Deep Synthesis and De-Identification of Large Data Sets: A Comparative Analysis

Eric White, Justin Whitlock, Andrew Turscak*, Paul Terwilliger, Daniel Miller, Mark Moreno*

Commander, Navy Reserve Forces Command

Norfolk, VA

***Corresponding Authors: Mark.A.Moreno1@navy.mil, Turscak_Andrew@bah.com**

ABSTRACT

Embracing the latest innovations in data analytics has become critical to compete in both the private and public sectors. Unfortunately, the data which yields the most meaningful results is often the most sensitive, introducing a potential risk if proper security measures are not taken. Moreover, this data is aggregated in large sums. With the increased use of sensitive data, particularly in machine learning applications, advanced data augmentation techniques introduce a means by which advanced analytics can be practiced securely and effectively in sensitive domains.

In this paper, we present two methods –Pseudonymization and Generative Adversarial Networks (GANs)– to de-identify data and protect the privacy of entities in data at rest on sensitive IT systems for secure use outside those systems. The former method is often used to test and optimize code by providing realistic values at the aggregate feature level. The latter can be used to train machine learning models and perform statistical analyses by learning the underlying distribution of specific observations in a manner that does not compromise the original records. The GANs in this study leverage Convolutional Neural Networks (CNNs) in the traditional arrangement of a generator and a discriminator, with a third CNN enforcing the syntactical relationship between features. We test our performance with a statistical evaluation of the underlying distributions of the synthetic features against the original feature vectors from which they were generated, visualizing high-dimensional relationships between data sets, and comparing supervised cross-validation scores on the synthetic data to those of models trained on the real data. Results showed strong statistical relationships between real and synthetic features but variable model performance across datasets.

ABOUT THE AUTHORS

Mr. Eric White (Booz Allen Hamilton) is a Data Engineer, providing analytical support and managing web presence. He specializes in building and optimizing scalable ETL pipelines from unstructured data and disaggregate systems. He holds a M.A. in Economics and Business Modelling and Simulation as well as a B.S. in International Business, both from Old Dominion University.

Mr. Justin Whitlock (Booz Allen Hamilton) is a Data Scientist. Professional interests include signal processing and analysis, natural language processing, modeling, simulation and linguistics. He holds a M.S. in Computer Science from Old Dominion University, a B.A. in International Studies from University of Nebraska at Omaha, and an A.A. in Arabic Language & Middle Eastern Studies from Defense Language Institute at Presidio of Monterey.

Mr. Andrew Turscak (Booz Allen Hamilton) is a Data Scientist, responsible for analytical support, model deployment, and R&D. Professional interests include natural language processing, network analytics, game theory, and simulation. He holds a M.S. in Computational Operations Research and B.A. in Economics, both from the College of William & Mary.

Mr. Paul Terwilliger (Booz Allen Hamilton) is a Data Scientist. He enjoys reading deep learning research papers and applying novel machine learning techniques to data science problems. He is a competitive chess player and builds statistical models for Chess.com in his spare time. He holds a B.A. in physics from the University of Pennsylvania.

Mr. Daniel Miller (Booz Allen Hamilton) is a project lead helping to empower clients to make data-driven decisions through the establishment of a data science capability. Professional interests include applied economics and ensemble learning. He holds a M.S. in Business Analytics at Indiana University and a B.A. in Economics and Business from the Virginia Military Institute.

Mr. Mark Moreno (CNRFC) is Deputy Director of Navy Reserve Force Analytics, code N36 responsible for establishing and growing the data analytics capability for the Navy Reserve Force and providing subject matter expertise and continuity for those efforts. He serves as a conduit between the contractors and military staff across all areas of the data analytics capability for each project. He holds a B.S. in Medical Technology from Univ. of WI Madison, retired from the Navy as a Submarine Full Time Support Officer, and has performed as a Program/Data Analyst at CNRFC since 2011.

Deep Synthesis and De-Identification of Large Data Sets: A Comparative Analysis

Eric White, Justin Whitlock, Andrew Turscak*, Paul Terwilliger, Daniel Miller, Mark Moreno*

Commander, Navy Reserve Forces Command

Norfolk, VA

***Corresponding Authors: Mark.A.Moreno1@navy.mil, Turscak_Andrew@bah.com**

INTRODUCTION

Problem Statement

Embracing the latest innovations in data analytics has become critical to compete in both the private and public sectors. In the DoD, failing to harness analytics causes the government to trail industry and makes the military vulnerable to technologically adept adversaries. Unfortunately, the data which yields the most meaningful results is often the most sensitive, introducing potential risk if proper security measures are not taken. The security risks of failing to use advanced analytics and failing to observe best security practices are equally unacceptable. With both increased use and increased protective measures necessary with sensitive data, other advanced techniques must be employed to ensure these methods can be practiced securely and effectively.

At Commander, Navy Reserve Forces Command, strict procedures govern where data can reside at rest based on the sensitivity of the information. When the proper technology is available on approved systems, this does not present any roadblocks to practicing secure advanced analytics. In practice, however, the latest technology is not always available, stunting projects that require the use of sensitive data. Furthermore, as attacks on data systems become more sophisticated, so must measures taken to protect the data. Fortunately, there are ways that advanced analytics itself can be harnessed to introduce these additional protections.

Traditional methods to protect data may include removing sensitive identifiers (e.g. PII) before randomly scrambling, sampling, offsetting, binning, or generalizing the original variables one-by-one. Whereas these methods have varying levels of security, all sacrifice aspects of the data that are necessary for critical portions of most analytic pipelines. Scrambled features, for example, will maintain the size and individual characteristics of descriptive factors that may be used for feature engineering, visualization, and code testing and optimization. They will not, however, preserve statistical integrity between records. This means that data subset on any scrambled feature value may lose characteristics that would become prevalent under that filter, and machine learning and statistical testing will likely produce invalid results. What's more, randomly replacing or removing sensitive identifiers altogether eliminates important information associated with multiple records in the data.

Approach

This paper explores Pseudonymization and a Generative Adversarial Network (GAN) as methods to generate de-identified datasets that are suitable for analysis while maintaining privacy of the individuals from the original dataset. Pseudonymization is easier to deploy, well-researched, and may be used for code architecture, optimization, and visualization purposes where only the distribution and structure of the individual variables matter. The specific method used in this study includes hashing and salting direct identifiers before replacing all associated features with randomly generated values. GANs may be employed where not only feature distributions, but consistency across records is necessary. It is also sometimes used when seeking model lift from synthetic data. For the GAN in this study, a generative network is used to produce synthetic data which is then evaluated for authenticity by a discriminator and classifier trained on a sensitive dataset.

Both the Pseudonymization and GAN were deployed on a Navy Reserve dataset derived from the one used in the paper "Feature Engineering and Ensemble Machine Learning in the Navy Reserve: Using Holistic Behavioral Profiles to Predict Mobilization Cancellation" (Milletich et al., 2019). The descriptive variables in the resulting pseudonymous and synthetic datasets were evaluated using the Anderson-Darling test to examine whether each synthetic feature maintained the statistical profile of the raw feature from which it was generated. Integrity between observations and the viability of the generated data for statistics and machine learning were tested in a series of machine learning experiments. Each experiment compared output from either a supervised or manifold learning algorithm trained on some subset of real data to that of the same model trained on synthetic data generated from that subset. Results show that while all individual features maintained their statistical profile in the synthetic data across experiments, output from models trained on synthetic data did not match that of models trained on original data. Each synthetic dataset did, however, possess useful properties that, when considered as a collective, satisfy most modeling needs without the necessity of raw data. Additionally, further tests on the synthetic features suggest that the GAN has a tendency to overfit by epoch, which may possibly be corrected with minimal adjustments to the model.

LITERATURE REVIEW

Many organizations use large datasets containing personally identifiable information (PII) which needs to be removed or obscured to protect an individual's privacy. A simple method of removing PII is to suppress features such as names, birthdates, or identification numbers by either removing them from the dataset or grouping them in a way that obscures individuals (Kelly, et al, 1992). The goal of these methods is to achieve k-anonymity, where individuals are obscured amongst at least k individuals, making reidentification more difficult. One drawback to suppression is the loss of usable data and the reduction of detail, leading to less precise predictions. Also, depending on the size of the population in the dataset, grouping attributes can still lead to identification of individuals or reveal information about an individual that would be considered private (Zayatz, et al, 2009). If the dataset lists the average salary of plumbers in a small town, and k-anonymity has been achieved, but every plumber works for the same company, then private financial information about that company has been inadvertently revealed causing information leakage. Information leakage is a serious problem, and it is a growing concern as the amount of data collected by companies and government agencies increases.

The need for highly granular data combined with the need for security has led to improved methods of obscuring data points through the addition of noise, adding offsets, or swapping values (Giessing, 2004, June). These methods of adding static to a data set may effectively obscure sensitive information, but security is exchanged for usefulness (Lubarsky, 2010). Finding the balance between security and utility can prove especially difficult for complex data types such as dates or locations (Garfinkel, 2015). The perturbation required to mitigate re-identification attacks can obscure the data and lead researchers to draw false conclusions. Pseudonymization replaces values with randomly generated pseudonyms, obscuring datapoints without suppression or loss of granularity (Corrales Compagnucci, 2019). The possibility of reidentification through reverse engineering the pseudonymization algorithm can be further complicated by using a secret key. However, reidentification is still possible when the dataset is compared to information which intersects the data at some point through non-direct identifiers (Lubarsky, 2010). The vulnerability of de-identified datasets has been proven again and again. Datapoints obscured using pseudonymization and other methods still represent personal information. This one-to-one relationship can be exploited to discover individuals, or even reverse engineer the algorithm to reveal the entire dataset (Lubarsky, 2010, Rocher, 2019). One concept that has been proposed is using synthetic data to substitute the original data, but until recently there has not been an effective method for generating synthetic data (Hermes, et al, 2012).

In the short time since Goodfellow et.al. introduced GANs in 2014, generative adversarial networks have been put to work in a variety of applications. The most recognizable of these includes advances in image generation, where GANs are used to create synthetic faces (Radford, et al, 2015) or create artwork that mimics the style of an artist or period (Elgammal, et al, 2017). A recent paper from Uber proposes using GANs to generate synthetic training images to accelerate learning for AI agents (Such, et al, 2019). The synthesized datasets clearly defined boundaries between classes, leading to faster and more accurate training of models. Synthetic data has also been used to augment real image sets to improve models where the size of a dataset was insufficient (Frid-Adar, et al, 2018). The use of GANs for creating completely synthetic images has implications for other data types, including tabular data. A compelling aspect of synthetic data is that by its very nature, it is secure against re-identification attacks (Park, et al, 2018). Unlike

pseudonymization methods, there is not an exploitable one-to-one relationship between synthetic and original datapoints that could lead to information leakage.

Applying GANs to the problem of PII in tabular data does come with its own issues. One of the strong points of machine learning is that underlying associations between features can be found and exploited. For the synthetic data to be useful for future modeling, the distributions and relationships between features need to be represented. In one approach, the authors use Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNN) with attention to maintain these relationships (Xu, Veeramachaneni, 2018). The resulting datasets outperformed conventional methods for anonymizing data and was robust to scaling. A second approach augments DCGAN (Radford, et al, 2015), an algorithm designed for generating images using Convolutional Neural Networks (CNN), for generating tabular data by adding a third CNN to act as a quality control (Park, et al, 2018). The purpose of the additional CNN is to maintain syntactic relations between features by classifying each synthetic data point and comparing the predicted label to the synthetic label. The authors found that the individual features were statistically similar to the original data, and that models built on the synthetic data performed better than models built using k-anonymous or perturbed datasets. One drawback to using this second method is a dependency on labeled data, which is contrary to the unsupervised learning aspect of most GAN applications.

ANALYTIC APPROACH

Data

Data were obtained from a larger sample of the Navy Reserve dataset used in the paper "Feature Engineering and Ensemble Machine Learning in the Navy Reserve: Using Holistic Behavioral Profiles to Predict Mobilization Cancellation" (Milletich et al., 2019). This dataset was an appropriate choice for its cleanliness, familiarity, and the substantial modeling and code base we have already built with it. Chiefly, however, it was chosen for representing the exact type of dataset that brings with it the concerns our methodology seeks to address in that records are associated with individual sailors whose information needs to remain protected. Data was collected and hashed using the SHA-256 cryptographic hashing algorithm to completely secure unique identifiers. In an ideal state, the entire process of data synthesis will be migrated to NMCI assets to eliminate dependency on secure standalone assets altogether.

The dataset consists of 10 years of career, demographic, and behavioral data from the full Navy Reserve population. The data was retrieved from the Navy Reserve Data Warehouse (NRDW), where updates to any feature value are stored daily. Our dataset captures these transactions by using a multi-delimited storage schema that includes each data value with the time stamp of its last update. Observations are labeled with a 0 or a 1 denoting whether a Reservist mobilized or canceled their mobilization within the window of interest. Of the observations in the complete mobilization dataset from which each individual experiment is derived, 11.8% are cancellations. If a SELRES did not have a mobilization in the window of a specific experiment, they were excluded from that experiment. Values for individual variables were feature engineered to capture factor values, numeric values, and temporal information as necessary to build a holistic profile of SELRES behavior in the Navy Reserve. Values were truncated prior to the mobilization event of interest so-as to eliminate data leakage. A full description of the data can be found in Milletich et al.

Pseudonymous Data

Pseudonymization is a standard approach used to de-identify data when the privacy of the entities needs to be protected, but the true values and underlying relationships of the features are not required. This technique replaces specified values with a randomly generated collection of characters meant to mimic the format of the original values allowing practitioners to work with pseudonymous data in a similar manner to the original data. The original values are stored in a table to ensure consistent mapping across all occurrences within the dataset. Because the generated pseudonyms are random, there is no mathematical relationship between the original and generated values. This enhances the privacy of the individuals by mitigating re-identification without access to additional information or the table. The consistent mapping of pseudonymous data allows for easy analysis of discrete data but may cause issues when analyzing ordinal or continuous data unless additional parameters are employed. Due to the urgent need

for protected data at CNRFC and because this approach is widely implemented, we began with a browser deployment of modified Pseudonymization technique as a proof of concept.

Prior to running the algorithm, direct identifiers are hashed using the SHA-256 algorithm and salted with a 64-bit secret key. Deployed in JavaScript, the module then iterates through the remaining values by character. Alphanumeric characters are concatenated until a non-alphanumeric character or the end of the cell is reached. The terminal value is then checked against a dictionary of previously read values. If the value is not found, the algorithm identifies the data type as numeric (integers, floats, dates) or non-numeric. A new pseudonym is generated by replacing each character with a randomly selected character of the same data type, that is, numbers for numbers within the range of the data and letters for letters. Additional parameters such as minimum/maximum values for dates can be enforced to prevent nonsensical values and enable more advanced analysis.

Synthetic Data (Previously GAN Methodology)

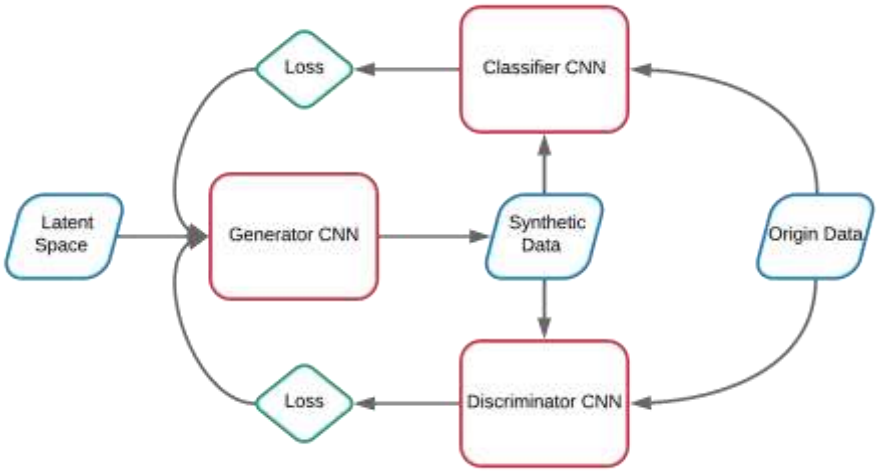

Figure 1. GAN architecture

Our GAN architecture is based on DCGAN with added classifier CNN employed by (Park, et al, 2018). The architecture is dependent upon a labeled dataset, which does limit its application, but was not a hinderance for the Naval Mobility dataset which has labels. In addition to the typical GAN architecture, which uses the discriminator loss to train the generator, the classifier loss is also a factor in training the generator. The classifier is trained on the origin data and makes predictions on synthetic data. The loss function for the classifier is calculated by comparing the predicted label for a datapoint to the synthetic label output from the generator. For a more detailed explanation of the implementation, please refer to their paper (Park, et al, 2018).

EXPERIMENTAL RESULTS

Synthetic datasets generated by each experiment were tested for statistical consistency between the original and synthetic data at the feature, label, and individual record levels. Random normally distributed data was also tested as a null baseline means of evaluation. At the feature and label levels, statistical consistency is important for feature engineering, data visualization, and any associated numeric or text operations on the synthetic data. This is because it enables value-agnostic coding operations on synthetic information to be applied to the real data with minimal modification and because it allows basic exploration of intra-feature relationships. At the individual record level, maintaining distributions across observations and labels is necessary to preserve enough signal between the synthetic features and synthetic labels to train a statistical or machine learning model.

Statistical similarity between features was computed using an Anderson-Darling for k-samples test. Anderson-Darling is a non-parametric test of whether the samples are drawn from the same population and is appropriate when testing across several samples of data. In our case, for each synthetic dataset, we tested every encrypted feature against its

original counterpart. Table 1 reports the p-values for the Anderson-Darling test across synthetic datasets generated from both experiments.

	mean p-value	range
Pseudonymous Data	0.001	[0.001, 0.001]
GAN Synthetic Data	0.025	[0.016, 0.034]

Table 1. Anderson-Darling p-value ranges for all variables

Table 1 shows that both models produced statistically similar features as those in the original data at $p \leq 0.05$ significance. This result means that for all synthetic datasets, the desired outcome of statistical consistency across features between datasets is satisfied.

Labels and records were tested by first training an XGBoost classifier on each set of synthetic data and comparing the results to the output from the real data. Similar AUC scores would indicate not only that important statistical properties persist through encryption, but that models can be trained and pickled from purely synthetic data. Table 2 shows results for an XGBoost classifier employed with out-of-the-box parameters.

AUC Scores	Orig. Model	GAN Synthetic Model	Augmented Model	HE Model	Random Model
Orig. Data	0.66	0.54	0.67	-	-
GAN Synthetic Data	0.51	0.90	-	-	-
Augmented Data	-	-	0.89	-	-
Pseudonymous Data	-	-	-	0.66	-
Random Normal Data	-	-	-	-	0.57

Table 2. AUC scores for all experiments with out-of-the-box XGBoost model

Table 2 shows that all models trained on synthetic data failed to produce similar output as those trained on real data. The most striking result from Table 2 is the ability of models trained on random normal data to generate comparable or better AUC scores than those trained on data generated by the GAN. This was because the labels generated by the GANs were singular and, as such, random normal data is statistically more likely to mimic the variance present in the real data than our singular synthetic values.

We hypothesized that the results in these experiments were likely due to overfitting between epochs when we discovered mixed labels being generated at intermediate steps. The overfitting occurs when the discriminator learns the behavior of the generator as it produces more records in one class and the generator reacts by producing more samples from the other class. The process then repeats. Simply put, the GAN oscillates between both classes as each model overcorrects for itself.

To test this hypothesis, we employed t-Distributed Stochastic Neighbor (TSNE) embedding to visualize the high-dimensional data relationships in two-dimensional space for a dataset of observations from two sequential batches. TSNE builds a probability distribution over the high-dimensional relationships in the data and maps them to points in lower dimensional space. Figure 2 shows the embedding for our data, with each point colored by the synthetic dataset it came from.

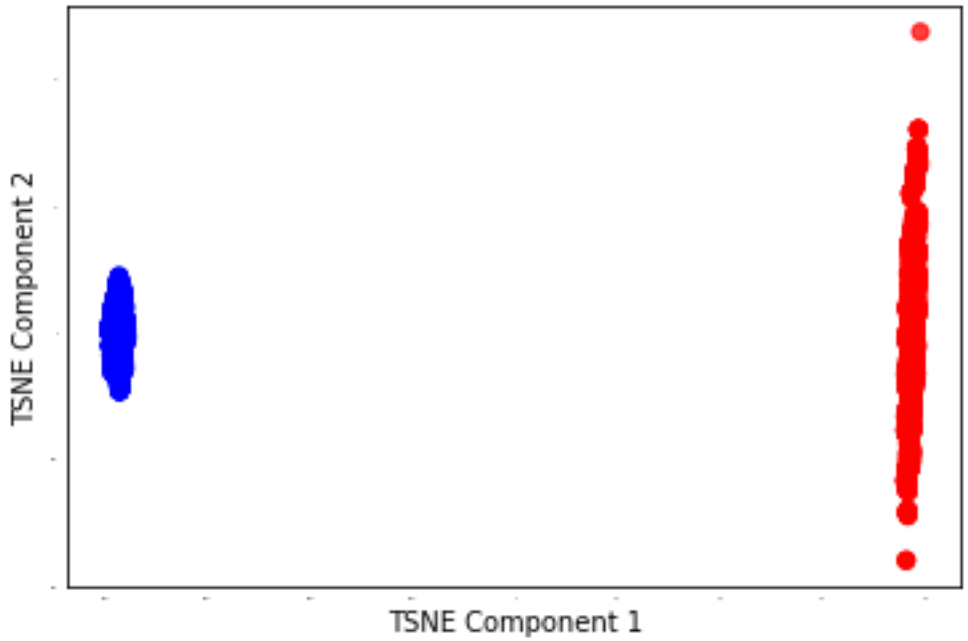

Figure 2. TSNE for sequential epochs

Figure 2 validates our hypothesis, as the separate synthetic datasets generated by each epoch clearly follow very different distributions, most likely following the label that they predict. Because we know observations in the real data do not adhere as strictly to these labels as the synthetic data, we added them to the two synthetic datasets and re-ran the algorithm.

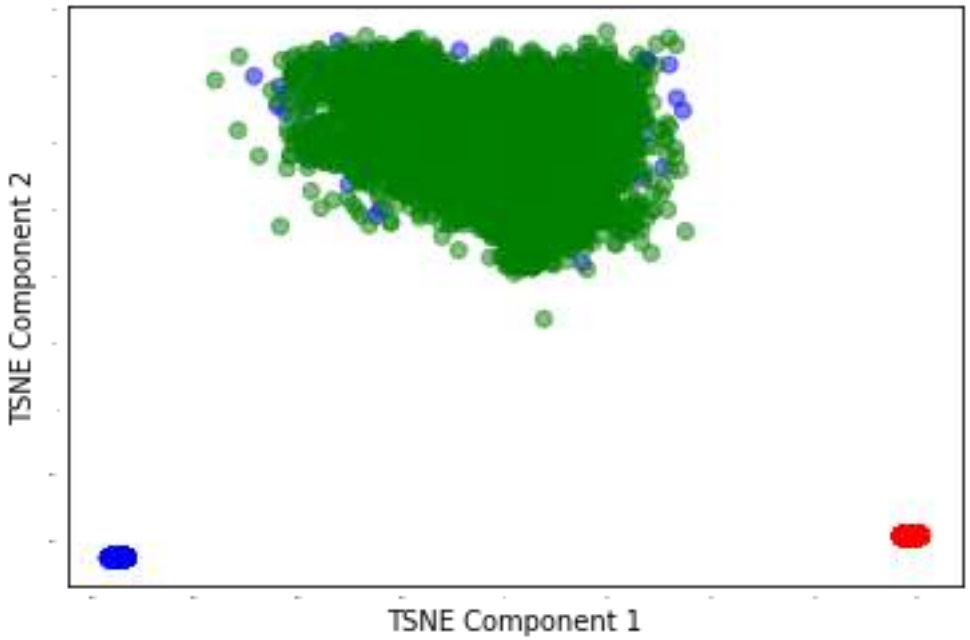

Figure 3. TSNE with real and synthetic data

A surprising result in Figure 3 is that a small sample of observations that came from one of the synthetic datasets more closely followed the high-dimensional behavior of the original data (indicated by the blue dots mixed in with the green) than that of their overfitted synthetic partners.

Just as imbalanced class sizes may result in singular observations in traditional classification, we hypothesized that the overfitting of the GAN could be due to the same. To correct for this, we randomly eliminated samples of the majority class in the original data until both class sizes were equal and re-ran the GAN using this dataset. Figure 4 shows that while a decent portion of the original data's (green) behavior still evaded the GAN, all synthetic observations (blue) appear to exhibit realistic high-dimensional behavior.

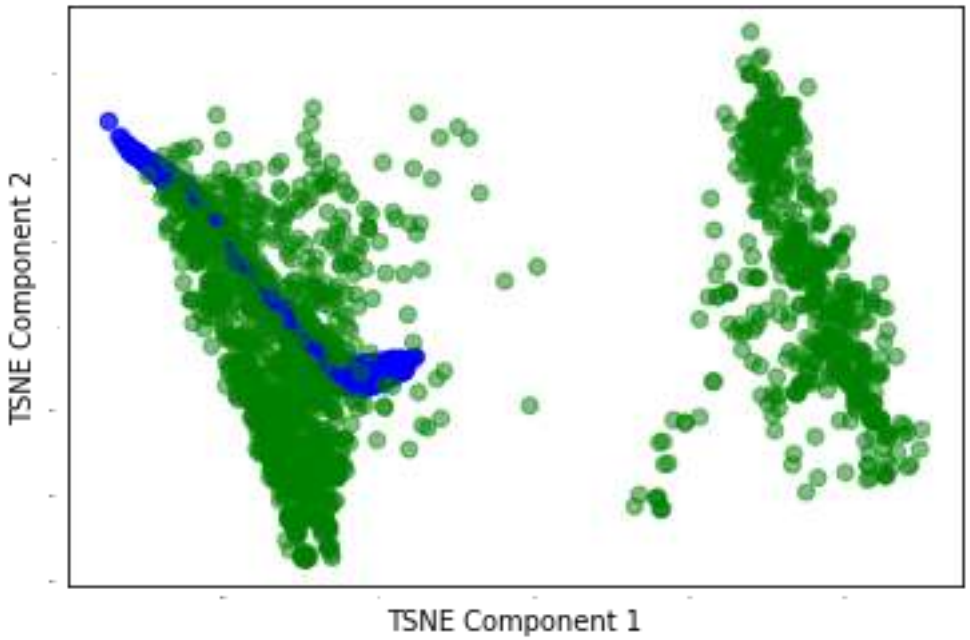

Figure 4. TSNE with balanced data

As a point of verification, the network was trained and tested on the same data set used by the authors (Park, et al, 2018). The resulting synthetic datasets were used to train an XGBoost Classifier model, and Table 4 shows the results compared to that of the original data, as well as an augmented dataset combining synthetic and real data.

AUC Scores	Origin Model	Synthetic Model	Augmented Model
Origin Data	1.00	0.50	1.00
GAN Synthetic Data	0.50	0.99	
Augmented Data			1.00

Table 4. AUC scores using the Adult dataset from (Park, et al, 2018).

The original data and the synthetic data both led to models that were highly accurate on data from the same pool. However, both models were random in their ability to accurately make predictions on data from their counterpart. Augmenting the data appears to have had neither a positive nor a negative impact on the outcome.

DISCUSSION

While no single set of generated data achieved all the qualities originally sought –that is, statistical integrity across records, features, and labels– all resultant datasets exemplify the cutting edge of data de-identification while possessing at least one of these qualities. For practical purposes, this means that experiments can be designed to procure secure synthetic datasets to address specific needs as they arise during development. Additionally, since each

desired property proved individually achievable, the results bode well for the possibility of producing a single dataset with all these properties by modifying the GANs.

Pseudonymization is not designed to lead to realistic results from statistical models or machine learning. The approach did, however, produce defensibly secure pseudonymous datasets with structural properties that empower the developer to complete expensive tasks outside of sensitive systems. These include but are not limited to unit testing, data cleaning and aggregation, runtime optimization, schema development, and user interface design. Additionally, all synthetic features followed the distribution of their original values with high statistical significance, allowing for additional tasks such as realistic visualization of individual features and limited exploratory data analysis.

The GANs resulted in synthetic datasets that allow a machine learning model to train to high accuracy in a purely synthetic environment. This is of value when tuning hyperparameters, optimizing algorithmic runtime, and exploring potential for statistics and machine learning in a sensitive dataset. There are, however, several noteworthy caveats to the usefulness of the datasets produced in these experiments. Whereas a machine learning model was able to produce excellent results on the synthetic data, this performance did not carry over when tested on the actual dataset, implying overfitting of the individual records to their labels by the GANs. This does not take away from the synthetic data's value in being used to build hyperparameter tuners and analytic algorithms, but it does necessitate retraining when deploying them to the original data. Results from t-Distributed Stochastic Neighbor Embedding highlight that the models trained on the synthetic data are unable to find a decision boundary that could recreate the spread of the real data, though this is mitigated with more balanced data. Finally, whereas the security of homomorphic encryption is well-researched and highly regarded in industry, the novelty of deploying GANs for this purpose means that there may still exist vulnerabilities that have yet to be exposed.

Future research should focus on increased data security and improving GAN output. With respect to security, researchers should take every effort to expose potential vulnerabilities in the synthetic data produced by the GANs to ensure that their data is secure. That said, passing such tests does not guarantee security, and until a larger body of academic research has accumulated, there is still some risk that information from the original data may be derived from the GAN output using a technique that has not yet been discovered.

This paper is based upon work funded by the Department of Navy under contract with Booz Allen Hamilton. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of the Navy.

REFERENCES

Corrales Compagnucci, M., Minssen, T., Arasilango, A., Ous, T., & Rajarajan, M. (2019). Homomorphic Encryption: The 'Holy Grail'for Big Data Analytics & Legal Compliance in the Pharmaceutical and Healthcare Sector?. Forthcoming, Special Issue on AI and ML of the European Pharmaceutical Law Review (EPLR).

Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. (2017). CAN: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. arXiv preprint arXiv:1706.07068.

Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018, April). Synthetic data augmentation using GAN for improved liver lesion classification. In 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018) (pp. 289-293). IEEE.

Giessing, S. (2004, June). Survey on methods for tabular data protection in ARGUS. In International Workshop on Privacy in Statistical Databases (pp. 1-13). Springer, Berlin, Heidelberg.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672-2680).

Hermes, K., & Poulsen, M. (2012). A review of current methods to generate synthetic spatial microdata using reweighting and future directions. Computers, Environment and Urban Systems, 36(4), 281-290.

Kelly, J. P., Golden, B. L., & Assad, A. A. (1992). Cell suppression: Disclosure protection for sensitive tabular data. Networks, 22(4), 397-417.

Lubarsky, B. (2010). Re-Identification of "Anonymized Data". UCLA L. REV, 1754(1701).

Milletich, R., Turscak, A., Miller, D., Moreno, M., White, E., Green, R., & Bergstrom, S. (2019). Feature engineering and ensemble machine learning in the Navy Reserve: Using holistic behavioral profiles to predict mobilization cancellation, presented at the MODSIM World Conference, Norfolk, 2019. Norfolk, VA.

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y. (2018). Data synthesis based on generative adversarial networks. Proceedings of the VLDB Endowment, 11(10), 1071-1083.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Rocher, L., Hendrickx, J. M., & De Montjoye, Y. A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. Nature communications, 10(1), 1-9.

Such, F. P., Rawal, A., Lehman, J., Stanley, K. O., & Clune, J. (2019). Generative Teaching Networks: Accelerating Neural Architecture Search by Learning to Generate Synthetic Training Data. arXiv preprint arXiv:1912.07768.

Xu, L., & Veeramachaneni, K. (2018). Synthesizing tabular data using generative adversarial networks. arXiv preprint arXiv:1811.11264.

Zayatz, L., Lucero, J., Massell, P., & Ramanayake, A. (2009). Disclosure avoidance for Census 2010 and American Community Survey five-year tabular data products. Statistics, 10.