BEWARE2 @ AIxIA, Rome, 6th Nov. 2023

# BRIO x Alkemy

## A bias detecting tool

G. Coraglia[1], F.A. D'Asaro[2], F.A. Genco[1], D. Giannuzzi[3], D. Posillipo[3], G. Primiero[1], C. Quaggio[3]

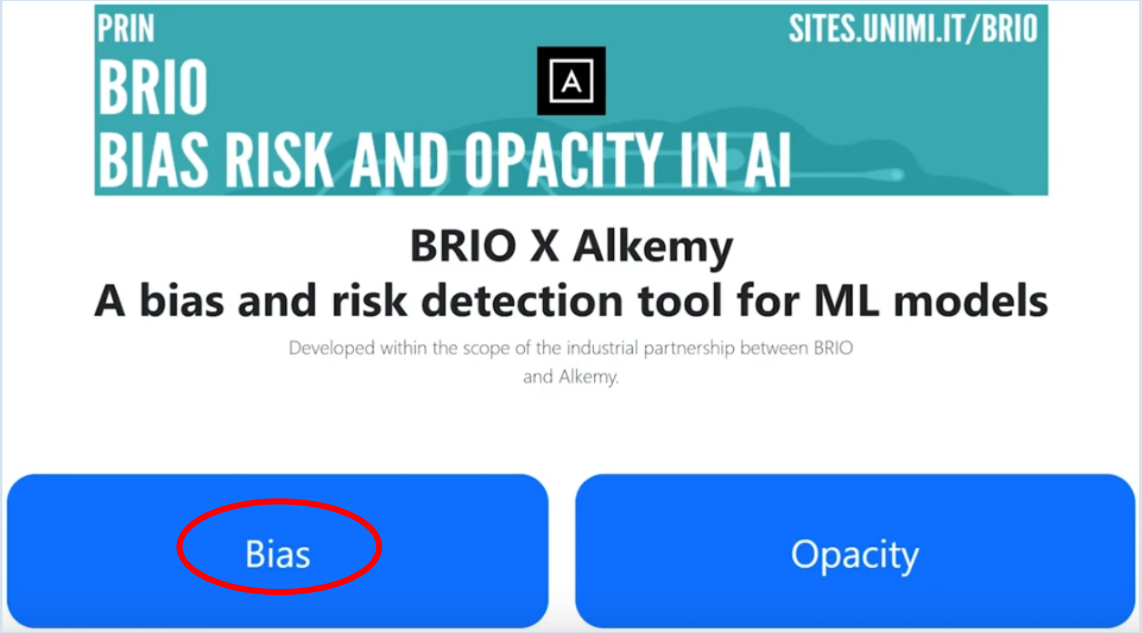
1 LUCI Lab, unimi
2 Ethos group, univr
3 Deeplearning & Big Data, Alkemy

Work funded by the PRIN project n.2020SSKZ7R BRIO (Bias, Risk and Opacity in AI).

BRIO

bias risk opacity (in AI)

collaboration w/ Alkemy to produce open source software




in here

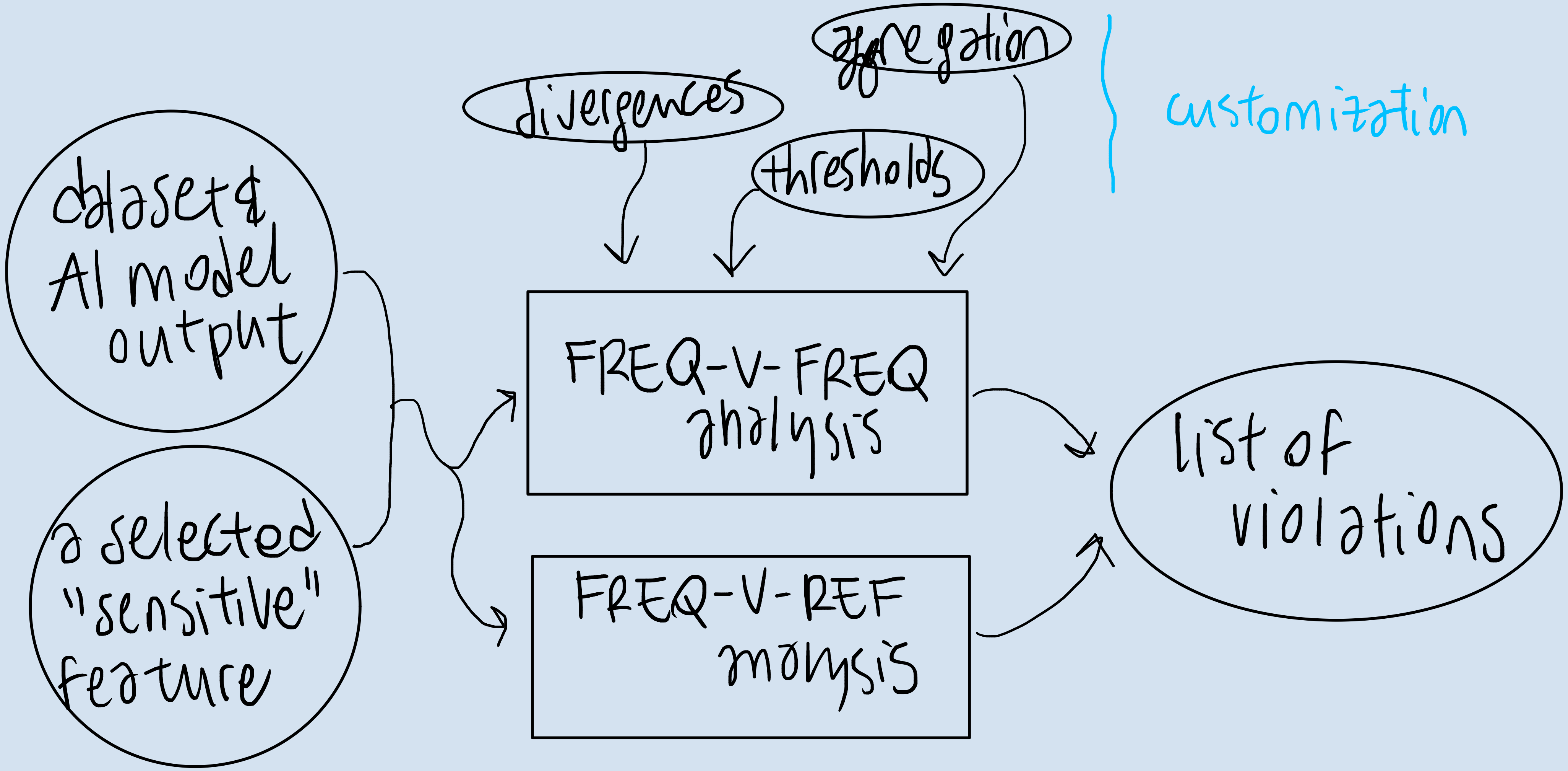
https://github.com/DLBD-Department/BRIO_x_Alkemy

# for this first iteration, we focus on bias detection

UNDERLYING THEORY

F. D'Asaro, G. Primiero, Probabilistic typed natural deduction for trustworthy computations
F. D'Asaro, F. Genco, G. Primiero, Checking trustworthiness of probabilistic computations in a typed natural deduction system
F. Genco, G. Primiero, A typed lambda-calculus for establishing trust in probabilistic programs

HIGH-LEVEL DESCRIPTION OF SOFTWARE



| WHAT THIS SOFTWARE IS | WHAT THIS SOFTWARE IS NOT |
| --- | --- |
| - a detection tool | - a correction tool<br>VIA<br>- optimization of a "loss" function<br>USING THE FACT THAT<br>- assumes that a "correct" label is known a priori |
| - "post-processing" | |
| - focuses on frequencies | |
| - blind to the model | - feature weighting    next module on opacity |


B. d'Alessandro, C. O'Neil, T. LaGatta, Conscientious classification: A data scientist's guide to discrimination-aware classification
R. Fu, Y. Huang, P. V. Singh, Artificial intelligence and algorithmic bias: Source, detection, mitigation, and implications

M. Hardt, E. Price, E. Price, N. Srebro, Equality of opportunity in supervised learning
G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, K. Q. Weinberger, On fairness and calibration
F. Kamiran, A. Karim, X. Zhang, Decision theory for discrimination-aware classification
I. Niño-Adan, D. Manjarres, I. Landa-Torres, E. Portillo, Feature weighting methods: A review

# WHAT THIS SOFTWARE IS

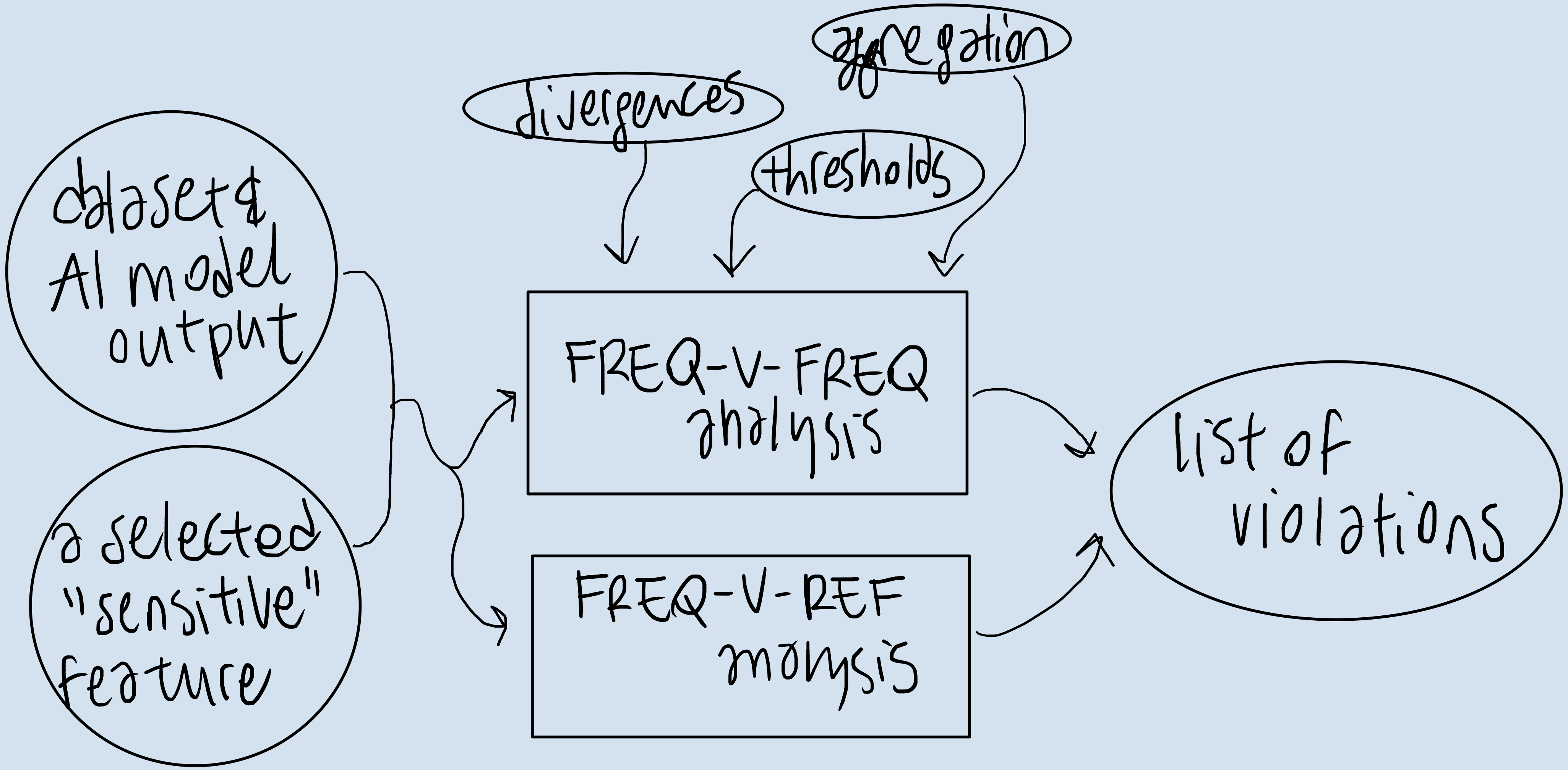
- a detection tool
- "post-processing"
- focuses on frequencies
- blind to the model

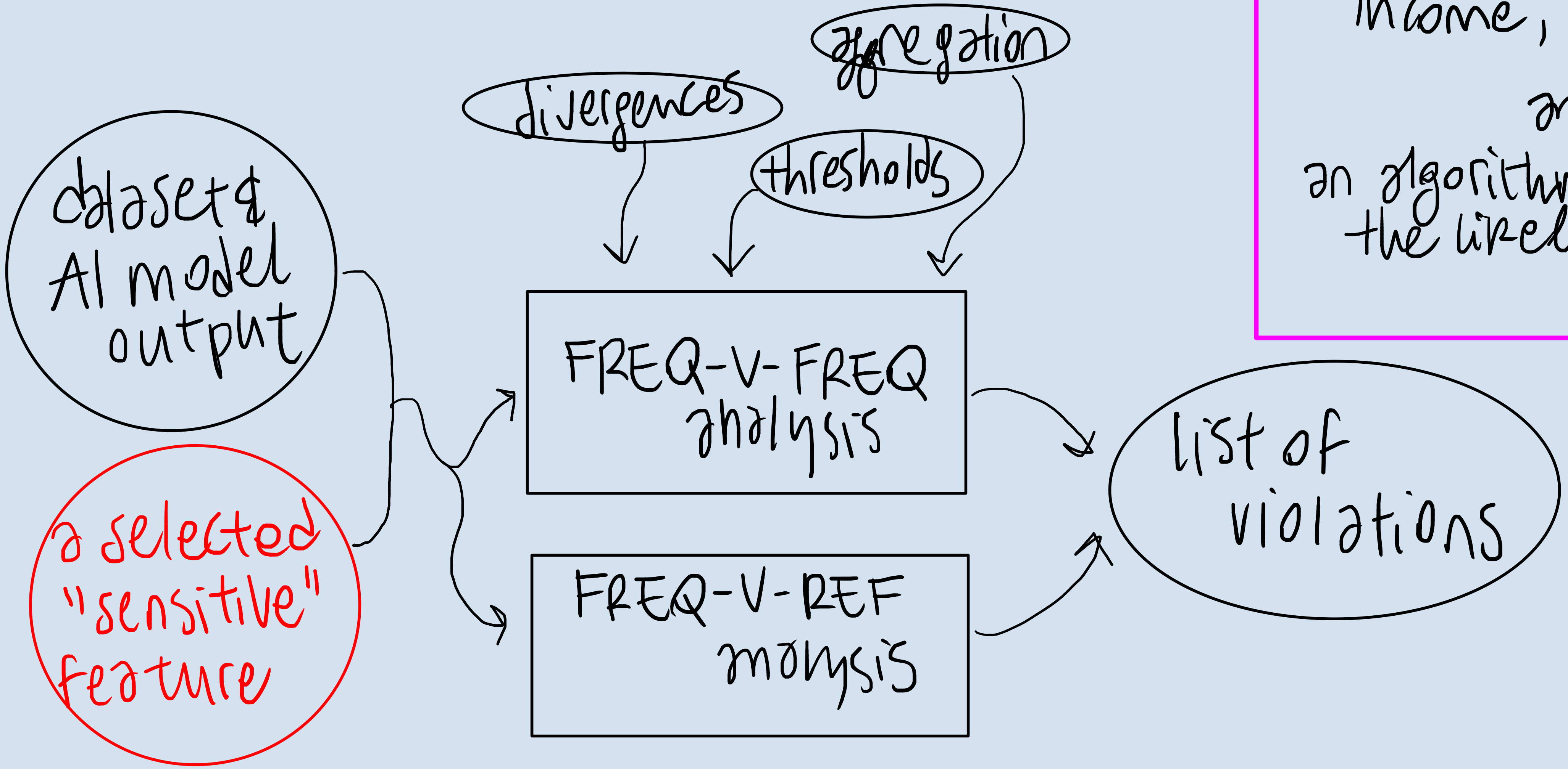
can be run locally

experts are encouraged to use it freely

B. d'Alessandro, C. O'Neil, T. LaGatta, Conscientious classification: A data scientist's guide to discrimination-aware classification
R. Fu, Y. Huang, P. V. Singh, Artificial intelligence and algorithmic bias: Source, detection, mitigation, and implications

M. Hardt, E. Price, E. Price, N. Srebro, Equality of opportunity in supervised learning
G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, K. Q. Weinberger, On fairness and calibration
F. Kamiran, A. Karim, X. Zhang, Decision theory for discrimination-aware classification
I. Niño-Adan, D. Manjarres, I. Landa-Torres, E. Portillo, Feature weighting methods: A review


dataset &
AI model
output
a selected
"sensitive"
feature
divergences
aggregation
thresholds
FREQ-V-FREQ
analysis
FREQ-V-REF
analysis
list of
violations




also "protected attribute"

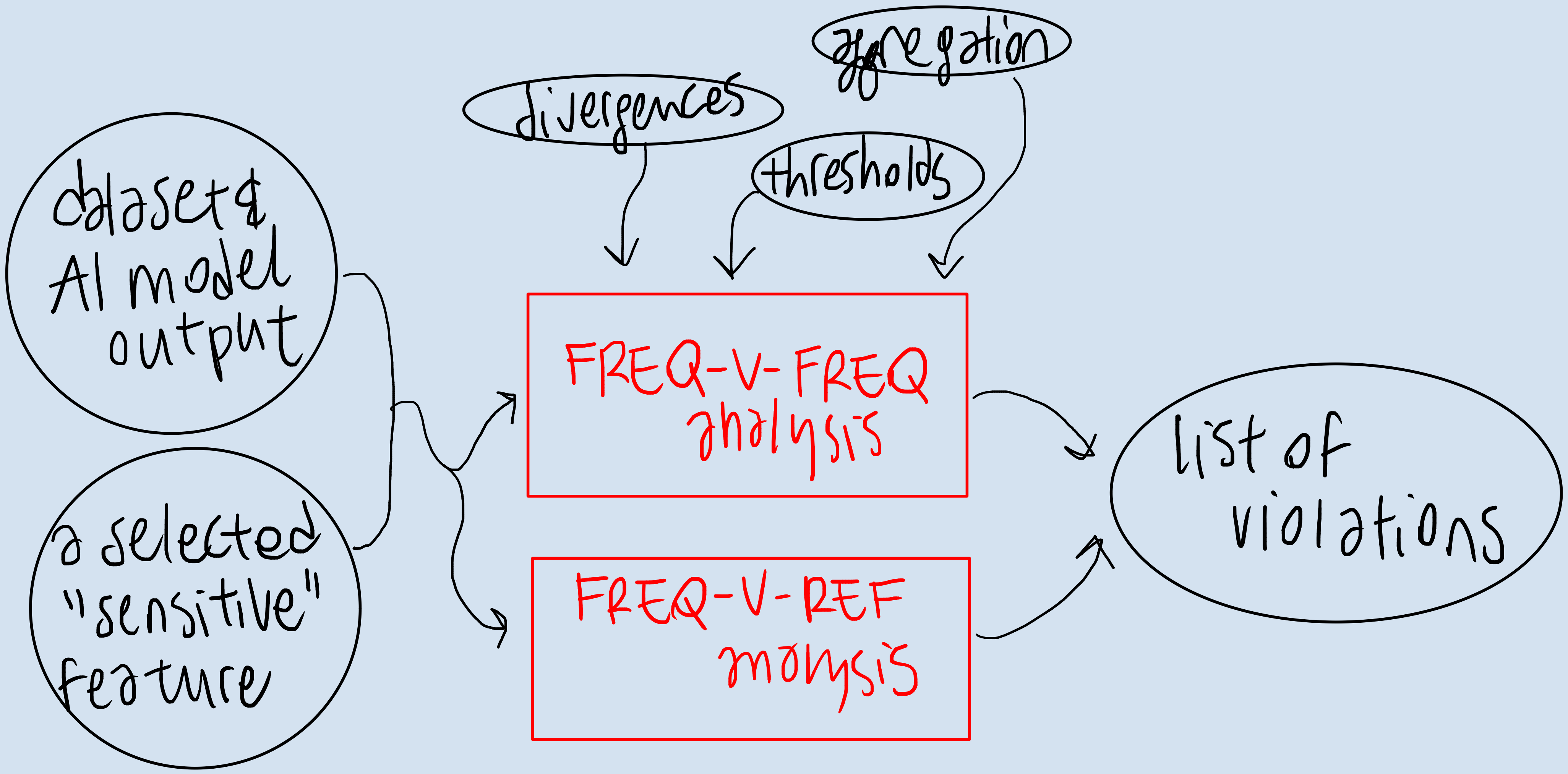
e.g. education

**RUNNING EXAMPLE**

a DB of individuals with their age, sex, income, education lvl

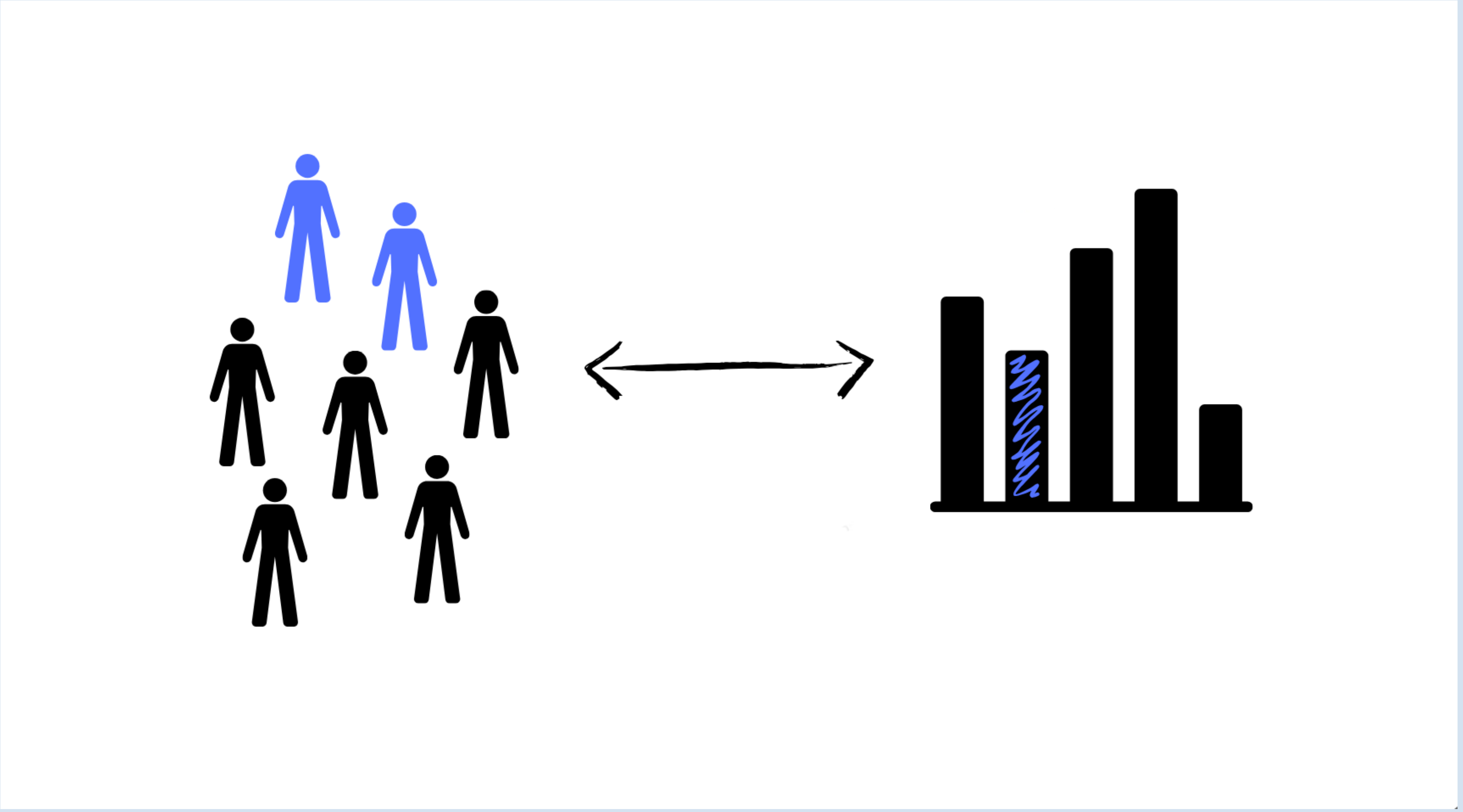
and

an algorithm predicting the likelyhood that they default


dataset & AI model output
a selected "sensitive" feature
divergences
aggregation
thresholds
FREQ-V-FREQ analysis
FREQ-V-REF analysis
list of violations

# FREQ-V-REF OPTION

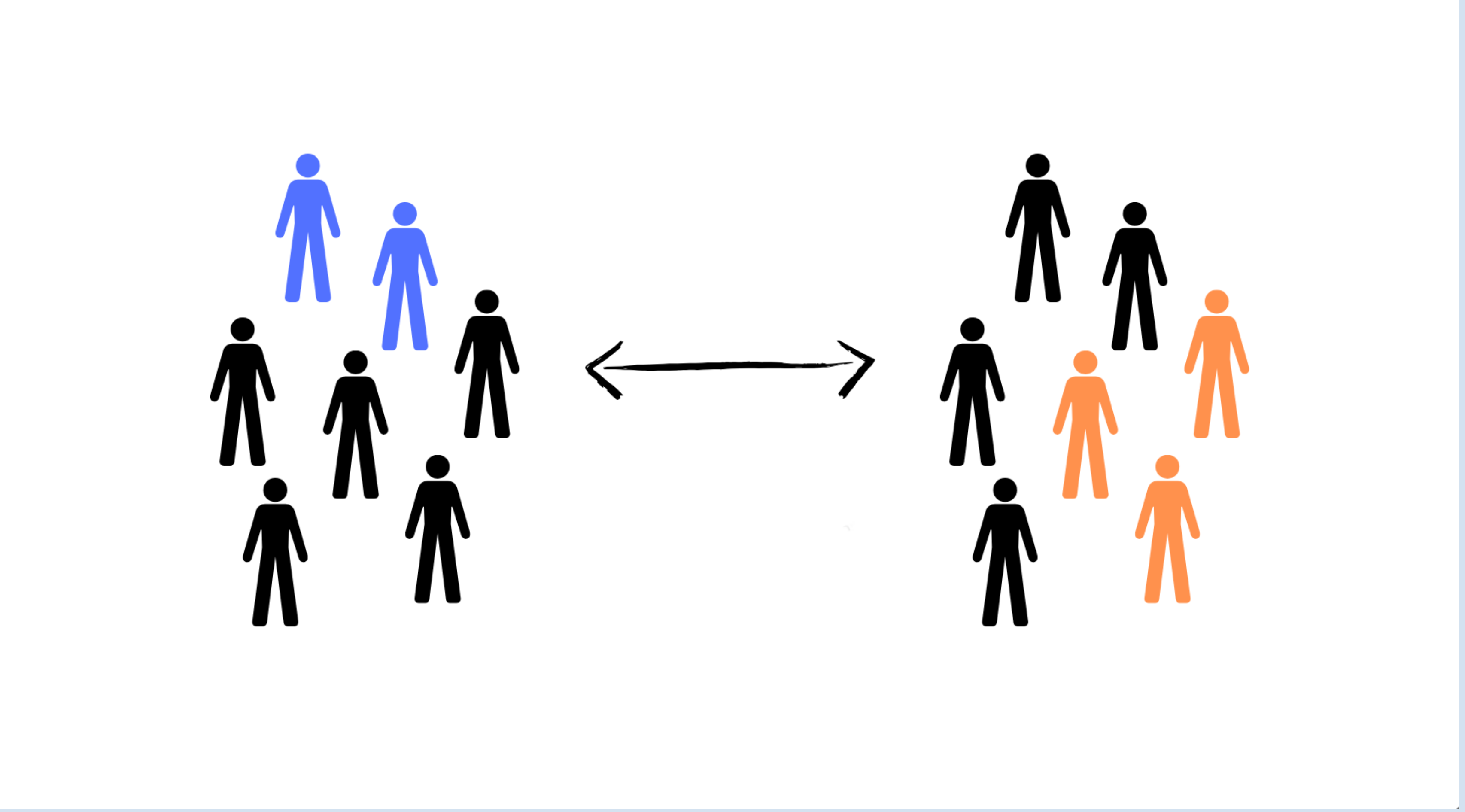
compares the frequency for a group with a known "optimal" behaviour



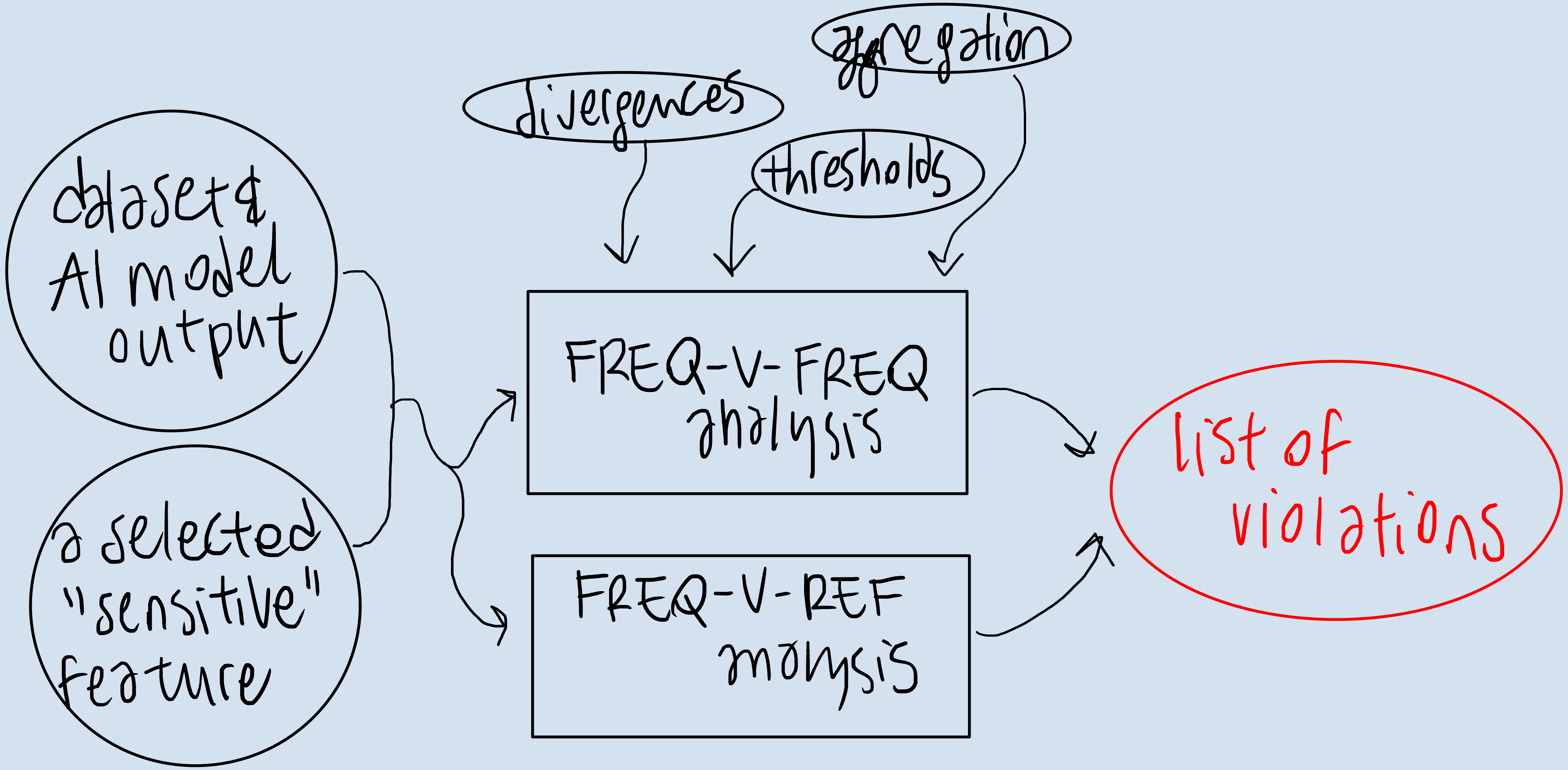
Ex we deem "sensitive" the variable recording the sex of a person, and see how much the algorithm's behaviour differs on an equal distribution of approval for men and women (under the hypothesis that they are in equal number in the dataset)

# FREQ-V-FREQ OPTION

compares frequencies for different groups



Ex we deem "sensitive" the variable recording the sex of a person, and see how much the algorithm's behaviour differs on fixed subsets of the database, for example its partition for education or marital status


dataset & AI model output
a selected "sensitive" feature
divergences
aggregation
thresholds
FREQ-V-FREQ analysis
FREQ-V-REF analysis
list of violations

screenshot of a run on our example database

- sensitive feature: sex
- option: FREQ-V-FREQ with conditioning on education and marital status

# Overall Result

(Distance, distance<=threshold, threshold, standard deviation)

(0.025269625352224545, **False** , 0.016368585412256314, None)

## Violations

Condition: (num observations, distance, distance<=threshold, threshold, standard deviation)

x3_education==5 : (75, 0.06772575250836121, False, 0.017549038105676658, None)
x4_marriage==3 : (95, 0.05425219941348974, False, 0.017404225646095797, None)
x3_education==2 & x4_marriage==3 : (56, 0.04678362573099415, False, 0.017753344485757386, None)
x3_education==3 & x4_marriage==2 : (616, 0.04200812107788854, False, 0.016703275417264275, None)
x3_education==2 & x4_marriage==2 : (2145, 0.03637291669292275, False, 0.016492906353361734, None)
x3_education==3 : (1499, 0.03290896164530149, False,

## Conditioned Results

Export CSV

| Condition applied | Result |
| --- | --- |
| x3_education==1 | (3119, 0.0183300648997351, False, 0.016451439592896744) |
| x3_education==3 | (1499, 0.03290896164530149, False, 0.016540570623988414) |
| x3_education==2 | (4250, 0.030006620324395897, False, 0.016422567122490656) |
| x3_education==4 | (40, 0.0, True, 0.018028781183844471) |
| x3_education==5 | (75, 0.06772575250836121, False, 0.017549038105676658) |
| x3_education==6 | (14, None, Not enough observations, ) |
| x3_education==0 | (3, None, Not enough |

screenshot of a run on our example database

- sensitive feature : sex
- option: FREQ-V-FREQ with conditioning on education and marital status

OPTIONS, WILL SEE LATER

## Overall Result

(Distance, distance<=threshold, threshold, standard deviation)

(0.025269625352224545, **False** , 0.016368585412256314, None)

75 people are in the edu==5 subset (no degree)

here

there seem to be a great(er than the threshold) distance between men & women

## Violations

Condition: (num observations, distance, distance<=threshold, threshold, standard deviation)

$\boxed{\text{x3\_education==5 : (75, 0.06772575250836121, False, 0.017549038105676658, None)}}$

x4_marriage==3 : (95, 0.05425219941348974, False, 0.017404225646095797, None)

x3_education==2 & x4_marriage==3 : (56, 0.04678362573099415, False, 0.017753344485757386, None)

x3_education==3 & x4_marriage==2 : (616, 0.04200812107788854, False, 0.016703275417264275, None)

x3_education==2 & x4_marriage==2 : (2145, 0.03637291669292275, False, 0.016492906353361734, None)

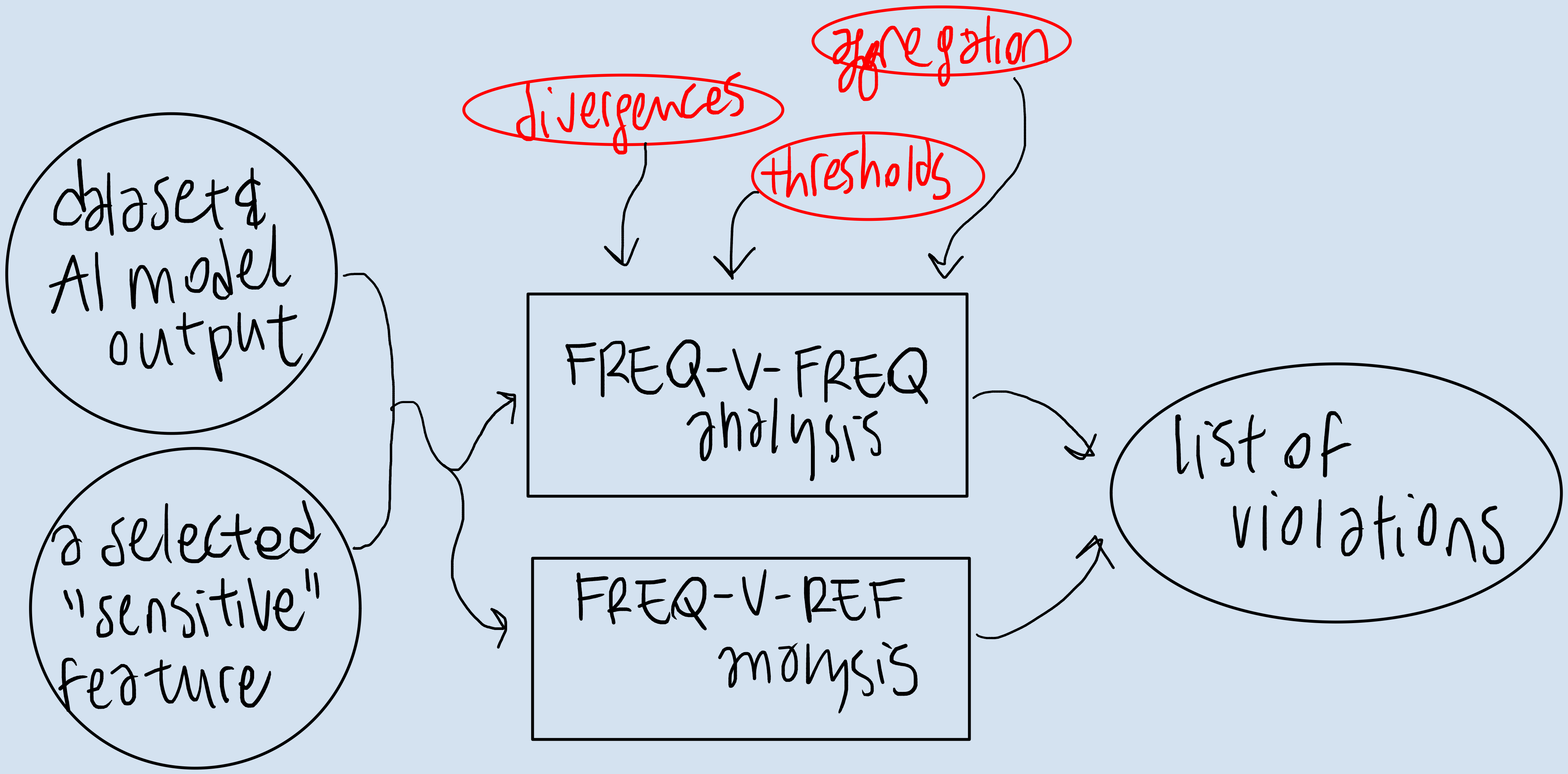
x3_education==3 : (1499, 0.03290896164530149, False,

## Conditioned Results

Export CSV

| Condition applied | Result |
|---|---|
| x3_education==1 | (3119, 0.0183300648997351, False, 0.016451439592896744) |
| x3_education==3 | (1499, 0.03290896164530149, False, 0.016540570623988414) |
| x3_education==2 | (4250, 0.030006620324395897, False, 0.016422567122490656) |
| x3_education==4 | (40, 0.0, True, 0.01802878118384471) |
| x3_education==5 | (75, 0.06772575250836121, False, 0.017549038105676658) |
| x3_education==6 | (14, None, Not enough observations, ) |
| x3_education==0 | (3, None, Not enough |

where do we get these numbers from?



need a notion of distance to compare behaviours

need to find a reasonable enough threshold

if the sensitive ft has more than 2 classes need an aggregating function

(NOT THE CASE HERE...)

# Overall Result

(Distance, distance<=threshold, threshold, standard deviation)

(0.025269625352224545, **False** , 0.016368585412256314, None)

## Violations

Condition: (num observations, distance, distance<=threshold, threshold, standard deviation)

x3_education==5 : (75, 0.06772575250836121, False, 0.017549038105676658, None)
x4_marriage==3 : (95, 0.05425219941348974, False, 0.017404225646095797, None)
x3_education==2 & x4_marriage==3 : (56, 0.04678362573099415, False, 0.017753344485757386, None)
x3_education==3 & x4_marriage==2 : (616, 0.04200812107788854, False, 0.016703275417264275, None)
x3_education==2 & x4_marriage==2 : (2145, 0.03637291669292275, False, 0.016492906353361734, None)
x3_education==3 : (1499, 0.03290896164530149, False,

## Conditioned Results

Export CSV

| Condition applied | Result |
|---|---|
| x3_education==1 | (3119, 0.0183300648997351, False, 0.016451439592896744) |
| x3_education==3 | (1499, 0.03290896164530149, False, 0.016540570623988414) |
| x3_education==2 | (4250, 0.030006620324395897, False, 0.016422567122490656) |
| x3_education==4 | (40, 0.0, True, 0.01802878118384471) |
| x3_education==5 | (75, 0.06772575250836121, False, 0.017549038105676658) |
| x3_education==6 | (14, None, Not enough observations, ) |
| x3_education==0 | (3, None, Not enough |

▷ options for DISTANCES (DIVERGENCES): ...

▷ options for AGGREGATING FUNCTIONS: ...

▷ options for the THRESHOLD

$\varepsilon$ selected manually

$$\varepsilon = f(\bar{r}, \underline{n_C}, \underline{n_D})$$

either HIGH: the feature is sensitive, be extra carefull

⇑

prefere false pos.

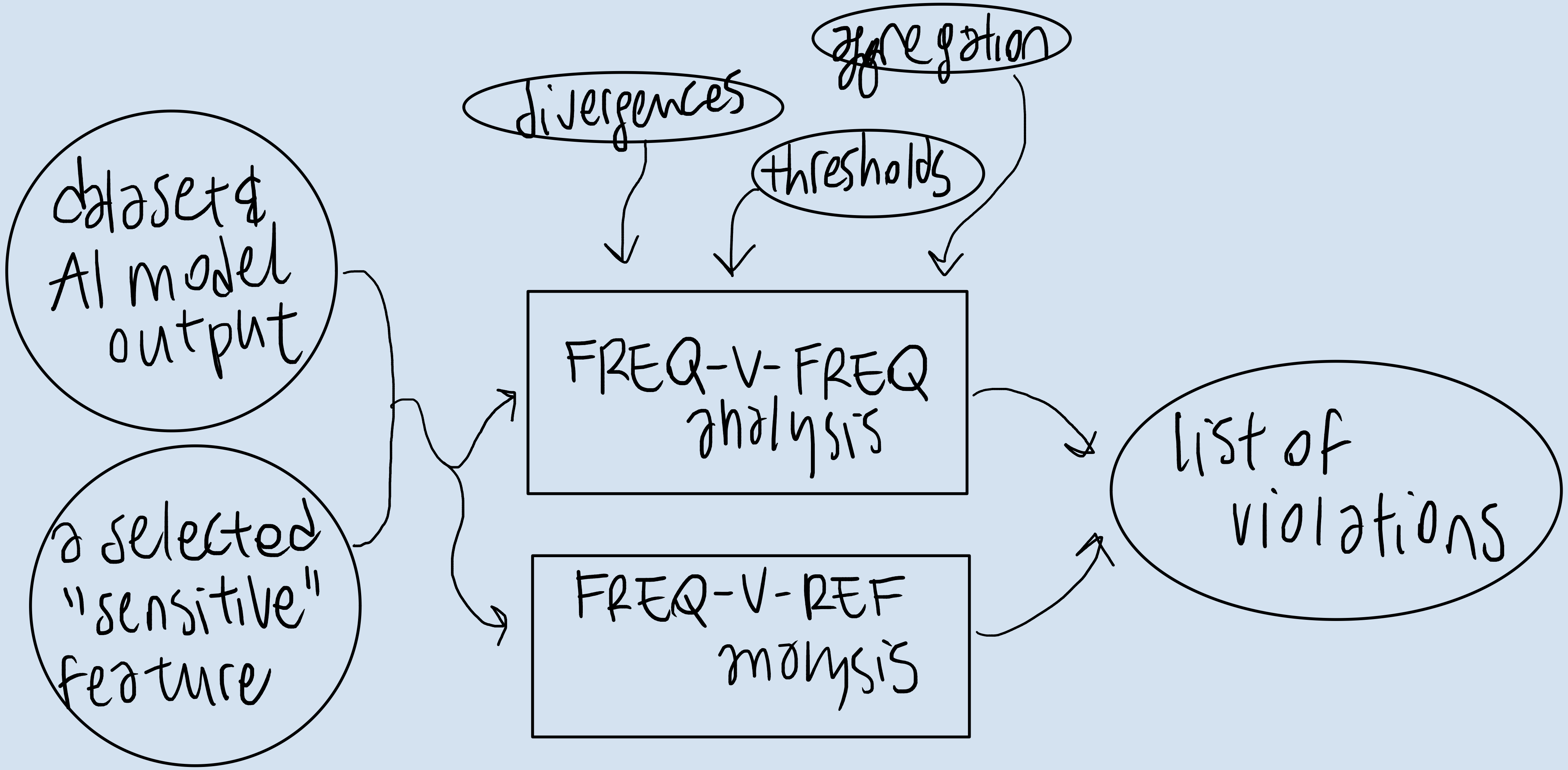
LOW: differences are relevant only if extreme

⇓

prefere false neg.

GRANULARITY
ie #{classes related to the sensitive features}

NUMEROSITY
ie #{individuals in the classes}




DOES THIS ACTUALLY WORK?

# VALIDATION

1) data from a cash and credit card issuer in Taiwan from [YL2009]
2) trained 3 models to predict default probability
3) tried 7 experiments with different options

results are available at [GIT] and they mostly! are in line with expectations

e.g. the more powerful the model, the more biased it is


[YL2009] I.-C. Yeh, C. Lien, The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients

[GIT] https://github.com/DLBD-Department/BRIO_x_Alkemy/tree/main/notebooks

FUTURE WORK WE PLAN TO DO

- implement a module for opacity
- implement a module using bias and opacity to evaluate risk
- do more experiments
- find more data

STILL the tool is already available & open source

So please feel free to

USE IT!

CONTRIBUTE TO IT!

THANK YOU FOR LISTENING ☺