Synthetic Data Does Not Reliably Protect Privacy, Researchers Claim

Originally published at https://www.unite.ai/synthetic-data-does-not-reliably-protect-privacy-researchers-claim/

A new research collaboration between France and the UK casts doubt on growing industry confidence that synthetic data can resolve the privacy, quality and availability issues (among other issues) that threaten progress in the machine learning sector.

Among several key points addressed, the authors assert that synthetic data modeled from real data retains enough of the genuine information as to provide no reliable protection from inference and membership attacks, which seek to deanonymize data and re-associate it with actual people.

Furthermore, the individuals most at risk from such attacks, including those with critical medical conditions or high hospital bills (in the case of medical record anonymization) are, through the 'outlier' nature of their condition, most likely to be re-identified by these techniques.

The paper observes:

> *'Given access to a synthetic dataset, a strategic adversary can infer, with high confidence, the presence of a target record in the original data.'*

The paper also notes that differentially private synthetic data, which obscures the signature of individual records, does indeed protect individuals' privacy, but only by significantly crippling the usefulness of the information retrieval systems that use it.

If anything, the researchers observe, differentially private approaches – which use 'real' information 'at one remove' via synthetic data – make the security scenario *worse* than it would have been otherwise:

> *'[Synthetic] datasets do not give any transparency about this tradeoff. It is impossible to predict what data characteristics will be preserved and what patterns will be suppressed.'*

The new paper, titled *Synthetic Data – Anonymisation Groundhog Day*, comes from two researchers at École Polytechnique Fédérale de Lausanne (EPFL) in Paris and a researcher from University College London (UCL).

The researchers conducted tests of existing private generative model training algorithms, and found that certain implementation decisions violate the formal privacy guarantees provided in the frameworks, leaving diverse records exposed to inference attacks.

The authors offer a revised version of each algorithm that potentially mitigates these exposures, and are making the code available as an open source library. They claim that this will help researchers to evaluate the privacy gains of synthetic data and usefully compare popular anonymization methods. The new framework incorporates two pertinent privacy attack methods that can be applied to any generative model training algorithm.

Synthetic Data

Synthetic data is used to train machine learning models in various scenarios, including cases where a lack of comprehensive information can potentially be filled in by ersatz data. One example of this is the possibility of using CGI-generated faces to provide 'difficult' or infrequent face photos for image synthesis datasets, where profile images, acute angles or unusual expressions are often seldom-seen in source material.

Other types of CGI imagery have been used to populate datasets that will eventually be run on non-synthetic data, such as datasets that feature hands and furniture.

In terms of privacy protection, synthetic data can be generated from real data by Generative Adversarial Network (GAN) systems that extract features from the real data and create similar, fictitious records that are likely to generalize well to later (unseen, real) data, but are intended to obfuscate details of real people featured in the source data.

Methodology

For the purposes of the new research, the authors evaluated privacy gains across five generative model training algorithms. Three of the models do not offer explicit privacy protection, while the other two come with differential privacy guarantees. These tabular models were chosen to represent a wide range of architectures.

The models attacked were BayNet, PrivBay (a derivation of PrivBayes/BayNet), CTGAN, PATEGAN and IndHist.

The evaluation framework for the models was implemented as a Python library with two core classes – *GenerativeModels* and *PrivacyAttacks*. The latter features two facets – a membership inference adversary, and a membership inference attack. The framework is also able to evaluate the privacy benefits of 'sanitized' (i.e. anonymized) data and synthetic data.

The two datasets used in the tests were the Adult Data Set from the UCI Machine Learning Repository, and the Hospital Discharge Data Public Use Data File from the Texas Department of State Health Services. The Texas dataset version used by the researchers contains 50,000 records sampled from patient records for the year 2013.

Attacks and Findings

The general objective of the research is to establish 'linkability' (the reassociation of real data with synthetic data that was inspired by it). Attack models used in the study include Logistic Regression, Random Forests and K-Nearest Neighbors classifiers.

The authors selected two target groups consisting of five randomly-selected records for 'minority' categories of the population, since these are most likely to be susceptible to a linkage attack. They also selected records with 'rare categorical attribute values' outside of that attributes 95% quantile. Examples include records related to high risk of mortality, high total hospital charges, and illness severity.

Though the paper does not elaborate on this aspect, from the point of view of likely real world attackers, these are exactly the kind of 'expensive' or 'high risk' patients most likely to be targeted by membership inference and other kinds of exfiltration approaches to patient records.

Multiple attack models were trained against public reference information to develop 'shadow models' over ten targets. The results across a range of experiments (as described earlier) indicate that a number of records were 'highly vulnerable' to linkage attacks aimed at them by the researchers. Results also found that 20% of all targets in the trials received a privacy gain of *zero* from synthetic data produced by GAN methods.

The researchers note that results varied, depending on the method used to generate synthetic data, the attack vector and the features of the targeted dataset. The report finds that in many cases, effective identity suppression through synthetic data approaches lowers the utility of the resulting systems. Effectively, such systems' usefulness and accuracy can in many cases be a direct index of how vulnerable they are to reidentification attacks.

The researchers conclude:

> *'If a synthetic dataset preserves the characteristics of the original data with high accuracy, and hence retains data utility for the use cases it is advertised for, it simultaneously enables adversaries to extract sensitive information about individuals.*
>
> *'A high gain in privacy through any of the anonymisation mechanisms we evaluated can only be achieved if the published synthetic or sanitised version of the original data does not carry through the signal of individual records in the raw data and in effect suppresses their record.'*

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai