Data Privacy in the Digital World

Tianen "Benjamin" Liu

4/30/2020

What's Wrong with Our Data?

- 2 pieces of information for example [1]: 1) there is only 1 person in a distant town A who has a certain disease; 2) Bob is from town A and checked in to the hospital
- Oftentimes, just removing sensitive information is not enough to protect the privacy of the data
- The sensitive information is identifiable when linked with another data set
- 87% of individuals living in the US can be uniquely identified by using 3 data features: birth date, zip code, and gender [2]

Introduction - Differential Privacy & Synthetic Data

- I'm doing a survey about mental health that requires my sensitive information
- 3 options for my college to release the data set for research
  - Option 1: Just remove sensitive information
- Still not private enough. My college will modify the data set
  - Captures the characteristics of the original data set while also making my information unidentifiable

Introduction - Differential Privacy & Synthetic Data

- I'm doing a survey about mental health that requires my sensitive information
- 3 options for my college to release the data set for research
  - Option 1: Just remove sensitive information
  - Option 2: Release a differentially private data set
  - Option 3: Release a synthetic data set

Healthy Minds Data Set

- Thanks to Dr. Denisha Champion and WFU Counseling Center
- Containing students' sensitive information: demographic, mental health disease diagnosis, self evaluations...
- Important for the Counseling Center and student wellbeing

Outline

- Definition of differential privacy
- Differential privacy method 1: Laplacian noise
- Synthetic data
- Differential privacy method 2: differentially private trees

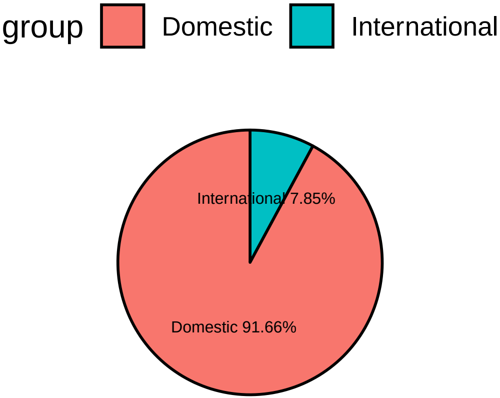
- Healthy Minds data analysis

Differential Privacy - Mathematical Definition

- Notations:
  - i: individual person. I: population
  - d_i: the information given by person i. $D_I = d_i | i \in I$: the whole data set
  - Q: the privatized query run on a data set
  - $R = Q(D_I)$: the resultant modified data set released publically using Option 2 and 3. r: one data point in R.

Differential Privacy - Mathematical Definition

- Notatins:
  - Q be the privatized query run on a data set
    - Number of international student responses: 128 (true value)
  - $R = Q(D_I)$ be the resultant modified data set released publically
    - Number of international student responses: 129 (128.7)



Differential Privacy - Mathematical Definition

- Unidentifiability makes sure that we don't know if a person is in the data set or not
- Then a single answer make no difference on the probability of getting the released data set
- we have

$$Q(D_{I-me}) = Q(D_I)$$

This should hold whenever, meaning the probability of $Q(D_{I-me})$ being equal to $Q(D_I)$ should be similar. Thus ϵ-differential privacy is defined as [3]:

$$\frac{Prob(Q(D_I) = r)}{Prob(Q(D_{I\pm i}) = r)} \leq e^{\epsilon}, \text{ for small } \epsilon \geq 0$$

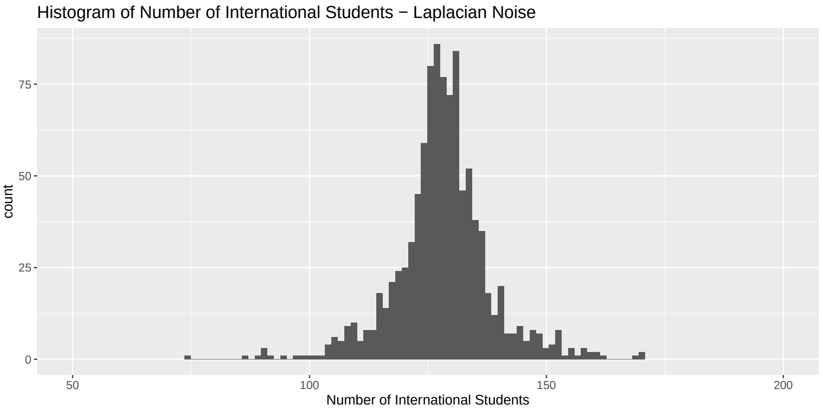
Differential Privacy Methods (Option 2)

Method 1: Laplacian Noise

- Add to the true answers noises drawn from the Laplace distribution: $\mathit{TrueValue} \pm \mathit{Noise}$
- Parameters: $\mathit{Noise} \sim \mathit{Lap}(\mu, b)$

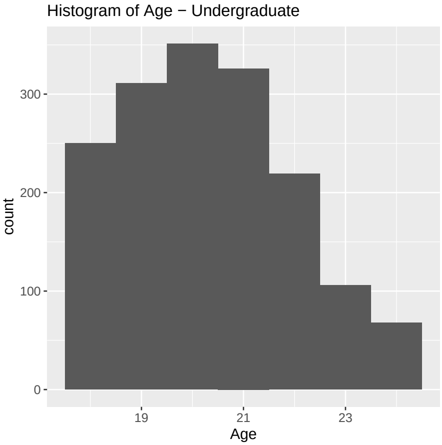
Method 1: Laplacian Noise - Count & Statistic

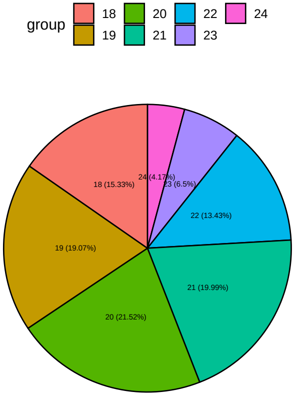
- Useful when releasing the count, mean, median,...
- Example: Query = count of international students in a data set (true count = 128)
- A lot of the noises are small



Method 1: Laplacian Noise - Numeric Data

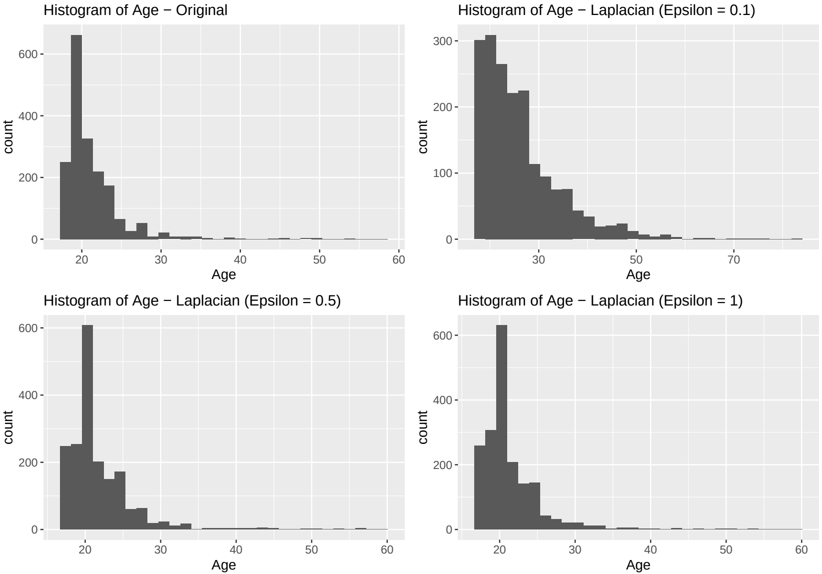
- Add noise to each age and round to a whole number
- Visualization of age distribution in Undergraduate responses





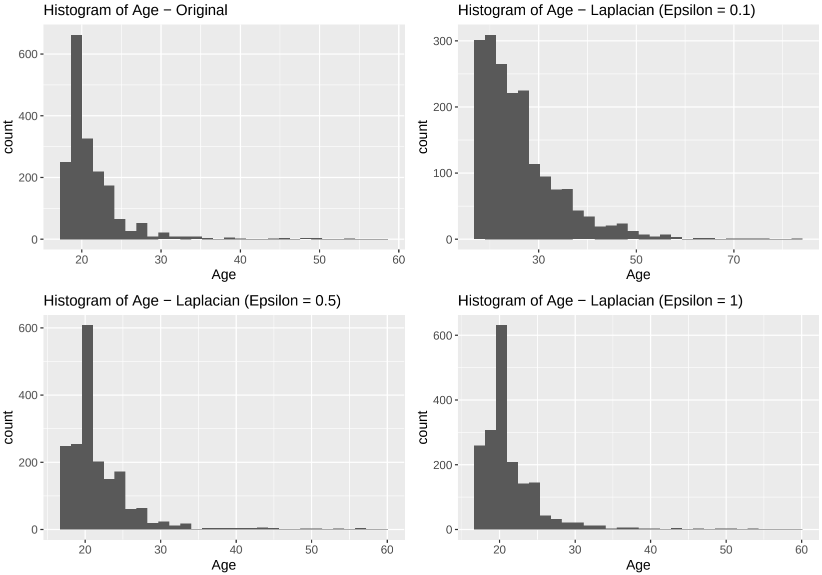
Method 1: Laplacian Noise - Numeric Data

- To visualize better, we use the age data from the full survey



Method 1: Laplacian Noise - Numeric Data

- Utility and Privacy Trade-off



Synthetic Data (Option 3)

Method 2: Synthetic Data

- Laplacian noise only creates a synthetic column (age, for example)
- Relationship breaks due to different age values
- Need other attributes to mimic the the original data
- Synthetic Data does not guarantee differential privacy, but concern is little.

	Age with Laplacian Noise	Original Age
Depression	0.089	0.164
Eating Disorder	0.011	-0.143
Neurodev. Disorder	0.205	0.0243

Table 1: Multinomial Regression: Mental Health Illness vs. Age

Method 2: Synthetic Data

- How do we create a synthetic data?
  - Similar to Multivariate Imputation by Chained Equations (MICE)
  - Column by column
  - Methods of generating new columns
  - Value constraints

Wake Forest University Healthy Minds Dataset

- Kristin Neff Self-Compassion Scale [4]
  - Q1: When I fail at something important to me I become consumed by feelings of inadequacy.
  - Q2: I try to be understanding and patient towards those aspects of my personality I don't like.
- How does the Neff Self-Compassion Scale impact Depression?
- Create a synthetic data for just `Depression` and Compassion Scale responses (denoted as `Comp_Scale` in models)
- Process:
  - Synthesize `Depression`, and use it as a predictor to synthesize `Comp_Scale_Q1`
  - Use both `Depression` and `Comp_Scale_Q1` as predictors to synthesize `Comp_Scale_Q2`...

Method 2: Synthetic Data

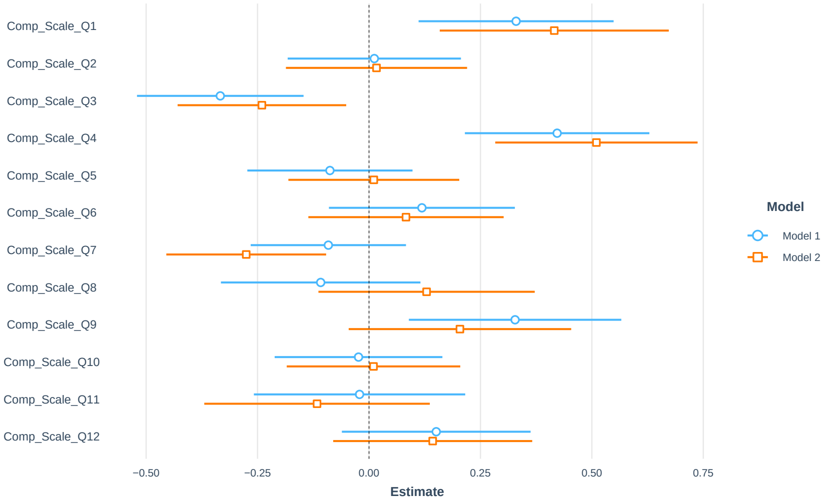
- Comparison between original and synthetic data

	Depression	Comp_Scale_Q1	Comp_Scale_Q2	Comp_Scale_Q3	Comp_Scale_Q4
2	0	NA	NA	NA	NA
3	0	3	2	3	2
4	0	4	1	3	2
5	0	3	4	4	3
6	1	1	4	4	1

	Depression	Comp_Scale_Q1	Comp_Scale_Q2	Comp_Scale_Q3	Comp_Scale_Q4
2	0	NA	NA	NA	NA
3	0	4	2	3	2
4	1	4	3	3	4
5	0	1	4	5	3
6	0	4	3	3	2

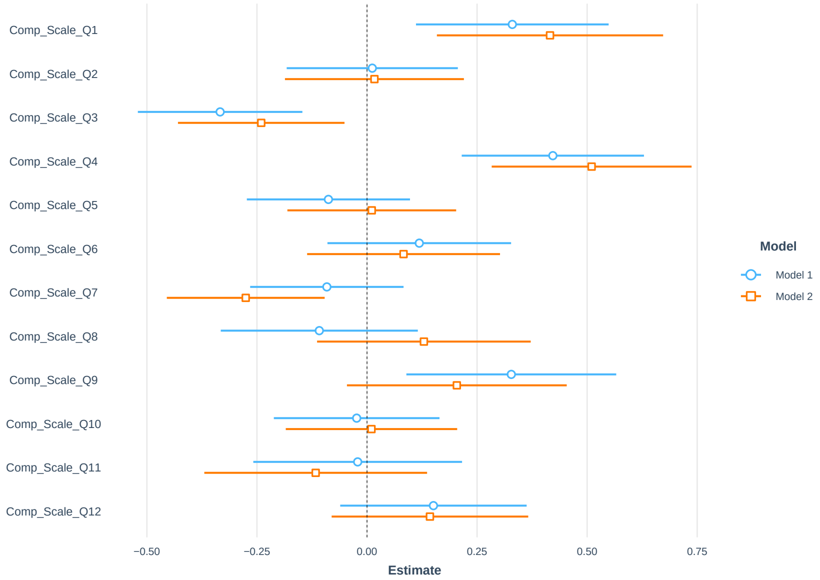
Method 2: Synthetic Data - Depression vs. Comp_Scale

- Confidence intervals of logistic regression parameters using original (Model 1) and synthetic (Model 2) data set



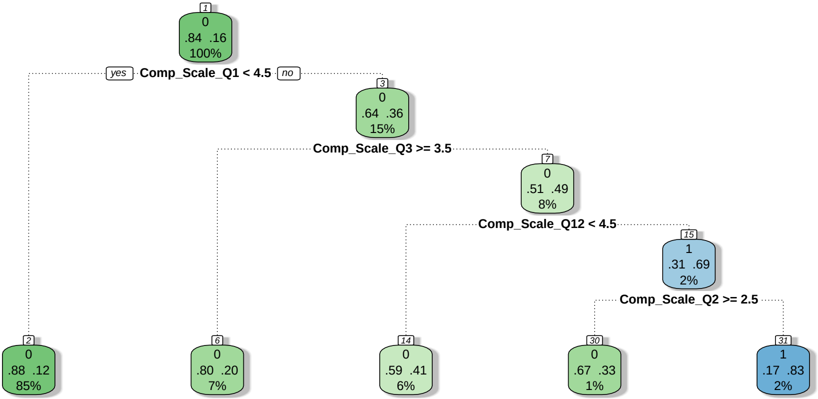
Method 2: Synthetic Data

- Does it help with our understanding of the relationship?



Tree Model

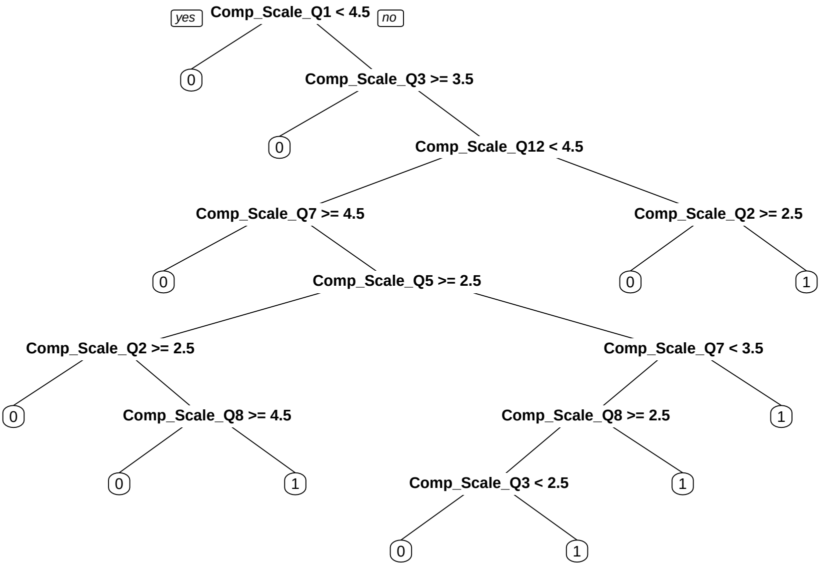
- Relationship between Kristin Neff Self-Compassion Scale and depression
- Tree model using original data (0: Not diagnosed with depression, 1: Diagnosed with depression)



"Comp_Scale_Q1": an answer to Question 1 of the Neff Compassion Scale short questions

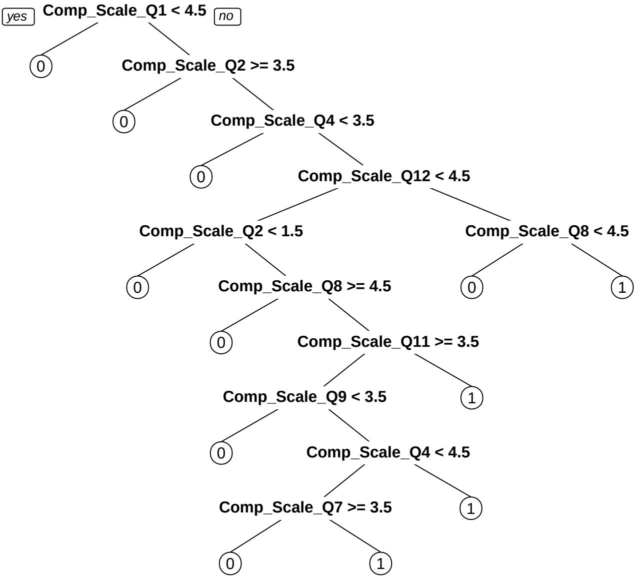
Tree Model

- Original data



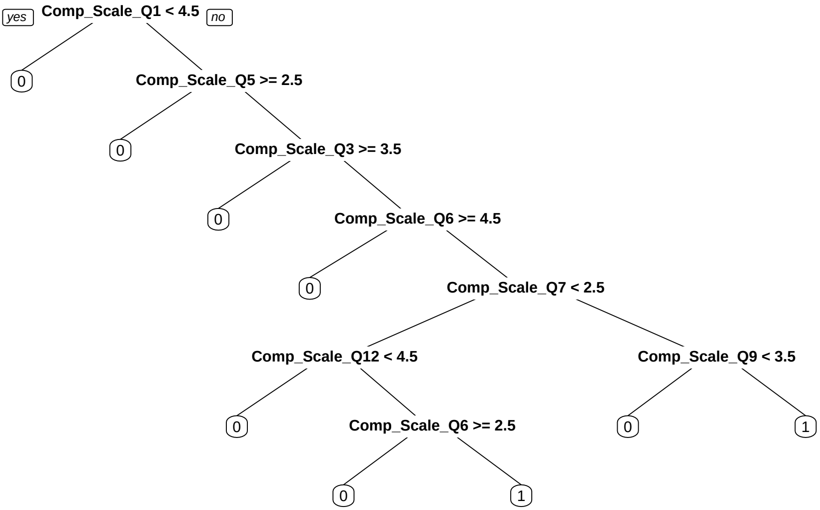
Tree Model

- Synthetic data



Tree Model

- Apply Laplacian noise on Compassion Scales (Explanatory Variable)

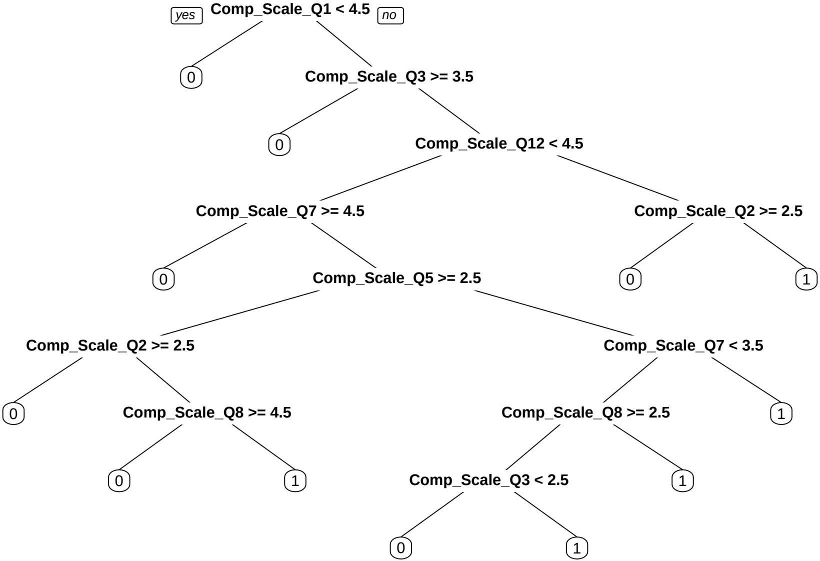


Method 3: Differentially Private Trees

- Take out each node and get the probability of 0 and 1 for `Depression`
- Apply Laplacian Noise on the probability of 1 and resample the response variable
- Synthetic column

Method 3: Differentially Private Trees

- In the new `Depression` column we generated 184 `Diagnosed` while true number is 230. New data resulted in the same tree.



Acknowledgements

- Dr. Nicole Dalzell
- Wake Forest University URECA Center & Wake Forest Research Fellowship
- Wake Forest University Department of Mathematics & Statistics
- Dr. Denisha Champion and Wake Forest University Counseling Center

Reference

- [1] Microsoft Corporation. Differential Privacy for Everyone. 2012.
- [2] L. Sweeney. Foundations of Privacy Protection from a Computer Science Perspective. Proceedings, Joint Statistical Meeting, AAAS, Indianapolis, IN. 2000.
- [3] Christine Task. Privacy-preserving Datamining: Differential Privacy And Applications. June 2014.
- [4] Raes, F., Pommier, E., Neff, K. D., & Van Gucht, D. (2011). Construction and factorial validation of a short form of the Self-Compassion Scale. Clinical Psychology & Psychotherapy. 18, 250-255.