



PEBAM

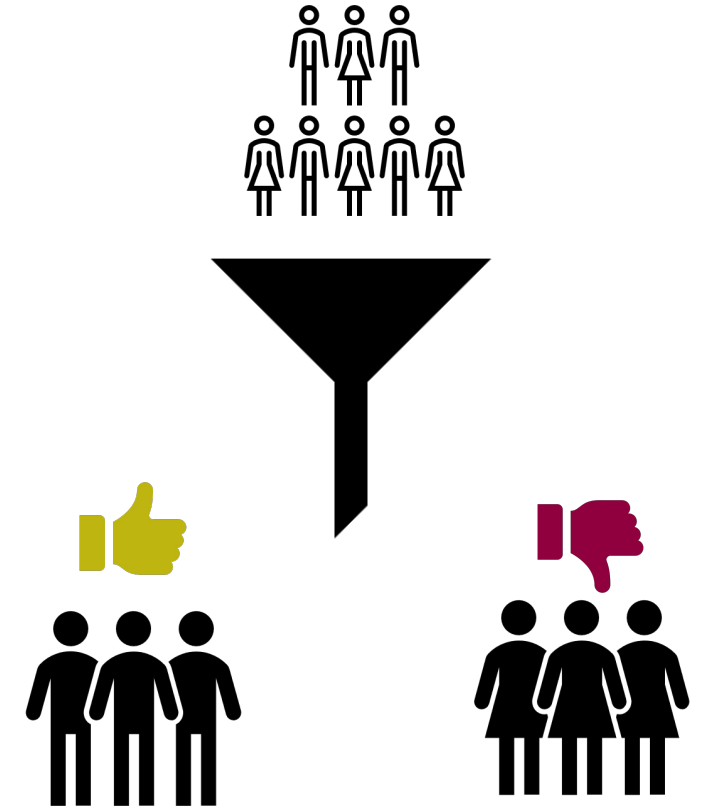
a Profile-based Evaluation method for
Bias Assessment on Mixed datasets

KI 2022: 45th German Conference on Artificial Intelligence, 22nd September

Mieke Wilms, Giovanni Sileno, Hinda Haned

Bias in Algorithmic Decision Making

The uptake of algorithmic decision-making (ADM) in an increasing number of human activities increases the risks of unfair, unacceptable discrimination.





Bias?

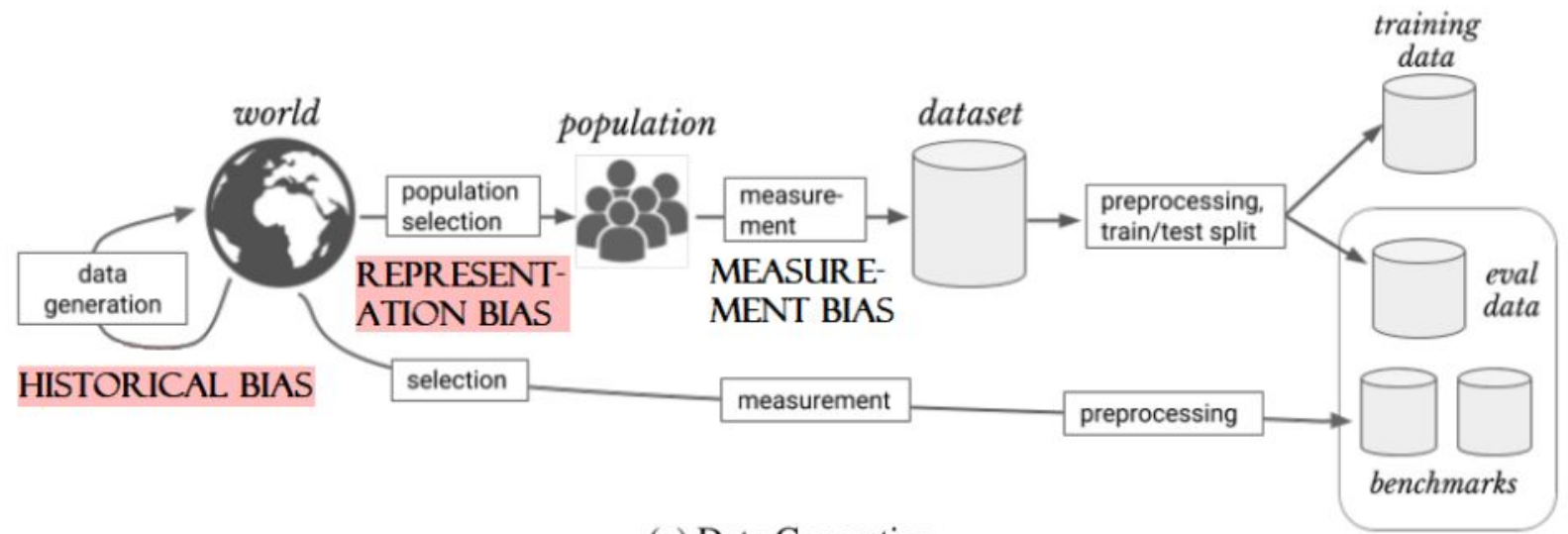
Common-sense use: *“A strong feeling in favour of or against one group of people, or one side in an argument, often not based on fair judgement.”*

Technical uses: eg. voltage bias, statistical bias (systematic difference between inferred and actual)

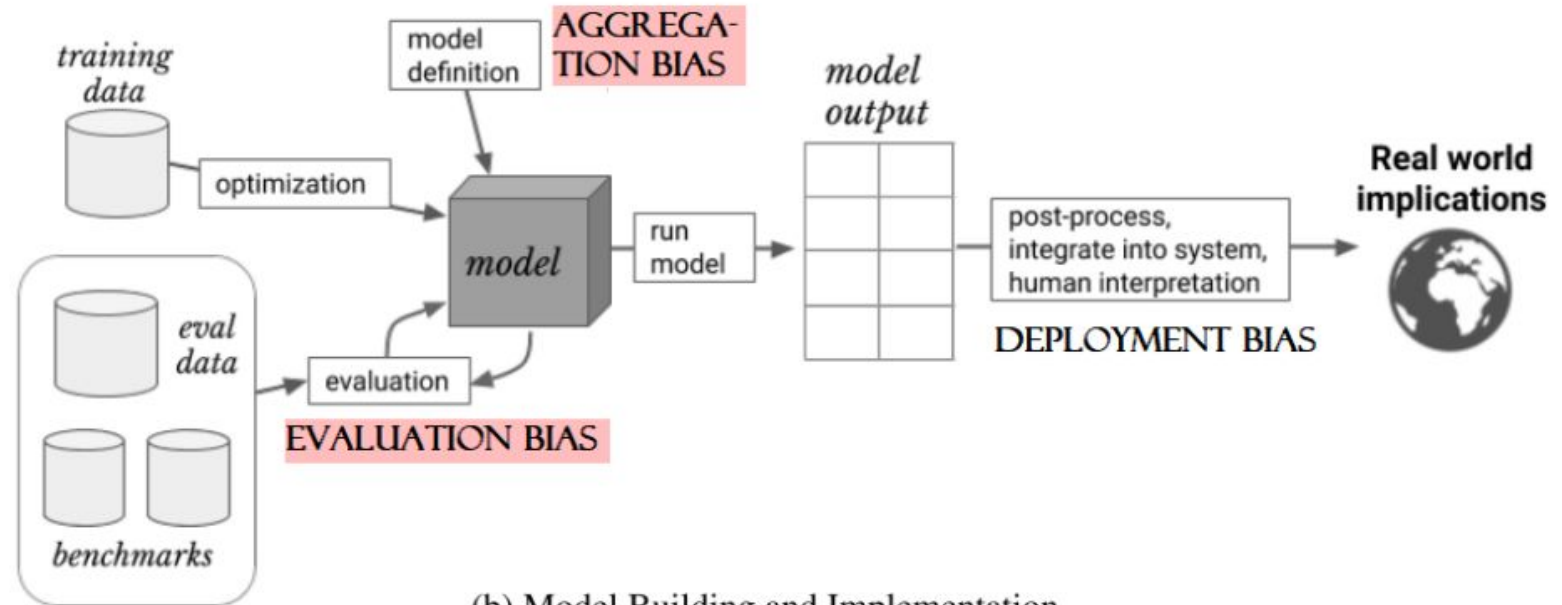
And in Algorithmic Decision-Making?

Bias?

In practice there is **not a single definition** of bias in Algorithmic Decision-Making, since there are different sources from where “bias” can arise during lifecycle of a machine learning model.



(a) Data Generation



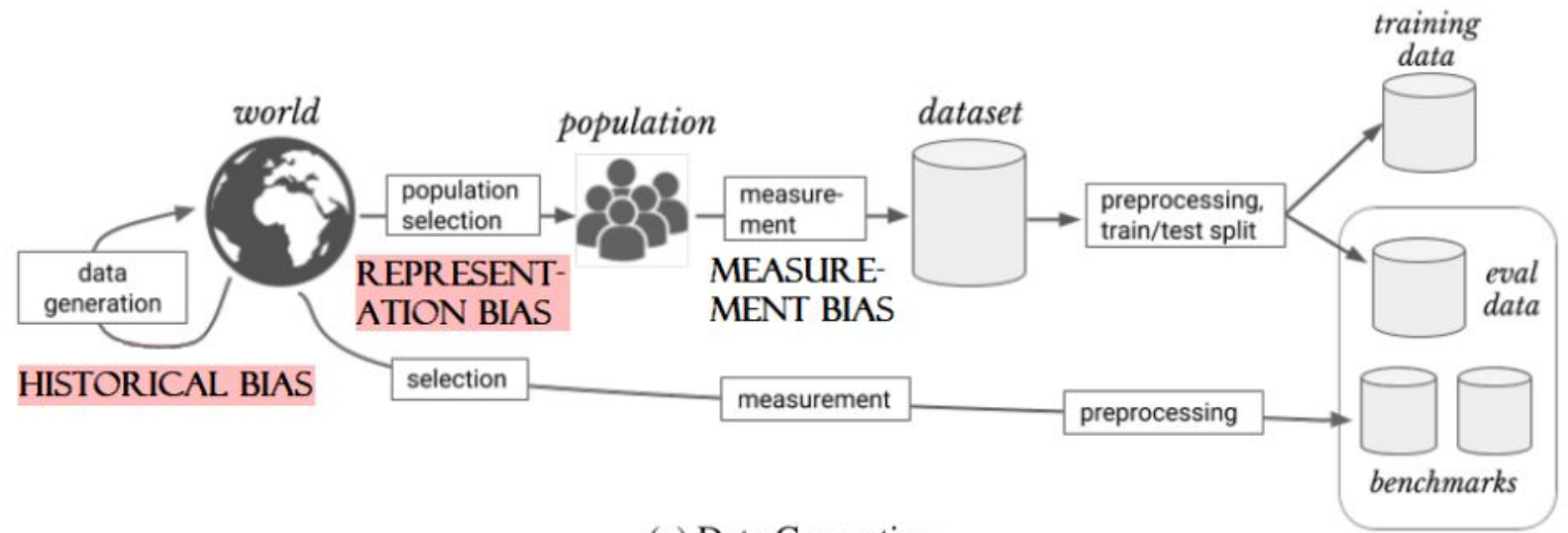
(b) Model Building and Implementation

Overview of the lifecycle of a machine learning model by H. Suresh and J.V. Gutttag in *A Framework for Understanding Unintended Consequences of Machine Learning*.

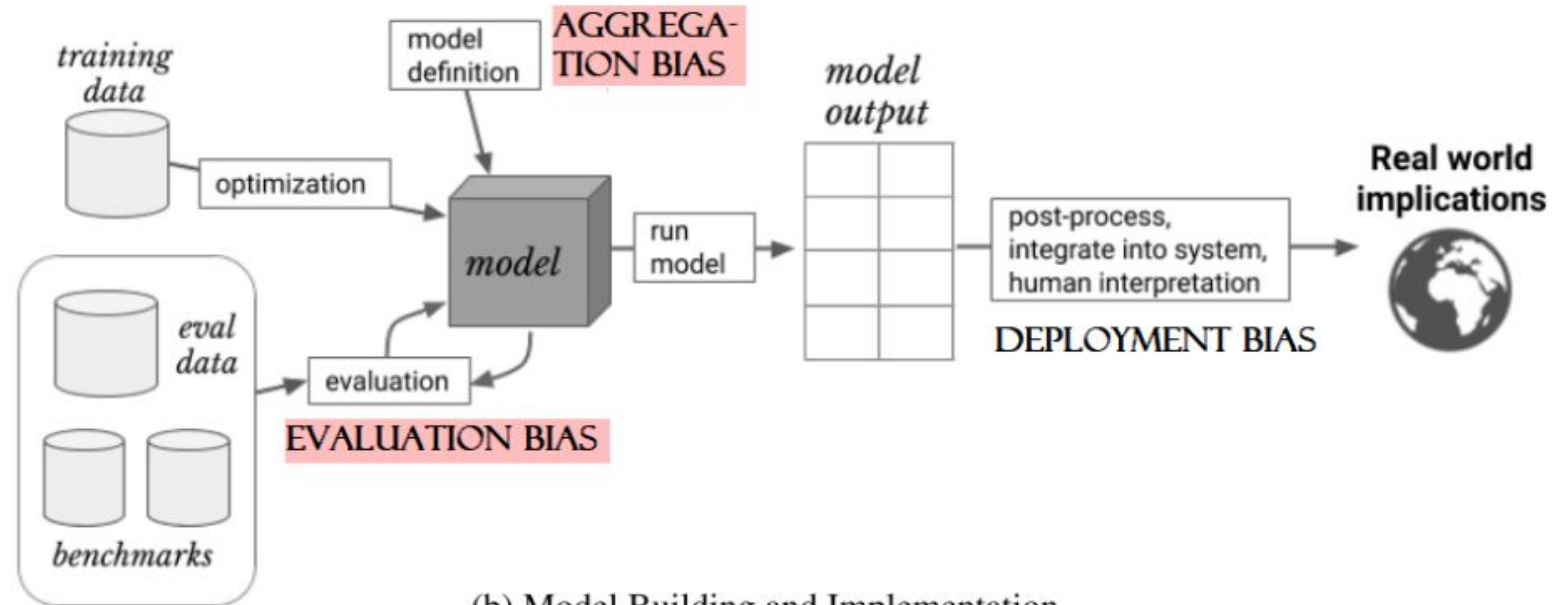
Bias?

In practice there is **not a single definition** of bias in Algorithmic Decision-Making, since there are different sources from where “bias” can arise during lifecycle of a machine learning model.

Not all biases are unacceptable, or unfair!



(a) Data Generation



(b) Model Building and Implementation



Fairness metrics

Algorithmic fairness is a research track aiming to measure fairness of algorithmic decision-making.

Several (and incompatible) definition of fairness have been presented.

Most fairness metrics focus on either **individual fairness** or **group fairness**, where a group is defined by a protected attribute (e.g. ethnicity, gender, etc).

In this work we elaborate on fairness with respect to **profiles**: groups of people that share similar (protected and non-protected) attributes?



Research Questions

- “ *How can **profiles** be defined?*
*How can we extract profiles from **mixed datasets**?*
*How can profiles be used to **assess bias**?*

Research Questions

“ *How can **profiles** be defined?*

*How can we extract profiles from **mixed datasets**?*

*How can profiles be used to **assess bias**?*

group of samples in a dataset that
share similar feature values

samples include numeric
and categorical features

in particular:
historical bias
representation bias
aggregation bias
evaluation bias

PEBAM

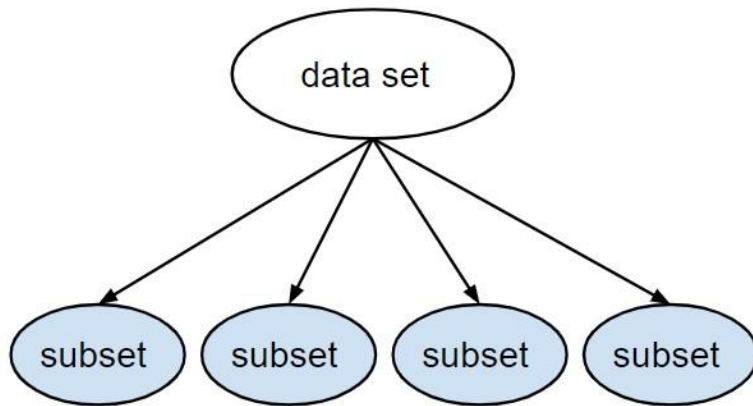
With this work, we introduce a **Profile-based Evaluation** method for **Bias Assessment** on **Mixed** datasets, PEBAM for short.

PEBAM consists of the following steps:

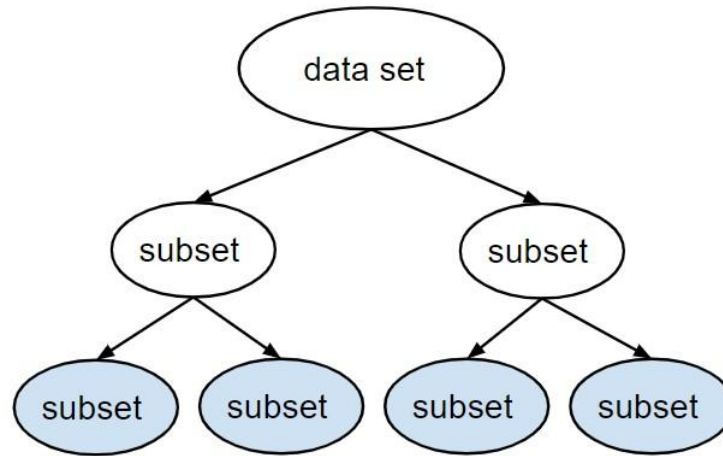
- i. **Selection of profiles** from the mixed datasets input data
- ii. **Evaluation** of the selected profiles in terms of **stability**
- iii. **Evaluation of bias(es) of a given classifier** with respect to those profiles

(i) Profile selection

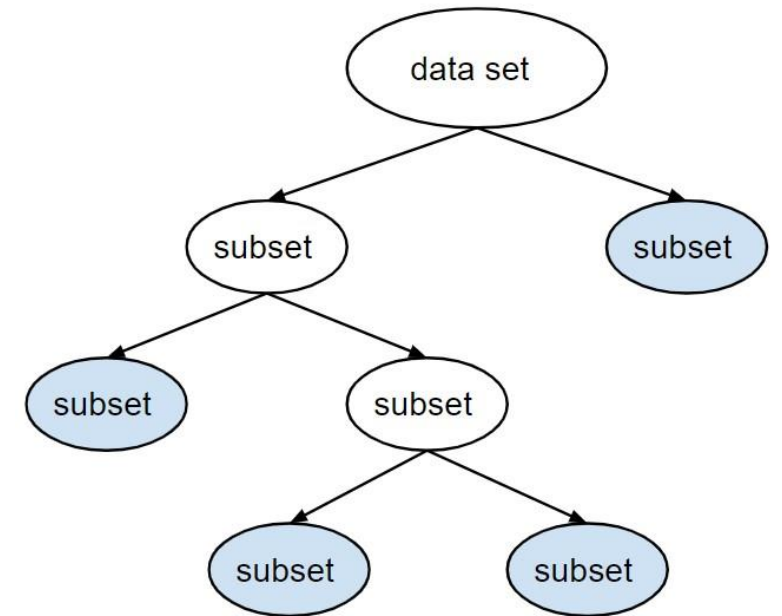
To extract profiles from datasets, we compare a baseline “simple” (a) and a “hierarchical” (b) clustering algorithms with a more advanced “semi-hierarchical” clustering algorithm (c).



(a)



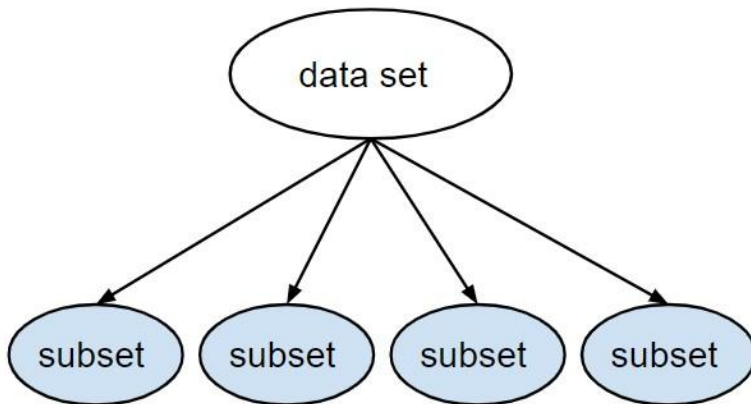
(b)



(c)

(i) Profile selection

To extract profiles from datasets, we compare a baseline “**simple**” (a) and a “hierarchical” (b) clustering algorithms with a more advanced “semi-hierarchical” clustering algorithm (c).



(a)

Average behaviour is not behaviour?

We make use of **k-medoids clustering** as basis to find groups of samples with similar attributes (profiles).

Mixed dataset?

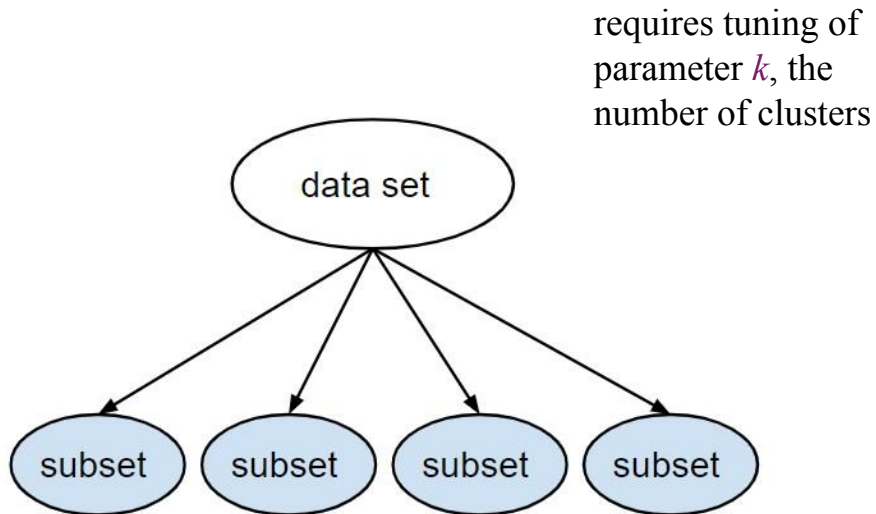
We use **Gower distance** for handling numeric and nominal features.

Tractability?

We use the **PAM** heuristics (Partitioning Around Medoids).

(i) Profile selection

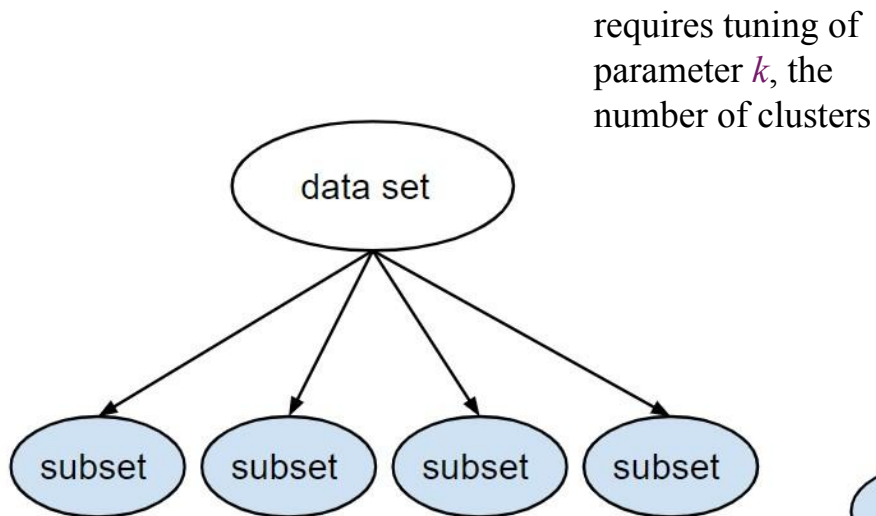
To extract profiles from datasets, we compare a baseline “**simple**” (a) and a “hierarchical” (b) clustering algorithms with a more advanced “semi-hierarchical” clustering algorithm (c).



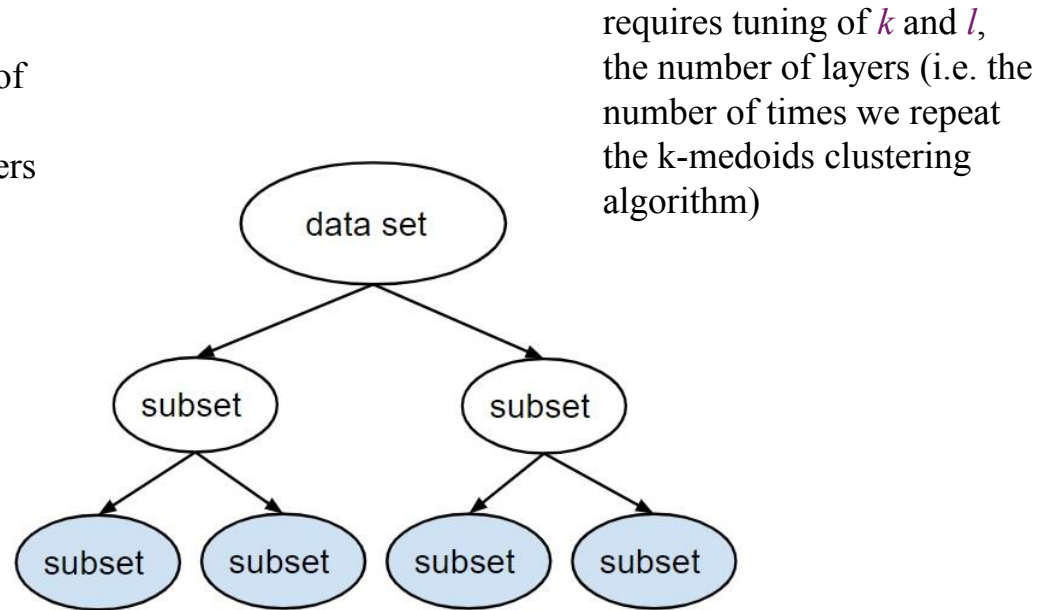
(a)

(i) Profile selection

To extract profiles from datasets, we compare a baseline “**simple**” (a) and a “**hierarchical**” (b) clustering algorithms with a more advanced “semi-hierarchical” clustering algorithm (c).



(a)



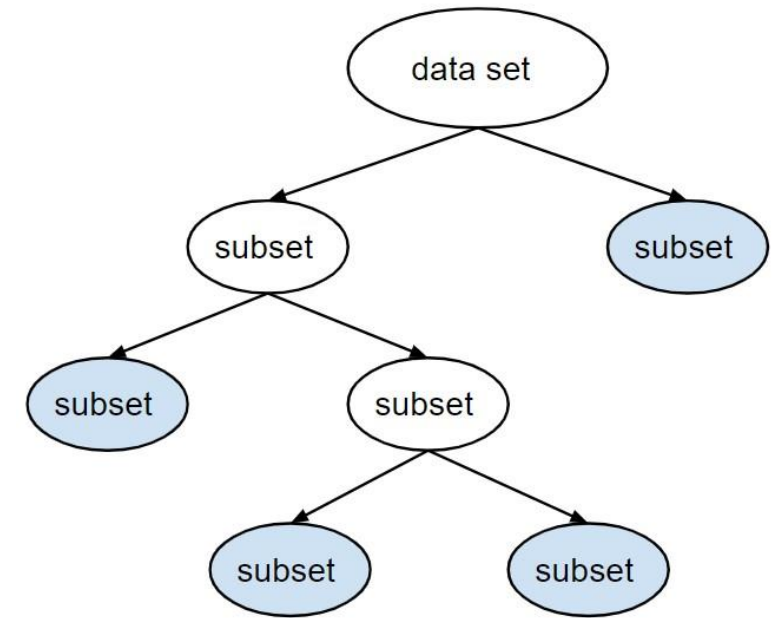
(b)

(i) Profile selection

To extract profiles from datasets, we compare a baseline “simple” (a) and a “hierarchical” (b) clustering algorithms with a more advanced “**semi-hierarchical**” clustering algorithm (c).

Clustering stops when

- **all** features are either *relevant* or *irrelevant* features,
- or when the cluster size is smaller than a fixed amount of elements n_{stop} .



(c)

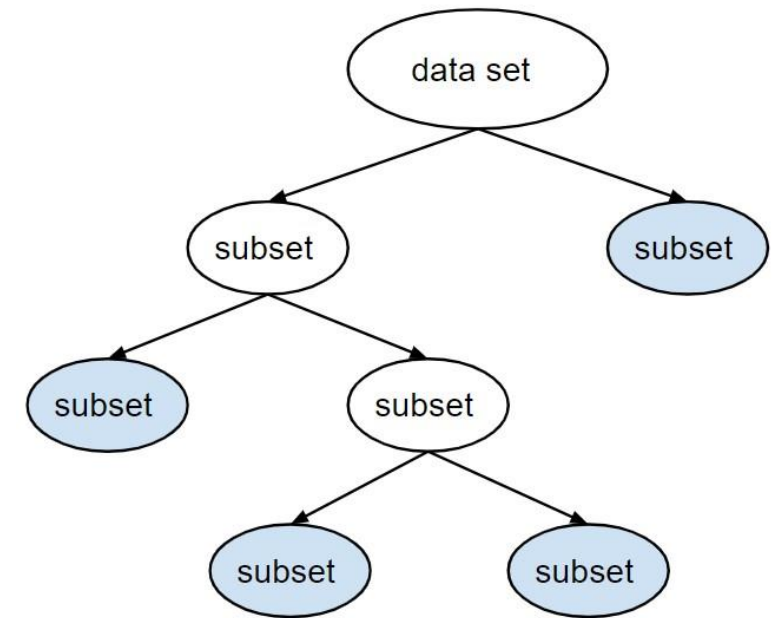
(i) Profile selection

To extract profiles from datasets, we compare a baseline “simple” (a) and a “hierarchical” (b) clustering algorithms with a more advanced “**semi-hierarchical**” clustering algorithm (c).

Clustering stops when

- **all** features are either *relevant* or *irrelevant* features,
- or when the cluster size is smaller than a fixed amount of elements n_{stop} .

Relevance of features is here determined by “**variability**”:



(c)

(i) Profile selection

To extract profiles from datasets, we compare a baseline “simple” (a) and a “hierarchical” (b) clustering algorithms with a more advanced “**semi-hierarchical**” clustering algorithm (c).

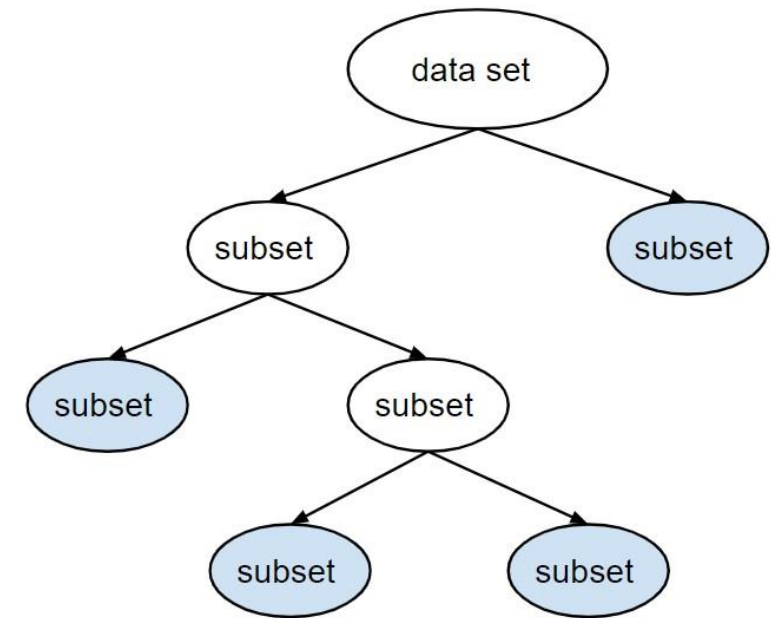
Clustering stops when

- **all** features are either *relevant* or *irrelevant* features,
- or when the cluster size is smaller than a fixed amount of elements n_{stop} .

Relevance of features is here determined by “**variability**”:

a numeric feature is

- *characteristic* of a profile if it **does not vary**
- *not characteristic* of a profile if it **approximates a uniform distribution**



(c)

(i) Profile selection

To extract profiles from datasets, we compare a baseline “simple” (a) and a “hierarchical” (b) clustering algorithms with a more advanced “**semi-hierarchical**” clustering algorithm (c).

Clustering stops when

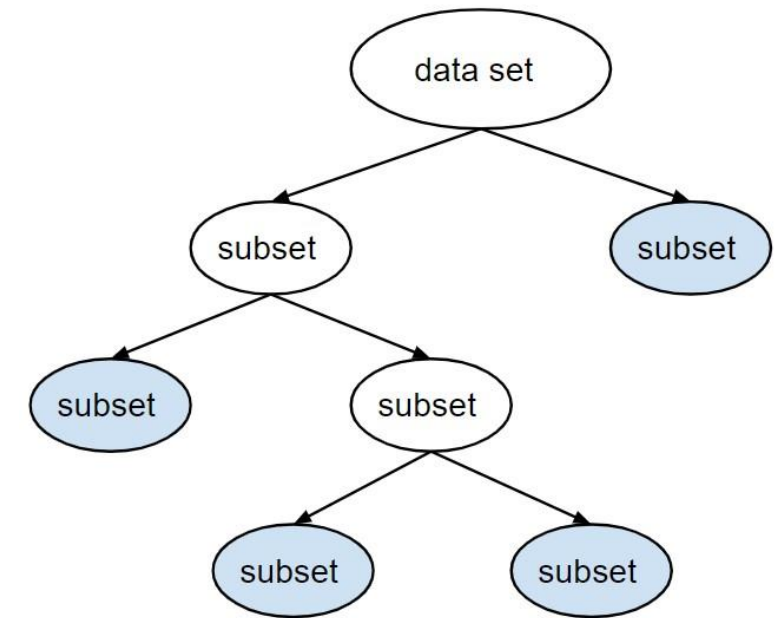
- **all** features are either *relevant* or *irrelevant* features,
- or when the cluster size is smaller than a fixed amount of elements n_{stop} .

Relevance of features is here determined by “**variability**”:

a numeric feature is

- *characteristic* of a profile if it **does not vary**
- *not characteristic* of a profile if it **approximates a uniform distribution**

→ we consider a normalized measure based on standard deviation



(i) Profile selection

To extract profiles from datasets, we compare a baseline “simple” (a) and a “hierarchical” (b) clustering algorithms with a more advanced “**semi-hierarchical**” clustering algorithm (c).

Clustering stops when

- **all** features are either *relevant* or *irrelevant* features,
- or when the cluster size is smaller than a fixed amount of elements n_{stop} .

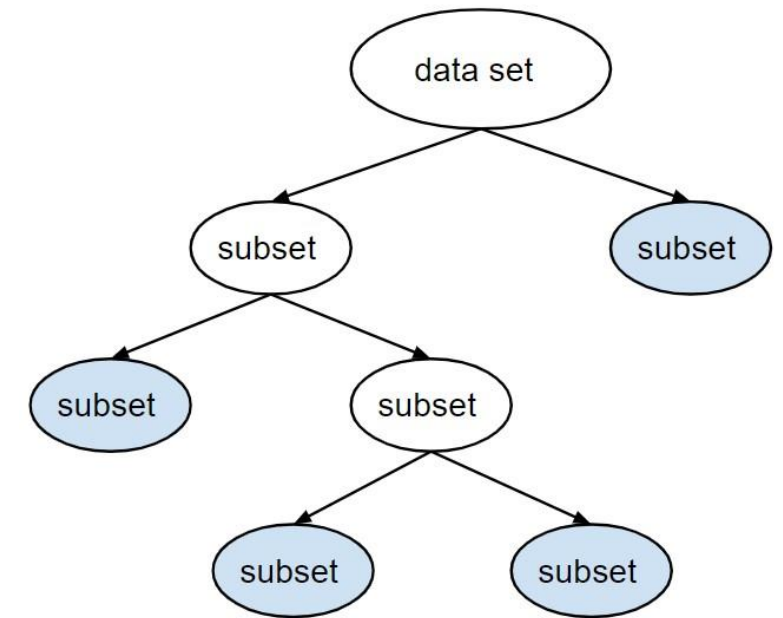
Relevance of features is here determined by “**variability**”:

a numeric feature is

- *characteristic* of a profile if it **does not vary**
- *not characteristic* of a profile if it **approximates a uniform distribution**

→ we consider a normalized measure based on standard deviation

for categorical features we consider a normalized measure based on the frequency of appearance of values



(c)

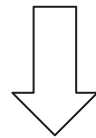
(i) Profile selection

To extract profiles from datasets, we compare a baseline “simple” (a) and a “hierarchical” (b) clustering algorithms with a more advanced “**semi-hierarchical**” clustering algorithm (c).

Clustering stops when

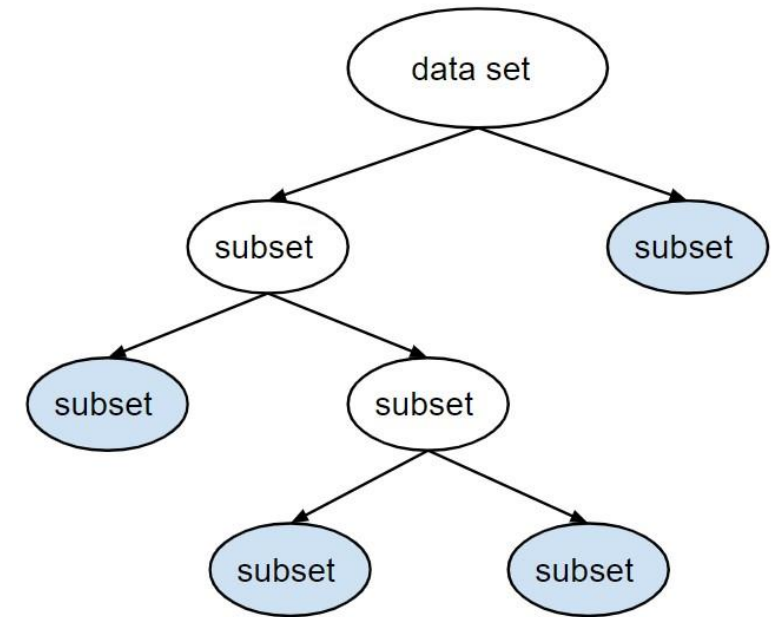
- **all** features are either *relevant* or *irrelevant* features,
- or when the cluster size is smaller than a fixed amount of elements n_{stop} .

Relevance of features is here determined by “**variability**”:



$$\begin{aligned} \text{variability} &> c_u \\ \text{variability} &< c_l \end{aligned}$$

feature is not descriptive for profile
feature is descriptive for profile



(c)



Example – German Credit Dataset

We use the **German Credit Dataset** as an example to showcase our method:

- real world credit risk dataset with risk-label ‘Good’/’Bad’
- 1000 individuals
- 8 attributes (both numerical and categorical)
 - age, sex, saving accounts, duration and purpose of the credit request, etc.
- used in several studies on algorithmic fairness



(ii) Evaluation of clustering method

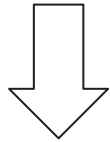
We use two metrics for the evaluation of the clustering methods:

1. Stability
2. Inter-clustering similarity

(ii) Evaluation of clustering method

We use two metrics for the evaluation of the clustering methods:

1. **Stability**
2. Inter-clustering similarity



we investigate **sensitivity towards the initialization of medoids**, running each algorithm 100 times with different (random selected) initial medoids.

algorithm	l	k	# clusters	# cluster comb.	freq. most occurring comb.	running time
simple	1	20	20	15	43	3.956 ± 0.376 s
simple	1	30	30	50	9	8.170 ± 0.777 s
simple	1	40	40	17	26	12.068 ± 0.875 s
hierarchical	2	4	16	31	24	1.119 ± 0.078 s
hierarchical	2	5	25	21	28	1.152 ± 0.077 s
hierarchical	2	6	36	18	32	1.298 ± 0.113 s
hierarchical	3	3	27	54	12	1.144 ± 0.047 s
hierarchical	4	2	16	18	29	1.450 ± 0.083 s
hierarchical	5	2	32	39	20	1.614 ± 0.071 s
semi-hierarchical	-	-	34*	47	23	2.264 ± 0.112 s

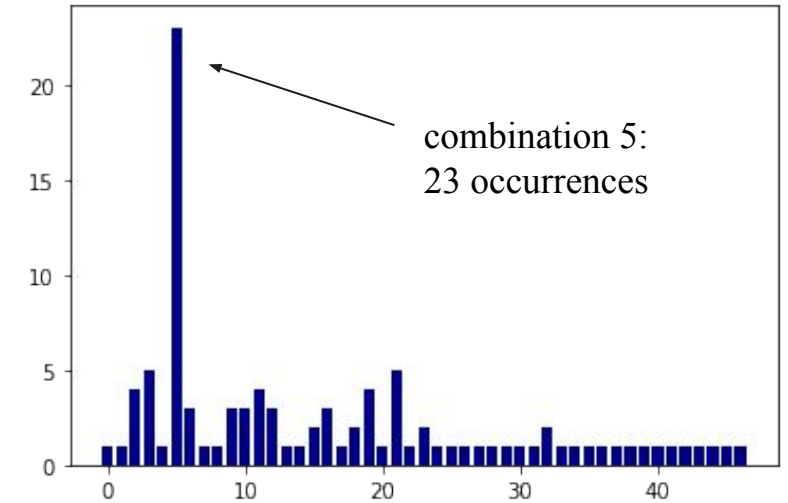
(ii) Evaluation of clustering method

We use two metrics for the evaluation of the clustering methods:

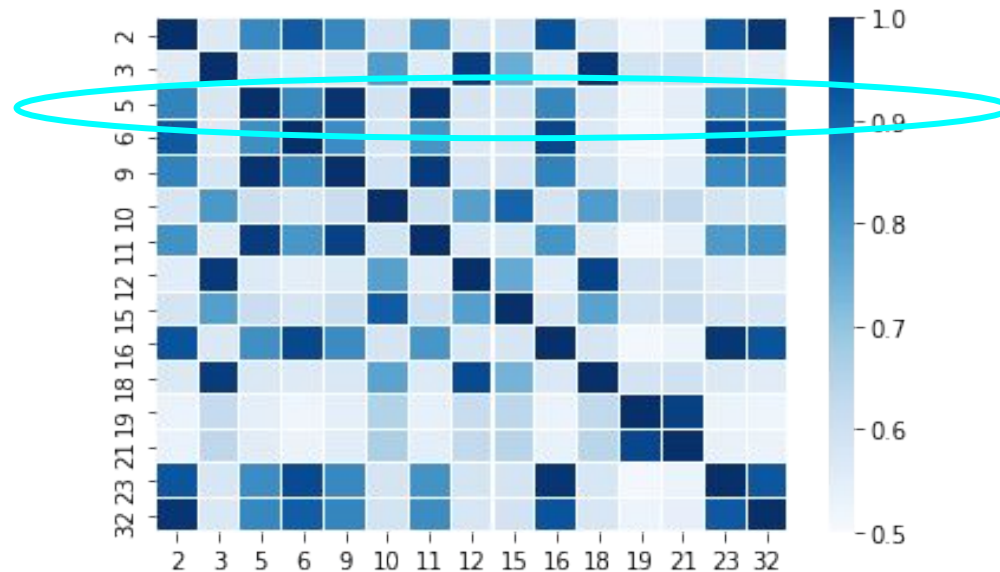
1. Stability
2. **Inter-clustering similarity**

Question: to what extent eg. the 47 different outcomes of running 100 times the **semi-hierarchical clustering** are different?

relative frequency



similarity map



(ii) Evaluation of clustering method

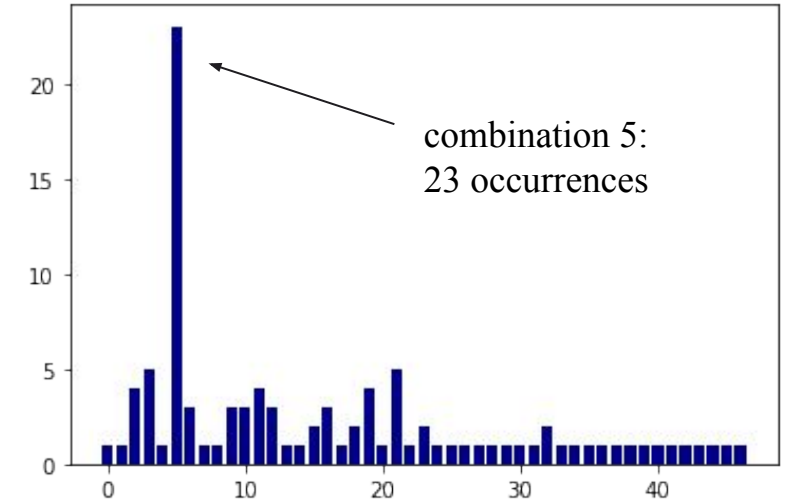
We use two metrics for the evaluation of the clustering methods:

1. Stability
2. **Inter-clustering similarity**

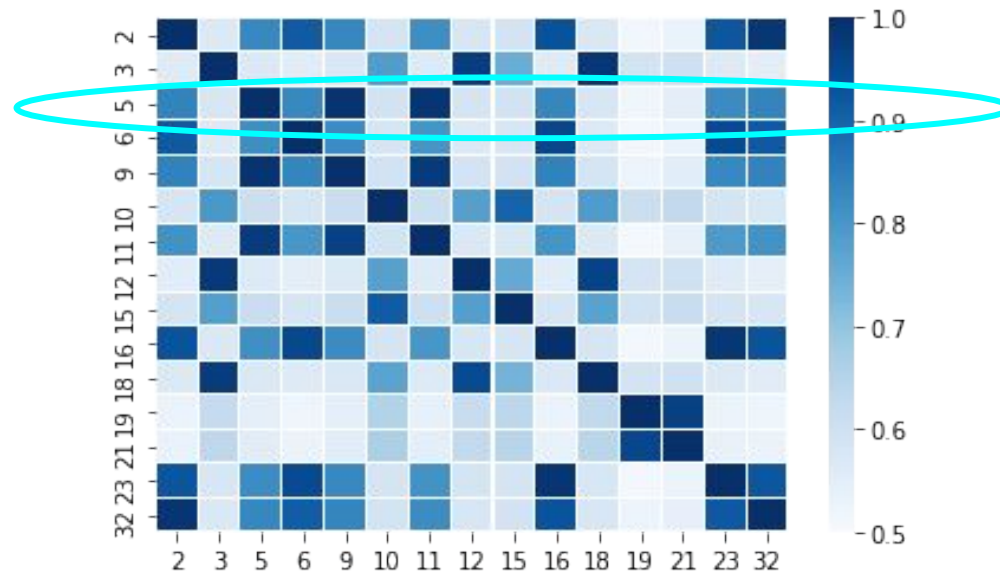
Question: to what extent eg. the 47 different outcomes of running 100 times the **semi-hierarchical clustering** are different?

➡ *rather similar to the most common combination..*

relative frequency



similarity map



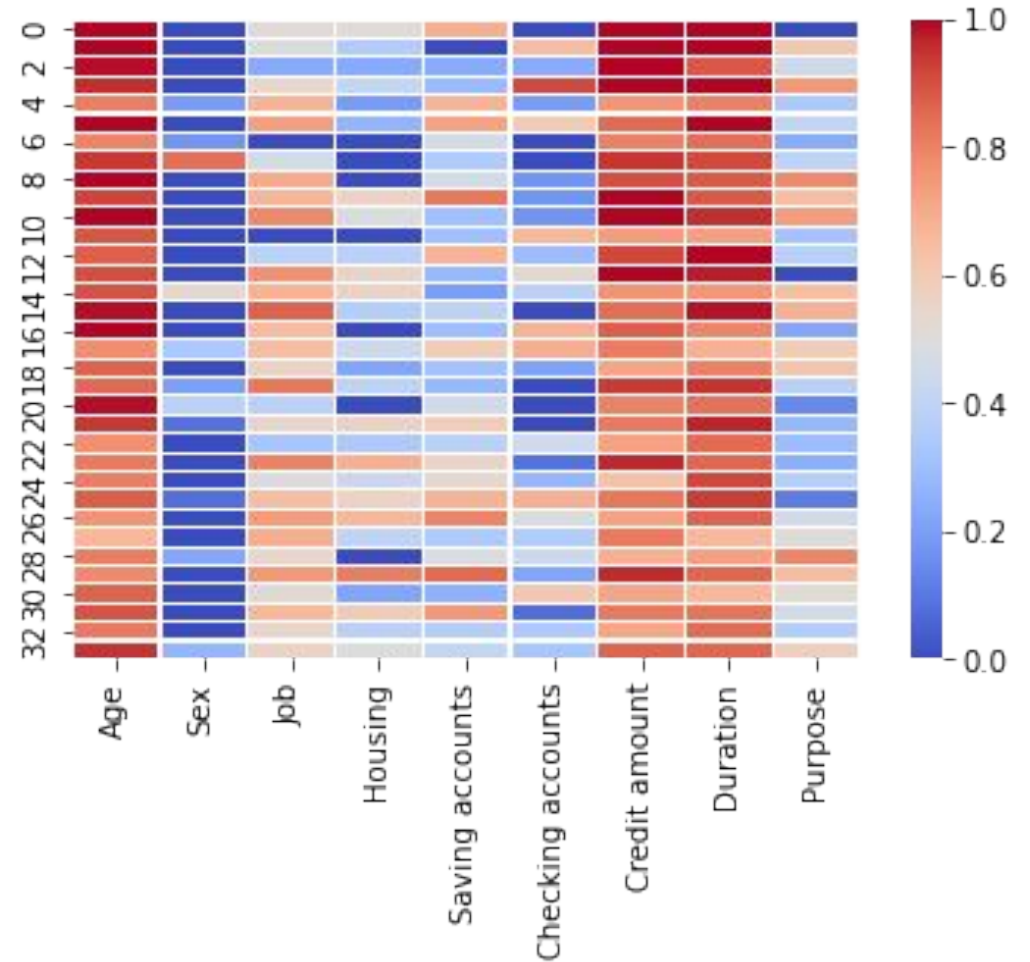
(ii) Evaluation of clustering method

We now take the clustering outcome that occurred most.

The **variability plot** shows
for each profile (vertical axis)
the variability score per feature (horizontal axis).

blue tile → **descriptive** feature.

red tile → **non-descriptive** feature.



(ii) Evaluation of clustering method

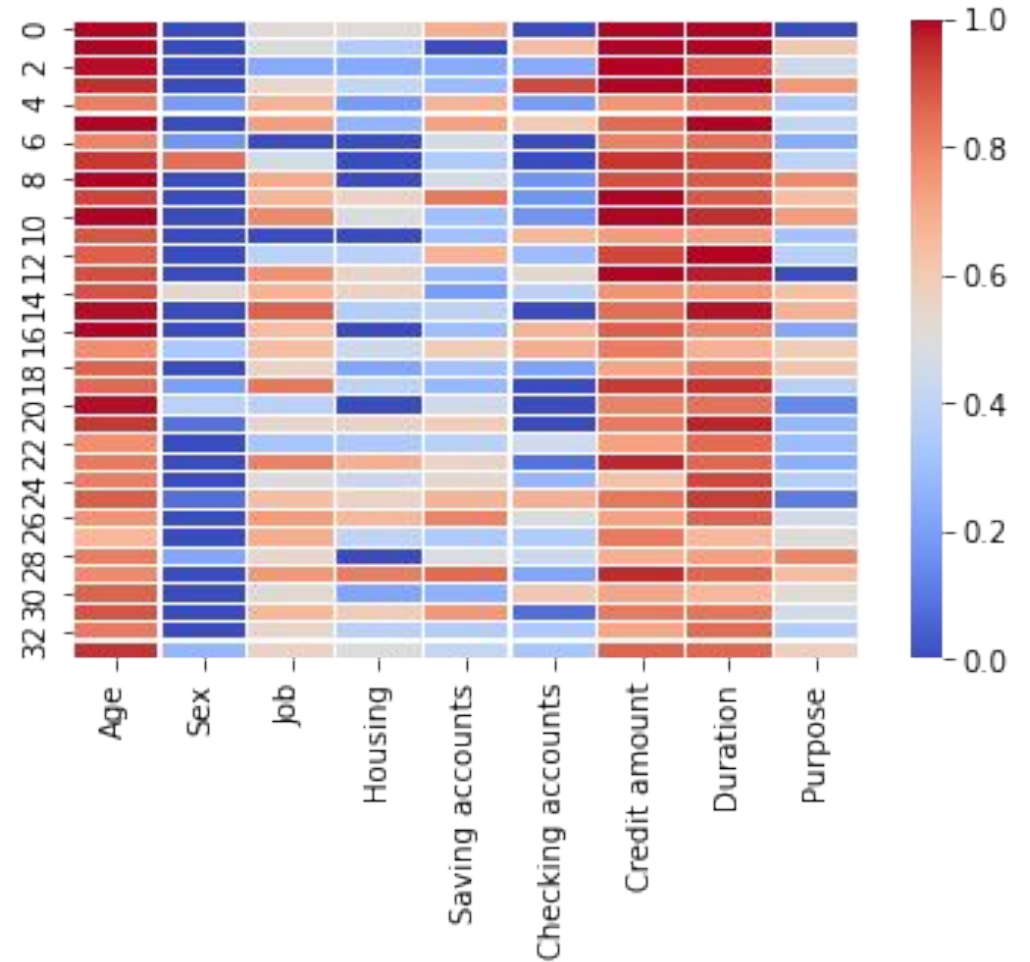
We now take the clustering outcome that occurred most.

The **variability plot** shows
for each profile (vertical axis)
the variability score per feature (horizontal axis).

blue tile → **descriptive** feature.

red tile → **non-descriptive** feature.

→ *most nominal features are descriptives...*





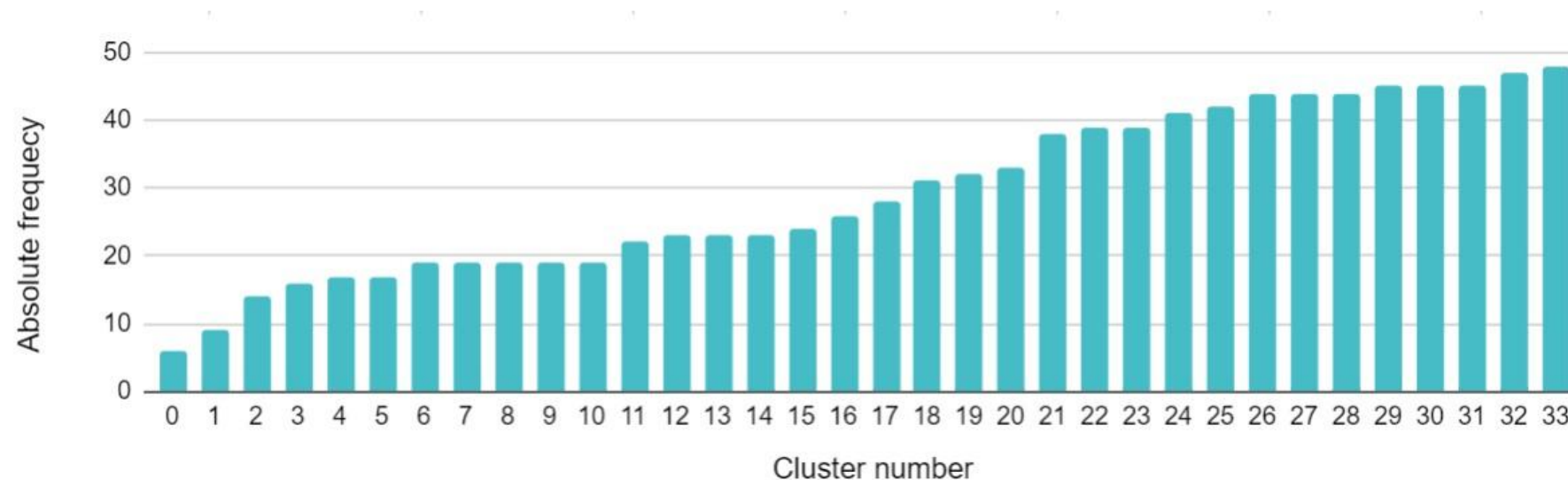
(iii) Profile-based Bias Evaluation

- We take the profiles obtained from the semi-hierarchical clustering
- We train three different classifiers on the German Credit dataset:
 - Logistic regression classifier, XGboost classifier, Support vector machine (SVM) classifier
 - following standard practice we remove protected attributes
- We perform bias assessment against the profiles

(iii) Representational Bias

Representational bias: What observed may be different of what is actual.

We look at the relative proportion of different profiles extracted from the German dataset, **expert knowledge** or **real-time data collection** can be used to test whether there is a bias.





(iii) Historical Bias towards protected class

A_r = proportion of people belonging to a protected class in the advantaged group and with attribute r ,

D_r = proportion of people of protected class in the disadvantaged group and with attribute r

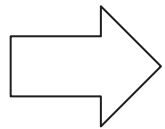
A decision making process exhibits **conditional demographic disparity** with respect to r iff: $D_r > A_r$

(iii) Historical Bias towards protected class

A_r = proportion of people belonging to a protected class in the advantaged group and with attribute r ,

D_r = proportion of people of protected class in the disadvantaged group and with attribute r

A decision making process exhibits **conditional demographic disparity** with respect to r iff: $D_r > A_r$



At **aggregate level**, the German Credit dataset exhibits demographic disparity (protected class = “female”)

$$\mathbf{D} = \#DG_f / \#DG_{f+m} = \mathbf{0.36} > \mathbf{0.29} = \#AG_f / \#AG_{f+m} = \mathbf{A}$$

*What if we do the same analysis at **profile level**?*

(iii) Historical Bias towards protected class

We test whether there is **demographic disparity for profiles**:

Let A_c denote the proportion of people belonging to the protected class (female) in the advantaged group of cluster c over all people in the advantaged group of cluster c and let D_c be vice versa the fraction of disadvantaged.

The table below shows A_c and D_c for each **profile** c .

c	D_c	A_c	c	D_c	A_c	c	D_c	A_c	c	D_c	A_c	c	D_c	A_c
0	0.00	0.00	7	0.36	0.25	~~14~~	~~0.00~~	~~0.19~~	21	0.00	0.03	28	0.96	0.90
1	0.00	0.00	8	0.00	0.00	15	1.00	1.00	22	0.00	0.00	29	0.00	0.00
2	0.00	0.00	9	0.00	0.00	16	1.00	1.00	23	0.00	0.00	30	1.00	1.00
3	0.00	0.00	10	1.00	1.00	17	1.00	0.86	24	0.00	0.00	31	1.00	1.00
4	0.00	0.07	11	0.00	0.00	18	0.00	0.00	25	0.00	0.03	32	0.00	0.00
5	1.00	1.00	12	0.00	0.00	19	0.06	0.06	26	1.00	1.00	33	0.00	0.09
6	0.00	0.08	13	0.00	0.00	20	0.24	0.00	27	0.00	0.00			

for some profiles *there is disparity*:

- it might be possible that non-protected attributes correlate with the protected attribute.
- it could also be reconstructed during the training process!



demographic disparity
against women
with low resources
when asking credit
for cars



Profile	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Sample
0	26	male	2	rent	moderate	unknown	3577	9	car	859
1	37	male	2	own	unknown	unknown	7409	36	business	868
2	39	male	3	own	little	unknown	6458	18	car	106
3	26	male	2	own	little	little	4370	42	radio/TV	639
4	31	male	2	own	quite rich	unknown	3430	24	radio/TV	19
5	38	female	2	own	unknown	unknown	1240	12	radio/TV	135
6	43	male	1	own	little	little	1344	12	car	929
7	36	male	2	rent	little	little	2799	9	car	586
8	39	male	2	own	little	little	2522	30	radio/TV	239
9	31	male	2	own	little	moderate	1935	24	business	169
10	33	female	2	own	little	little	1131	18	furniture/equipment	166
11	26	male	1	own	little	moderate	625	12	radio/TV	220
12	23	male	2	own	unknown	moderate	1444	15	radio/TV	632
13	42	male	2	own	little	little	4153	18	furniture/equipment	899
14	29	male	2	own	unknown	unknown	3556	15	car	962
15	37	female	2	own	little	moderate	3612	18	furniture/equipment	537
16	27	female	2	own	little	little	2389	18	radio/TV	866
17	26	female	2	rent	little	unknown	1388	9	furniture/equipment	582
18	29	male	2	own	little	unknown	2743	28	radio/TV	426
19	53	male	2	free	little	little	4870	24	car	4
20	36	male	2	own	little	little	1721	15	car	461
21	38	male	2	own	little	unknown	804	12	radio/TV	997
22	29	male	2	own	little	moderate	1103	12	radio/TV	696
23	43	male	2	own	unknown	unknown	2197	24	car	406
24	27	male	2	own	little	little	3552	24	furniture/equipment	558
25	30	male	2	own	little	moderate	1056	18	car	580
26	24	female	2	own	little	moderate	2150	30	car	252
27	34	male	2	own	little	unknown	2759	12	furniture/equipment	452
28	24	female	2	rent	little	little	2124	18	furniture/equipment	761
29	34	male	2	own	little	moderate	5800	30	car	893
30	34	female	2	own	little	unknown	1493	12	radio/TV	638
31	30	female	2	own	little	unknown	1055	18	car	161
32	35	male	2	own	little	unknown	2346	24	car	654
33	35	male	2	own	unknown	unknown	1979	15	radio/TV	625

(slight)
demographic disparity
against women
with low resources
when asking credit
for furniture/equipment



(iii) Historical Bias

We now look at how much informative a **profile** w.r.t. a decision by taking out the protected attribute:

Let A_c^* (D_c^*) be the fraction of (dis)advantaged individuals in profile c over the total individuals within that profile:

c	D^*	A^*	c	D^*	A^*	c	D^*	A^*	c	D^*	A^*	c	D^*	A^*
0	0.00	1.00	7	0.58	0.42	14	0.09	0.91	21	0.05	0.95	28	0.52	0.48
1	0.11	0.89	8	0.58	0.42	15	0.62	0.38	22	0.36	0.64	29	0.40	0.60
2	0.14	0.86	9	0.47	0.53	16	0.38	0.62	23	0.15	0.85	30	0.09	0.91
3	0.62	0.38	10	0.53	0.47	17	0.25	0.75	24	0.37	0.63	31	0.22	0.78
4	0.12	0.88	11	0.18	0.82	18	0.19	0.81	25	0.20	0.71	32	0.13	0.87
5	0.12	0.88	12	0.35	0.65	19	0.50	0.50	26	0.45	0.55	33	0.06	0.94
6	0.32	0.68	13	0.52	0.48	20	0.52	0.48	27	0.14	0.86			

Around half of the profiles produce both positive and negative outcomes.

In the other half, there are profiles for which the majority of the samples is clearly advantaged or clearly disadvantaged (hints to a *decision-making rationale?*)

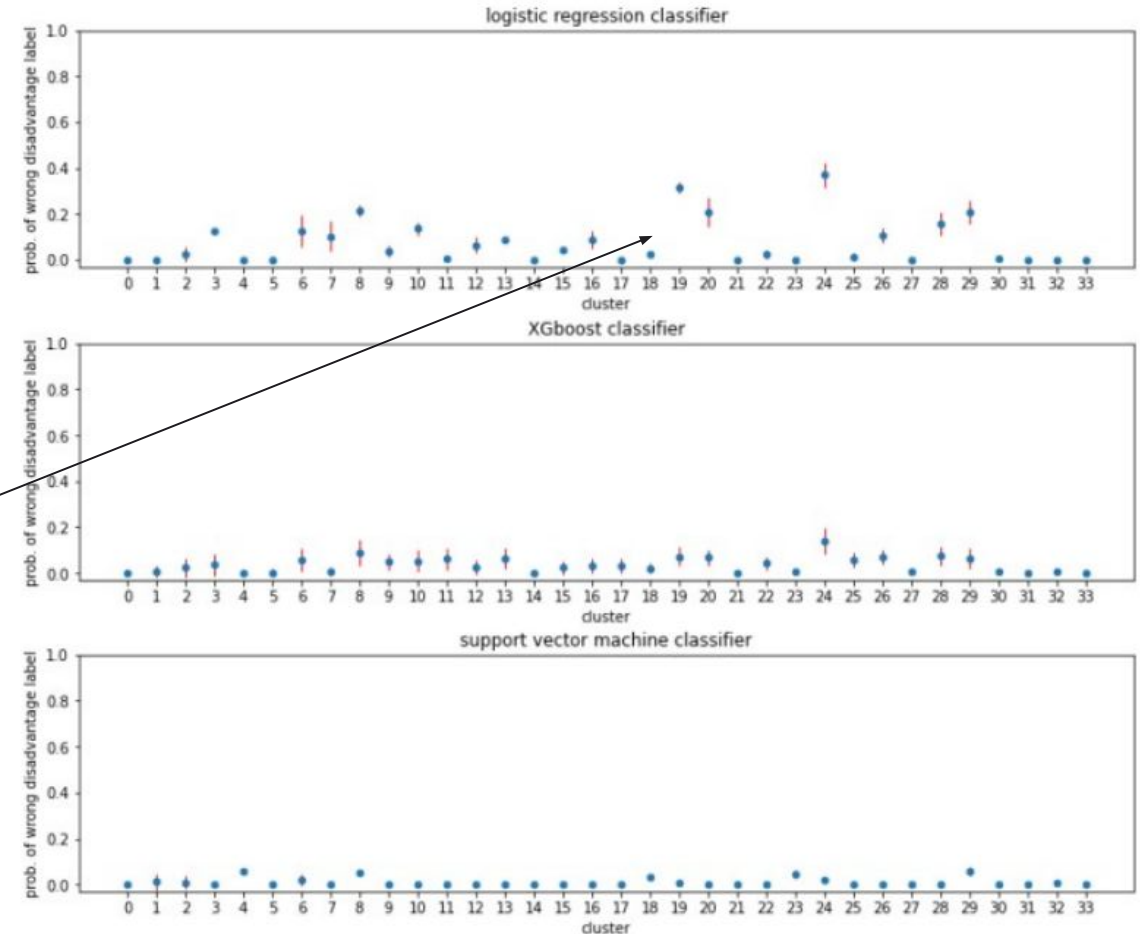
(iii) Aggregation Bias

We study for each profile the **probability of getting a disadvantage label** ('bad') **when the true label is the advantage** ('good') **label**.

We run all three ADM algorithms 100 times with different train-test split.

For **aggregation bias** we look at the mean performance.

For the logistic regression classifier there are profiles for which the **false negative** probability is way higher than for others.



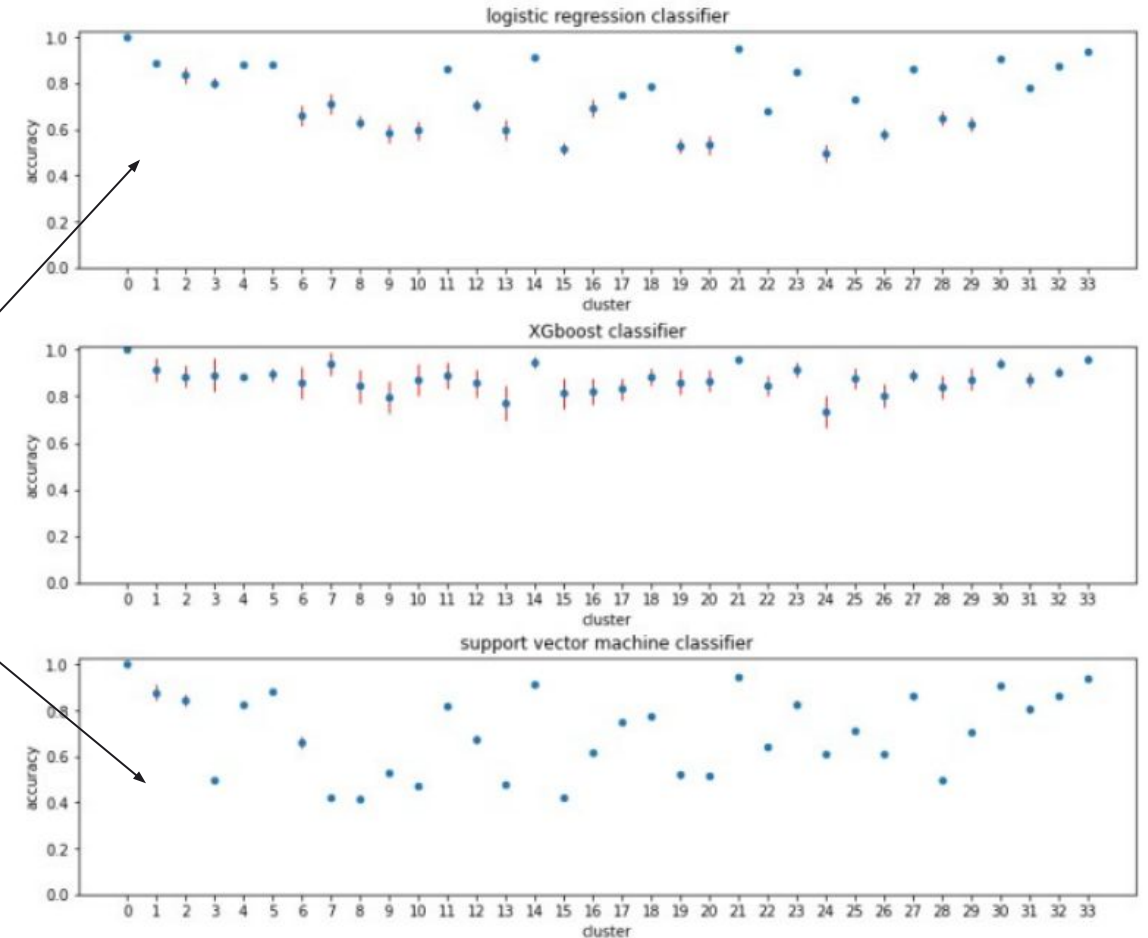
(iii) Aggregation Bias

We study for each profile the **probability of getting a correct prediction label**.

We run all three ADM algorithms 100 times with different train-test split.

For **aggregation bias** we look at the mean performance.

For the logistic regression classifier and the SVM classifier there are profiles that score significantly lower than others.



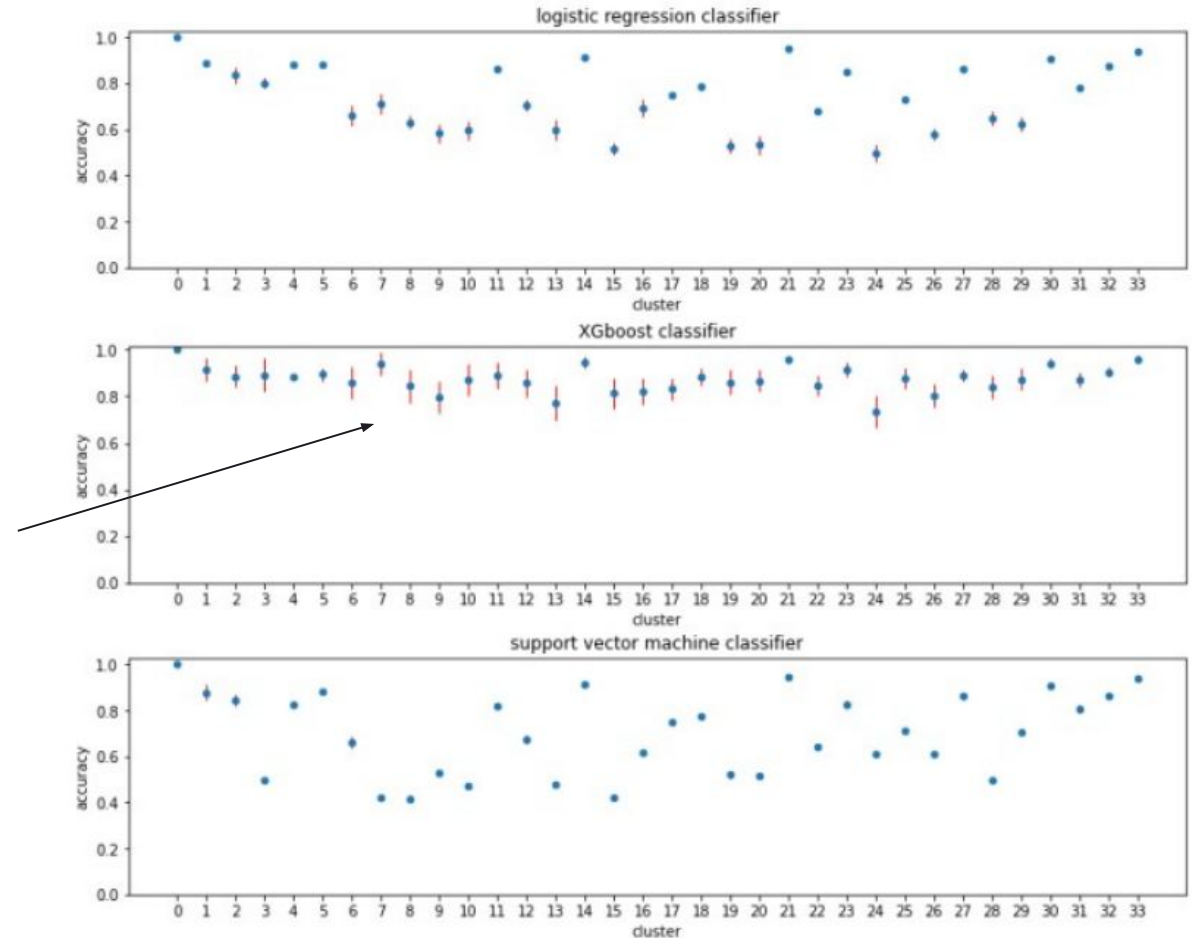
(iii) Evaluation Bias

We study for each profile the **probability of getting a correct prediction label**.

We run all three ADM algorithms 100 times with different train-test split.

For **evaluation bias** we look at the variance.

We see the XGboost classifier has the highest standard deviation for a lot of profiles. This means that this algorithm is the **most sensitive towards the train-test split** that has been used.





Future research...

- *testing the method on more datasets,*
- *integrating/comparing*
 - *other clustering methods (eg. entropy-based),*
 - *other machine-learning methods*
- *looking at additional types of biases*

code available at: **`https://github.com/mcwilms/PEBAM`**



PEBAM

a Profile-based Evaluation method for
Bias Assessment on Mixed datasets

KI 2022: 45th German Conference on Artificial Intelligence, 22nd September

Mieke Wilms, Giovanni Sileno, Hinda Haned