TU Graz

HCI-KDD

Interactive Machine Learning for improving k-anonymity

...

Feiertag Tamara
Waltl Christine
Wolf Julian

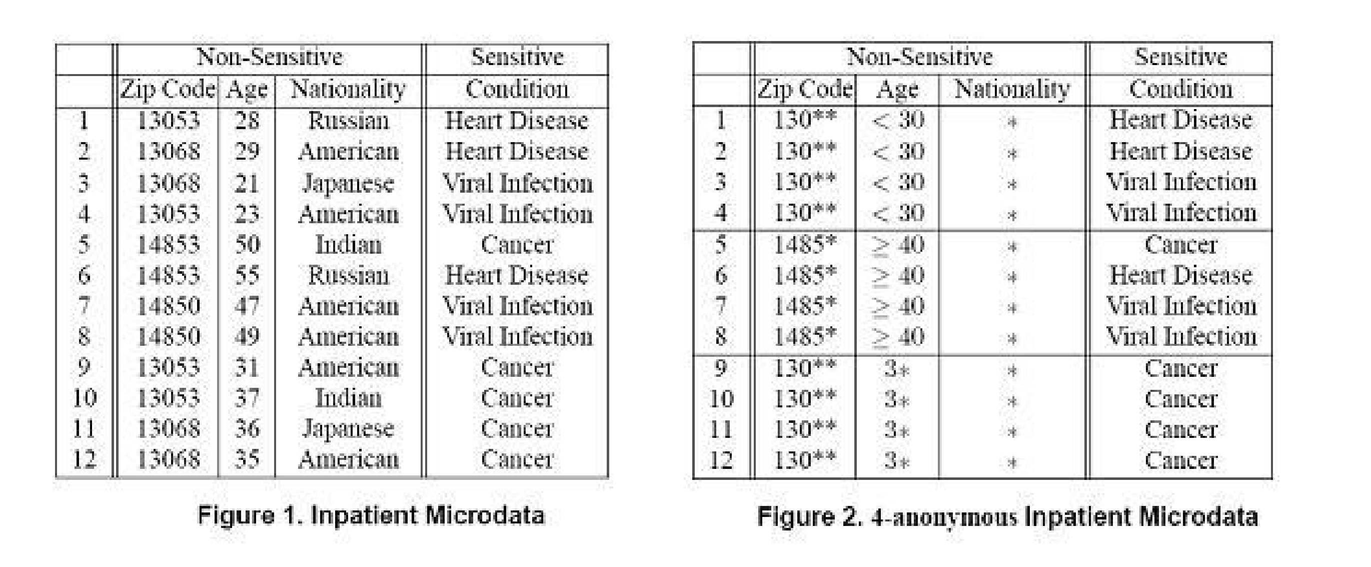
Anonymization

Name	Age	Zip	Gender	Disease
Alexa	26	41070	Female	Allergies
…	…	…	…	…

- increase privacy
- DELETE identifiers (name, email, phone nr., SSN)
- KEEP sensitive data (e.g. medical diagnosis)
- GENERALIZE quasi-identifiers / together retrieve identity (e.g. age, zip, gender,...)

k-anonymization

for every entry in the DS, there must be at least k-1 entries with identical quasi-identifiers



	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

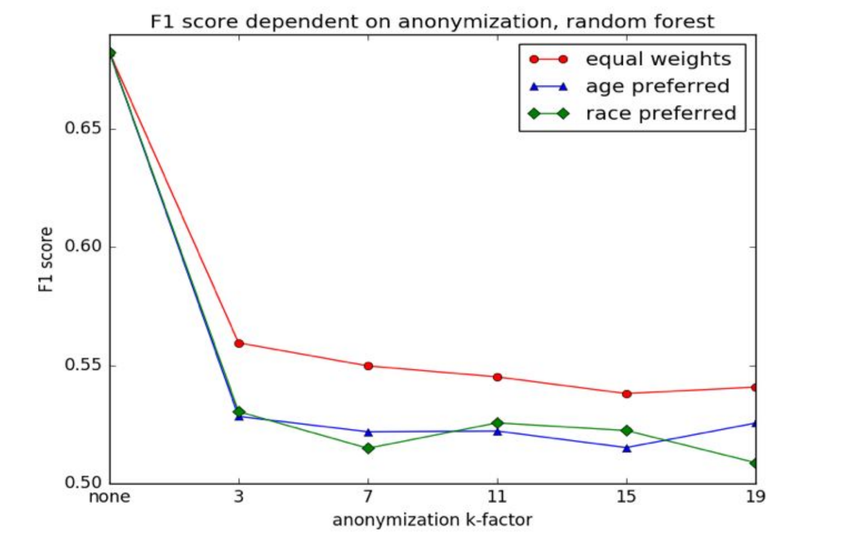
	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

Figure 2. 4-anonymous Inpatient Microdata

Source: http://www.opengardensblog.futuretext.com

Problem

- big data utility => low privacy
- high privacy => small data utility



- possible solution: interactive machine learning
  - user input influences learning algorithm

Our task

[51 - 76]	*	North_America	Male	*	Married-civ-spouse
[51 - 76]	*	North_America	Male	*	Married-civ-spouse
[51 - 76]	*	North_America	Male	*	Married-civ-spouse

52	Private	United-States	Male	White	Married-civ-spouse

[48 - 70]	Private	America	Male	White	*
[48 - 70]	Private	America	Male	White	*
[48 - 70]	Private	America	Male	White	*

Our task

- use user input to adapt weight vectors of our quasi-identifiers
- compare results
- improve our strategy

How we did it …

Technologies

AngularJS

Package Manager:

- NPM, Bower

Packages:

- anonymization.js
    - Input CSV -> Weights -> K-Factor -> anonymized data

Web Development partial automatisation:

- Gulp (automatize web development processes)
- Sass (for generating the css stylesheets)

Some elements from:

- Bootstrap

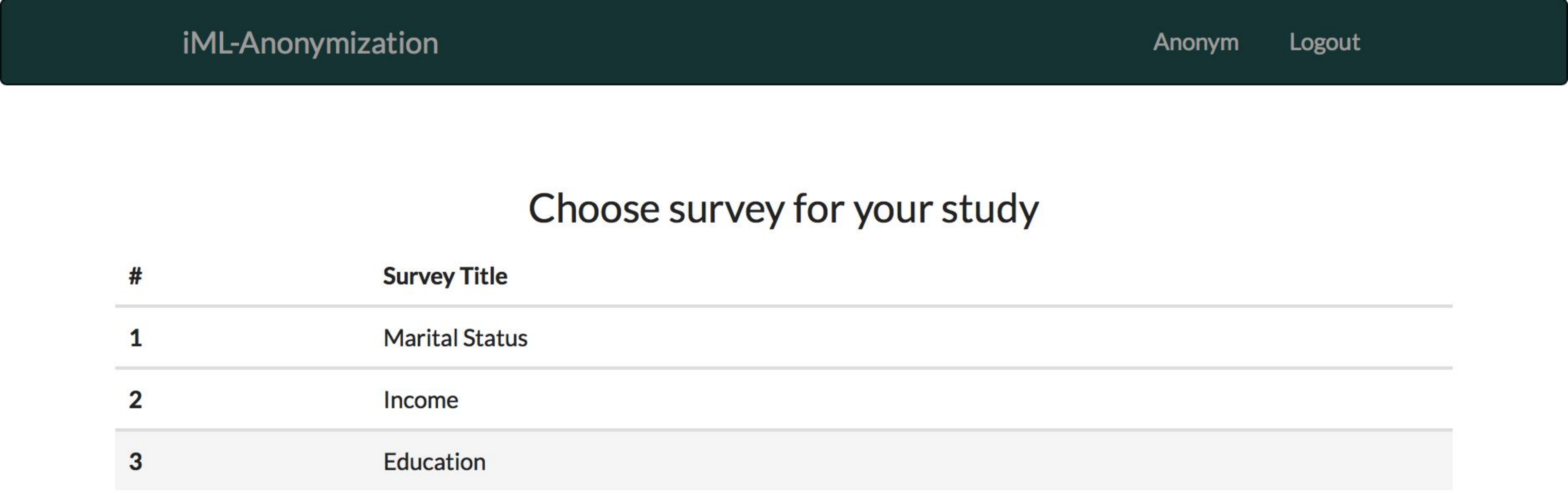
Our algorithm

configurable:

- start k-factor (default 2)
- end k-factor (default 7)
- => rounds (default 5)
- cases per round (default 3), same weights
    1. average user's choices
    2. recalculate weights
    3. next round

Changes State-of-the-Art Presentation -> Now

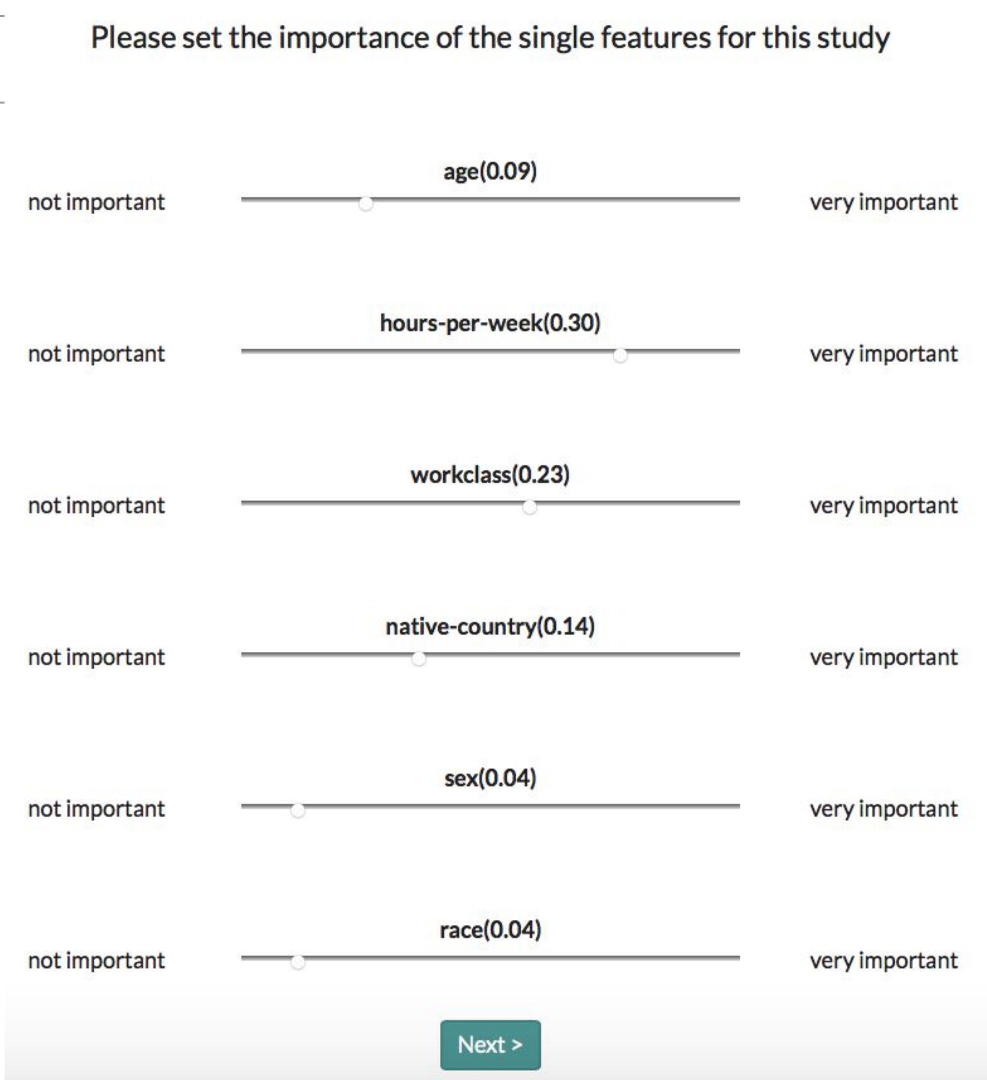
- new design (SCSS, flexbox)
- CSS autoprefixer (cross browser support)
- progress bar
- improved sliders' UI
- minor bugfixes
- major code clean-up



Choose survey for your study

#	Survey Title
1	Marital Status
2	Income
3	Education

Example for Weight-Selection



Please set the importance of the single features for this study

age(0.09)
not important — very important

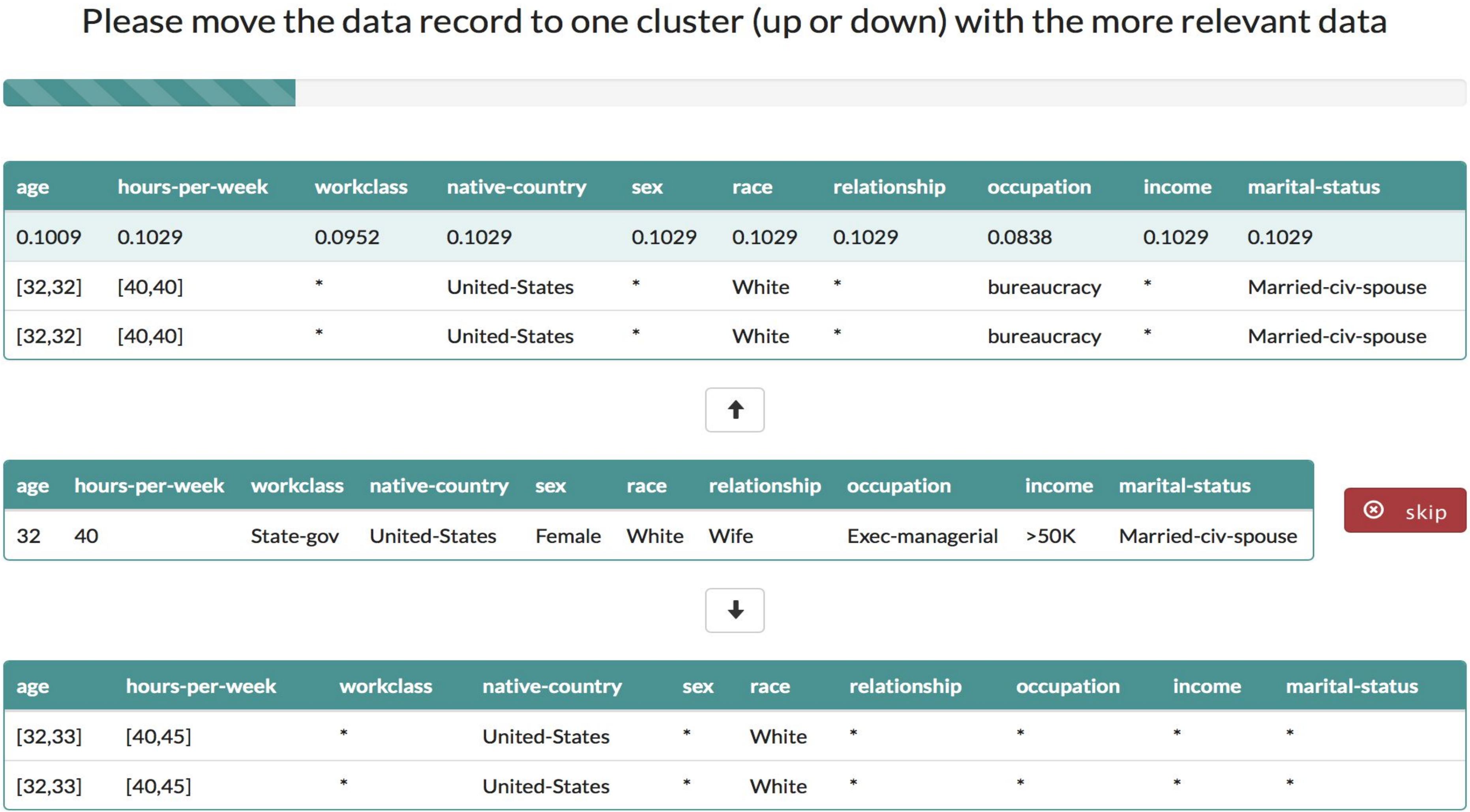
hours-per-week(0.30)
not important — very important

workclass(0.23)
not important — very important

native-country(0.14)
not important — very important

sex(0.04)
not important — very important

race(0.04)
not important — very important



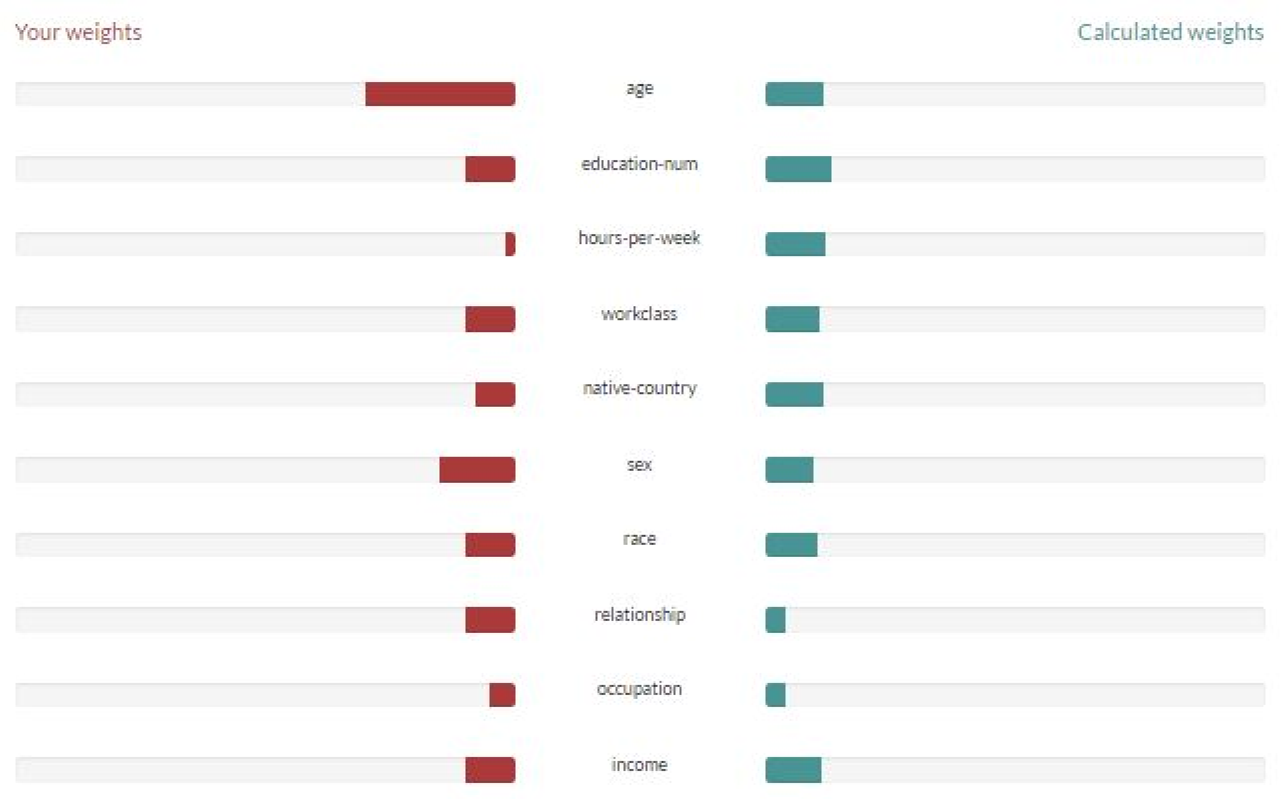
Please move the data record to one cluster (up or down) with the more relevant data

age	hours-per-week	workclass	native-country	sex	race	relationship	occupation	income	marital-status
0.1009	0.1029	0.0952	0.1029	0.1029	0.1029	0.1029	0.0838	0.1029	0.1029
[32,32]	[40,40]	*	United-States	*	White	*	bureaucracy	*	Married-civ-spouse
[32,32]	[40,40]	*	United-States	*	White	*	bureaucracy	*	Married-civ-spouse

age	hours-per-week	workclass	native-country	sex	race	relationship	occupation	income	marital-status
32	40	State-gov	United-States	Female	White	Wife	Exec-managerial	>50K	Married-civ-spouse

age	hours-per-week	workclass	native-country	sex	race	relationship	occupation	income	marital-status
[32,33]	[40,45]	*	United-States	*	White	*	*	*	*
[32,33]	[40,45]	*	United-States	*	White	*	*	*	*

Weight Comparison User - iML



Live Demo

...

https://github.com/tamarafeiertag/iML-Anonymization