Information/Tutorial Pages

Engineering Privacy in Software

DATA ANONYMIZATION: WESTERN PENNSYLVANIA REGIONAL DATA CENTER

TEAM MEMBERS
HANA HABIB
PREETHI JOSEPHINA MUDIALBA
SIDDHARTH NAIR
QUAN(BILL) QUAN
CHELSE SWOOPES [Project Manager]
LIDONG WEI

Index

I. Goal of Tool

Our goal is to protect the privacy of data subjects included in a particular dataset of interest to the WPRDC, as well as other datasets containing similar data. Attributes that our tool can anonymize are those which could lead to re identification because of personally identifiable information (PII).

II. Data Privacy

PII is any information which can either be used as is or combined with additional information to identify an individual. PII can be used maliciously to negatively affect the life of an identified individual [1]. Besides an outside party knowing private information about the data subject, the outside party could use the information to negatively affect the identified data subject.

E-mail addresses, dates of birth, addresses, and phone numbers, or combinations of attributes may lead to reidentification of individuals. Depending on the type of data included in the dataset, anonymization will reduce the risk of both objective harms, including identity theft or credit card fraud, and subjective harms, including unwanted attention and discomfort associated with the fear of being reidentified from the release of a dataset.
To further emphasize the importance of dataset privacy, see the following video:
http://www.tripwiremovie.net

III. Suppression

When using **unKnown**, users have the option to select whether an attribute is 'Sensitive,' 'Insensitive,' or 'Identifying.'
When a user opts to select an attribute as 'Identifying,' **unKnown** suppresses the information associated with that attribute. Suppressing an attribute removes private information to protect an individual or group from being inferred or identified.
EXAMPLE: After suppression, rows 1 and 3 are identical and rows 2 and 4 are identical [2].

Original Dataset

first	last	age	race
Harry	Stone	34	Afr-Am
John	Reyser	36	Cauc
Beatrice	Stone	34	Afr-Am
John	Delgado	22	Hisp

Suppressed Dataset

first	last	age	race
*	Stone	34	Afr-Am
John	*	*	*
*	Stone	34	Afr-Am
John	*	*	*

[2]

IV. The Theory of K-Anonymity [FAQ]

1. Who created k-anonymization? Why?

The theory of k-anonymity was developed by Dr. Latanya Sweeney. It is based on the the observation that 87% of the U.S. population is uniquely identified by date of birth, gender, postal code [3].

2. Why not use just suppression?

As shown in the example above, suppressing an attribute/column completely redacts information about that particular attribute. This can reduce the quality and usability of the dataset [3].

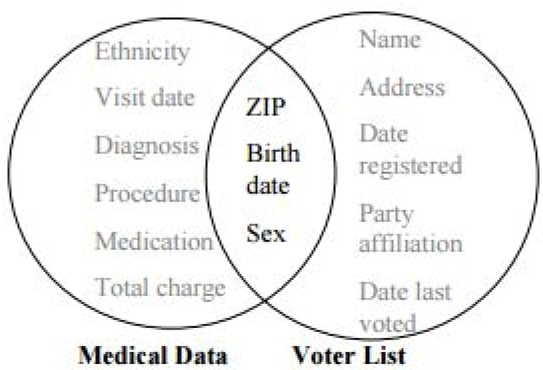
3. What is the purpose of K-Anonymity?

The purpose is to reduce linking of data subjects across different datasets while also keeping a useful dataset [3].

4. What causes linking?

Quasi-identifiers - “set of data elements in entity-specific data that in combination associates uniquely or almost uniquely to an entity and therefore can serve as a means of directly or indirectly recognizing the specific entity that is the subject of the data” [3]

<u>EXAMPLE:</u> From medical data and voter list, the quasi-identifiers that can be used to link an individual to the data include zip code, birthdate, and sex [3]



[3]

5. How does k-anonymity work?

Let’s say you have a table, represented as RT, with $A_1,\ldots,A_n$ attributes. The quasi-identifiers associated with your table are represented as QI_{RT} [4].

k-anonymity is satisfied if and only if “each sequence of values in $RT[QI_{RT}]$ appears with at least k occurrences in $RT[QI_{RT}]$” [4].

EXAMPLE: A table that satisfies k-anonymity: k=2 and QI= {Race, Birth, Gender, Zip} [4]

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

[4]

Notice: For each tuples in the table [T], the values of the tuple that comprise the quasi-identifier appear at least twice in the table [4].

→ t1[QI_T] = t2[QI_T], t3[QIT] = t4[QI_T], t5[QI_T] = t6[QI_T], t7[QI_T] = t8[QI_T] = t9[QI_T], and t10[QI_T] = t11[QI_T] [4]

V. Technical Requirements

Software or Hardware	Usage	Source
Development Computer(s)	Tool Use	User's discretion
Dataset	Anonymization	Any CSV Dataset From User
Git & Github Repo	Source Code README	https://github.com/hhabib/anonymization-tool

Note: Users are expected to know their dataset well and be able to identify personally identifiable information(PII) in the dataset.

VI. Installation: How to Run Our Tool

1. Docker Compose Installation

NOTE: Docker must be installed on your computer/server for this tool. For Windows users, Docker requires Windows Pro 10 and Hyper V; an alternative is to download docker via virtual machine.

Docker Installation Instructions: here.

*Operating Systems

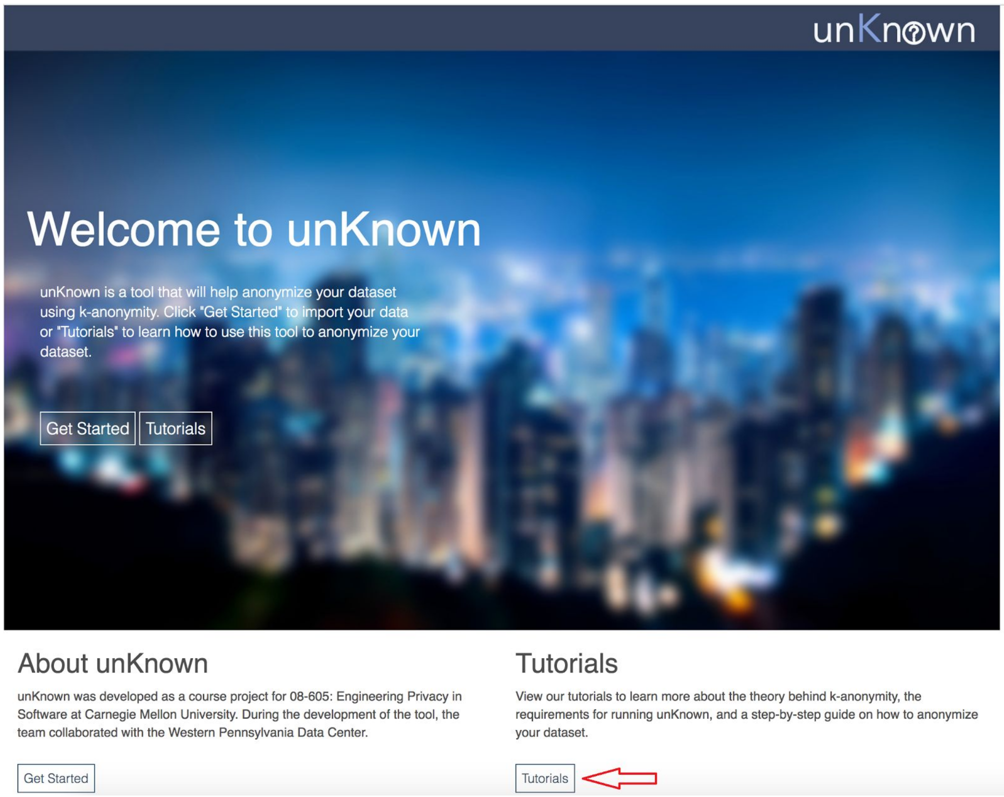
1. If you are a Mac or Windows user:
2. You are done! Docker compose in already installed along with your docker.
3. If you are a Linux user:
4. Docker-compose Installation Instruction: here. (Only for Linux User)

2. Launch Our Tool

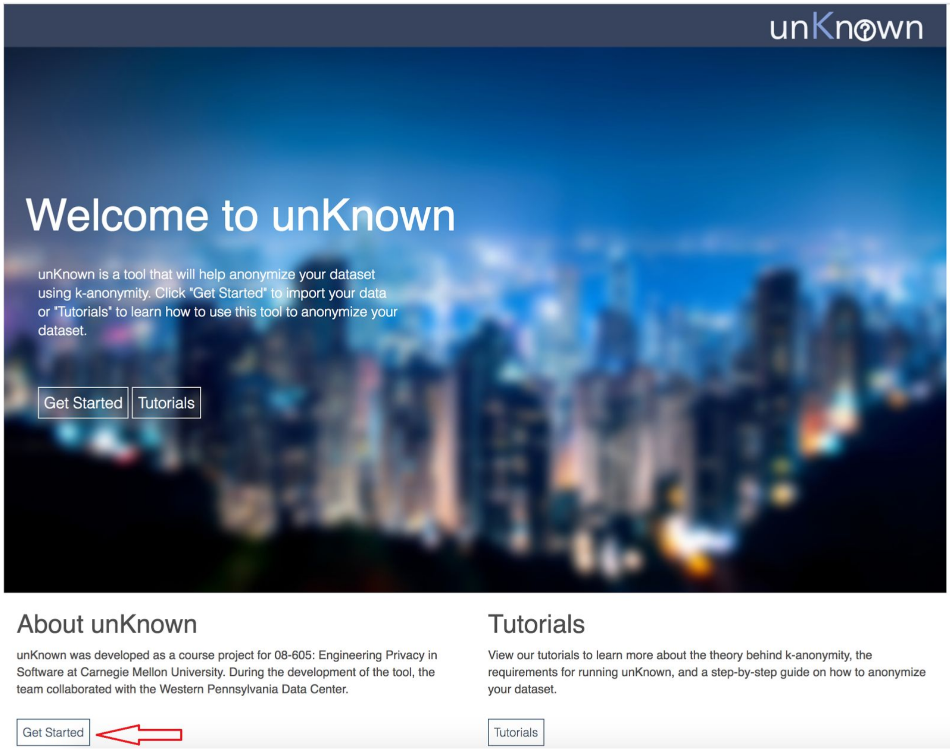
1. Make sure you have docker-compose installed
2. Download the code from our github repository
3. Go the the directory with `docker-compose.yml`, run `docker-compose up`
4. Note that the initial attempt of running this command will take several minutes to build the image. After the first build, it will be faster.
5. Visit our tool at `localhost:8000` with your browser.

VII. Tool Tutorial

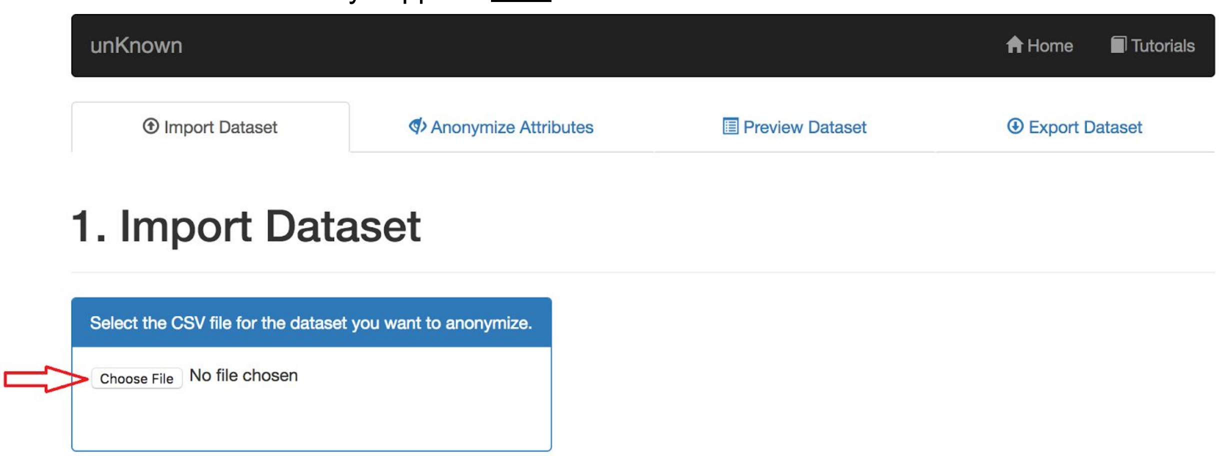
Step 1: Click ‘Tutorials’ to download a PDF copy of the “Information/Tutorial Pages.”



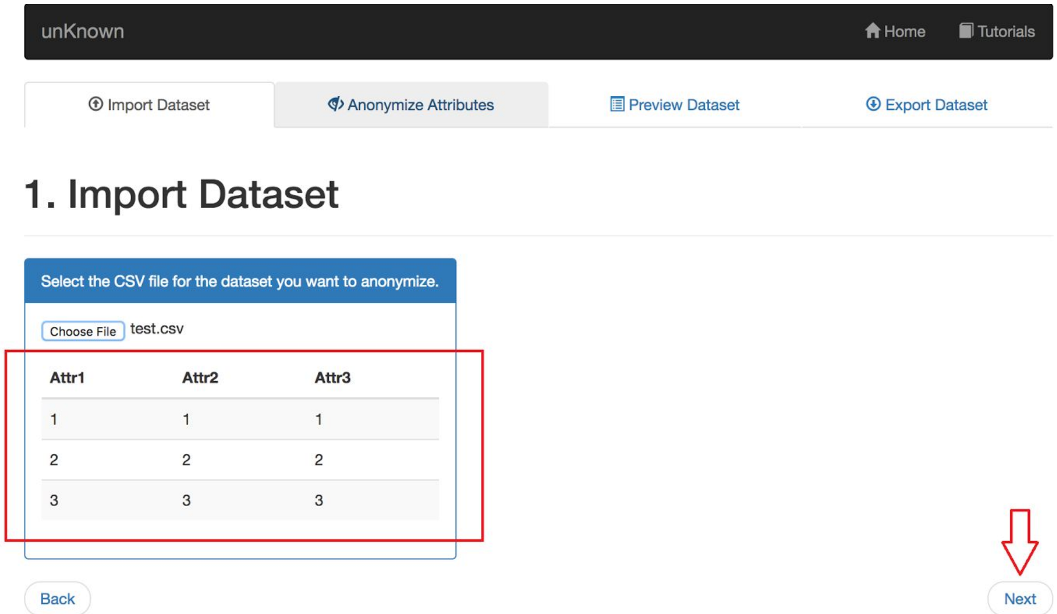
Step 2: Click “Get Started” to begin the process of uploading your dataset.



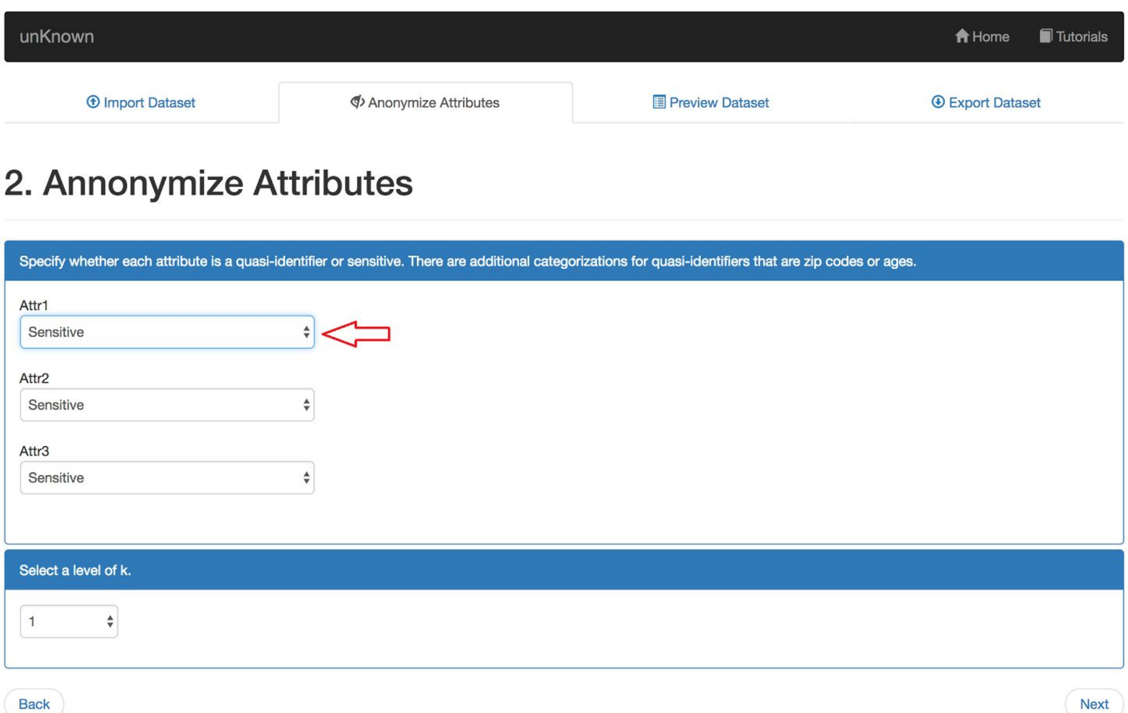
Step 3: Upload the dataset by choosing the file from your computer.
***Note: This tool only supports **<u>CSV</u>** files.



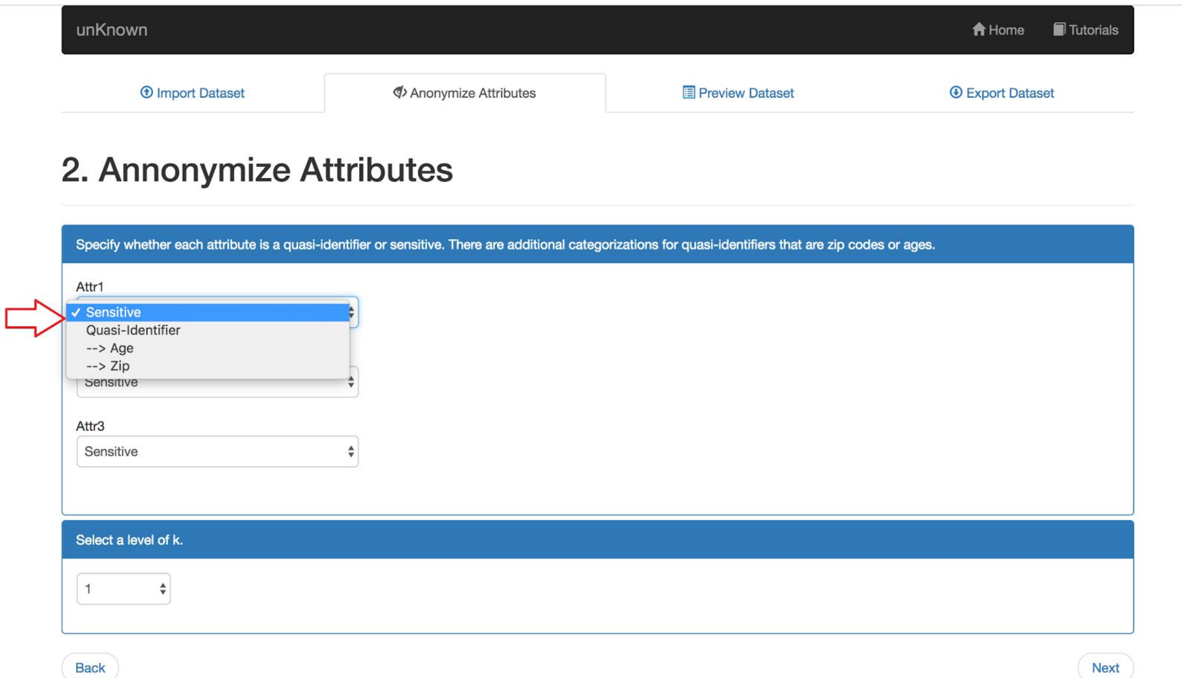
Step 4: Your dataset is now uploaded. Click “Next” to continue.



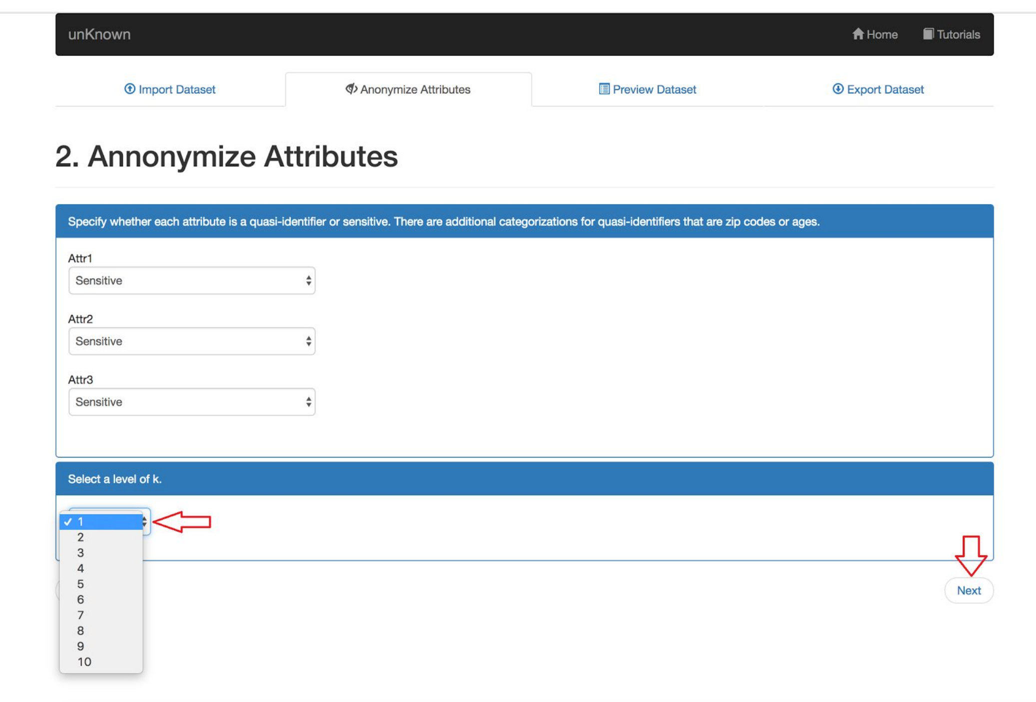
Step 5: All attributes are now listed with a drop down menu.



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.

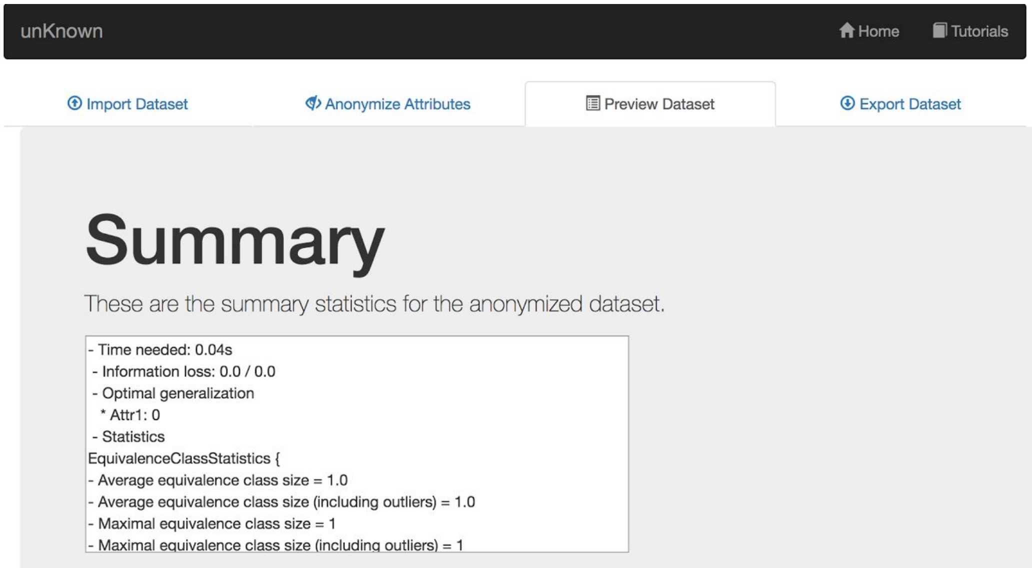


Step 7: Select a level of ‘k’ that you would like to apply to your dataset, then click ‘Next.’

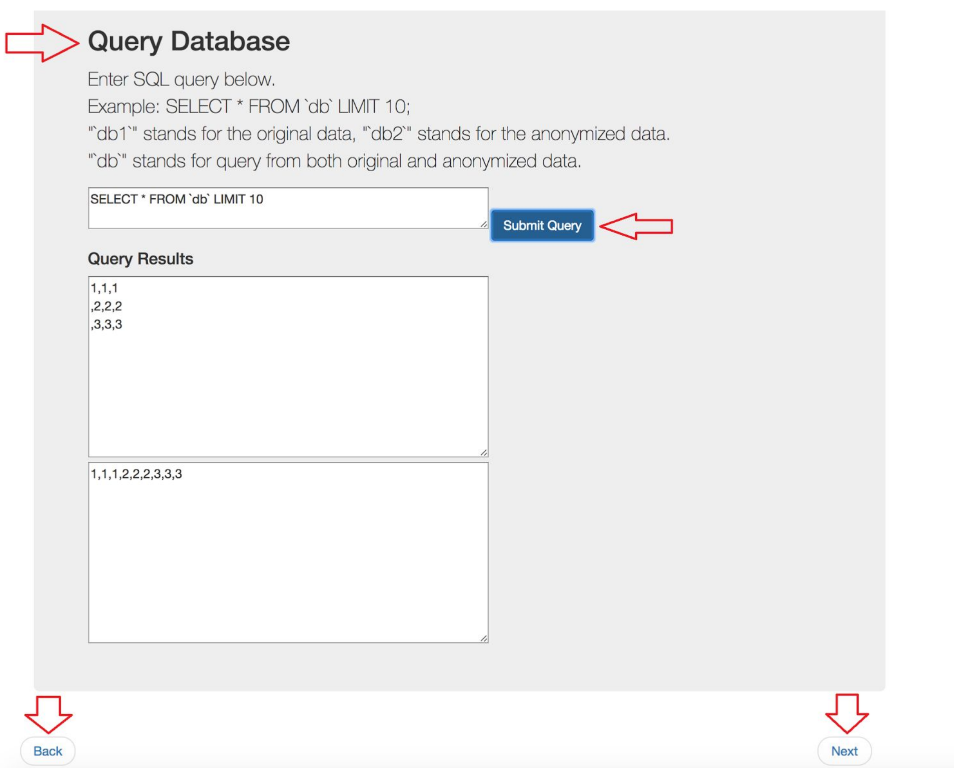


Step 8: Summary statistics for your anonymized dataset are shown.

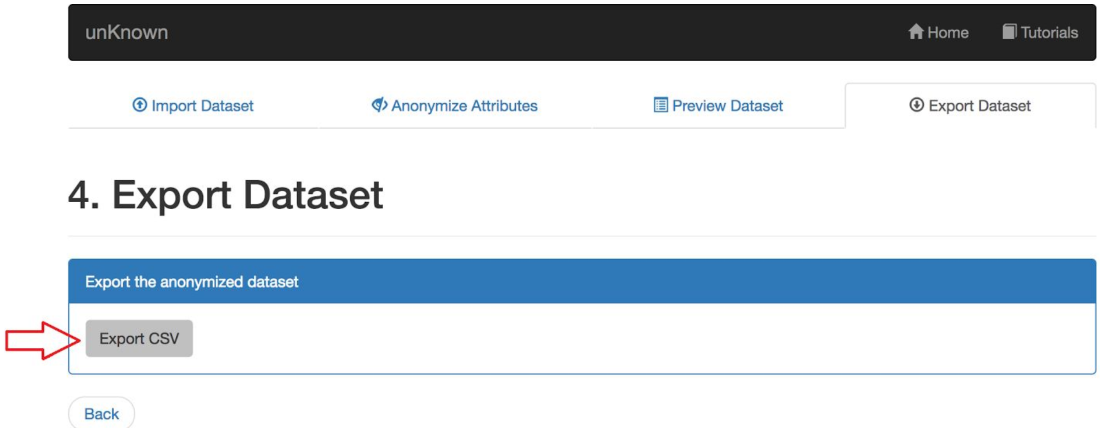
If you are satisfied with the statistics, proceed to step 9. If you would prefer more or less anonymization/usability, please select 'Back' to re-select a value for 'k.'



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



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



VIII. References

1. B. Raghunathan. The Complete Book of Data Anonymization From Planning to Implementation. Chapter 1. Published 2013. http://www.odbms.org/wp-content/uploads/2014/03/The-Complete-Book-of-Data-Anonymization_Chap_1.pdf

2. Author Unknown. “K-Anonymity”. Published 2007. https://www.cs.cmu.edu/~jblocki/Slides/K-Anonymity.pdf

3. L. Sweeney. “Simple Demographics Often Identify People Uniquely. Published 2000. https://dataprivacylab.org/projects/identifiability/paper1.pdf

4. L. Sweeney. “k-anonymity: a model for protecting privacy”. Received May 2002. https://epic.org/privacy/reidentification/Sweeney_Article.pdf

For additional information on data privacy, suppression, and k-anonymity, please see the following references.

- P. Samarati and L. Sweeney. “Generalizing Data to Provide Anonymity when Disclosing Information”. Publish Date NA. https://pdfs.semanticscholar.org/5c15/b11610d7c3ee8d6d99846c276795c072eec3.pdf

- L. Ohno-Machado, S. Venterbo, and S. Dreiseitl. “Effects of Data Anonymization by Cell Suppression on Descriptive Statistics and Predictive Modeling Performance”. Published Nov 2002. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC419433/