Data Synthesis:
A Tool for Responsible Data Sharing

Khaled El Emam

16th June 2021
Agenda

1. Introduction to Synthesis
   General description of what synthetic data is and general use cases

2. Privacy and Utility
   An examination of privacy risks and the utility of synthetic data

3. Methods
   A brief look at methods for the generation of synthetic data
Synthetic Data Uses

• Data Sharing and Data Access
  • AI and data science projects
  • Software testing
  • Proof of concept and technology evaluations
  • Open data/open science
  • Hackathons and data competitions/challenges

• Data Amplification and Data Augmentation
  • Amplifying small datasets
  • Correct bias
The Synthesis Process

![Diagram of the Synthesis Process]

<table>
<thead>
<tr>
<th>COU1A</th>
<th>AGECAT</th>
<th>AGELE70</th>
<th>WHITE</th>
<th>MALE</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>33.75155</td>
</tr>
<tr>
<td>United States</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>39.24707</td>
</tr>
<tr>
<td>United States</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>26.5625</td>
</tr>
<tr>
<td>United States</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>40.58273</td>
</tr>
<tr>
<td>United States</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>24.42046</td>
</tr>
<tr>
<td>United States</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>19.07124</td>
</tr>
<tr>
<td>United States</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>26.04938</td>
</tr>
<tr>
<td>United States</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>25.46939</td>
</tr>
</tbody>
</table>
Data Simulator

Allows generation of synthetic data without direct access to real data
Simulator Exchange

Data Consumers
Two Synthesis Strategies

Full Synthesis
Synthesize all variables

Partial Synthesis
Synthesize quasi-identifiers
Identifiability Spectrum

Identifiability Threshold

Identifiable Data (Probability=1)

Not Identifiable Data (Probability=Ø)

Personal Information

Not Personal Information
## Privacy Risks

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Fully Synthetic Data</th>
<th>Original Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Washington Hospital Data</td>
<td>0.0197</td>
<td>0.098</td>
</tr>
<tr>
<td>Canadian COVID Data</td>
<td>0.0086</td>
<td>0.034</td>
</tr>
</tbody>
</table>

A commonly used risk threshold = 0.09
Privacy-Utility Tradeoff

![Diagram showing the tradeoff between privacy protection and data utility. The graph illustrates that as privacy increases, utility decreases, with an ideal situation and acceptable tradeoff points.]
Distribution Comparisons

Real - Date_Reported and Synthetic - Date_Reported

Real - BMI and Synthetic - BMI
Mortality Over Time

![Graph showing mortality over time with two datasets: Real and Synthetic. The x-axis represents Date Reported, and the y-axis represents Mean Predicted Probability of Death. The graph compares the two datasets over time, highlighting differences and trends.](image-url)
Mortality By Age

![Graph showing mortality by age group with data points for real and synthetic datasets.](image-url)
Utility Framework

- An important concern of data users is the data utility
- Utility has multiple dimensions to it
- Synthetic data may be optimized on multiple utility dimensions simultaneously to meet the needs of multiple users, or on single dimensions to address the needs of limited users
Risk-based Approach

Data Transformations + Controls
Risk-based Approach

- Generalization
- Suppression
- Addition of noise
- Microaggregation
Risk-based Approach

- Security controls
- Privacy controls
- Contractual controls
The Erosion of Trust

The New York Times

Your Data Were ‘Anonymized’? These Scientists Can Still Identify You

Computer scientists have developed an algorithm that can pick out almost any American in databases supposedly stripped of personal information.

Opinion | THE PRIVACY PROJECT

Twelve Million Phones, One Dataset, Zero Privacy

By Stuart A. Thompson and Charlie Warzel
DEC. 10, 2019

the guardian

'Anonymised' data can never be totally anonymous, says study

Findings say it is impossible for researchers to fully protect real identities in datasets

You’re very easy to track down, even when your data has been anonymized

A new study shows you can be easily re-identified from almost any database, even when your personal details have been stripped out.

by Charlotte Jee
Jul 23, 2019

ACM TECHNEWS

'Anonymized' Data Can Never Be Totally Anonymous, says Study

by The Guardian

Online Profiling and Invasion of Privacy: The Myth of Anonymization

02/20/2013 12:23 pm ET | Updated Apr 22, 2013
Skill Set

- The skills needed to create de-personalized datasets are very specialized, take time to develop, and generally difficult to find cost-effectively.
- This limits the ability to scale.
- Synthesis requires minimal skills in practice – it is a computationally intensive process.
Regulatory Questions

• Is synthetic data considered non-identifiable information?

• Does the act of converting identifiable information into non-identifiable synthetic information require additional consent or authorization?

• Can a data custodian outsource the creation of synthetic data?

• Can synthetic data be used for any purpose?
Sequential Synthesis

![Diagram of Sequential Synthesis]

- **Real Data**
- **Synthetic Data**
- **Synthesizer**
- **Generator**
- **Discriminator**
- **Evaluation Results**
Variational Auto Encoder (VAE)
Generative Adversarial Network (GAN)
TOOLBOX OF TECHNIQUES
QUESTIONS
References