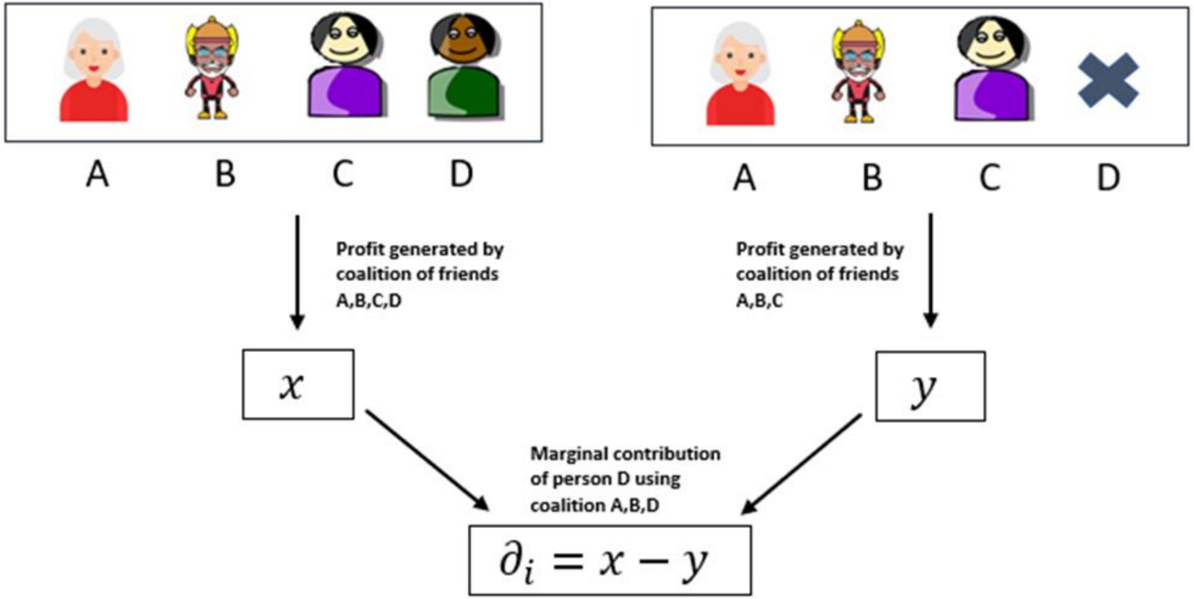
Translate methods

– Binary classification vs Regression



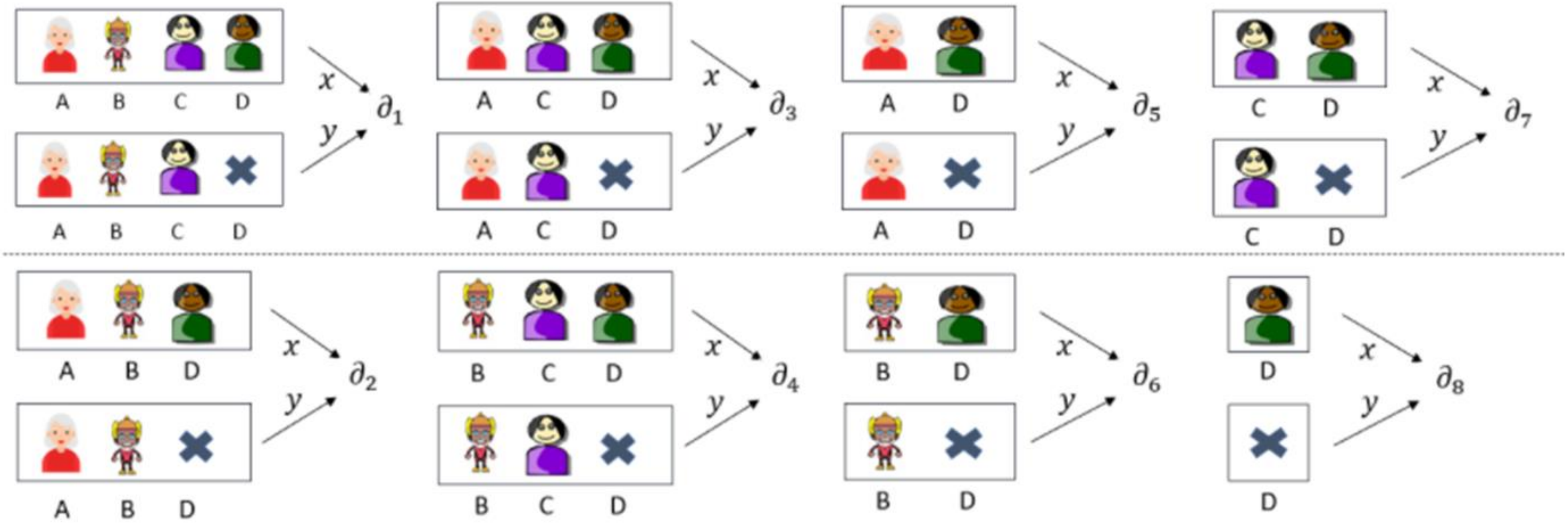
– Binary Classification vs Multiclass.
One-vs-All. One-vs-One.

SHAP – Regression



Adapted from https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3 by Fernando Lopez

SHAP – Regression (2)



The shapley value for person D is therefore: $\Phi_D = \frac{\delta_1+\delta_2+\delta_3+\delta_4+\delta_5+\delta_6+\delta_7+\delta_8}{8}$

SHAP – Multiclass

Features

Petal length (cm)

Petal width (cm)

Sepal length (cm)

Sepal width (cm)

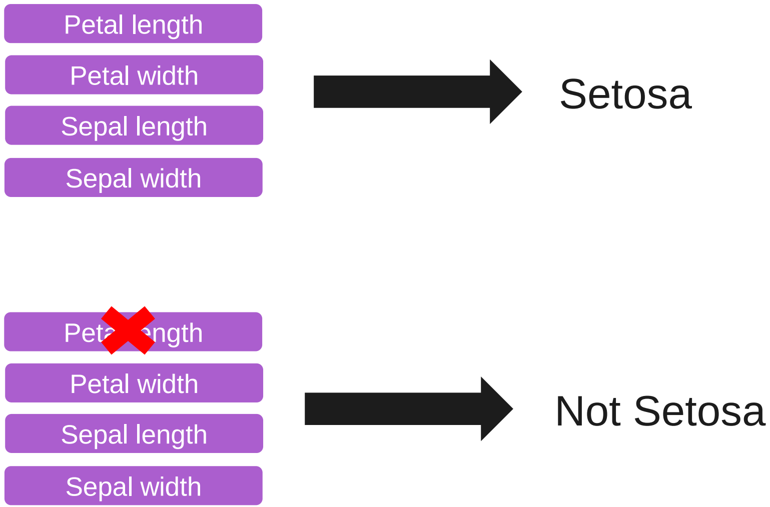
Virginica

Versicolor

Setosa

SHAP – Multiclass

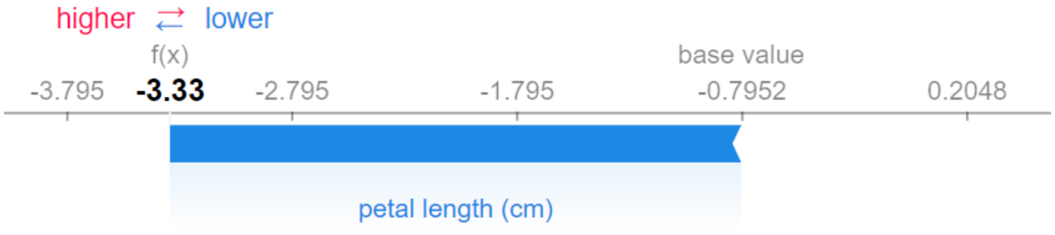
For one sample:



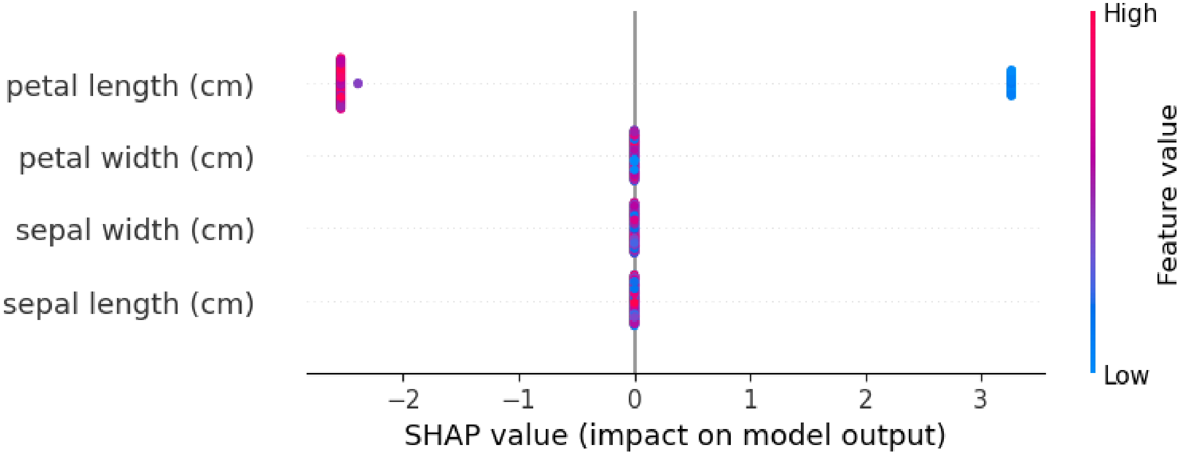
Several binary classification problems ==> Shapley values for each one, and then combining the results

SHAP – Multiclass

One sample/one class:

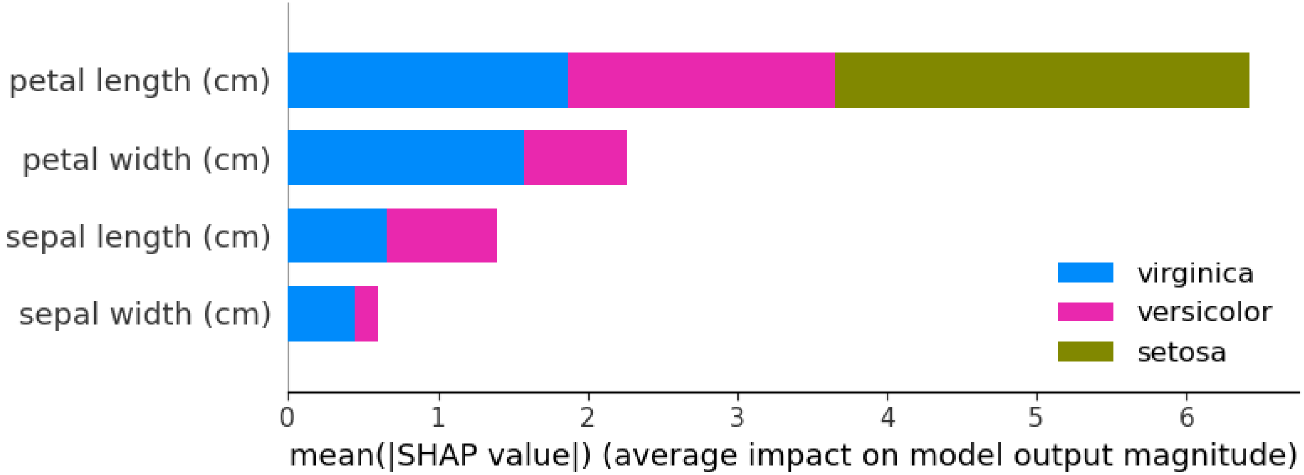


All samples/one class:



SHAP – Multiclass

All classes:



Interaction with Fairness

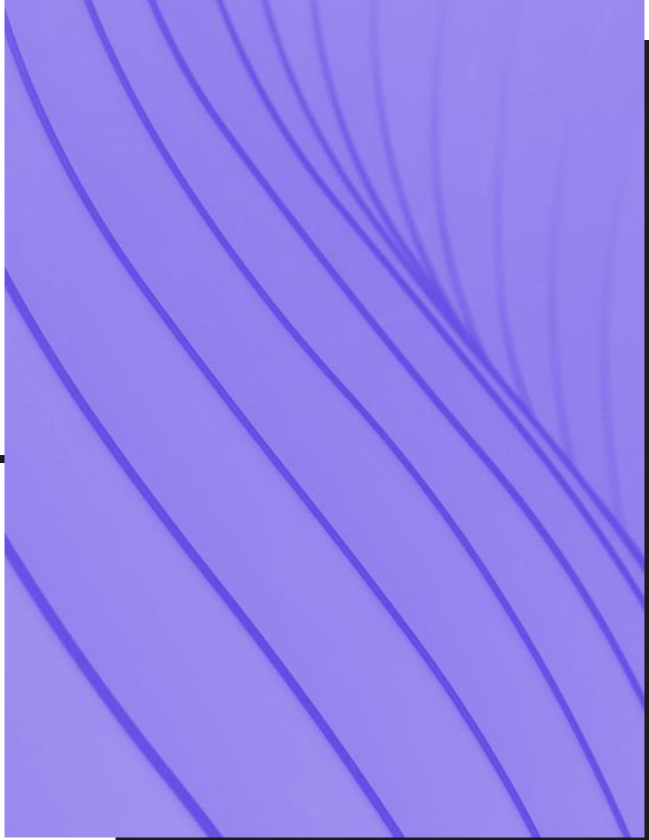
Example questions one can answer:

– Are the most influential factors reasonable? Are they the same across different groups ?

Instead of explaining output => explain fairness metric

II – Robustness

- Motivation
- Methods Example
- Interactions with Fairness



What is Robustness?

Robustness & Safety

- Resilience to attack and security (e.g. adversarial training)
- Fallback plan and general safety
- Accuracy
- Reliability and Reproducibility

In practice:

- Resistance to outliers
- Small changes in input -> small changes in output

EU-HLEG. (2019). Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Motivation

- Least-square errors sensitive to outliers. $$\sum_{i=1}^{n}(y_i - \overline{y})^2$$

- Regression to predict health insurance premium. An outlier could spoil the regression.

Regression

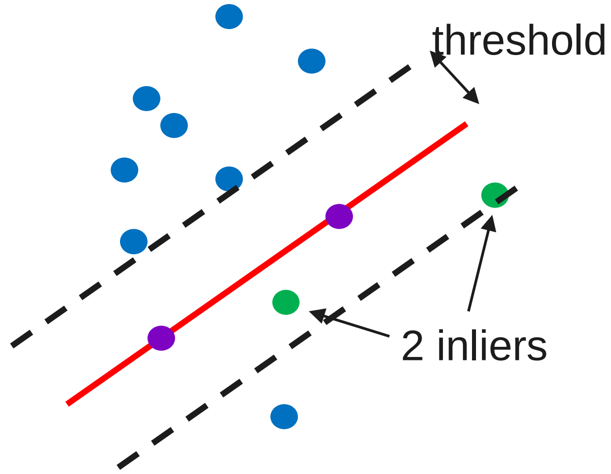
- Least-square errors (L2-norm) --> Least Absolute Deviation (L1-norm)

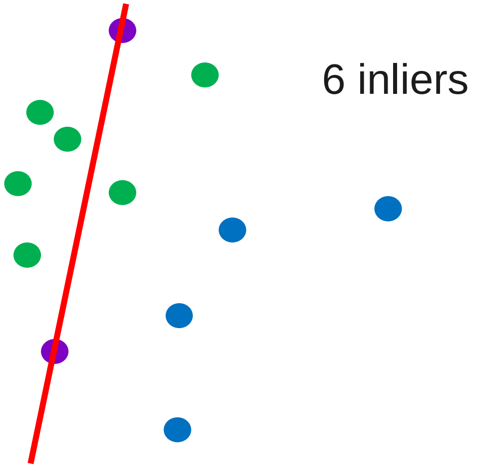
$$\sum_{i=1}^{n} \left(y_i - \hat{y_i}\right)^2 \qquad\qquad \sum_{i=1}^{n} \left|y_i - \hat{y_i}\right|$$

- RANSAC (Random Sample Consensus) algorithm, which fits a model to a subset of the data, and then uses this model to identify inliers and outliers, and refits the model using only the inliers.

RANSAC
(Random Sample Consensus)

- Subset data randomly (minimum number of points to find parameters)
- Fit model on subset
- Remaining data points -> inliers or outliers
- Select highest scoring models and keep inliers





Distributionally Robust Optimization (DRO)

- Works for any supervised algorithm (binary, regression, multiclass)
- Alternative to ERM = **Empirical Risk Minimization**. Instead of minimizing average loss, minimize worst case
- Better when the data-generating distribution P is NOT representative of the overall population of interest

Distributionally Robust Optimization

– **Empirical Risk Minimization.** "Classic" way of training. Miminize average empirical loss.

$$\min_{\theta} \mathbb{E}_P[\ell(\theta; X)]$$

Model parameters

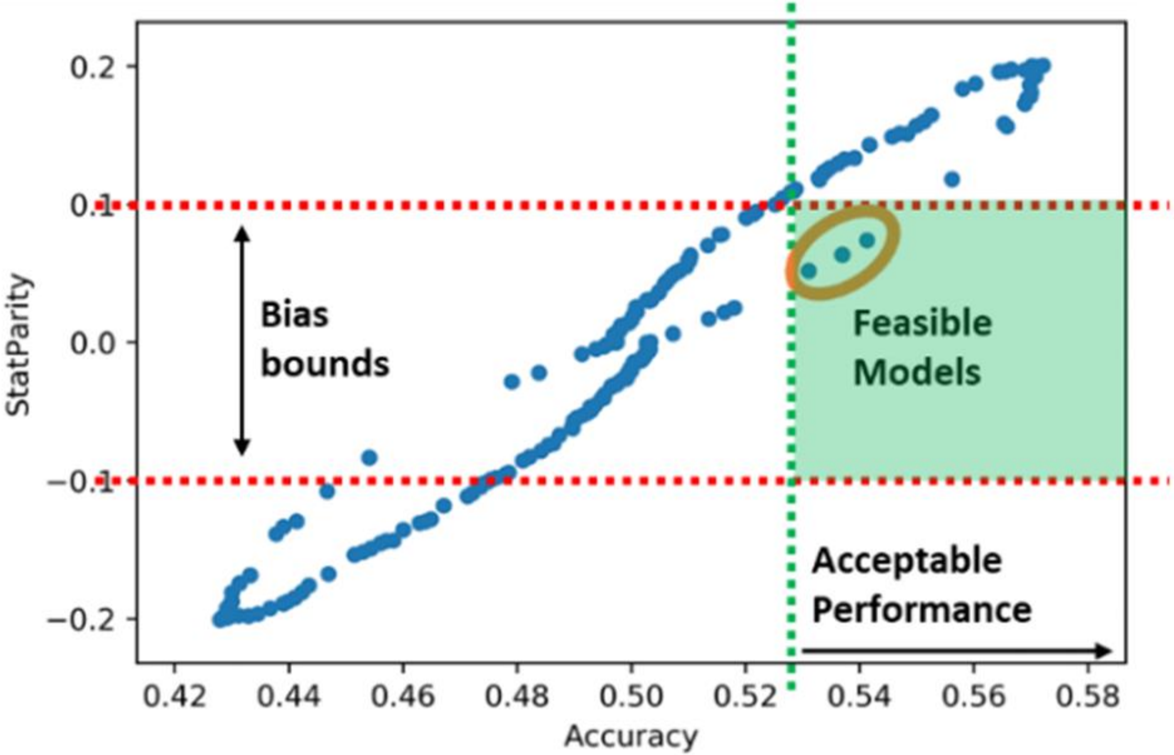
Random empirical samples X~P

– **Distributionally Robust Optimization**

$$\min_{\theta} \sup_{Q \in \mathbb{Q}} \mathbb{E}_Q[\ell(\theta; X)]$$

distributional uncertainty set of this DRO problem (which is composed of probability models which govern the distribution of X - should represent realistic distributional shifts)

Trade-offs: Accuracy vs Fairness



Trade-offs:
Adversarial robustness vs Fairness

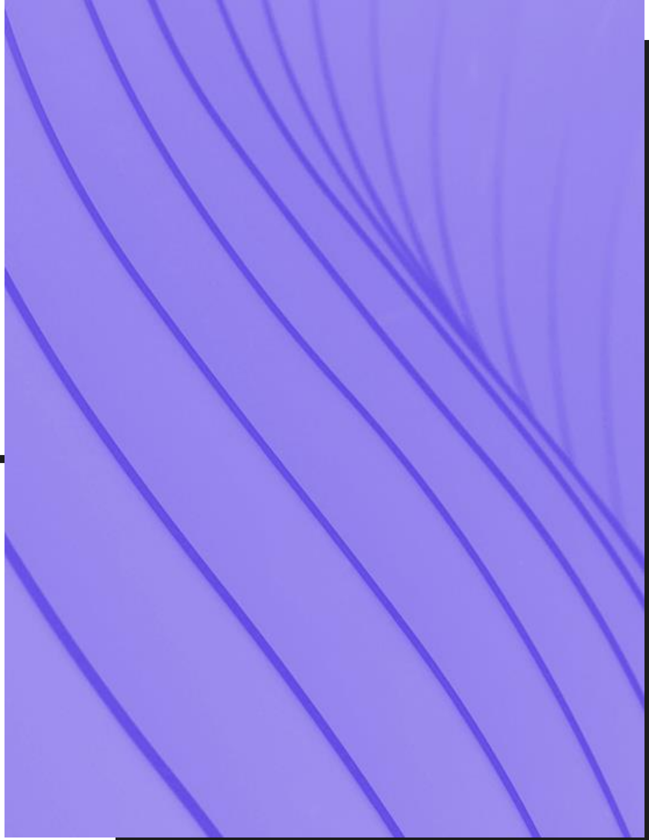
– Paper by Xu et al. 2021

– “robust fairness” problem of adversarial training: large disparity of accuracy and robustness among different classes (not observed in natural training)

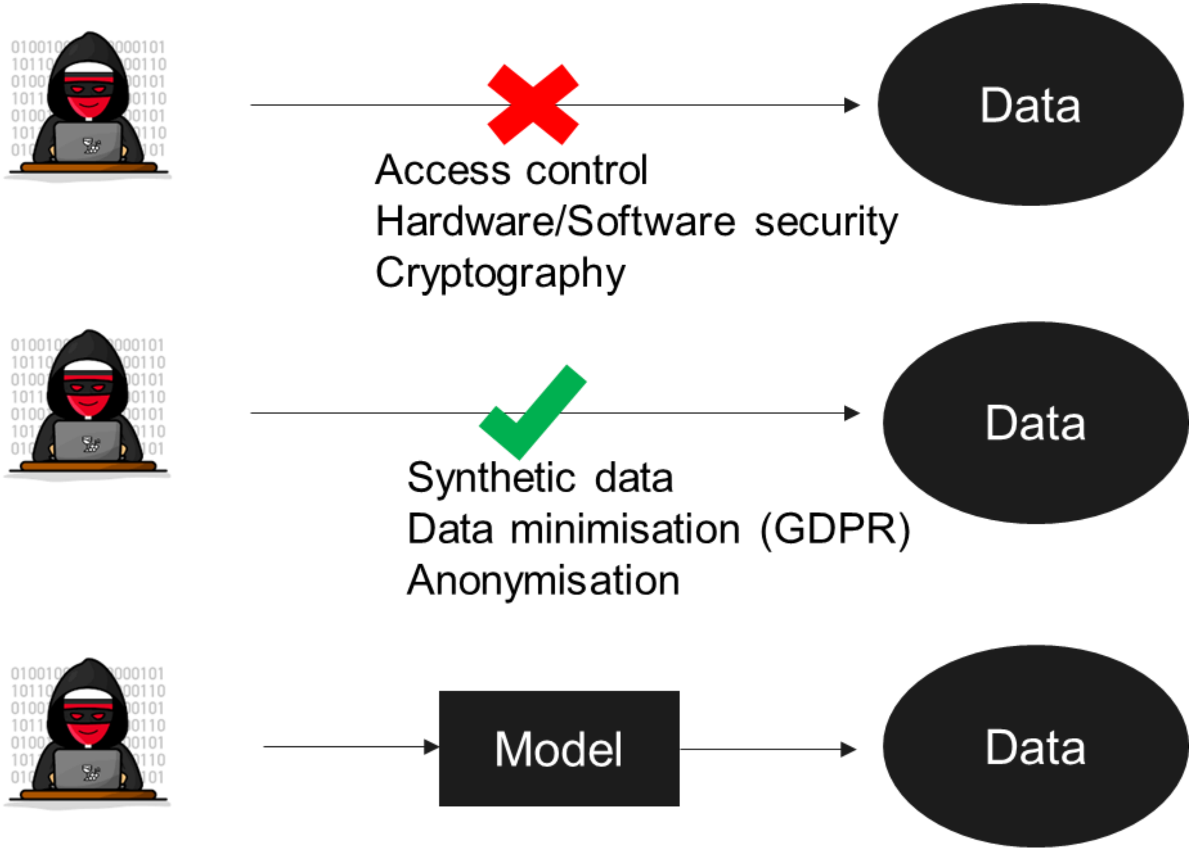
– adversarial training -> tendency to "favor the accuracy of the classes which are “easier” to be predicted."

III – Privacy

- Motivation
- Attacks factors
- Interactions with Fairness



Attacker access



Motivation

- **Membership inference.** Medical study about Alzheimer disease, if hospital records are used to train a model for such a study, one could potentially infer if a particular patient has been used in the study, potentially divulging that the patient may have dementia.

- **Model extraction.** If able to query the model a lot, one can create a mock up model, which can then be used for adversarial attacks.

Membership inference

- Classification & Regression (Gupta et al, 2021)
- Overfitting is the main factor (correlated with increased generalisation error)
- Naive Bayes are less susceptible to membership inference attacks than decision trees or neural networks (Rigaki & Garcia, 2021)
- The more classes, the more signals about the internal state of the model are available to the attacker (Shokri et al, 2017)

Model extraction

– Classification & Regression

– Linear regression/classifiers easy to "reverse engineer" contrary to deep neural networks

– Overfitting prevents attack (opposite for Membership Inference)

– Higher number of classes may lead to worse attack performance (Liu & al, 2021)

Interactions

- **With Fairness.** Sensitive information: sex, gender, religion, ethnicity, etc. overlaps with information required to measure/mitigate group fairness (Chang & Shokri, 2021)

- **With Robustness.** Robust model training (e.g. adversarial training) makes models more susceptible to membership inference attacks as increase generalization error (Raghunatha et al, 2019)

References & Further readings

- "Assessing and Mitigating Bias and Discrimination in AI" Turing course, Milestone 5 (https://github.com/alan-turing-institute/bias-in-AI-course)
- **Explainability:** https://evgenypogorelov.com/multiclass-xgb-shap.html
- **Robustness:** Chen, Ruidi, Boran Hao, and Ioannis Paschalidis. "Distributionally Robust Multiclass Classification and Applications in Deep CNN Image Classifiers." *arXiv preprint arXiv:2109.12772* (2021)
- **Privacy:** Rigaki, Maria, and Sebastian Garcia. "A survey of privacy attacks in machine learning." *arXiv preprint arXiv:2007.07646* (2020).

Conclusion