DataSifter Workflow

**Step 1**: The data we use includes structured numerical/categorical columns and unstructured text

**Step 2**: Identify sensitive outcome(s) to protect

**Step 3** [Yiming]: LightGBM on the text feature/cell to identify predictive keywords for the sensitive outcomes in Step 2 separately.

**Step 4** [Yitong]: Build a “semantic radius” around each keyword to establish a semi-quantitative distance metric to anchor different levels of obfuscation. For example, the top 10 keywords for the outcome “Marital Status” are: 'wife', 'husband', 'married', 'alone', 'daughter', 'widowed', 'son', 'lives', 'sex', 'she'. Based on semantic meanings, we can extract two “semantic clusters”: 1: 'wife', 'husband', 'daughter', ‘son’; 2: 'married', 'alone', 'widowed'. This step is to guarantee the readability of obfuscated text.

Input 1 (for obfuscation): # keywords (higher = more obfuscation)

Input 2 (for obfuscation): radius around each keyword using word2vec (further away = more obfuscation)

For example:

*small*=top 3 keywords, 0-20% distance from each keyword

*medium*=top 5 keywords, 20-40% distance from each keyword

*large*=top 10 keywords, 40–60% distance from each keyword

# keyword=k

# words on the semantic radius=N

Extra columns/features for SDV = k + k\*N = k\*(N+1)

RECIPES FOR STEP 5

* KEYWORD: wife → 'husband', 'married',[0-33%] 'alone', 'daughter', 'widowed',[34-66%] 'son', 'lives', 'sex', 'she'[67-100%]
* *semantic\_radius(keyword,#words on the radius=N)*
* word2vec\_small(*keyword*=wife, *semantic\_radius=*search words within dist of 2 from wife,1 to *N/3*)
* word2vec\_medium(*keyword*=wife, *semantic\_radius*=search words within dist of 2-4 from wife,N/3 to 2N/3)
* word2vec\_large(*keyword*=wife, *semantic\_radius=*search words within dist of 4-6 from wife,2N/3 to N)

N words on the spectrum small-medium-large

word2vec: <https://en.wikipedia.org/wiki/Word2vec#:~:text=Word2vec%20is%20a%20technique%20for,words%20for%20a%20partial%20sentence>. Training on the current data in order to vectorize - DATA DEPENDENT

GloVe: <https://en.wikipedia.org/wiki/GloVe_(machine_learning)> Training on the current data in order to vectorize - DATA DEPENDENT

BERT: <https://en.wikipedia.org/wiki/BERT_(language_model)> UNIVERSAL

**Step 5**: Apply Datasifter to obfuscate Data: swap the keywords within “semantic clusters”, may use BERT to further improve readability?

**Step 6** [Yitong]: recast text (before and after obfuscation) into tabular (e.g., incident matrix)

**Utility metric**: apply LightGBM on the text feature/cell to identify predictive keywords for the sensitive outcomes in Step 2 and after Step 5 (after semantic radius has been ap[plied and keywords swapped).

**Step 7** [Yitong]: apply SDV Privacy-Utility metrics to the recasted datasets (before and after obfuscation)

Husband: {wife, daughter, son, child, name, soccer,

{wife, daughter, son, child,....}