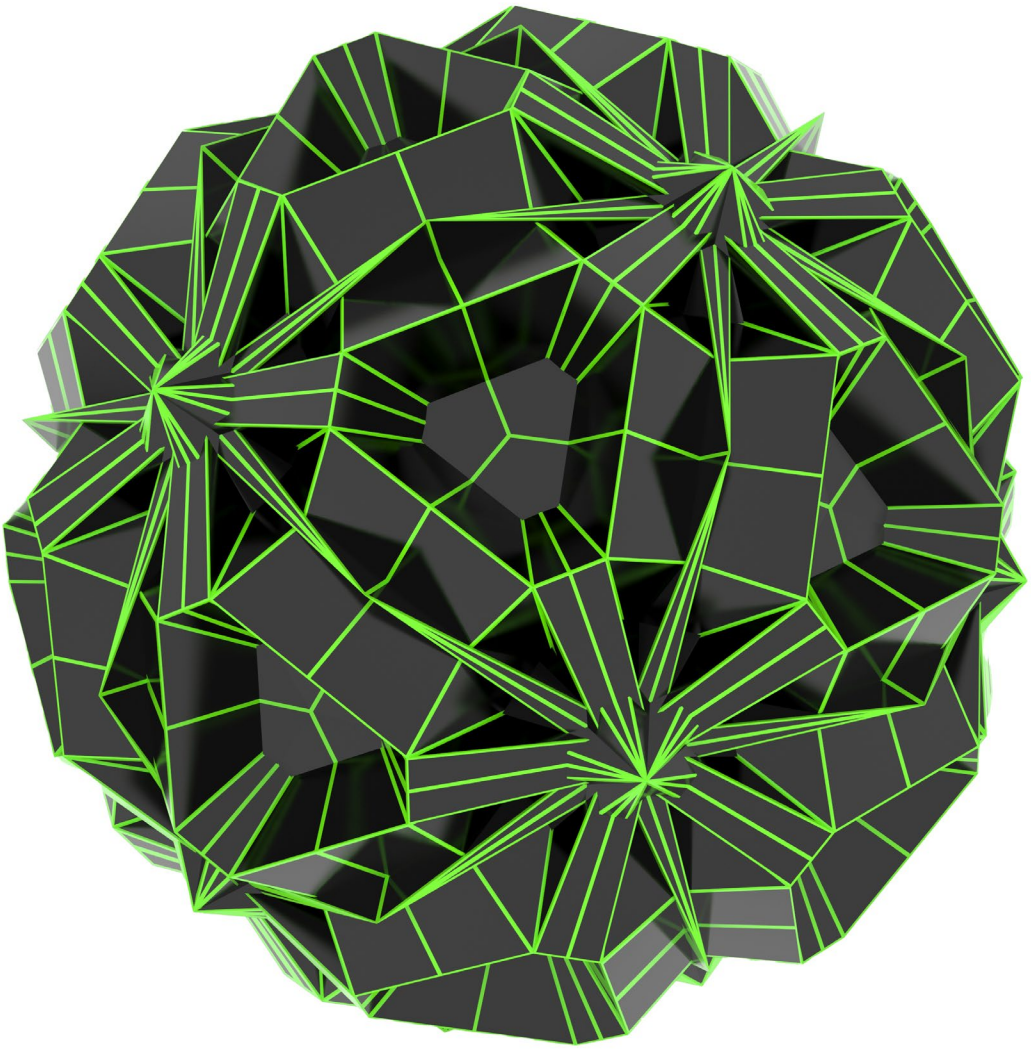
Deloitte.



Privacy, Part 2
Differential Privacy and Synthetic Data

Executive Summary

Facing regulatory penalties and reputational damage, companies and institutions around the world are highly motivated to protect the sensitive information of individuals from unauthorized access and disclosure. They can be made liable for unintentional leaks or targeted theft if data was not sufficiently protected and subsequently published. Many privacy-enhancing techniques can help protect sensitive information. This paper will explore the theoretical concepts and practical application of differential privacy and synthetic data, highlighting their relative strengths and limitations and the advantages of applying them in combination with traditional anonymization techniques.

MAKING AN IMPACT THAT MATTERS since 1845

Introduction

In the digital age, data is collected, used, and shared at an ever-increasing rate, a substantial portion of which is personal or otherwise. Periodic data breaches and privacy violations have nonetheless sensitized authorities and the general public to the acute need for improved protection measures. The Article 29 Working Party, an independent body advising the European Commission, has identified three criteria strongly related to re-identification risk:

Singling out an individual in a dataset

Linking two records within a dataset (or between two separate datasets)

Inferring any information in the dataset

The Deloitte Whitepaper "Anonymization of Sensitive Data" introduced K-Anonymity as a generalization approach for anonymizing data, which bundles individuals with similar characteristics into homogeneous groups, thereby creating a new privacy-enhanced dataset. However, K-Anonymity suffers from some deficiencies, such as the possibility of re-identification by savvy adversaries using background knowledge to infer personal information about individuals in the dataset. While techniques such as l-diversity and t-closeness propose to address some of these deficiencies, they still depend on assumptions about the adversary. They cannot protect against all possible attacks. Moreover, K-Anonymity and its related approaches reduce the uniqueness of each row in the dataset, reducing its utility as a basis for accurate analysis.

An effective anonymization solution must prevent all three of these risks. This paper examines the ability of anonymization techniques, such as differential privacy and synthetic data to mitigate privacy risks while maintaining data utility, with a focus on real-world applications in credit risk assessment, where sensitive information is often unavoidable, yet must by law be afforded certain protections.

A more sophisticated masking approach "differential privacy" adds statistical noise to datasets containing personal data. This minimizes re-identification and other privacy risks to a theoretical mathematical minimum while maintaining data utility. Differential privacy is a privacy-enhancing technique that does not make any assumptions about the adversary's access to background data. However, it can become complex, and a naïve implementation could still be vulnerable to attacks.

An alternative to anonymizing historical data is to algorithmically create representative datapoints that mimic the original distribution and relationships of the historical data without containing any identifying information. This "synthetic data" is achieved by generative models that learn the distribution of attributes within the original dataset and draw artificial samples to create new data that breaks the 1:1 relation with the original records. The process is irreversible: Unlike pseudonymization, no "key" exists by which the original identities may be recovered or "re-identified." Synthetic data has the potential to address the limitations of traditional anonymization methods, such as K-Anonymity. However, synthetic data generation has limits: ironing out wrinkles in the data, such as outliers or original biases, which can alter the characteristics of the data it aims to substitute. Additionally, validating the accuracy of synthetic data can be challenging, and inconsistencies may arise when replicating complexities from the original dataset.

In essence, the challenge with privacy-enhancing techniques is balancing data privacy and utility. Differential privacy and synthetic data that each offer different advantages and limitations. While differential privacy provides a rigorous mathematical foundation and the ability to protect against strong adversaries, synthetic data provides privacy-preserving data while maintaining high utility. Ultimately, the choice of approach will depend on the specific use case and the trade-offs between privacy and utility.

Differential Privacy

Differential privacy allows companies to share, publish, or train an AI model on private information by adding statistical noise to the data to mask the original value. It is a technique for obtaining useful information from data sets that contain personal data.

When a data source is queried, the response is first modified by adding a specified amount of noise to obscure the individual identity before it can be viewed by the analyst. Differential privacy does not ensure sensitive data will not be exposed; it only guarantees that the presence of an individual data subject will not be disclosed within the privacy risk appetite specified by the organization to balance between privacy and utility.

Differential privacy is a rigorous privacy technique with statistical guarantees of privacy. It allows for the quantification of privacy through a loss parameter noted using the Greek letter ε (epsilon value is widely known as privacy budget). Determining a suitable value for ε is critical. Generally, there exists a consensus that ε should be set to a small value, which Wood et al.[1] argue should be less than 1. However, no precise setting exists for ε, and it is highly use-case-specific. It is also a matter of perspective and motives – the individual ideally seeks perfect privacy; however, the data analyst seeks maximum accuracy. Large technology providers have promised to implement differential privacy but have so far declined to report the value of ε. This lack of transparency demonstrates the importance of this value in determinng the degree of privacy and thereby the reluctance of companies to disclose it. Anonymization is achieved for numeric and non-numeric values when the added noise complies with either the Laplace[2] or Exponential Mechanism[3] respectively. The Laplace Mechanism independently perturbs each coordinate of the output with Laplace noise scaled to the sensitivity of the function. The technique adds sufficient noise to hide the contribution of any single individual, no matter what the dataset was. The idea behind the Exponential Mechanism is to make high-quality outputs exponentially more likely at a rate that depends on the sensitivity of a quality score and the privacy parameter ε.

Two different modes exist:

- Centralized differential privacy (CDP), and
- Local differential privacy (LDP).

In centralized differential privacy, the data is stored centrally before executing the differentially private algorithm. In contrast, local differential privacy algorithms[4] execute before the data leaves the participant, i.e., when a participant enters data into a website, the client runs the differential algorithm locally, such that the data is perturbed before sending to a web server. Consequently, all sensitive data is not stored in one location. The drawback to this technique is that the total noise applied is much larger than in the central algorithm, thus affecting its utility.

Differential privacy allows companies to share, publish, or train an AI model on private information by adding statistical noise to the data to mask the original value.

[1] A. Wood et al., "Differential Privacy: A Primer for a Non-Technical Audience". In: Vanderbilt Journal of Entertainment & Technology Law 21 (2018), p. 209.
[2] Introduced by Dwork et al. (2006), the Laplace Mechanism is used for numeric attributes and adds noise drawn from a Laplace distribution to the private data.
[3] Introduced by McSherry and Talwar (2007), the Exponential Mechanism is a privacy preserving technique for also non-numeric attributes and preserves more general sets of properties by selecting the "best" element from a set while preserving differential privacy.
[4] S. Kasiviswanathan et al., "What Can We Learn Privately?" In: 49th Annual IEEE Symposium on Foundations of Computer Science, 2008, pp. 531–540.

To implement an effective privacy guarantee, the following must be considered:

1. **For whom privacy should be provided.**
   Using the illustrative example from the bank, client-level protection is required if a borrower's identity must be protected. Since a borrower can have more than one loan, this replication of personal information must be accounted for in the model. In some cases, to improve the utility of the data, the analyst might settle for only hiding certain characteristics, for instance, the default status or income.

2. **How differential privacy should be deployed.**
   Continuing with the banking example, the main advantage of local differential privacy is that each borrower adds noise on their end of the process (input) instead of centralized differential privacy. Implementing local differential privacy would be unrealistic for borrowers' personal data, as the personal data is stored within the local bank.

3. **Choosing the "right" value of ε.**
   The parameter ε must be chosen carefully to properly balance privacy with utility. Comparing the mean of the original and the differentially private data sets shows that the greater the value of ε, the smaller the difference between the mean and standard deviation between the two datasets. The standard deviation differs throughout due to the random nature of superimposed noise.

4. **Population Statistics, no specific information on individuals.**
   Differential privacy is designed to generalize statistical queries that make predictions about the population, not to infer information about individuals.

The overarching concept of differential privacy is to obscure the original identity of the data subject and its characteristics by adding noise – analogous to blurring a face or a license plate in a photograph. Aside from the risk of re-identification (or "unblurring"), a significant challenge with differential privacy is preserving the usefulness of the obscured data for analysis purposes. The two risks are diametrically opposed: adding less noise retains utility at the price of re-identification risk, which could be achieved by augmenting the "blurred" data with supplementary information.

Synthetic Data

Synthetic data is algorithmically created data that resembles and behaves like real data. Generative models learn the distribution across all attributes (i.e., maintaining the relationship between them) within the original dataset and draw artificial samples from it to develop synthetic data. The synthetic data generation method breaks the 1:1 relation between the initial and artificial records.

Synthetic data can become a crucial feature for future business development as it can be generated to meet specific needs or conditions unavailable in the original data set. This new data can be useful in cases such as:

When data protection and its resulting privacy requirements limit data availability

When data is needed for testing the robustness of a model, but such data either does not exist or is not available to the testers due to access restrictions

When data is biased or unbalanced (a word of caution: using synthetic data to "balance" a dataset could effectively introduce another form of bias in the data set.)

One major drawback of using deep learning models for synthetic data is that they might memorize features learned in the training phase. Consequently, some reproduced data may contain sensitive information, leading to privacy leaks. Experts define privacy-sensitive leakage of a model as the information that an adversary can learn about the targets from the model[5]. For instance, the attacker's objective in a banking inference attack is to infer if a particular individual data record was included in the training dataset.

Various scenarios exist where companies use synthetic data to make information available for processing when regulations or other privacy concerns restrict access to the original data. For instance, processing customer data within a GDPR regime requires adherence to strict compliance and governance rules. In such cases, synthetic data is used as a liability avoidance method that gives companies more agility and freedom to process data safely within and between institutions.

Furthermore, real-world data can be difficult or expensive to acquire. Research and innovation rely on the ability to access and analyze granular and statistically representative data, the fuel for Machine Learning (ML) models. Often, synthetic data may be easier to produce than collecting an adequate amount of original data – and it is easier to meet the regulator's expectations. It also allows the training of models on various situations that real-world data might not capture.

When determining the best method for creating synthetic data, it is essential first to consider what type of synthetic data is needed:

- **Fully synthetic data**
  This data does not contain original data; all attributes are still fully available. The risk of re-identification is low.

- **Partially synthetic data**
  Only sensitive data is replaced with synthetic data. This method requires a heavy dependency on the imputation model. This manipulation leads to decreased model dependence but does mean that some disclosure is possible owing to the true values remaining within the dataset.

Often, synthetic data may be easier to produce than collecting an adequate amount of original data – and it is easier to meet the regulator's expectations.

[5] Rigaki and Garcia (2021): A Survey of Privacy Attacks in Machine Learning.

Experiments

To show the potential of generating innocuous data, the following experiments investigate the general properties and performance of different state-of-the-art generative models and techniques in a realistic situation. Five methods face off against one another:

- Generative Pre-trained Transformer (GPT)
- Generative Adversarial Network (GAN)
- Wasserstein Generative Adversarial Network (WGAN)
- Conditional Tabular Generative Adversarial Network (CTGAN)
- A baseline "perturb, shuffle, sample with replacement" technique (implemented in the aiStudio tool "De-Identify") using different epsilon values (3.67, 9.25, and 12.2)

All models are trained on the "Give me some credit"[6] dataset to generate representative synthetic data. The primary objective of the exercise is to augment the representation of infrequent defaults in the data, balancing out the dataset to increase the accuracy of low default portfolios.

In addition to the implementation of the fifth synthetic data technique, the assessment function of De-Identify was used to subsequently evaluate all methods against numerous metrics in order to assess the quality of the resultant synthetic data. In this case, the model essentially re-created the original dataset (overfitting), thus failing to address privacy concerns.

Visual, statistical, and model-based tests consistently confirm the conclusion. Visual tests, focused on the distribution and column-wise correlations, conveniently expose patterns between original and synthetic data. Principal Component Analysis (PCA) reveals that the synthetic data is a near replica of the original set. PCA is particularly telling in that it does not discard any samples or characteristics of the attributes. It achieves this by reducing the many features into a few comprehensive dimensions, or "principal components," representing the dataset. These describe the varied influences, or "loadings" of the original characteristics which provide convenient markers to isolate effects producing differences among clusters.

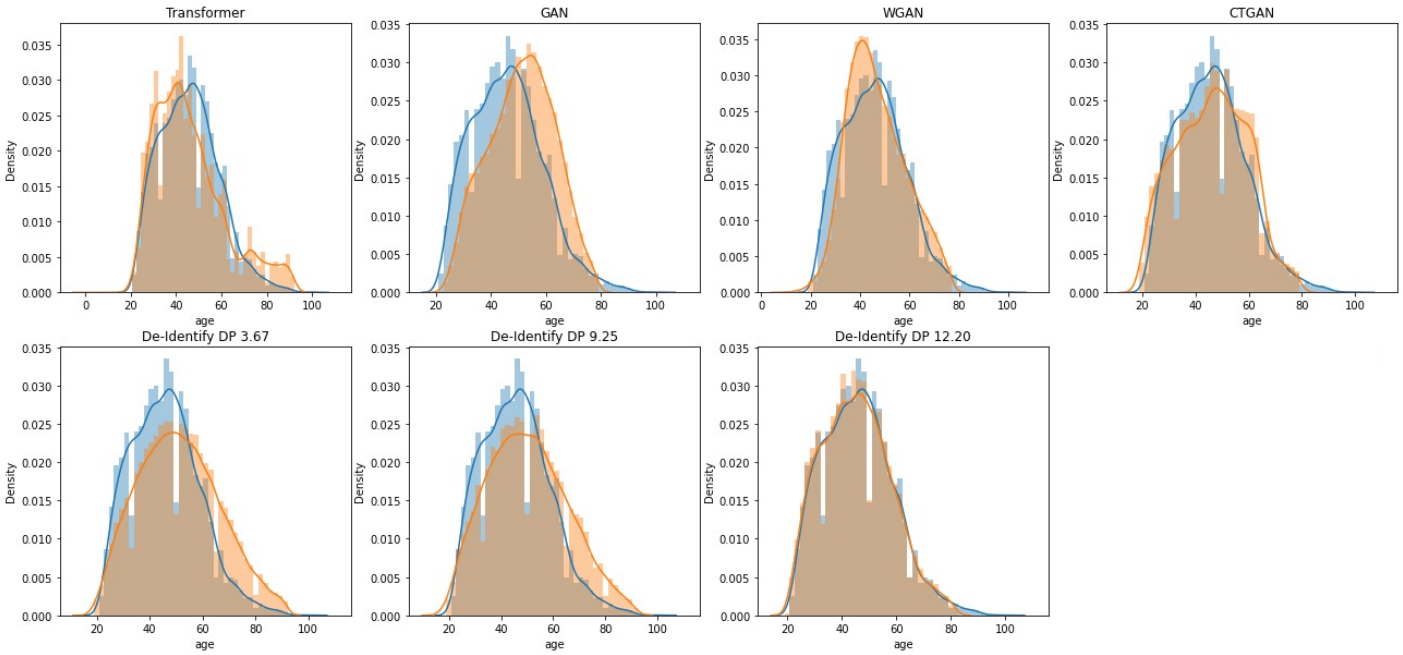
Two statistical tests compare synthetic vs original datasets from multiple viewpoints. The Jensen-Shannon (JS) Divergence Test measures the distance between two data distributions. JS divergence is applied to each feature independently; it is not designed as a covariant feature measurement but rather a metric that shows how each feature has diverged independently from the baseline values. A higher value of JS divergence indicates greater dissimilarity between distributions, while a value closer to zero indicates greater similarity. The two-sample Kolmogorov-Smirnov (KS) test reveals whether two samples originate from a population with the same distribution, the null hypothesis. A logistic regression considers whether classifier model performance changes when replacing the real with synthetic data.

[6] Kaggle Dataset (https://www.kaggle.com/c/GiveMeSomeCredit) downloaded on 12.09.2022

Results

Visual analysis of chosen indicative features – one discrete and one continuous feature – summarizes the behavior of the synthetic data. Figure 1 shows the distribution plot for the discrete feature, the age of all lenders. All models capture the distribution of the discrete feature age with a slight degree of skew and deviant modes. It also shows that the differential privacy approach outperforms the deep-learning approaches.

Fig. 1 – Distribution of discrete attribute (Age)



Source: the re-identification risk tool "De-Identify" from the Deloitte aiStudio

Tab. 1 – Jensen-Shannon Divergence and Kolmogorov-Smirnov-Test for the discrete attribute (Age)

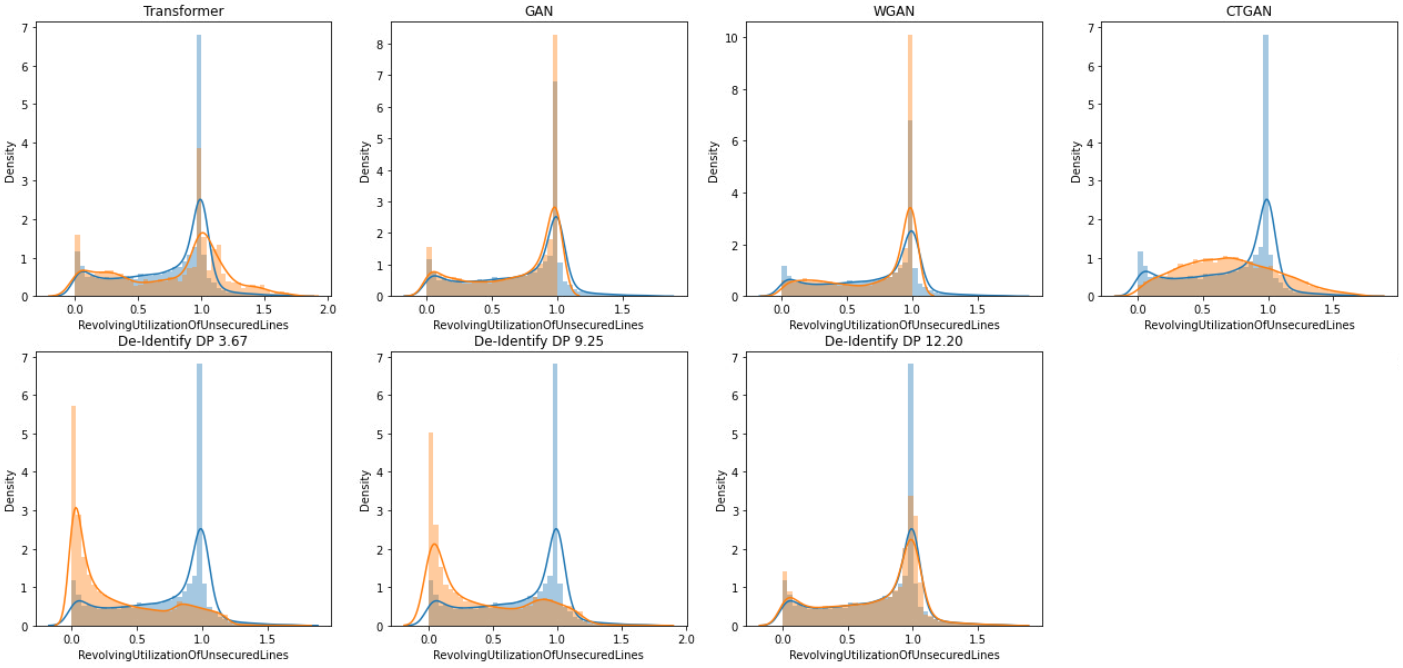
Model	JS Divergence	p-value for KS-Test
Transformer	0.43	0.00
GAN	0.02	0.00
WGAN	0.02	0.00
CTGAN	**0.02**	**3.83**
De-Identify DP 3.67	0.02	0.00
De-Identify DP 9.25	0.02	0.00
De-Identify DP 12.20	0.02	0.36

Table 1 shows the results of the two statistical tests for the discrete attribute (Age). The conclusion of the combined JS Divergence and KS-Tests is that only CTGAN and the differential privacy approach with the highest epsilon can adequately model the real data, owing to the high degree of skew and deviations in the mode and tails of the distribution within the generated data compared to the original dataset.

Figure 2 illustrates the corresponding diagrams for the continuous feature, in this case, the balance on credit cards and personal lines of credit except real estate and installment debt like car loans divided by the sum of credit lines.

All models capture the individual distribution for the continuous attribute except the differential privacy approaches with low epsilon and the CTGAN.

Fig. 2 – Distribution of continuous attribute (Revolving Utilization of Unsecured Lines)



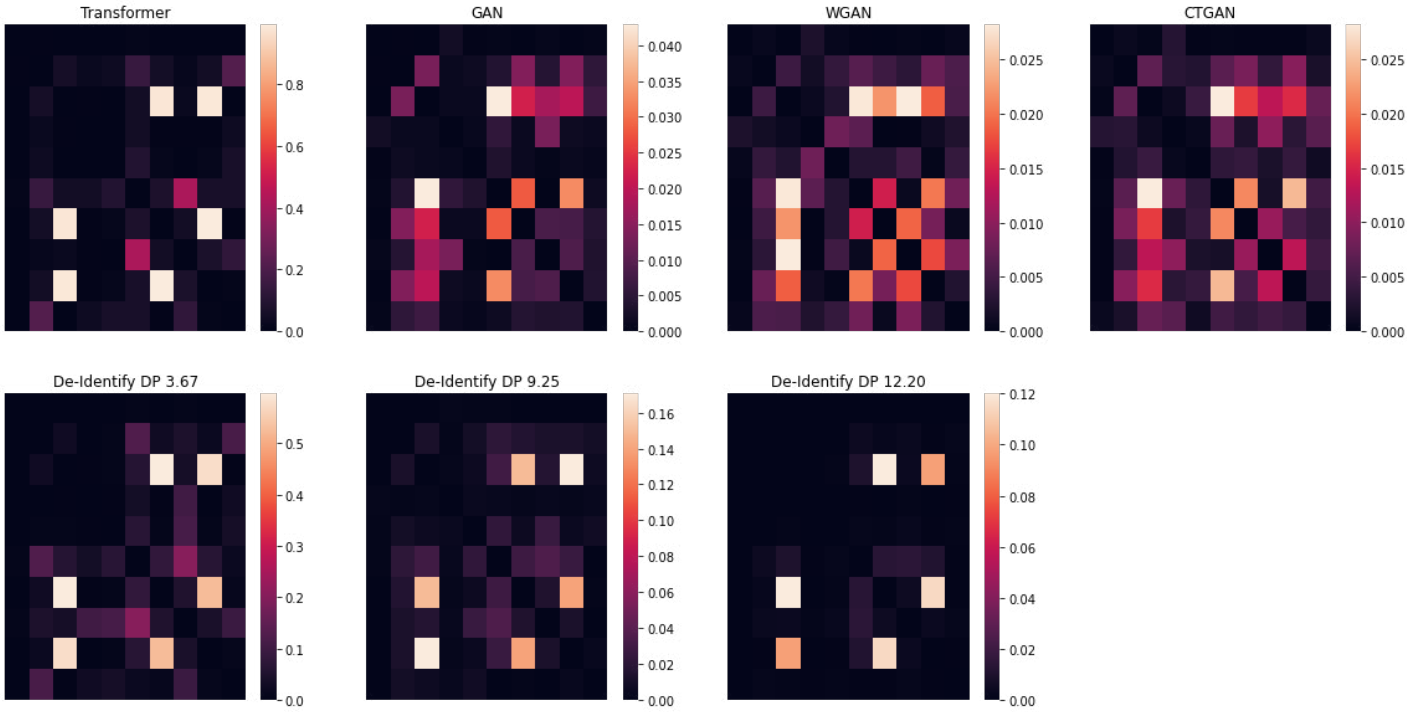
Source: the re-identification risk tool "De-Identify" from the Deloitte aiStudio

Tab. 2 – Jensen-Shannon Divergence and Kolmogorov-Smirnov-Test for the continuous attribute (Revolving Utilization Of Unsecured Lines)

Model	JS Divergence	p-value for KS-Test
Transformer	0.52	0.00
GAN	0.49	0.00
WGAN	0.49	0.00
CTGAN	0.49	0.00
De-Identify DP 3.67	0.65	0.00
De-Identify DP 9.25	0.64	0.00
De-Identify DP 12.20	0.59	0.00

Table 2 shows the results of the two statistical tests for the discrete attribute (Age). The conclusion of the combined JS Divergence and KS-Tests is that **no model** captures the distribution of the real data. The rather high JS Divergence firmly rejects the null hypothesis that no distribution of the synthetic data is identical to the original. It arrives at a similar conclusion to the analysis of the discrete attribute: a high degree of skew, deviations in the mode and tails of the generated data distribution compared to the original dataset.

Fig. 3 – Differences of Correlation between the Original and Synthetic Data



Source: the re-identification risk tool "De-Identify" from the Deloitte aiStudio

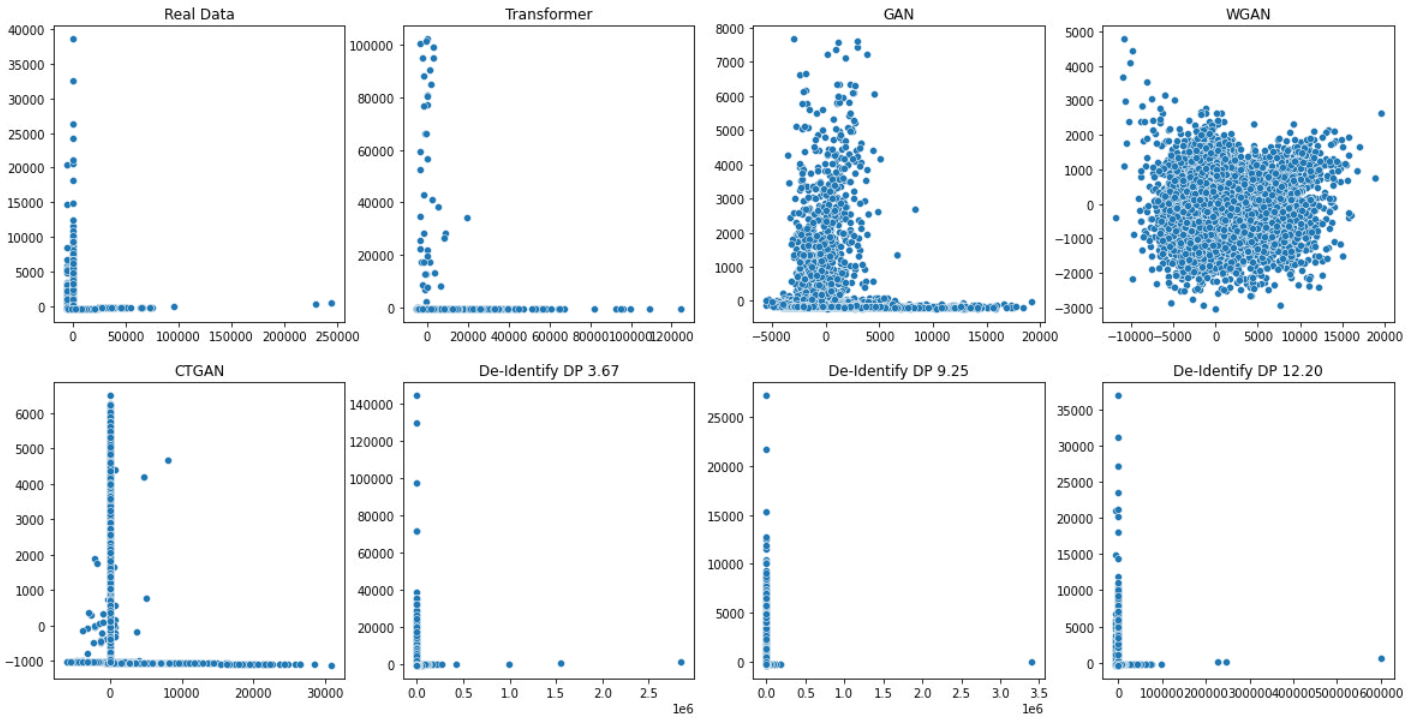
Figure 3 illustrates the preservation of the relationship between the attributes of the synthetic data compared to the original data. It evaluates the column-wise correlation between every pair of fields by calculating the average absolute difference between these values across all fields. The heatmap (fig. 3) depicts the differences between the training data and the synthetic data: the higher the epsilon chosen, the closer the distribution of the synthetic distribution to the original gets.

Visual inspection indicates that all the models can adequately capture the correlations between features. The Transformer model and all differential private approaches capture the correlation of the original dataset well. However, other models also preserve the correlation with some added noise. In other words, the Transformer generates a high utility and low privacy dataset. In contrast, the other models manage a dataset with a moderate-to-high utility and moderate privacy.

Often, synthetic data may be easier to produce than collecting an adequate amount of original data – and it is easier to meet the regulator's expectations.

Again, Principal Component Analysis verifies the statistical integrity of the multi-dimensional datasets, applied first to the original data and then again to the synthetic data in figure 4. The similarity between the synthetic data generated by CTGAN and the Transformer is immediately apparent.

Fig. 4 – PCA Analysis



Source: the re-identification risk tool "De-Identify" from the Deloitte aiStudio

Tab. 3 – Accuracy of Logistic Regression Model Using Datasets Generated by Different Models

Model	AUC
Real Data	0.82
Transformer	0.78
GAN	0.77
WGAN	0.76
CTGAN	0.70
DP 3.67	0.58
DP 9.25	0.80
DP 12.20	0.81

Training challenger logistic regression models from the various synthetic datasets and comparing them to a model trained on the original data demonstrates their relative effectiveness at the classification task, and thereby the utility of each synthetic data approach. Performance is evaluated against the Area-Under-the-Curve (AUC) metric, compared side-by-side in Table 3.

The logistic regression trained on the synthetic dataset (higher epsilon) performed as well as the original data. This is unsurprising, as the high epsilon implies a low degree of privacy, implicitly explaining the high fidelity vis-à-vis the model trained on the original data. Where performance is defined through fidelity and privacy, the slightly lower accuracy on the Transformer is an acceptable trade-off, followed by the GAN approach. In cases where privacy is not critical, the differential privacy approach with a moderate epsilon would be a reasonable choice, as it retains a high discriminatory power comparable to the original model despite having a less accurate continuous attribute.

Interpretation

Synthetic data can achieve a higher fidelity compared to models trained on the original dataset. The differences come where the discriminatory power of the classifier and the degree of privacy are included in the performance metric.

Applying deep learning to generate synthetic data retains high data utility and a reasonable level of privacy, suggesting this approach is appropriate for publishing data while preserving privacy. Slight differences in approach and parameter settings can shift the balance between utility and privacy in either direction.

If privacy is the main concern, then generating samples from a Transformer model holds the most promise, creating synthetic data with discrete and continuous attributes, little noise, and adequately preserving correlations for classification tasks.

Re-Identification Risk

Synthetic data and differential privacy reduce the risk that personal or sensitive information be re-derived from anonymized data to a theoretical mathematical minimum. The resultant synthetic data is resilient against identity theft attacks. Differential privacy uses noise to mask the presence of any particular individual in the input data. While it has many advantages over K-Anonymity, its protection is not perfect, although no method is 100% effective.

Even entirely synthetic data could still reveal the identity of individuals within its training set when combined with supplemental information or through sophisticated re-identification techniques, such as Pattern Recognition and Behavioral Analysis. It is important to note that the choice of measurement method depends on the specific characteristics of the dataset. Size: a large dataset will require more computationally efficient methods than smaller datasets. Content: the presence and nature of quasi-identifiers require more thorough masking. Goals: privacy requirements of some applications may be more stringent than others. Situational utility: subsequent operations, such as forecasting or advanced analysis, introduce additional dependencies, such as the correlation between subsequent datapoints. A sensitive dataset (employee data) illustrates the trade-off between privacy protection and remaining utility after applying various privacy techniques.

The formal properties of K-Anonymity, L-Diversity and T-Closeness demonstrate the degree to which privacy has been preserved. By adding noise into the data combined with standard methods such as suppression, pseudonymization & generalization, the anonymized dataset shows a strong K-Anonymity, average L-Diversity and weak T-Closeness (fig. 5), specifically:

- **K-Anonymity**
  Each record in the dataset is indistinguishable from at least k-1 other records based on attributes such as Resources, Start Date and Offering. For this scenario only non-anonymized quasi-identifiers have been used to calculate the degree.

- **L-Diversity**
  The diversity of sensitive attributes are considered within each equivalence class of the dataset based on the selected quasi-identifiers. The sensitive attributes in this dataset are skillcategory and skillname of the employees.

- **T-Closeness**
  The distribution of sensitive attributes within each equivalence class differs from the overall distribution in the dataset.

Fig. 5 – Re-identification Risk Assessment

Source: the re-identification risk tool "De-Identify" from the Deloitte aiStudio

Implications

- Re-identification risk is generally lower with strong K-Anonymity because it obscures the specific individual within a group of indistinguishable records.

- L-Diversity extends K-Anonymity and ensures that there is a moderate level of diversity in terms of sensitive attributes. While L-Diversity adds another layer of protection, the effectiveness depends on the actual diversity within each equivalence class. As the dataset has limited variations in each attribute and utility wants to be preserved, average L-Diversity can be seen as sufficient.

- The weak T-Closeness indicates that the distributions between the equivalence classes are not similar, potentially allowing for information leakage. This could introduce a higher risk of sensitive attribute disclosure, potentially increasing the risk of re-identification in certain scenarios. In figure 7, the distance is calculated by comparing the distributions of different groups within the dataset and the distribution to the whole dataset for each sensitive attribute. Here, the distributions of the skill-category and skillname in particular differ significantly. Improving slightly the degree of T-Closeness would already lead to high information loss, therefore it can be seen as a trade-off for preserving information, and the weak T-Closeness can be accepted.

One straightforward method to evaluate the preservation of utility in the synthetic data is to observe how relationship between attributes may have changed. The first step is to calculate the correlation between features within its respective dataset (original or synthetic), depicted as a heatmap (fig. 6: tables A and B). The second step subtracts one heatmap from the other to reveal any deviations, the stronger of which would indicate a loss in utility. The comparison heatmap shows only marginal deviation between original and synthetic feature correlations, which indicates utility has largely been preserved within the synthetic data.

Fig. 6 – Correlation Difference = Subtraction of Heatmap Tables B from A

Table A: Original Data

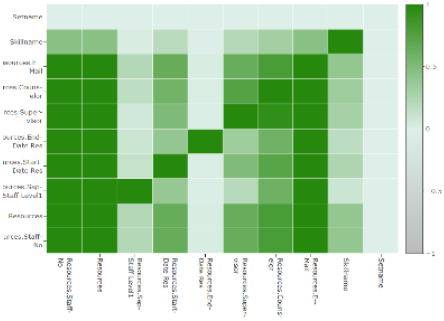
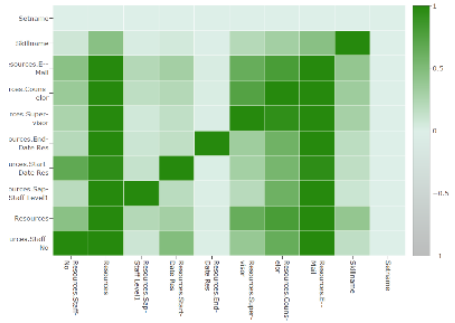
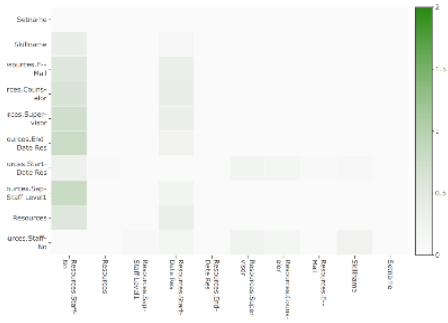


Table B: Synthetic Data

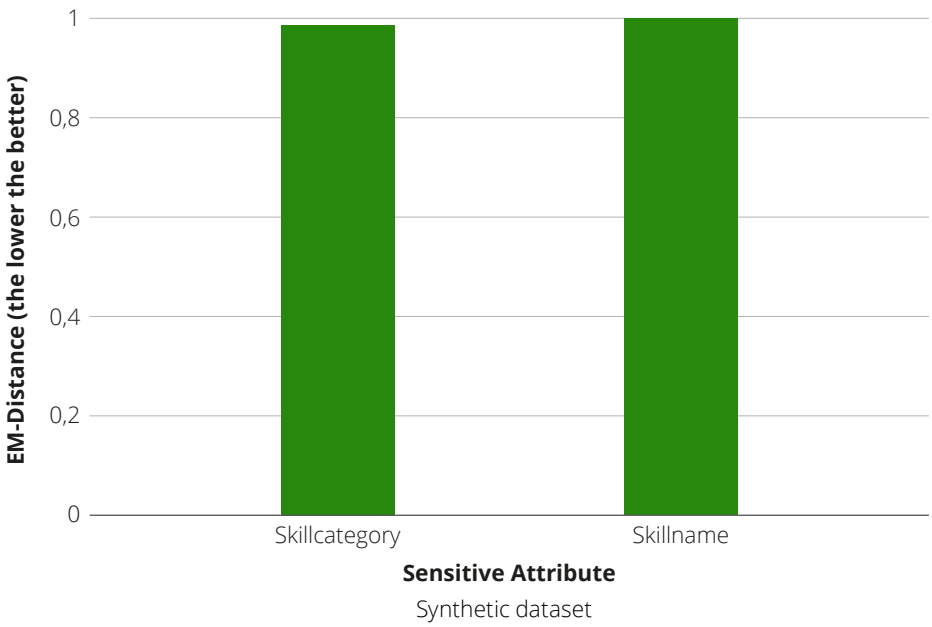


Comparison: Tables B - A



Source: the re-identification risk tool "De-Identify" from the Deloitte aiStudio

Fig. 7 – Distance of dstribution of the sensitive attributes between the original and the synthetic dataset



Source: the re-identification risk tool "De-Identify" from the Deloitte aiStudio

Overall, it appears that a fair compromise has been found between utility and privacy. K-Anonymity is bolstered by anonymizing all personal and quasi-identifiers, making it difficult to isolate individuals. More aggressive anonymization might increase T-Closeness, but at the expense of utility by reducing the correlation between anonymized and original datasets.

It is nevertheless important to note that achieving a balance between these privacy metrics requires careful consideration of context, notably the nature and sensitivity of the anonymized information, and potential adversary knowledge. Additionally, continuous monitoring and updates to the anonymization techniques may be necessary to address emerging re-identification risks. While strong K-Anonymity provides a good baseline protection, the effectiveness of L-Diversity and T-Closeness depends on the specific characteristics of the data and the nature of the sensitive attributes.

Conclusion

The experiments demonstrate that the Transformer model and the differential privacy approach with the moderate epsilon (ε) perform the best with regard to capturing the distribution of discrete and continuous attributes while preserving the utility of the data (correlation as well as the performance of the logistic regression trained on the generated data). It is important to emphasize that no single method applied in the experiments outperformed the others in consideration of all the metrics.

Despite its many advantages, differential privacy may not be suitable in all cases. Loss in accuracy of models trained on anonymized data is inevitable and therefore important for practitioners to understand well before application. Yet its rigorous mathematical foundation and ability to protect against severe adversaries make a strong case for its application to many (ML) modeling situations where privacy is a prime concern.

Through the lens of the credit decision classifier model case study, this paper highlights the following challenges when dealing with the generation of synthetic data:

- **Missing outliers**
  Synthetic data can, at best, imitate real-world data. It may not contain outliers that also characterize the original data. This omission may present a significant limitation. In some situations, outliers may even be more critical than data points that remain within expected intervals, evoking the musings of Nassim Nicholas Taleb[7], "the inability to predict outliers implies the inability to predict the course of history."

- **Quality of the original data**
  The quality of synthetic data is correlated with the quality of the input data and the data generation model. Synthetic data may also reflect the biases in source data.

- **Quality assurance**
  Especially in complex datasets, comparing synthetic data with known accurate or human-annotated data is an essential control step. There could be inconsistencies in synthetic data when trying to replicate complexities within original datasets.

- **User acceptance**
  Synthetic data is an emerging concept that may not be accepted as valid by users – here, appropriate change management is strongly needed.

Differential privacy in general and synthetic data in particular provide more robust protection against re-identification than traditional masking methods through a well-balanced compromise between utility and privacy for data processing. Synthetic data is a promising technology, with a wide variety of applications. Not only can it help fill gaps in situations of data sparsity, it can also help avoid the substantial costs for breaches of privacy or related fairness considerations. With "built-in anonymity", synthetic data unshackles developers of AI models, allowing them to use the data freely and focus on model performance without fear of infringing on data privacy. Nevertheless, caution is advised in generating synthetic data, striking the right balance between utility and avoiding the risk of re-identification. Deloitte's De-Identify can both generate synthetic data and, importantly, evaluate the residual re-identification risk through application of leading privacy metrics. This facilitates the iterative process of balancing between privacy and utility, the optimium of which will depend on the particular data and its application. Its masking and scrambling functions on categorical data allow stress-testing models for bias, i.e., validating whether the same results can be obtained by changing the values of attributes such as race or gender.

The field of differential privacy and synthetic data is dynamic, with new methods and tools periodically arriving on the scene. For example, Generative AI capabilities increase the potential to create viable synthetic data, positioning it to become a powerful means to train targeted Machine Learning models. Assessment methods are also evolving. Beyond those evaluated in this paper, it is always advisable to scan for the emerging privacy-preserving techniques, but also not to blindly trust them without performing due diligence. Experiments such as those discussed in this paper can provide an example for how to evaluate upcoming techniques.

[7] Taleb, Nassim Nicholas, The Black Swan: the Impact of the Highly Improbable. New York: Random House, 2007.

Contacts

David Thogmartin
Director
Risk Advisory
dthogmartin@deloitte.de

Patrick Spitzer
Manager
Risk Advisory
pspitzer@deloitte.de

Richard Leaton
Senior Manager
Technology Strategy
rleaton@deloitte.com

Akilavan Ganeshkumar
Consultant
Risk Advisory
akganeshkumar@deloitte.de

Deloitte.

Issue 04/2024