Sharing Knowledge.Preserving Privacy.

UnKnown: A Data Anonymization Tool

Carnegie Mellon University

Engineering Privacy in Software

Team Members: Hana Habib, Preethi Josephina Mudialba, Siddharth Nair, Bill Quan, Chelse Swoopes, Lidong Wei
Project Advisor: Rebecca Balebako

Step 2: Click “Get Started” to begin the process of uploading your dataset.

Step 3: Upload the dataset by choosing the file from your computer.

***Note: This tool only supports **CSV** files.

Step 4: Your dataset is now uploaded. Click “Next” to continue.

unKnown

Home Tutorials

Import Dataset | Anonymize Attributes | Preview Dataset | Export Dataset

1. Import Dataset

Select the CSV file for the dataset you want to anonymize.

Choose File od2016_mod.csv

Age	Sex	Race	Overdose	Incident Zip
37	Female	Black	Cocaine	15211
26	Male	Black	Acetyl Fentanyl	15203
39	Male	Black	Heroin	15209
56	Male	Black	Alcohol	15208
65	Male	Black	Alcohol	15212

Success! Displaying 5 of 420 data points.

Back Next

Step 5: All attributes are now listed with a drop down menu.

Step 6: Click the drop down to specify whether each attribute is a quasi-identifier or sensitive. Note: There are additional categories for quasi-identifiers that are age and zip code.

Step 7: Select a level of ‘k’ that you would like to apply to your dataset, then click ‘Next.’

Step 8: Summary statistics for your anonymized dataset are shown.
If you are satisfied with the statistics, proceed to step 9. If you would prefer more or less anonymization/usability, please select ‘Back’ to re-select a value for ‘k.’

3. Preview Dataset

Summary statistics about the anonymization applied.

Legend:
- Information loss - definition for statistic
- Optimal generalization - definition for statistic
- Equivalence class - definition for statistic

General Statistics	
- Time needed	0.24s
- Information loss	0.320 / 0.320
Optimal Generalization	
- Incident Zip	4
- Sex	0
- Race	0
- Overdose	26
- Age	3
Equivalence Class Statistics	
- Average equivalence class size	105.0
- Average equivalence class size (including outliers)	105.0
- Maximal equivalence class size	252
- Maximal equivalence class size (including outliers)	252

Step 9: (Optional) If you would like to query your dataset, follow the on-screen directions and click 'Submit Query'. Otherwise, skip the on-screen directions and click 'Next'.

Query the original and anonymized datasets by entering a SQL query below.

Info:
- `db1` queries the original dataset
- `db2` queries the anonymized dataset
- `db` queries both datasets

```
SELECT * FROM `db` LIMIT 5;
```

Submit Query

Original Data:

37	Female	Black	Cocaine	15211
26	Male	Black	Acetyl Fentanyl	15203
39	Male	Black	Heroin	15209
56	Male	Black	Alcohol	15208
65	Male	Black	Alcohol	15212

Anonymized Data:

[16-85]	Female	Black	**********	*
[16-85]	Male	Black	**********	*
[16-85]	Male	Black	**********	*
[16-85]	Male	Black	**********	*
[16-85]	Male	Black	**********	*

Back Next

Step 10: To retrieve a copy of your anonymized dataset, click 'Export CSV.'

Sharing Knowledge.Preserving Privacy.

UnKnown: A Data Anonymization Tool