Three Preprocessing Approaches to Fairness Correction

Visualization-Empowered Human-in-the-Loop AI Focus Period

Vladimiro González-Zelaya

19 May 2025

Linköping University — Campus Norrköping

Talk Outline

Newcastle University

Fairness Notions

Parametrised Data Sampling

Fair and Private Data Correction

Genetic Pipeline Optimisation

Fairness Notions

Protected Attribute (PA)

Definition

Feature of a dataset that is prone to an unjustified discriminatory decision

Examples

- Race or Skin Colour
- Sex or Gender
- Age
- Income Level
- Education
- Nationality

Protected Attribute (PA)

Newcastle University

Definition

Feature of a dataset that is prone to an unjustified discriminatory decision

Examples

- Race or Skin Colour
- Sex or Gender
- Age
- Income Level
- Education
- Nationality

Population Subgroups

Protected Attribute

- Favoured
- Unfavoured

Class

- Positive
- Negative

$U+ \quad F+$

$U- \quad F-$

Population Subgroups

Protected Attribute

- Favoured
- Unfavoured

Class

- Positive
- Negative

$U+ \quad F+$

$U- \quad F-$

Population Subgroups

Protected Attribute

- Favoured
- Unfavoured

Class

- Positive
- Negative

$$U+ \quad F+$$

$$U- \quad F-$$

Some Fairness Definitions

Individual Fairness

Similar individuals should be treated in a **similar way**

Demographic Parity

Same **positive rate** across *PA* groups

Equalised Odds

$\hat{Y}$ and *PA* are **independent**, conditional on Y

Some Fairness Definitions

Individual Fairness

$$d(x_1, x_2) \leqslant \delta \Rightarrow \hat{Y}(x_1) \approx \hat{Y}(x_2)$$

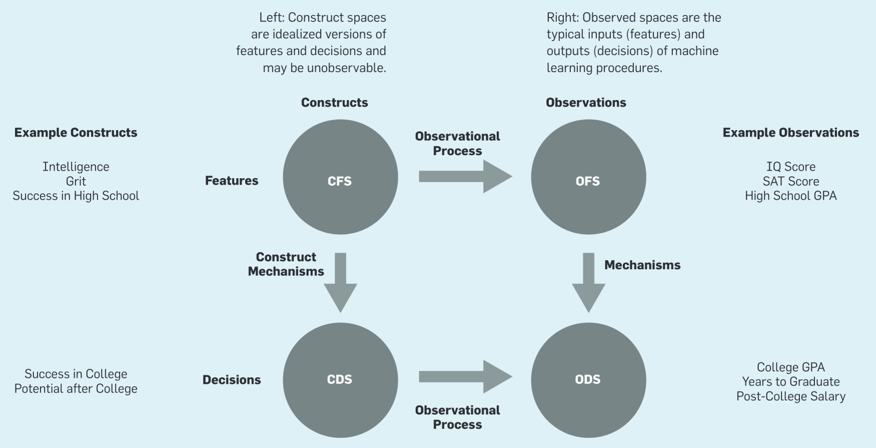
Demographic Parity

$$P(\hat{Y} = 1 \mid PA = 0) = P(\hat{Y} = 1 \mid PA = 1)$$

Equalised Odds

$$P(\hat{Y} = 1 \mid PA = 0, Y = y) = P(\hat{Y} = 1 \mid PA = 1, Y = y), \quad y \in \{0, 1\}$$

The (Im)possibility of Fairness[1]



[1]Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian (2021). "The (Im)possibility of Fairness". In: *Communications of the ACM*.

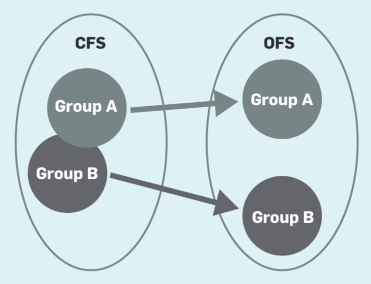
The (Im)possibility of Fairness[1]

WYSIWYG

Construct space and observed space maintain the relative position of individuals w.r.t. the task. Aligns with individual fairness.

We're All Equal

Within a given construct space all groups are essentially the same. Aligns with group fairness.



[1] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian (2021). "The (Im)possibility of Fairness". In: *Communications of the ACM*.

Ways to Correct Unfairness

Pre-Processing Modify the training set to “sample from a better world”

In-Processing Add *constraints* or *regularisation terms* to improve fairness

Post-Processing Adjust the predictions after fitting the model

Ways to Correct Unfairness

Newcastle University

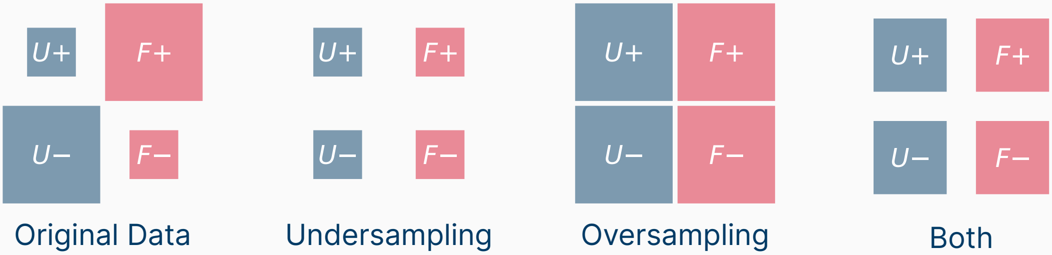
Pre-Processing Modify the training set to “sample from a better world”

In-Processing Add *constraints* or *regularisation terms* to improve fairness

Post-Processing Adjust the predictions after fitting the model

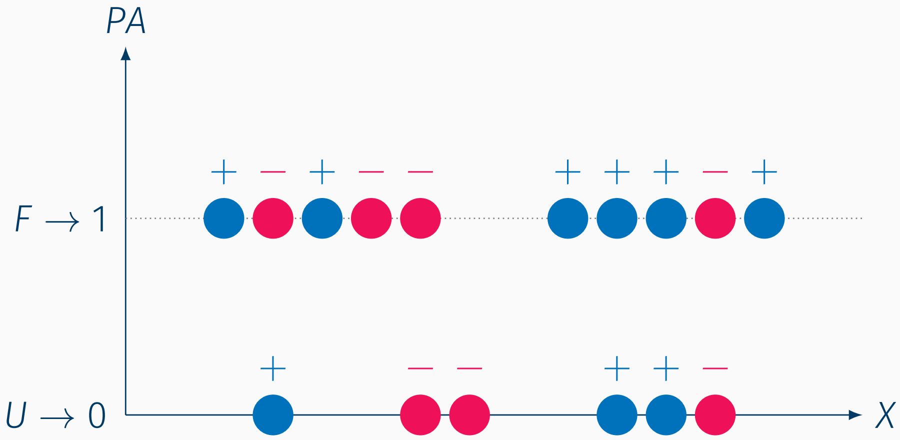
Parametrised Data Sampling

Fairness Correction Through Resampling[2]

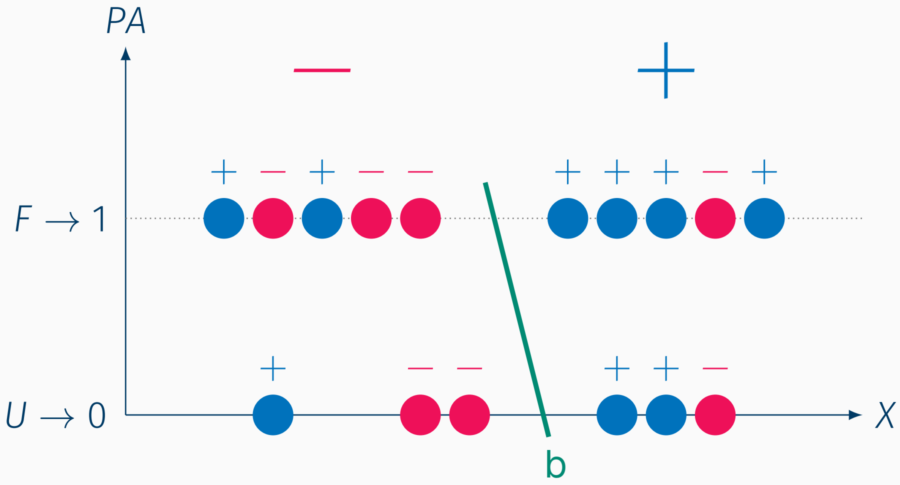


[2] Vladimiro González-Zelaya, Julián Salas, Dennis Prangle, and Paolo Missier (2021). "Optimising Fairness through Parametrised Data Sampling". In: *International Conference on Extending Database Technology*.

An Optimal Linear Classifier (for Accuracy)



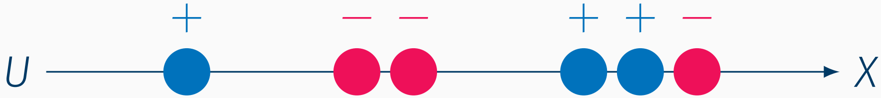
An Optimal Linear Classifier (for Accuracy)



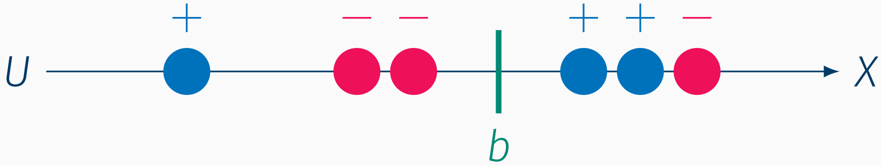
The positive predictions for U increase by:

- Undersampling negative instances
- Oversampling positive instances

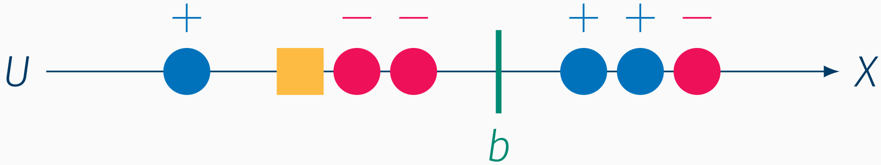
Predictions for U



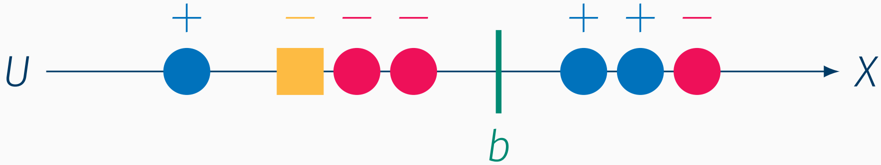
Predictions for U



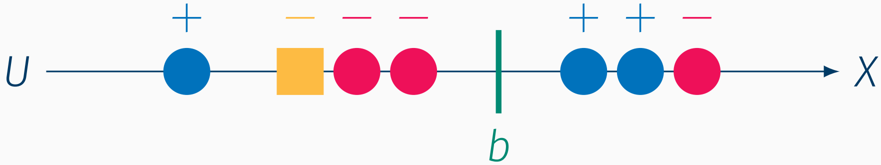
Predictions for U



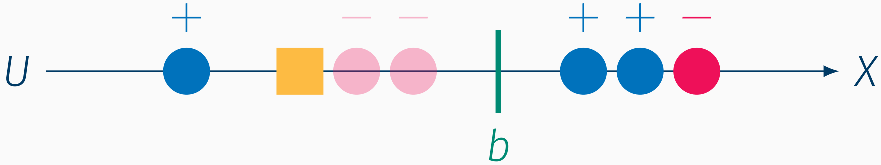
Predictions for U



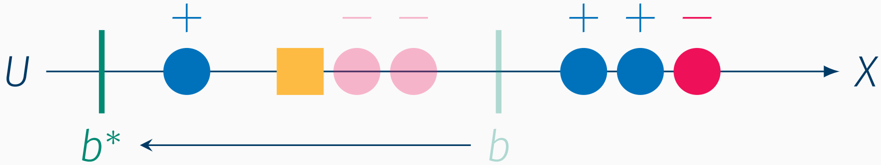
Negative Undersampling Left of *b*



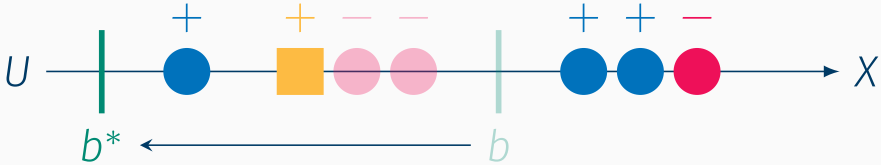
Negative Undersampling Left of b



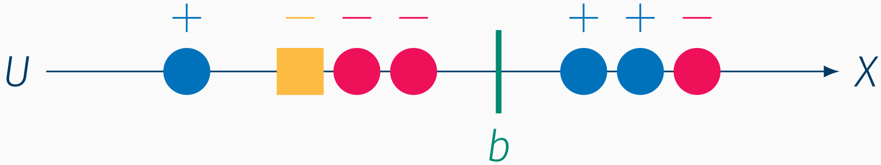
Negative Undersampling Left of b



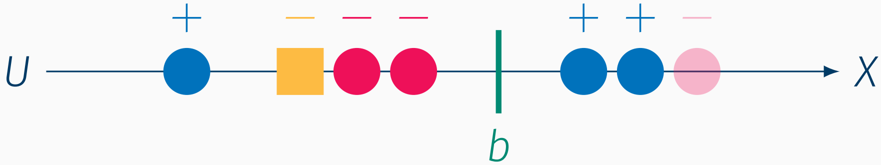
Negative Undersampling Left of b



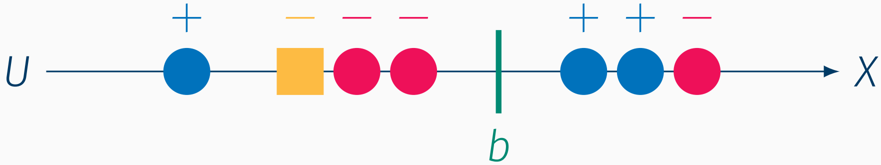
Negative Undersampling Right of b



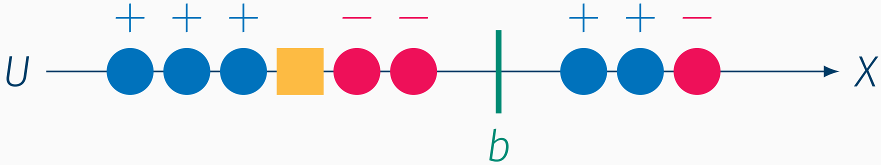
Negative Undersampling Right of b



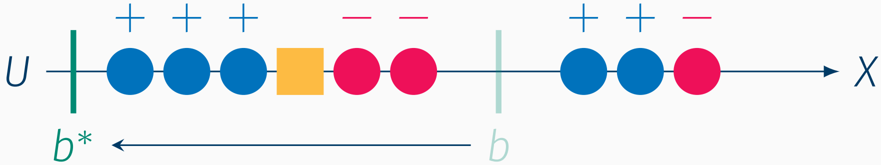
Positive Oversampling Left of b



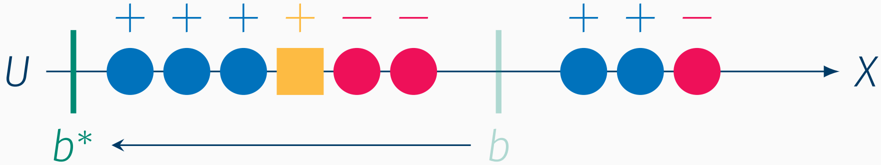
Positive Oversampling Left of b



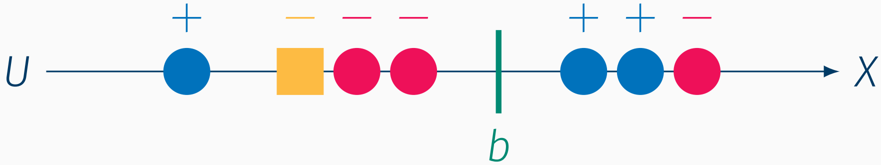
Positive Oversampling Left of b



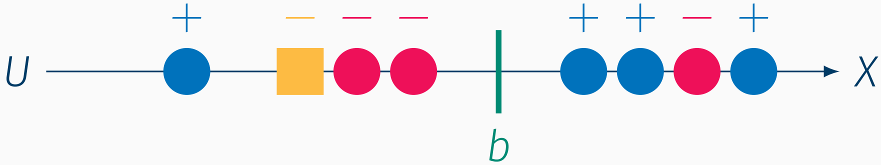
Positive Oversampling Left of b



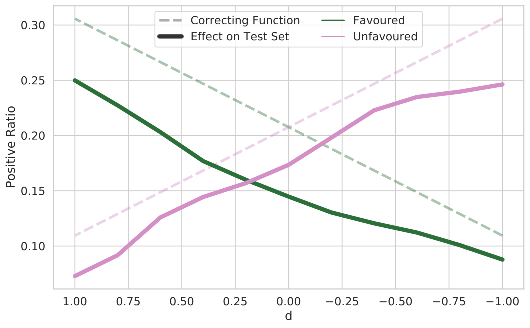
Positive Oversampling Right of b

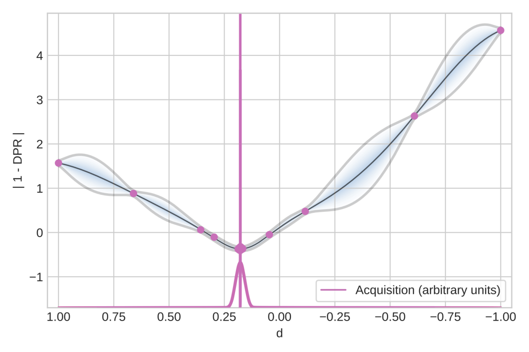


Positive Oversampling Right of b



The Optimal Resampling Parameter





Dealing with Multiple PAs

	Age	Country	Gender	Ethnicity	Combined PA
	(20-30]	Mexico	Male	Latin	Unfavoured
	(30-40]	Canada	Female	White	Favoured
Subgroup PR					
Dataset PR	0.3	0.3	0.3	0.3	
Difference					

Dealing with Multiple PAs

Age	Country	Gender	Ethnicity	Combined PA
(20-30]	Mexico	Male	Latin	Unfavoured
(30-40]	Canada	Female	White	Favoured

Subgroup PR				
Dataset PR	0.3	0.3	0.3	0.3
Difference				

Dealing with Multiple PAs

	Age	Country	Gender	Ethnicity	Combined PA
	(20-30]	Mexico	Male	Latin	Unfavoured
	(30-40]	Canada	Female	White	Favoured
Subgroup PR	0.2	0.3	0.4	0.1	
Dataset PR	0.3	0.3	0.3	0.3	
Difference					

Dealing with Multiple PAs

Newcastle University

	Age	Country	Gender	Ethnicity	Combined PA
	(20-30]	Mexico	Male	Latin	Unfavoured
	(30-40]	Canada	Female	White	Favoured
Subgroup PR	0.2	0.3	0.4	0.1	
Dataset PR	0.3	0.3	0.3	0.3	
Difference					

Dealing with Multiple PAs

	Age	Country	Gender	Ethnicity	Combined PA
	(20-30]	Mexico	Male	Latin	Unfavoured
	(30-40]	Canada	Female	White	Favoured
Subgroup PR	0.2	0.3	0.4	0.1	
Dataset PR	0.3	0.3	0.3	0.3	
Difference	-0.1	$+0.0$	$+0.1$	-0.2	

Dealing with Multiple PAs

	Age	Country	Gender	Ethnicity	Combined PA
	(20-30]	Mexico	Male	Latin	Unfavoured
	(30-40]	Canada	Female	White	Favoured
Subgroup PR	0.2	0.3	0.4	0.1	
Dataset PR	0.3	0.3	0.3	0.3	
Difference	-0.1	$+0.0$	$+0.1$	-0.2	Sum $= -0.2$

Dealing with Multiple PAs

	Age	Country	Gender	Ethnicity	Combined PA
	(20-30]	Mexico	Male	Latin	Unfavoured
	(30-40]	Canada	Female	White	Favoured
Subgroup PR	0.2	0.3	0.4	0.1	
Dataset PR	0.3	0.3	0.3	0.3	
Difference	-0.1	$+0.0$	$+0.1$	-0.2	Sum $= -0.2$

Dealing with Multiple PAs

	Age	Country	Gender	Ethnicity	Combined PA
	(20-30]	Mexico	Male	Latin	Unfavoured
	(30-40]	Canada	Female	White	Favoured
Subgroup PR					
Dataset PR	0.3	0.3	0.3	0.3	
Difference					

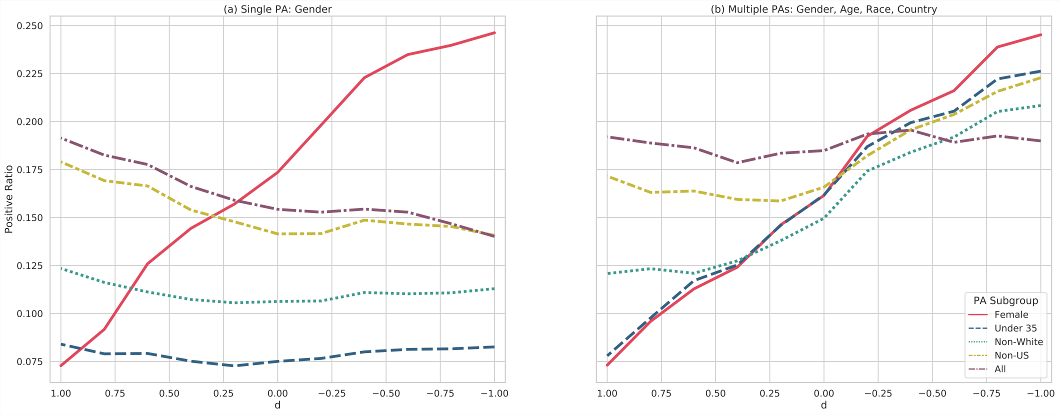
Dealing with Multiple PAs

	Age	Country	Gender	Ethnicity	Combined PA
	(20-30]	Mexico	Male	Latin	Unfavoured
	(30-40]	Canada	Female	White	Favoured
Subgroup PR	0.4	0.4	0.1	0.4	
Dataset PR	0.3	0.3	0.3	0.3	
Difference	+0.1	+0.1	−0.2	+0.1	Sum = +0.1

Dealing with Multiple PAs

	Age	Country	Gender	Ethnicity	Combined PA
	(20-30]	Mexico	Male	Latin	Unfavoured
	(30-40]	Canada	Female	White	Favoured
Subgroup PR	0.4	0.4	0.1	0.4	
Dataset PR	0.3	0.3	0.3	0.3	
Difference	$+0.1$	$+0.1$	-0.2	$+0.1$	Sum $= +0.1$

Multiple PA Correction



Fair and Private Data Correction

Fairness and Privacy

Both aim at concealing sensitive information while preserving data utility:

Fairness To prevent *classifier* behaviours related to sensitive data

Privacy To protect sensitive data from disclosure to *adversaries*

Privacy Definitions

Quasi-Identifiers Collection of features such that their values may be used to **re-identify** an individual

***k*-Anonymity** Dataset records' are **indistinguishable** from at least $k-1$ other records w.r.t. QIs

***k*-Group** Set of **indistinguishable** records in a k-anonymous dset

***t*-Closeness** The **PA distributions** of the k-groups are **similar** to the whole dataset's

Privacy Definitions

Quasi-Identifiers Collection of features such that their values may be used to **re-identify** an individual

***k*-Anonymity** Dataset records' are **indistinguishable** from at least $k-1$ other records w.r.t. QIs

***k*-Group** Set of **indistinguishable** records in a k-anonymous dset

***t*-Closeness** The **PA distributions** of the k-groups are **similar** to the whole dataset's

Privacy Definitions

Quasi-Identifiers Collection of features such that their values may be used to **re-identify** an individual

k-Anonymity Dataset records' are **indistinguishable** from at least $k-1$ other records w.r.t. QIs

k-Group Set of **indistinguishable** records in a k-anonymous dset

t-Closeness The **PA distributions** of the k-groups are **similar** to the whole dataset's

Privacy Definitions

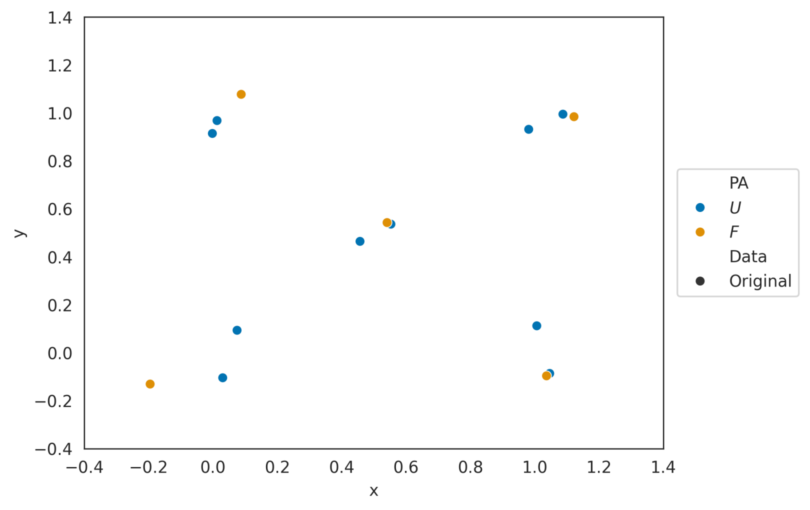
Quasi-Identifiers Collection of features such that their values may be used to **re-identify** an individual

***k*-Anonymity** Dataset records' are **indistinguishable** from at least $k-1$ other records w.r.t. QIs

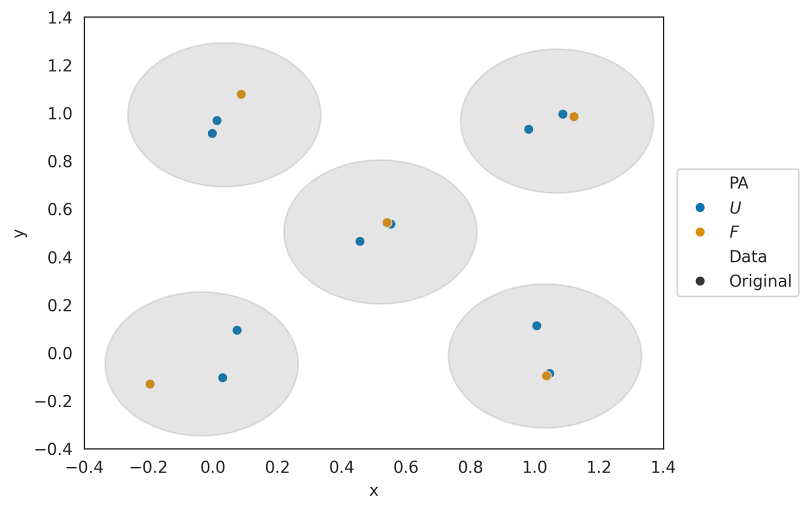
***k*-Group** Set of **indistinguishable** records in a k-anonymous dset

***t*-Closeness** The **PA distributions** of the k-groups are **similar** to the whole dataset's

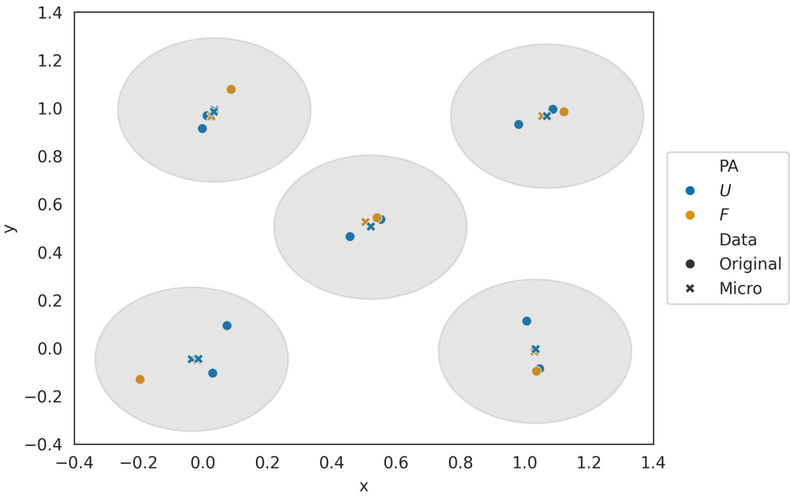
t-Closeness and k-Anonymity Example



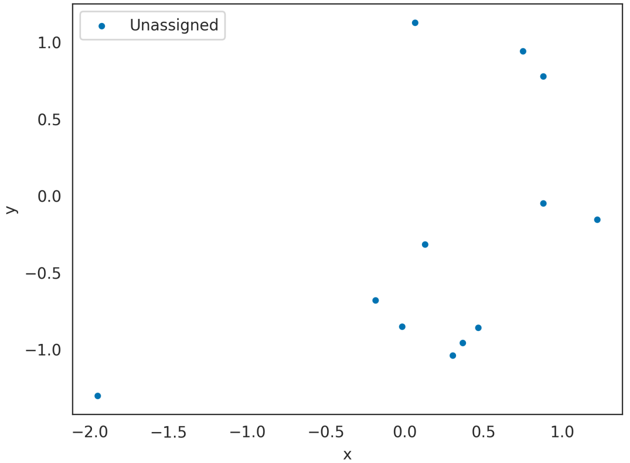
t-Closeness and k-Anonymity Example



t-Closeness and k-Anonymity Example

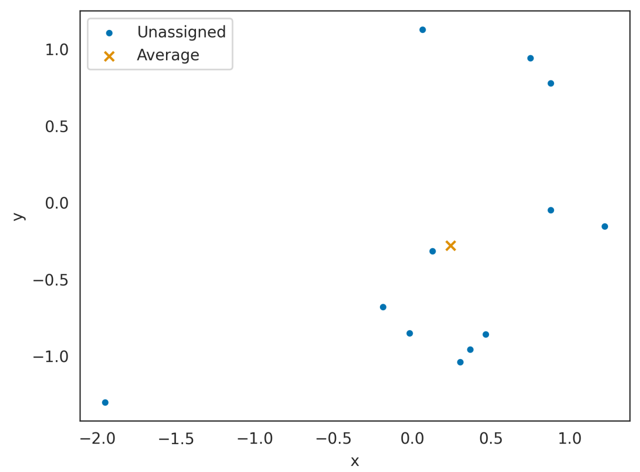


MDAV — Maximum Distance to AVerage[3]



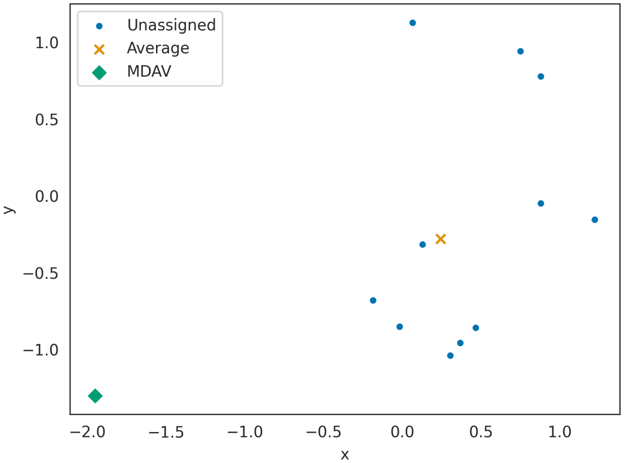
[3] Josep Domingo-Ferrer and Vicenç Torra (2005). "Ordinal, Continuous and Heterogeneous k-Anonymity through Microaggregation". In: *Data Mining and Knowledge Discovery* 11, pp. 195–212.

MDAV — Maximum Distance to AVerage[3]



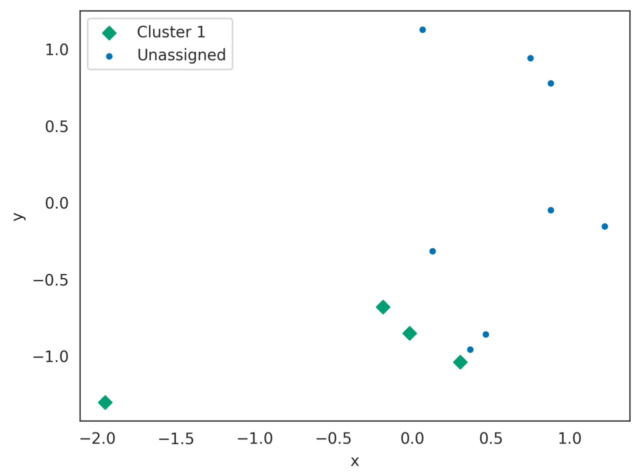
[3] Josep Domingo-Ferrer and Vicenç Torra (2005). "Ordinal, Continuous and Heterogeneous k-Anonymity through Microaggregation". In: *Data Mining and Knowledge Discovery* 11, pp. 195–212.

MDAV — Maximum Distance to AVerage[3]



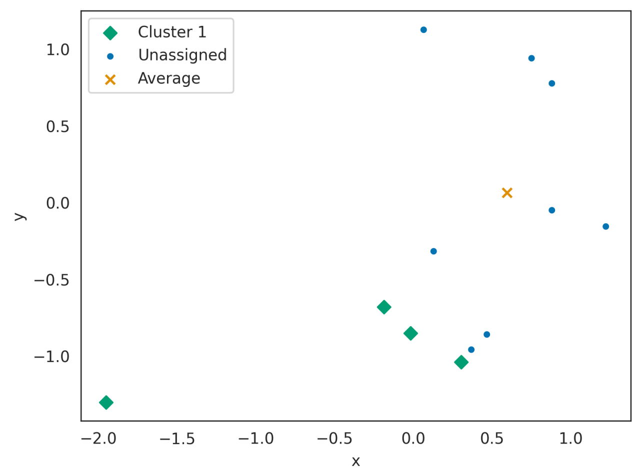
[3] Josep Domingo-Ferrer and Vicenç Torra (2005). "Ordinal, Continuous and Heterogeneous k-Anonymity through Microaggregation". In: *Data Mining and Knowledge Discovery* 11, pp. 195–212.

MDAV — Maximum Distance to AVerage[3]



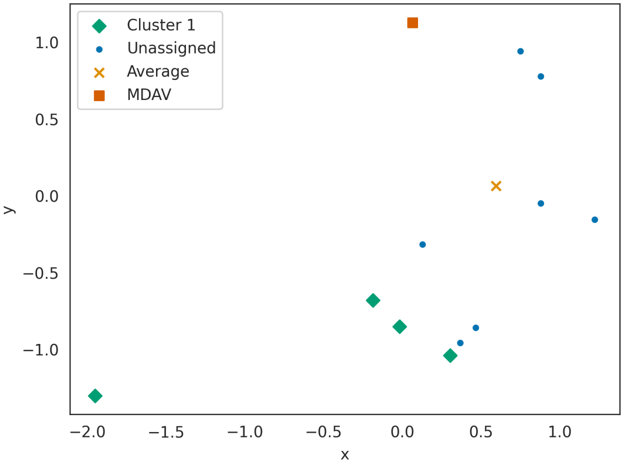
[3] Josep Domingo-Ferrer and Vicenç Torra (2005). "Ordinal, Continuous and Heterogeneous k-Anonymity through Microaggregation". In: *Data Mining and Knowledge Discovery* 11, pp. 195–212.

MDAV — Maximum Distance to AVerage[3]



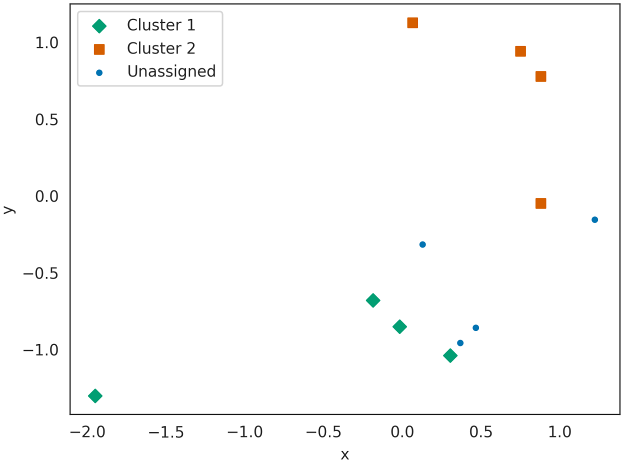
[3] Josep Domingo-Ferrer and Vicenç Torra (2005). "Ordinal, Continuous and Heterogeneous k-Anonymity through Microaggregation". In: *Data Mining and Knowledge Discovery* 11, pp. 195–212.

MDAV — Maximum Distance to AVerage[3]



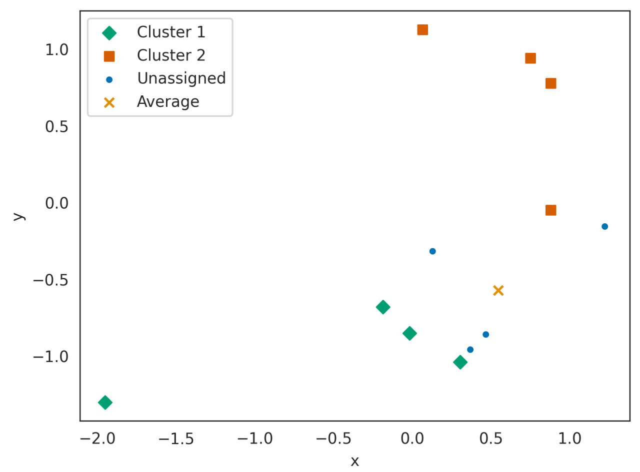
[3] Josep Domingo-Ferrer and Vicenç Torra (2005). "Ordinal, Continuous and Heterogeneous k-Anonymity through Microaggregation". In: *Data Mining and Knowledge Discovery* 11, pp. 195–212.

MDAV — Maximum Distance to AVerage[3]



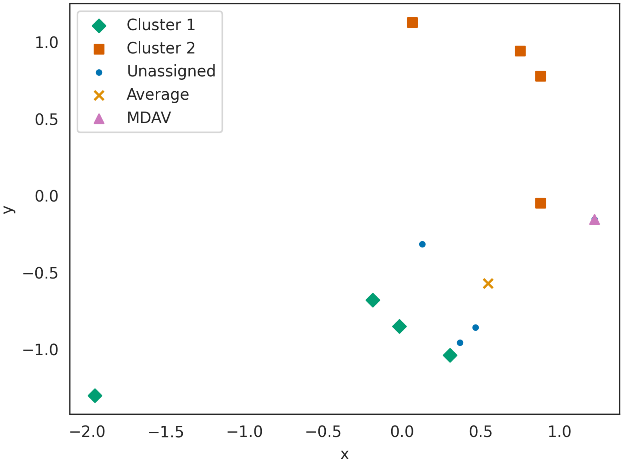
[3] Josep Domingo-Ferrer and Vicenç Torra (2005). "Ordinal, Continuous and Heterogeneous k-Anonymity through Microaggregation". In: *Data Mining and Knowledge Discovery* 11, pp. 195–212.

MDAV — Maximum Distance to AVerage[3]



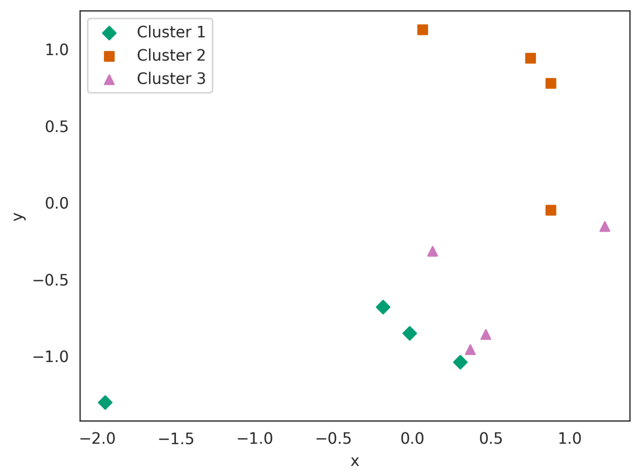
[3] Josep Domingo-Ferrer and Vicenç Torra (2005). "Ordinal, Continuous and Heterogeneous k-Anonymity through Microaggregation". In: *Data Mining and Knowledge Discovery* 11, pp. 195–212.

MDAV — Maximum Distance to AVerage[3]



[3] Josep Domingo-Ferrer and Vicenç Torra (2005). "Ordinal, Continuous and Heterogeneous k-Anonymity through Microaggregation". In: *Data Mining and Knowledge Discovery* 11, pp. 195–212.

MDAV — Maximum Distance to AVerage[3]



[3] Josep Domingo-Ferrer and Vicenç Torra (2005). "Ordinal, Continuous and Heterogeneous k-Anonymity through Microaggregation". In: *Data Mining and Knowledge Discovery* 11, pp. 195–212.

Fair-MDAV[4] Overview

1. **Cluster** D into (m, n)-*fairlets* (sets of m unfavoured and n favoured records) using the MDAV algorithm, where

$$\frac{m}{n} \approx \frac{|U|}{|F|}, \text{ subject to } m + n = k$$

2. **Microaggregate** the feature values with their corresponding fairlet's mean/mode, except for PA and Class, whose original values are kept
3. **Locally correct the fairness** of each fairlet by relabelling its records depending on their PA values, so that

$$\frac{|U^+|}{|U|} \geqslant \tau \cdot \frac{|F^+|}{|F|}, \text{ where } \tau \text{ modulates the correction}$$

[4] Vladimiro González-Zelaya, Julián Salas, David Megías, and Paolo Missier (2023). "Fair and Private Data Preprocessing through Microaggregation". In: *ACM Transactions on Knowledge Discovery from Data*.

Fair-MDAV[4] Overview

1. **Cluster** D into (m, n)-*fairlets* (sets of m unfavoured and n favoured records) using the MDAV algorithm, where

$$\frac{m}{n} \approx \frac{|U|}{|F|}, \text{ subject to } m + n = k$$

2. **Microaggregate** the feature values with their corresponding fairlet's mean/mode, except for PA and Class, whose original values are kept
3. **Locally correct the fairness** of each fairlet by relabelling its records depending on their PA values, so that

$$\frac{|U^+|}{|U|} \geqslant \tau \cdot \frac{|F^+|}{|F|}, \text{ where } \tau \text{ modulates the correction}$$

[4] Vladimiro González-Zelaya, Julián Salas, David Megías, and Paolo Missier (2023). "Fair and Private Data Preprocessing through Microaggregation". In: *ACM Transactions on Knowledge Discovery from Data*.

Fair-MDAV[4] Overview

1. **Cluster** D into (m, n)-*fairlets* (sets of m unfavoured and n favoured records) using the MDAV algorithm, where

$$\frac{m}{n} \approx \frac{|U|}{|F|}, \text{ subject to } m + n = k$$

2. **Microaggregate** the feature values with their corresponding fairlet's mean/mode, except for PA and Class, whose original values are kept
3. **Locally correct the fairness** of each fairlet by relabelling its records depending on their PA values, so that

$$\frac{|U^+|}{|U|} \geqslant \tau \cdot \frac{|F^+|}{|F|}, \text{ where } \tau \text{ modulates the correction}$$

[4] Vladimiro González-Zelaya, Julián Salas, David Megías, and Paolo Missier (2023). "Fair and Private Data Preprocessing through Microaggregation". In: *ACM Transactions on Knowledge Discovery from Data*.

Fair-MDAV Example ($k = 3, \tau = 1$)

Example Data

id	X	PA	Class
a	1	F	1
b	2	U	0
c	3	U	1
d	11	F	0
e	12	F	0
f	13	F	1
g	14	F	1

→

(1, 2)-Fairlets

id	X	X_{ma}	PA	Class	Fair Class
a	1	4.67	F	1	1
b	2	4.67	U	0	1
c	3	9.33	U	1	1
d	11	4.67	F	0	0
e	12	9.33	F	0	0
f	13	9.33	F	1	1
g	*Dropped*				

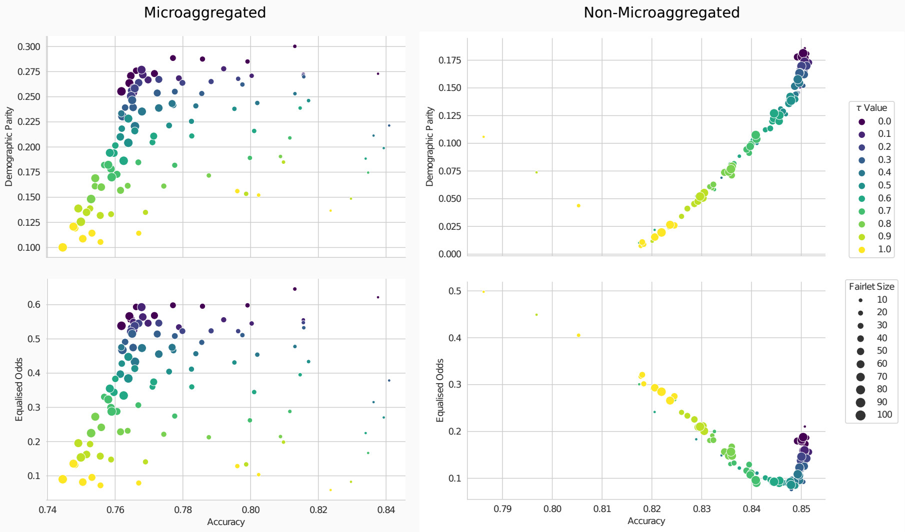
Experimental Parameter Values

The following parameter values were tested over three benchmark datasets [5]:

Description	Parameter	Values
Fairlet Size	m	$\left\lceil k \cdot \frac{\lvert U \rvert}{\lvert D \rvert} \right\rceil$ for $k \in \{10, 20, \cdots, 100\}$
	n	$k - m$ for $k \in \{10, 20, \cdots, 100\}$
Fairness Correction	τ	$0, 0.1, \cdots, 1$
Microaggregation	ma	`True`, `False`

[5] *Census Income*, *COMPAS*, and *German Credit*, available online

Fairness/Accuracy Trade-Off (Census Income Dataset)



Genetic Pipeline Optimisation

Preprocessing Pipelines

- Steps that transform the raw input data into its final form as a training set
- Some are *required* by the classification framework:
  - Encoding categorical variables
  - Imputing missing data
- Others may *optionally* be deployed:
  - Class balancing
  - Feature selection
  - Feature scaling
- Steps usually combined into *pipelines* based on best-practice considerations, with model *performance* as the main objective

Preprocessing Pipelines

- Steps that transform the raw input data into its final form as a training set
- Some are *required* by the classification framework:
  - Encoding categorical variables
  - Imputing missing data
- Others may *optionally* be deployed:
  - Class balancing
  - Feature selection
  - Feature scaling
- Steps usually combined into *pipelines* based on best-practice considerations, with model *performance* as the main objective

Preprocessing Pipelines

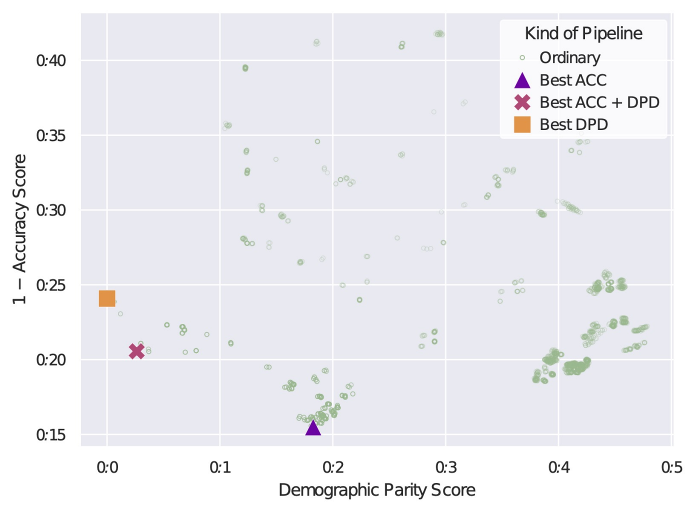
- Steps that transform the raw input data into its final form as a training set
- Some are *required* by the classification framework:
  - Encoding categorical variables
  - Imputing missing data
- Others may *optionally* be deployed:
  - Class balancing
  - Feature selection
  - Feature scaling
- Steps usually combined into *pipelines* based on best-practice considerations, with model *performance* as the main objective

Preprocessing Pipelines

Newcastle University

- Steps that transform the raw input data into its final form as a training set
- Some are *required* by the classification framework:
  - Encoding categorical variables
  - Imputing missing data
- Others may *optionally* be deployed:
  - Class balancing
  - Feature selection
  - Feature scaling
- Steps usually combined into *pipelines* based on best-practice considerations, with model *performance* as the main objective

Pipeline-Space Fairness/Accuracy and Pareto Front



The FairPipes Algorithm[6]

Preprocessors and Ordering

Gene	Tag	Values
Encoder	enc	one-hot, ordinal, target, leave one out (LOO), weight of evidence (WoE), count
Imputer	imp	mean, median, most frequent
Sampler	samp	none, under, over
Scaler	scale	none, maxabs, minmax, norm, quantile
Feature Selector	sel	none, k-best
Order		(samp, scale, sel), (samp, sel, scale), ...

(a) Generate *n* random pipelines

id	**Random Pipelines**
1	one-hot_enc, mean_imp, under_samp, norm_scale, k-best_sel
2	ordinal_enc, mean_imp, under_samp, k-best_sel, norm_scale
3	one-hot_enc, median_imp, k-best_sel, norm_scale, no_samp
4	count_enc, mean_imp, over_samp, no_sel, norm_scale

(b) Preprocess data, learn classifiers and sort pipelines by objective metric; top one becomes the *elite* [id 3]

id	**Sorted Pipelines**	Objective	P_{parent}	$P_{survive}$
3	one-hot_enc, median_imp, k-best_sel, norm_scale, no_samp	0.276	4/10	1
4	count_enc, mean_imp, over_samp, no_sel, norm_scale	0.406	3/10	3/6
1	one-hot_enc, mean_imp, under_samp, norm_scale, k-best_sel	0.689	2/10	2/6
2	ordinal_enc, mean_imp, under_samp, k-best_sel, norm_scale	0.783	1/10	1/6

(c) Choose parents with P_{parent} [ids 3, 4]

id	**Parent Pipelines**
3	one-hot_enc, median_imp, k-best_sel, norm_scale, no_samp
4	count_enc, mean_imp, over_samp, no_sel, norm_scale

(d) Choose a gene with uniform probability [Sampler] and crossover the parents' gene-values [ids 3, 4 → 5, 6]

id	**Crossover Pipelines**
5	one-hot_enc, median_imp, k-best_sel, norm_scale, **over_samp**
6	count_enc, mean_imp, **no_samp**, no_sel, norm_scale

(e) Append a *non-elite* pipeline with $P_{survive}$ [id 1]

id	**Mutable Pipelines**
5	one-hot_enc, median_imp, k-best_sel, norm_scale, over_samp
6	count_enc, mean_imp, no_samp, no_sel, norm_scale
1	one-hot_enc, mean_imp, under_samp, norm_scale, k-best_sel

(f) Each gene *may* mutate with $P(\text{mutate}) = m$ [Order in id 1 → 7]

id	**Mutated Pipelines**
5	one-hot_enc, median_imp, k-best_sel, norm_scale, over_samp
6	count_enc, mean_imp, no_samp, no_sel, norm_scale
7	one-hot_enc, mean_imp, **k-best_sel, under_samp, norm_scale**

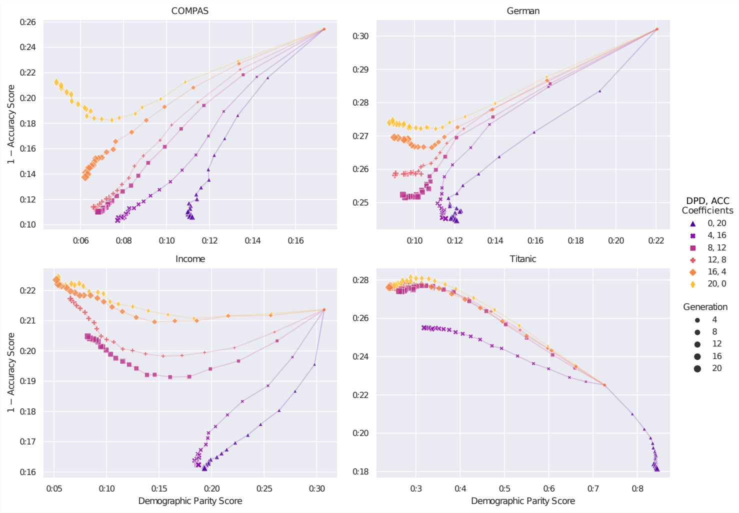
(g) Append the non-mutated *elite* [id 3]

id	**First Generation Pipelines**	Objective
5	one-hot_enc, median_imp, k-best_sel, norm_scale, over_samp	0.458
6	count_enc, mean_imp, no_samp, no_sel, norm_scale	0.232
7	one-hot_enc, mean_imp, k-best_sel, under_samp, norm_scale	0.520
3	one-hot_enc, median_imp, k-best_sel, norm_scale, no_samp	0.276

For subsequent generations, repeat from (b)

[6] Vladimiro González-Zelaya, Julián Salas, Dennis Prangle, and Paolo Missier (2023). "Preprocessing Matters: Automated Pipeline Selection for Fair Classification". In: *MDAI 2023*.

Evolution Toward Optimal Solutions



Conclusions

- There are multiple incompatible fairness definitions
- Definitions should align with world-views
- Possible to correct fairness through data pre-processing
- Can use fairness-specific methods or optimise pipeline choices
- Fairness and privacy can be achieved simultaneously

Conclusions

- There are multiple incompatible fairness definitions
- Definitions should align with world-views
- Possible to correct fairness through data pre-processing
- Can use fairness-specific methods or optimise pipeline choices
- Fairness and privacy can be achieved simultaneously

Conclusions

- There are multiple incompatible fairness definitions
- Definitions should align with world-views
- Possible to correct fairness through data pre-processing
- Can use fairness-specific methods or optimise pipeline choices
- Fairness and privacy can be achieved simultaneously

Conclusions

- There are multiple incompatible fairness definitions
- Definitions should align with world-views
- Possible to correct fairness through data pre-processing
- Can use fairness-specific methods or optimise pipeline choices
- Fairness and privacy can be achieved simultaneously

Conclusions

Newcastle University

- There are multiple incompatible fairness definitions
- Definitions should align with world-views
- Possible to correct fairness through data pre-processing
- Can use fairness-specific methods or optimise pipeline choices
- Fairness and privacy can be achieved simultaneously