FINANCIAL CONDUCT AUTHORITY

Using Synthetic Data in Financial Services

The Synthetic Data Expert Group

March 2024

About the Synthetic Data Expert Group

The Synthetic Data Expert Group (SDEG) is a specialised sub-group of the Innovation Advisory Group (IAG) and was established in February 2023 by the FCA Innovation department. It operates under the guidance of the IAG Terms of Reference and chaired by the FCA. The SDEG helps to foster collaboration across industry, regulators, academia, and civil society to advance the responsible use of synthetic data to shape digital markets to achieve good outcomes and digital transformation at the FCA.

The SDEG was launched in March 2023 and plans to run until November 2024.

The SDEG is comprised of 21 members who were selected against a set criterion, in an open and competitive process. The group explore issues surrounding the use of synthetic data in UK financial markets by identifying relevant use cases, key theoretical challenges and sharing practical experiences of using synthetic data. Additionally, the group has provided valuable feedback on FCA projects involving synthetic data. More information on SDEG membership can be found in the Appendix.

Contents

Disclaimer

This report has been collectively authored by members of SDEG and colleagues across the FCA. The contents of this report reflect the practical experiences members of the SDEG have encountered when generating or using synthetic data. The report is designed to help regulators and industry practitioners better understand the opportunities and challenges of synthetic data.

The contents of this report do not represent the views of the FCA or any participating organisation. It does not endorse or condemn the use of synthetic data and does not imply compliance with UK data protection law.

This report and the applications, discussions and outputs of the Synthetic Data Expert Group should not be taken as an indication of recommendations, guidance or future policy.

Executive Foreword

Jessica Rusu,
CDIIO, FCA

Data is pivotal to financial services and is needed to help build intelligent systems that drive forward transformation. The FCA data strategy highlights the critical role data can play in developing innovative solutions that help to address key challenges and unlock new opportunities in financial services.

I am pleased to introduce the Synthetic Data Expert Group's report, which explores how synthetic data can be used to overcome data challenges and sheds light on practical applications in financial services. The Synthetic Data Expert Group (SDEG) is a sub-group of the Innovation Advisory Group (IAG), tasked with exploring the use of synthetic data in financial markets.

This report reflects a diverse range of perspectives, expertise and skills. It is a culmination of extensive research and collaboration within the SDEG, focusing on three key themes across the data lifecycle; data augmentation and bias mitigation, system testing and model validation, and internal and external data sharing for fraud controls. These themes represent key areas where synthetic data offers transformative implications while remaining cognisant of the associated risks and pitfalls inherent in these advancements.

Synthetic data is one of the many privacy enhancing technologies that can expand data usage and support data sharing without revealing underlying sensitive information contained in the data. Although there are still open questions which are being researched, synthetic data has the potential to help contribute to some of the large public policy issues in financial services, such as financial crime and fraud, and drive societal good through fostering a fairer financial landscape.

For example, synthetic data can serve as a tool for enhancing the robustness of fraud detection models and improve their adaptability to evolving threats. The exploration of synthetic data in credit scoring highlights its potential to mitigate biases, cultivating a fairer and more inclusive financial landscape. As technology advances, rigorous system testing and model validation becomes increasingly important. The Groups findings show

how synthetic data can help simulate diverse scenarios, supporting the resilience and accuracy of financial systems. The report also addresses the complexities of internal and external data sharing for fraud and anti-money laundering controls. Synthetic data emerges as a promising approach to facilitate collaboration while safeguarding sensitive information. Striking the right balance between information sharing and data protection is crucial, and this report provides valuable insights into this.

The deep expertise offered by SDEG members in this report provides valuable insights into some practical applications of synthetic data that can help shape the future landscape of data usage in the financial sector.

In conclusion, I would like to commend the SDEG's extensive efforts in producing this report and thank all its members alongside FCA colleagues. I believe that the insights and conclusions provided will serve as a cornerstone for industry practitioners, policymakers, and regulators, and provide a new mechanism for cross-sector collaboration. Further understanding the potential benefits and drawbacks of synthetic data is important for advancing the resilience, fairness, and efficiency of the financial services sector.

Sincerely,

Jessica Rusu, Chief Data, Information and Intelligence Officer, FCA

Chapter 1

Introduction

1.1 As financial services become more digital, an increasing volume of data is recorded. Vast data pools can help firms to better understand business operations and reporting, test and develop new products and/or services, improve decision making and lead to consumer benefit through more personalised services and innovative financial products. Access to data enables institutions to use more advanced modelling techniques and train artificial intelligence (AI) models more effectively. Despite a rise in the amount of data that is generated, challenges remain for institutions to access and share data that could drive societal benefit.

1.2 Synthetic data is one of the many Privacy Enhancing Techniques (PETs) that can be used to mitigate against the privacy risks of data sharing. Synthetic data is a privacy-preserving technique that can be used to address the challenges associated with sharing sensitive data such as personal or financial data. It works by generating statistically realistic but artificial data that can be used to create advanced modelling techniques and train AI models without compromising individual privacy or data protection laws.

1.3 The FCA's feedback statement on the Synthetic Data Call for Input identified data availability, quality and regulatory uncertainty as some of the key challenges industry are currently facing when trying to generate and use synthetic data. In response to the feedback statement, in March 2023 the FCA set up the Synthetic Data Expert Group (SDEG), bringing together 21 experts from across industry to help overcome barriers to adoption relating to synthetic data in financial services and regulatory circles.

1.4 The generation of synthetic data has the potential to help industry and regulatory bodies address pressing societal challenges and perennial issues in financial services, such as fraud and financial crime. This work supports the FCA's three-year strategy by leveraging shared expertise and providing insights to help shape digital markets to achieve good outcomes and digital transformation at the FCA.

Aims of the report

1.5 This report provides insight into the experiences of the SDEG members in generating and applying synthetic data in the context of financial services. It aims to help industry and regulatory practitioners to develop a comprehensive understanding of the techniques, tools, practical challenges and opportunities associated with synthetic data to contribute to the effective and safe deployment of synthetic data.

1.6 The insights of this report will be of interest to industry participants including financial services, regulators and policymakers internationally. The key findings will serve as a helpful guide by explaining the steps to consider when creating and using synthetic data,

and the types of problem statements where synthetic data can be useful. Whilst the use cases in this report are explored in relation to financial services, the insights can also be applied to other sectors.

1.7 Applications of synthetic data are diverse. Responses to the Feedback Statement and wider research indicate themes are beginning to emerge around how synthetic data can be used. Based on these themes and the expertise of members, we have selected three elements of the data lifecycle to illustrate the utility of synthetic data.

1. **Data augmentation and bias mitigation:** The transformation of data to expand and/or reduce the inherent bias associated with the underlying data for model generation.
2. **Systems testing and model validation:** The generation of synthetic data to rigorously test the robustness of AI, machine learning systems and validate their performance under diverse scenarios.
3. **Internal and external data sharing:** The responsible sharing of synthetic data and associated models within an organisation (internal) and/or to support external facing financial services.

1.8 Each theme serves as an illustration of the different opportunities and challenges that can arise when using synthetic data effectively and responsibly in financial services and how elements such as privacy, utility and fidelity are considered in practice.

1.9 Within each of the three themes, SDEG members identified two use cases where synthetic data has proved beneficial in financial services and provide learnings and insights on each. The use cases include fraud detection, credit scoring, open banking, authorised push payment (APP) fraud and anti-money laundering (AML). A full overview of the themes and use cases are found on page 12.

1.10 Through exploring different applications across the data lifecycle, the report shows examples of how synthetic data can be used to drive societal good through fostering a more inclusive and fair financial landscape that is underpinned by responsible and robust modelling techniques.

Chapter 2

Context and background

Synthetic data and privacy enhancing technologies (PETs)

2.1 Synthetic data can be challenging to define because it is currently used across a broad spectrum of sectors in addition to financial services and for a wide range of activities. Accounting for this, the Royal Society define synthetic data as:

"...data that has been generated using a purpose built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)."

2.2 There are different generation techniques that can be used to create synthetic data. These often include the use of agent-based modelling, econometric techniques, deep learning architectures such as Generative Adversarial Networks (GANs), Variational Auto-encoders (VAEs), gaussian copulas or a mixture of techniques.

2.3 The chosen approach is often dictated by a range of factors such as the complexity of the data inputs, the use case/problem statement in question and the desired outcome and results. Synthetic data is not the only technique that can be used to protect data privacy. There are other PETs, such as differential privacy or homomorphic encryption, that may be more appropriate to use than synthetic data. It is also common to use several PETs in-combination with each other. The box below provides more information on different types of PETs.

Privacy Enhancing Technologies (PETs) at a glance

PETs are a group of emerging technologies that enable the sharing of datasets whilst preserving privacy.

The Information Commissioner's Office (ICO) define PETs as "...technologies that can help organisations share and use people's data responsibly, lawfully, and securely, including by minimising the amount of data used and by encrypting or anonymising personal information"

Examples of PETs include:

- **Homomorphic encryption**: an encryption method that enables computational operations on encrypted data.
- **Secure multi-party computation:** is a tool that enables distributed computation, where analysis can occur on combined data sets without different parties revealing the private source data to one-another.
- **Federated Learning:** a machine learning approach where a model is trained on individual devices without sharing the raw data.

- **Differential privacy:** differential privacy adds 'statistical noise' to a dataset to mitigate the privacy risk, and can be used to statistically quantify the privacy risk of a dataset.
- **Zero-Knowledge proof:** a method whereby one party can prove to another party that a given statement is true without revealing the statement's contents.
- **Synthetic data:** data generated using data synthesis algorithms, replicating the patterns and statistical properties of real data (which may be personal information).

More information on the PETs above can be found here.

Box 1: Information on PETs has been summarised from ICO Final Guidance on Privacy Enhancing Technologies (ICO, 2023) and the FCA's feedback Statement on Synthetic Data Call for Input (FCA, 2023).

Evaluating privacy, utility and fidelity

2.4 A common challenge when generating synthetic data is validating that the data that has been created is fit for purpose, with the appropriate characteristics for the use case in question, yet private enough to protect the underlying data. The FCA, ICO and Alan Turing Institute held a joint roundtable in March 2023 to explore the emerging techniques that are used to validate synthetic date and industry/academic perspectives on validation approaches.

2.5 The roundtable identified three key elements to consider when evaluating synthetic data; privacy, utility, and fidelity. This means the synthetic data that has been created is private enough to protect the individuals/firms behind the data, has enough utility for the task at hand and has the desired statistical properties of the real data. Any data sharing technology, including synthetic data, is subject to a trade-off between privacy, utility and fidelity.

Definitions only for the purpose of this report

Privacy: Measures the risk that specific individuals (or other sensitive data) can be re-identified from the synthetic dataset.

Utility: A synthetic dataset's 'usefulness' for a given task or set of tasks, for example for training AI or Machine Learning models.

Fidelity: Refers to measures that directly compare the synthetic dataset with the real dataset i.e., the statistical similarity of the synthetic dataset to the input real data.

Box 2: Concepts to consider when validating synthetic data. Source: Synthetic Data - what, why and how? (Jordan et al. 2022)

2.6 Measuring the privacy, utility, and fidelity of synthetic data sets is crucial. Assessment of privacy safeguards preserves confidentiality and ensures legal, regulatory, and ethical standards are met. Evaluating the utility and fidelity ensures the relevance and accuracy of the synthetic data against unwanted access, and directly impacts the reliability of machine learning models trained on the data for real-world applications.

2.7 There are different approaches one can take to evaluate the privacy, utility, and fidelity of synthetic data. The selection of metrics and qualitative judgments may depend on factors such as the problem statement, use case and end application of the data.

2.8 For each of the use cases in this report, the authors provide an overview of the trade-offs and key considerations of evaluating the privacy, utility, and fidelity of the problem statement in question. Validating these properties are essential to the effective and safe deployment of synthetic data. As such, we expect this to be an ongoing area of research and continue to see developments in how these characteristics can be evaluated.

Market initiatives

2.9 There is increasing awareness of synthetic data's effectiveness in addressing data challenges within the financial sector. The Information Commissioner's Office (ICO) actively contributed to this landscape by disseminating guidance on PETs and exploring the practical applications of synthetic data for achieving data minimisation objectives. This commitment is further emphasised by the Responsible Technology Adoption (RTA) unit's PETs Adoption Guide, highlighting the integration of privacy-enhancing technologies within regulatory frameworks. In the UK, academic institutions including the Alan Turing Institute and the Royal Society, actively contribute to the ongoing dialogue on privacy-preserving research.

2.10 The relevance of synthetic data is underscored by innovation challenges such as the UK-US PETs Challenge led by governmental, standard-setting and research institutions in the United Kingdom and the United States, and research initiatives from bodies like the OECD and the United Nations. A notable recent contribution by the Bank for International Settlements (BIS) Innovation Hub's Nordic Centre, Project Aurora, explored how PETs could be used to address AML challenges through collaboration analytics and learning. This project demonstrated the complexity associated with using PETs, machine learning models and network analysis to counter money laundering and enhance detection models. In the US, the National Institute of Standards and Technology (NIST) has a comprehensive work programme supporting synthetic data generation, including the creation of a suite of open source tools and prize challenges to advance data sharing.

2.11 In the financial sector, there are a growing number of organisations using synthetic data to develop algorithms for generating realistic datasets. Topics of interest often focus on areas such as anti-money laundering, customer journey analysis, payment processes, documentation, and equity market data. In industry, many technology firms are employing synthetic data techniques to navigate complex data sharing challenges.

2.12 For instance, within Microsoft's AI Lab project, specifically the synthetic data generator, has been devised to safeguard privacy during the sharing and analysis of sensitive datasets. Amazon Science has published work on generating synthetic data to overcome challenges associated with expensive and complex data collection. Specifically, Amazon's approach involves employing a Large Language Model (LLM)-based "teacher" model to generate synthetic training data for a specific task, subsequently fine-tuning a smaller "student" model with the generated data. These initiatives reflect a broader industry trend, signalling the exploration of synthetic data beyond financial services.

2.13 In summary, the increasing level of research and experimentation of synthetic data across a diverse range of domains indicates its potential in fostering beneficial innovation and support developments in financial services.

Chapter 3

Theme and use cases

3.1 The table below provides an overview of the use cases explored in this report. The use cases below are authored by SDEG members. More detail can be found in Appendix 1.

Theme 1: Data augmentation and bias mitigation

Use Case Title	Focus
1. Transaction sequences used in fraud detection machine learning models	The availability of fraudulent financial transaction data to train a fraud detection machine learning model
2. Reject inference in credit scoring	Generating synthetic data to mitigate selection bias in credit scoring training data

Theme 2: Systems testing and model validation

Use Case Title	Focus
1. Synthetic Open Banking data for model testing	Generating synthetic transaction data labels to enhance training sets and transactional data to resemble consumer patterns and behaviours
2. Cross-sector synthetic data for Authorised Push Payment Fraud	Creation of synthetic data relating to individual, banks and fraud typologies to complement real data and be used in the FCA's APP Fraud TechSprint

Theme 3: Internal and external data sharing

Use Case Title	Focus
1. Common data sets for community research into societal challenges	The US-UK Challenge using Privacy Enhancing Technologies to enable data sharing to improve financial crime prevention and pandemic responses
2. Data sharing to increase efficiency and effectiveness of anti-money laundering controls	Using cross-organisational and international transaction data to develop machine learning models to recognise illicit payment patterns

About the use cases

The following themes explore different financial services use cases where synthetic data can be used across the data lifecycle. Within each theme, SDEG members detail the **use case, methodological decisions, trip hazards** and **lessons learned** from using synthetic data.

The use cases are designed to aid practitioners to use synthetic data effectively and responsibly.

Note, given input from different authors, the tone and language may vary across use cases.

Chapter 4

Theme 1: Data augmentation and bias mitigation

Overview of theme

4.1 For modern machine learning pipelines, the quality of the output they produce is directly dependent on the quality of the data used to train the models, both personally identifiable information (PII) and non-PII. If the data used during training is of poor quality, incomplete, or biased the outcomes of the models trained on such data sets will be impacted. The key idea driving the use of synthetic data for data augmentation is to enhance the performance of data-driven machine learning pipelines by enhancing the training data. Furthermore, the goal of a machine learning model is not just to perform well on the training data but also post-deployment on the data that the model hasn't been exposed to before. While synthetic data is not the only technique that can help data-driven machine learning perform well, it provides a unified data-centric solution to this challenge. However, synthetic data is necessarily a distorted version of the real data. Therefore, any inference performed on synthetic data comes with additional model risks or caveats.

4.2 Here we will focus on data used to develop predictive statistical or supervised machine learning models since the performance of these models are dependent on the quantity and quality of the data used to train them.

4.3 We define data augmentation as increasing the amount and quality of data available for use. This is particularly beneficial for machine learning models, as they rely on rich data to extract patterns from. In cases where there is insufficient volume or diversity in the data, the performance of the model might experience limitations.

4.4 In the data augmentation use case we discuss the availability of fraudulent financial transaction data (for example, debit or credit card payments) to train a fraud detection machine learning models. The volume of fraudulent transactions held by a financial organisation may be very low, typically make up fewer than 0.2% of transactions. Augmenting the real-world fraud data with synthetic fraudulent transactions may be used to improve detection performance.

4.5 Robust model performance also relies on high quality training data, representative of that on which it will be used. In particular, training data should be free of material biases such as under-representation or omission of key segments. A common source of bias is when data is filtered by a selection process, such as consumer lending decisions for credit. The second use case explores the use of synthetic data to mitigate bias resulting from a lack of information on the payment performance of previously rejected credit applicants when developing credit scores.

Theme 1 in the data lifecycle

4.6 Augmenting data and mitigating bias generally occur as part of the data preparation phase, but their necessity may only become apparent after an initial model has been developed and analysed, leading to an iteration of the lifecycle.

Current practices for data sharing and cross sector applicability

4.7 The techniques discussed herein are applicable in the financial services sector but are also applicable everywhere that data augmentation and bias mitigation are needed, such as for image recognition, autonomous vehicles, home and auto insurance claims, and healthcare. To use healthcare as an example, bias in synthetic data could perpetuate existing biases, such as where certain types, classifications, illnesses, etc. are lacking. Data augmentation could be useful to combat data scarcity and potential imbalances in the data, but could also perpetuate existing biases.

Use case 1: Transaction sequences used in fraud detection machine learning models

The problem statement

Financial services companies issuing cards, or providing other payment services, have a duty to protect their customers from third party fraud, whereby criminals steal payment instrument details and transact as if they were the customer. Typically, payment service providers will implement automated systems, often including machine learning models, for detecting and blocking attempts by fraudsters to make such transactions.

To train such models effectively, a sufficient volume and variety of historical fraudulent and genuine transaction attempts are required; however, attempted fraudulent transactions typically make up a very small percentage of all transactions, and the number of such transactions may be insufficient to train an effective and accurate model.

Generating synthetic data replicating fraudulent transaction patterns and including it in the training data can be a viable way to improve the model detection rate and reduce false positives (genuine transactions incorrectly classified as fraud). To achieve a model performance improvement, these synthetically produced fraudulent transactions need to be as realistic as possible.

Synthetic data methodology

One option for such synthetic data generation is an instance-based machine learning method using historical real fraudulent transactions to produce a set of statistically similar, synthetic fraudulent transactions. We outline the procedure below.

Consider a hypothetical set of historical fraudulent card transaction data, with three key fields: Transaction Amount (£s), Merchant Type (e.g. electronic goods retailer, food retailer, etc) and Channel (e.g. online, in-store, etc). These data will have joint distributions among each of the fields which can be used in the creation of synthetic data. To generate the first synthetic record, the process selects one field at random and picks a random value according to the distribution of that field e.g. Transaction Amount of £2,000. Next, another field is selected at random, and a value selected base on its distribution conditional on the first feature e.g. Merchant Type of electronic good retailer, given the Transaction Amount of £2,000. And so on for Channel, and any other relevant fields.

Transaction Amount	Merchant Type	Channel	Time
£2,000	Electronics	Online	10:22 GMT
£142	Food	In-store	12:48 GMT
£617	Digital Goods	Online	21:01 GMT

Once values for each field have been generated, the synthetic data record should be analysed for privacy, to ensure that the process did not accidentally recreate a record from the original dataset (or one that used to reidentify any of the original data) and discarded if necessary. This process may then repeat until sufficient synthetic data has been generated.

Evaluating privacy, utility and fidelity

Generally, if privacy of the original data is of concern, the synthetic data should protect the privacy of the individuals or organisations whose data was used to generate it. This use case, however, assumes that the synthetic data will be used to augment the original data, and therefore the privacy of the synthetic data with respect to the original data is not as concerning as utility and to some extent, the fidelity.

In this use case the utility of the data is met by synthesising fraudulent transaction data that is statistically similar to the training data in the ways described above. For utility, many different tests could be performed. We use the maximum mean discrepancy (MMD), Chi-squared, the Kolmogorov-Smirnov test and the Mann-Whitney test. To test fidelity, we compare the performance of the original and synthetic data across numerous machine learning models, including a classification comparison using the Fowlkes-Mallows index, and a clustering comparison performed using a combination of an intrinsic measurement (Calinski-Harabasz score) and an extrinsic measurement (a particular formulation of mutual information).

Trip hazards and risk of poor practice

Lawful basis for processing

The data for this use case is pseudonymised personal data, and as such, a lawful basis for processing should be established for any processing, including creating synthetic data in the way described. Only fields which are truly necessary for the generation of

synthetic data should be used, and in this case no fields which individually constitute personal information (such as names) are necessary and should be removed before analysis commences.

Model Performance

That aside, a primary concern is to ensure that a model trained on the real and synthetic data exhibits improved performance. It is important that this performance improvement is seen not just on the composite set of synthetic and real training data (which is self-fulfilling, assuming model training has been performed without error), but on a holdout sample unseen by the model, and ideally on the real data alone, even if the volume is low.

Model Validation

It is also important to record the data generation process in sufficient detail that users can understand potential strengths and weaknesses. This can be achieved via auditable 'data cards' and is particularly important if the synthetic data is to be released to other data processors, in which case the data card should include details of the type of statistical information from the real data that the synthetic data may reveal.

The risk of revealing information about the real data is usually higher if a large volume of synthetic data is to be released to other data processors, and care should be taken to ensure that the synthetic data do not allow inferences to be made about the real data. This risk decreases the higher the dimensionality of the synthetic data.

Lessons learned

Using synthetic data to augment the availability of fraudulent transactions works well, especially in the context where the synthetic fraudulent transaction data is used to augment (and not replace) the original fraudulent transaction data, as long as there is a lawful basis for making synthetic data from the original data.

Use case 2: Reject inference in credit scoring

The problem statement

Determining the credit worthiness of applicants for consumer credit is an important activity for lenders. For decades, automated processes based on statistical models ('credit scores') have been the norm, and whilst there exists a body of best-practice in the industry, this is evolving in the light of new technology and heightened expectations around bias, fairness and privacy.

Credit scores are developed to predict the likelihood of default, based on a number of factors such as an applicant's previous repayment history and credit utilisation. This is then used to make a lending decision.

Whist many lenders may have sufficient data from their previous lending, they suffer from a significant bias: the data does not contain repayment information for credit

applications which did not result in lending, either because the application was rejected, or was accepted but not drawn down. The latter are generally fewer and low risk, so can be safely excluded from the development. The rejects, however, may make up a significant proportion of applications and are systematically biased: they were selected by the incumbent lending strategy, not at random, and have a higher risk of default. It is therefore vital that their risk is not underestimated (to avoid lending irresponsibly) or indeed too overestimated (to avoid financial exclusion).

One solution to this challenge involves creating synthetic data on previous rejected applicants i.e. inferring whether they would have repaid the loan in the counterfactual situation that they were accepted for credit. This is known as 'reject inference'. Whilst it is possible to source payment performance on similar loans granted by other lenders from Credit Reference Agencies (CRAs) – known as 'reject referencing' – this doesn't guarantee complete or representative coverage, so some inference using synthetic data generation is always required.

Synthetic data methodology

Statistical modelling such as linear and logistic regression is traditionally used to generate synthetic payment performance for reject inference. However, the methodology is not straightforward due to the systematic biases described above. Techniques to mitigate this can be divided into two classes: those that rely on data with known payment performance alone, and those that augment with external performance data, typically from CRAs.

Using performance on accepts alone requires a methodology to mitigate the systematic bias: a model developed on accepted applicants is likely to underestimate risk when applied to rejected applicants. Based on analysis of the inferred risk of segments of rejects, expert judgement can be applied to adjust it to match expectations.

Empirical studies have been performed using variations of this methodology, by treating a subset of accepted applicants as if they were rejects and masking payment performance, applying the methodology, then using the masked performance to assess the accuracy of the inferred risk. They show that more systematic approaches given the best results. For example, intuitively, one would expect the reject inference model to underpredict the most on applicants with a profile (e.g. credit history) that is most different to the accepted population. This difference in profile can be estimated using an 'Accept/Reject' model, which predicts the likelihood that an applicant will be accepted. Those with a lower probability of being accepted have a profile most different to the accepts, so should have their inferred probabilities increased the most.

Whilst reject inference without external data can give acceptable results, using applicant performance on similar credit granted at a similar time with other lenders is generally an improvement. There do, however, remain a variety of methodological choices to make such as: which credit granted at other lenders is similar enough to serve as a reference; how close in time should it be to the original application; how should rejected applicants that didn't open another credit account soon after be treated? There is no formula for these choices, they must be made on a case-by-case basis.

Evaluating privacy, utility, and fidelity

The nature of this use case – attaching synthetic counterfactual estimates of default risk to actual historical credit applicants – results in a different formulation of the privacy, utility, fidelity trade-off compared to generating completely new synthetic records from a benchmark dataset.

Lenders have been permitted for many decades to use consented, pseudonymised data on their previous credit applicants to develop credit scores, and the sharing of consumer credit data via CRAs is a well-established and secure process. As long as lenders and CRAs follow these good practices, generally generating synthetic data for reject inference on rejected applicants should not further impact their privacy.

Fidelity for this use case is not well defined because for a counterfactual outcome there is no 'ground truth' to compare to. That said, approximate benchmarks may be obtained via reject referencing, or masking the performance of some accepts, and it is good practice to use them where possible.

Ultimately, the utility of the reject inference is determined by how accurate and reasonable the predictions of the final model developed on actual and inferred outcomes are. Consideration should be given to whether the model is sufficiently conservative on previously rejected applicants, whether it correlates well with other risk indicators (such as income and credit history) and whether it is intuitive across sub-populations.

Trip hazards and risk of poor practice

The key risk of poor reject inference is the under- or overestimation of risk on applicants with a similar profile to those rejected historically. The best way to mitigate this is to conduct extensive validation to ensure the credit score has a reasonable behaviour on rejects. Key areas to be aware of are:

- Performance of the final model should be split by accepts and rejects to ensure that it performs well on both;
- The size of the reject population is important: for example, if 80% of historical applications were rejected, the inferred performance will dominate the model;
- Usually, the model should have better discrimination on the whole population than on accepts alone, since the rejects should be easier to rank.

Lessons learned

Validating synthetic data representing counterfactual outcomes requires extensive analysis and subject matter expertise to ensure the outputs are reasonable and relevant to the use case. A balance must be struck between overly complicated approaches which do not allow the overlay of expertise, and overly simplistic approaches which do not effectively differentiate the risk of applicants. Above all, one cannot create information from nothing. The use of external data for reference rejects can significantly improve the utility of synthetic data and consequently the performance of the credit score.

Regulatory considerations for data augmentation and bias mitigation

These use cases typically require the processing of pseudonymised personal data, so as usual a lawful basis for processing must be established. The most common basis is consumer consent, which is typically obtained as part of the terms and conditions of a credit product. This generally covers the processing and sharing of data for analytics in support of responsible lending and fraud prevention, including the creation of synthetic data.

Any resulting models must also conform to relevant legislation and regulation governing their use, such as The FCA Consumer Credit Sourcebook and the Equality Act 2010. Given that use of synthetic data in this way has an impact on model risk, the PRA's supervisory statement on management of model risk (SS1/23) is also relevant for banks.

Note: this box provides an overview of the important regulatory considerations relating to use case one and two above from SDEG members. These considerations are not a comprehensive list and should not be taken as indication of a policy position or guidance.

Chapter 5

Theme 2: System testing and model validation

Overview of theme

5.1 Systems testing and validation is a critical use case for synthetic data. Using real data can have significant privacy, security, and permission implications. In some cases, real or sample data may lack the coverage needed for testing obscure edge cases or validating new data concepts and scenarios. Synthesising data, among other techniques, can act as an alternative approach to using real data or complement real data. In this theme, we explore the use of synthetic data for system testing and model validation in two use cases:

1. Testing whether synthetic data can augment/replace real-world training data in a bank transaction classification engine, and
2. Providing synthetic cross-sector banking, telco, identity, and crime data to explore and validate new data concepts for combatting authorised push payment (APP) Fraud

Theme 2 in the data lifecycle

5.2 Testing and validation needs manifest all through the data process/product lifecycle, for example in developing propositions, scaling to production, or in software update test-cycles. At all these touchpoints, it may be necessary to test ideas and implementations with data. Because many data processes and products utilise sensitive or personal data, testing will naturally require data with similar properties, which may conflict with best practise, for example data minimisation, or regulatory/legal requirements.

5.3 In addition, assuming that compliance needs have been met, using real data can require additional steps to be taken, including ensuring that real data is sufficiently de-sensitised, relevant, and diverse for testing. This can be an on-going overhead for organisations to manage. It is time consuming to keep up to date with evolving compliance and product context which may lead to sub-optimal outcomes using older, or heavily amended data due to availability/time pressure.

5.4 Synthetic data potentially supports this use case in the following ways:

- Mitigates many privacy and security concerns
- May in some cases be generated more frequently with shorter lead time than real data can be retrieved, screened, and redacted
- Amplifying absent/rare edge cases, and generating at much larger scales

Current practices

5.5 Data anonymisation/testing tools are available in the market and are effective in certain scenarios for generating test data for system updates. These include validating Extract, Transform, Load (ETL) and data model pipelines, or for 'smoke-testing' data and developing processes that involve machine learning. To be effective, they necessarily remove coherent linkage between datasets for data subjects, mask time-dependence and other correlations between variables which may be desirable to preserve in certain scenarios. Synthetic data, among other techniques, may offer a solution where these characteristics can be retained, while also addressing privacy issues. However, as with other data sharing techniques, the trade-off between utility, fidelity and privacy will still need to be considered.

Use case 1: Synthetic Open Banking data for model testing

Problem statement

By its nature, transactional data is highly sensitive and time relevant. Express consumer consent is an important aspect of Open Banking and transactional data. In addition, a common challenge with applications using data that can potentially include Personal Identifiable Information (PII) is the security aspects of storing this information and sharing test files with organisations looking to evaluate those models/applications against a range of potential scenarios and edge cases.

Whilst recognising the limitations of synthetic data and that synthetic data cannot substitute real-customer data, especially in areas such as credit decisioning, the analytics research and development functions within organisations might also want to have a first-hand view of the potential of synthetically generated data as well as evaluate and quantify the downstream impacts on model performance and system testing. This use case focuses on a project to create synthetic transaction descriptions (text) to augment and enhance the existing training sets. Although this use case is focused on transactional data, it is expected that these approaches could be generalised for other types of consumer data.

The research project explored in this use case was a proof of concept, and therefore, the stakeholder group involved in the project was limited to the analytics team and Data Scientists that would be generating the synthetic data. It also included product representatives to gain input on the product proposition, and input from legal and commercial teams to review existing data contracts and conditions.

Synthetic data methodology

There were two aspects when looking to generate synthetic transactional data:

a. Generating individual transaction descriptions, a text field containing information for the transaction with distributions similar to those found in the real-life data such as Tesco, London, etc.
b. Generating arrays of synthetic transactional data that, when aggregated to customer level, they would resemble the consumer patterns and behaviours from the real-life data.

These two models are independent of each other and are focusing on either creating bespoke synthetic transactions or synthetic customer profiles for income and expenditure.

For the individual transaction descriptions, a model was used to generate text, one token at a time, predicting the next token based on the context of the previous tokens, and a variant allowed users to guide / control the generation of text by providing specific input or context. Therefore, this such allowed for the introduction of additional information or constraints during the generation of the text descriptions.

For the generation of arrays of synthetic data, the data sample used to create the synthetic data was selected to reflect customers from specific product segments, ensuring some homogeneity in the transactional patterns and spending trends.

Evaluating privacy, utility, and fidelity

The key driver of the research described in this use case was to test existing applications and machine learning models. Privacy is both a motivator behind the use of synthetic data instead of real-consumer transactional data, and a key consideration when evaluating the privacy and fidelity of the synthetically created data, due to legal requirements. Privacy is less of a challenge for synthetically created individual transactions as, by definition, it is not easy to reverse-engineer and identify an individual from a single transaction. Similarly, the aggregated profiles are removed from a detailed list of all of the individual transactions from a unique consumer, thus reducing the risks associated with privacy.

One key challenge was understanding the impact to models when for contractual, regulatory, or system requirements, part of the underlying model training data might need to be deleted while ensuring the ongoing stability of predictive models/capabilities. In this instance fidelity was an important dimension of the evaluation as by varying the requirements for similarity of the synthetic data to the "real-world" data, the range of impacts to the models was evaluated.

For the testing component, fidelity was not as much of a driver compared to privacy or utility. Due to the sensitive nature of transactional data, testing existing Open Banking applications and systems might require large volumes of transactional data that are

similar enough to what will be seen in real life to fully evaluate and test the end-to-end capability of a system but not to be able to be linked back to specific consumers and their behaviours.

Trip hazards and risk of poor practice

System testing is considered one of the highest priority use cases as before organisations onboard into the Open Banking journey, it is important to ensure that the end-to-end processes and systems are ready for use. Furthermore, it is important to establish the capabilities to do extensive system testing by feeding large volumes of transactional data to understand both capacity/speed of processing, as well as accuracy of outputs and ability to handle edge cases. There are risks if the limitations are not well understood and organisations dive into the creation of synthetic data for credit or customer management decisions without evaluation of the potential negative impacts.

Lessons learned

For the individual text description creation, the success of the similarity of the synthetically created data was evaluated by assessing the impact on model performance, for models currently using transactional data obtained by a range of different real-life sources. The final accuracy of challenger models trained on different datasets, by adjusting the ratio of real-life data, and synthetically created data, was compared to the accuracy of the predictive model trained on only real-life data.

It is not always the case of choosing whether to use synthetic data or real-life data. It is important to also understand the potential implications of having to remove some of the real-life data from existing training or validation sets, and evaluating the impacts on model predictiveness and accuracy.

Data source	Description	Accuracy change
Real-life data only	Only real-life data used for training	Benchmark
Synthetic data + real-life data	50% synthetic data, 50% real data used	-2.5% of the benchmark
Synthetic data + real-life data	70% synthetic data, 30% real data used	-5.0% of the benchmark
Synthetic data only	Only synthetic data used for training	-32.0% of the benchmark

In summary, the preliminary findings indicated that for the research in this use case, a threshold of at least 30% real data or an optimisation of the ratio of real-world and synthetic data was needed to maintain strong model accuracy. Further research could be undertaken to compare the accuracy using different synthetic data generation tools to identify whether certain approaches or configurations may lead to better results.

For the creation of arrays of data, a 70% overall data quality was achieved with an approximately 60% similarity of synthetic data vs real data in terms of marginal distributions. The results also suggested that over 90% of the synthetic data rows were not copies of real data, an important question in the beginning of this research project. Finally, the model managed to capture specific underlying conditions in the real data and avoided creating arrays that would have no likelihood in appearing in a real-life dataset.

An alternative evaluation of the impacts on model performance and model accuracy was to potentially consider using as a benchmark the smaller dataset of real data against the enhanced dataset. This would align with the analytical motivation of maintaining model performance and model accuracy in cases where a proportion of real-life data was needed to be deleted for regulatory, governance, consent, or system migration reasons.

Both points above support the potential of using synthetically generated data for software and system testing as a first step prior to fully relying on synthetic data for training and validation of certain types of predictive models and use cases.

Use case 2: Cross-sector synthetic data for authorised push payment fraud

Problem statement

The FCA and PSR ran an authorised push payment (APP) fraud TechSprint in September 2022. A key driver for this data was to help the participants create solutions to tackle the TechSprint problem statements:

1. Real-time APP fraud prevention using new and existing technologies: what are the barriers and limitations to current (real-time) APP fraud prevention technologies and processes, and how might we encourage the firms we regulate to improve and adopt them?
2. Enhanced Data Sharing: how financial services firms and multiple sectors can share data and relevant analytics, securely in real-time, to spot and prevent fraud?
3. Spotting fraud at source: communicating them to those in the chain, including PSPs so that they can take affirmative action and protect consumers.

The FCA on-boarded real pseudonymised banking data to be used by participants in the TechSprint. It had several limiting factors: Redaction of certain fields like the transaction narrative; No interoperability between different banks' data; A 'stand-alone' perspective in a multi-sector problem space encompassing banking, payments, telco, and fincrime; No labelled fraud examples for concept development/validation.

To overcome these obstacles, the FCA and PSR collaborated with a third party to design and then build a synthetic dataset to complement the real data, with:

- Synthetic identity, account, and transaction data from a banking perspective
- The ability to track payments between account holders in each synthetic bank

- Synthetic identity, account, and call and short message service metadata capturing communication between the synthetic people and businesses
- Multiple typologies of APP Fraud scams
- 'Ground truth' of attempted and successful scams, identifying the scammer and victim

Synthetic data methodology

Generating synthetic 'double' data using artificial intelligence/machine learning approaches would require real data that has the full scope of the synthetic data requirements:

1. Unredacted original data
2. Referential integrity across multiple bank and telco providers
3. Accurately and completely labelled fraud attempts and successes

Achieving this task within the three-month timeline leading up to the TechSprint was unrealistic given various challenges related to data regulation, risk, and other factors. Furthermore, uncertainty surrounding the scalability of artificial intelligence and machine learning methods raised question about their ability to effectively scale across multiple tables, and how large a dataset would be possible to generate.

For these reasons, an agent-based simulation approach was used. This eliminates the need for real data as an input, however, it does require expert judgement and development in creating the scenario, baseline agent behaviours, and specific properties (e.g. fraud typologies) to meet the dataset requirements. Synthetic consumers, merchants, and other organisations (e.g. employers) were created to interact with each other within the simulation over a 2-year time period. As the focus was consumer fraud, most businesses were only 'semi-simulated', i.e. Business-to-Business interaction was almost entirely out of scope, and merchants and employers were assumed to have infinite stock and cash to meet consumer need. The exception to this was that businesses set up for scams had a few extra behaviour types, as per the scam activity they initiated.

Once the 'baseline' simulation was established, simulated fraudsters were introduced to attempt scams according to the typical typology execution path, for example initiating a text message exchange with a potential victim while impersonating a family member. Each agent in the simulation had a 'vulnerability' attribute which influenced the success rate of each scam attempt. If a scam was successful, then it would initiate payments from the victim to scammer in line with the scam typology.

The interactions between all the actors in the simulation were logged, and then processed into meaningful formats for the TechSprint participants, i.e. flat file tables with a recognisable schema aligned with what would be found in real-world banking and telco datasets.

The participants were then able to 'see' the holistic data picture of dominant 'business as usual' banking and telco behaviour, with data revealing scam activity also present. The

scam behaviour was calibrated so that the activity was more prevalent than in the real world, enabling easier proposition development to meet the TechSprint objectives. In a similar fashion, data problems like quality, matching, and labelling were 'solved' in the data generation process so that the focus would be on the TechSprint questions, and not solving real-world data logistics issues.

For example, in many data-driven initiatives require time-consuming data permission, transformation, linkage, and quality problems to be solved – often on the critical delivery path. By using synthetic data to explore problems, test hypotheses, and create solutions, the value of the potential delivered outcomes can be assessed before (or in parallel with) addressing expensive real data tasks. Furthermore, if the synthetic data is sufficiently high utility and fidelity, the assets developed when processing the synthetic data may be readily recycled for use on the real data. This 1) mitigates 'wastage' of effort and 2) may also better inform the upstream real data task requirements, leading to even better outcomes than might originally have been realised.

Evaluating privacy, utility, and fidelity

In this use case, privacy barriers were non-existent, as no real-world data about individuals or organisations was used, and so the problem reduces to a direct utility-fidelity trade-off. During development, the priority was agreed to be utility. There was high fidelity/low utility data already available in the form of the redacted banking data, and so the synthetic dataset was calibrated to enable analysis of APP fraud across multiple data perspectives:

1. Ability to analyse across different companies in same sector, for example tracking payments from accounts in one bank to another.
2. Ability to analyse across different sectors, for example correlating a scam victim's banking data with their phone or text message records.

That said, it was important that the data appeared realistic and contained recognisable and detectable patterns, and manifestations of key behaviours. For example, for the typical income and spending data, key behaviours like common purchases, bill payments and salaries were included. Networks of synthetic people who knew each other were generated, and individuals more frequently communicated and sent payments within them. Behaviours and communication types were parameterised with a combination of expert knowledge, and reference public domain data.

With these baseline behaviours, the scam pathologies could be implemented. These were designed and calibrated with subject matter experts, as well as quantitatively benchmarked against industry data, such as the UK Finance Annual Fraud report, which defines each distinct scam typology, and the typical frequency and amounts of each. Utility and Fidelity were then evaluated with expert feedback during development, and user feedback during and after the TechSprint.

Trip hazards and risk of poor practice

Agent Simulation is a less common approach to generating synthetic data. Although there are generic off the shelf simulation toolsets available, specific data synthesis tooling is less common (versus artificial intelligence/machine learning methods) at this point in time, and so there is a requirement for very strong data science/engineering expertise to adapt them to this usage. Furthermore, there is a strong reliance on expert knowledge to calibrate and validate the behaviours and output data and achieve the required fidelity and utility.

The focus in this use case has been on utility, and so while the data can clearly be used to build and develop tools like machine learning predictive models, their predictive value in the 'real world' requires careful testing and assessment. The main use case for synthetic data made in this way is for innovating and testing new data concepts where the real data is prohibitively difficult to obtain or create.

Lessons learned

Users continue to provide feedback, which will form the roadmap for future improvements. The range of Business as Usual (BaU) behaviours has been kept relatively simple in the initial release following feedback in the TechSprint, and the option remains to adjust the volume and complexity of the population and their possible actions as users interact with the data. Similarly, the signal/noise ratio of fraud may be adjusted depending on user feedback.

By creating data in this way, with no 'real world' training data and hence zero risk of unintended disclosure of sensitive personal and/or commercial data, it has been possible to release the data to an extended user group, and to then 'crowd source' the potential improvements. Because of this, the utility and fidelity will grow over time as the user group guides the priority enhancements needed.

Regulatory considerations for systems testing and model validation

For the Transaction Categorisation research project, the data sample used was approved for research and analytical purposes and pre-existing scripts were used to remove any potential Personal Identifiable Information (PII).

In live applications/products, depending on the source of the data, limitations on the time windows to retain this data might apply. A potential concern could be around the suitability of using historical data when the economy and consumer behaviours might have significantly shifted from those patterns. This was deemed to be more a generic discussion on building predictive models on historic data – there are approaches to minimise the potential risks – and less a regulatory concern for the use of synthetic data per se.

In the TechSprint use case, the objective of the data synthesis was not to meet a regulatory requirement, nor was it generated/governed within regulatory

guidance. Instead, the purpose was to generate a dataset that would be currently very difficult or impossible to exist either as 'real' data, or as a synthetic double that mimicked real data. The motivation for doing so was to test potential new regulatory, legal, and ethical issues, arising from the potential usage of data in this way. By bringing together academic, industry, regulator, advisory, law enforcement, consumer and other subject matter experts, the analysis of these concepts was made more practical and less abstract and a usable and 'real enough' dataset was generated, helping to drive discussions on these important issues.

Note: this box provides an overview of the important regulatory considerations relating to use case one and two above from SDEG members. These considerations are not a comprehensive list and should not be taken as indication of a policy position or guidance.

Chapter 6

Theme 3: Internal and external data sharing

Overview of theme

6.1 The final stage of the data lifecycle centres on data sharing, exploring how synthetic data can be used to overcome intra- and inter-organisational boundaries that inhibit data sharing. We discuss these questions in the context of two use cases, which use synthetic data to increase the efficiency and effectiveness of public goods. This is accomplished via the provisioning of common data sets for community research into societal challenges and anomaly detection controls for financial crime (fraud and AML).

6.2 In both use cases, we highlight the advantages and challenges of synthetic data. Synthetic data tends to have a lower privacy risk than real data but cannot guarantee privacy; an implementation-specific risk assessment would generally be required.

6.3 Another primary challenge concerns trade-offs between privacy and fidelity (which in turn can affect utility). In practice, public deployments of synthetic data have erred on the side of privacy—for instance, by generating datasets that are based on simulators designed using domain expertise (e.g. simulation and agent-based modelling), rather than based on machine learning models trained on data. This contrasts with techniques often seen in the private release of real data which is utility-led. One notable counter-example is the U.S. census data release. For this, the U.S. Census Bureau used a differential privacy parameter that was larger than that used in the demonstration (that is, adding less noise, thereby providing less privacy but increasing the fidelity and ultimately utility of the release). The increase was justified on grounds that it was necessary to "best balance between the need to release detailed, usable statistics from the 2020 Census with our statutory responsibility to protect the privacy of individuals' data". This illustrates some of the challenges that arise regarding external data sharing.

Current practices

6.4 There have been several regional cross-bank information sharing pilot initiatives conducted in recent years. These include programmes to support better financial crime detection. Example case studies are presented in section four of Financial Action Task Force report on information sharing. These include programmes to support better financial crime detection. However, data privacy concerns and constraints remain an obstacle to wider use. The emerging field of synthetic data has the potential to reduce these obstacles and facilitate wider cross bank sharing of data models. This will help increase community understanding of the best approaches to detect financial crime such as fraud and money laundering and allow the development of more accurate tools.

6.5 The use cases described below suggest several important aspects to considers when using synthetic data for data sharing:

1. Synthetic data can be low fidelity while still having high utility, e.g., for a TechSprint or to compare different downstream models. Hence, data holders should decide what is the minimum data fidelity required to achieve their downstream goals; this determination can guide the choice of synthetic data models and techniques.
2. It is essential to understand if and how the privacy or data use restrictions on the original data also apply to synthetic data. For instance, if the original data is subject to UK/EU GDPR restrictions, the data holder should demonstrate that generated synthetic data cannot be linked to any individual from the source data.
3. There are a lack of industry-standard metrics for evaluating the three important properties of the synthetic data that is generated: privacy, utility and fidelity. As such, it is important for synthetic data deployments to justify their choice of metrics for evaluating these properties. This justification may be driven by the particular use case.

Use case 1: Common data sets for community research into societal challenges

Problem statement

Synthetic data is being increasingly used by government, industry, academic stakeholders to support community understanding of societal challenges and potential solutions. For example, popular machine learning platforms like Kaggle are increasingly featuring synthetic datasets of fraud, and several major banks have documented their use of synthetic data to improve fraud models, among other objectives (See market initiatives section for some examples).

Building on the TechSprint example, we have also recently seen an uptake in collaborative initiatives, such as the US-UK PETs Prize Challenge, which explored how privacy enhancing technologies could be used to drive innovative solutions that enhance artificial intelligence models without revealing sensitive data and maintaining end-to-end privacy guarantees. The Challenge focused on societal issues including pandemic planning and forecasting, and financial crime prevention. The US-UK challenge involved government agencies, regulatory bodies, organisations and assessors from both nations.

Synthetic data methodology

The synthetic data was generated using a multi-layered technique with off-the shelf tools. The initial synthetic data set was then further manipulated using a combination of manual and algorithmic techniques to increase privacy. Data elements with highest sensitivity involving personal data were exclusively created artificially. This layered approach enabled the benefits of specific tools to be utilized that on their own would be insufficient to meet privacy requirements when sharing data externally.

Evaluating privacy, utility, and fidelity

For the PETs Prize challenge, the data is used for benchmarking the effectiveness of solutions, by a diverse community of researchers external to the data provider. As such data fidelity was deprioritised. High fidelity representations of personal data were not considered relevant for such purposes and would have unnecessarily increased data privacy risk. Hence, as the data set was not derived from real, sensitive data, there was no need to quantify the final dataset's privacy using common metrics like differential privacy. While fraud was a component of the data set, the primary objective of the challenge was research into privacy enhancing technologies, including synthetic data, rather than fraud detection accuracy and therefore a high level of fidelity was not essential.

For the PETs Prize Challenge, utility and fidelity were not measured using quantitative metrics. Rather, the data's utility was determined by the contestants' ability to evaluate and compare different technologies and techniques to the data; this is a qualitative notion of utility.

Similarly, the PETs prize data set did not require quantitative measures of privacy. The data set was generated under a principle of 'proportionality' that states that data sets should not contain a higher data privacy risk than is necessary for the specific purpose of the use case.

Trip hazards and risk of poor practice

One challenge that stems from using low-fidelity synthetic data (as in the PETs challenge) is generalisability. That is, what conclusions can one draw from synthetic data that was generated by a (known or unknown) process. For example, for low fidelity approaches such as the synthetic data used for the PET prize challenge, using such data to draw conclusions on the volume or value of fraud a specific solution may detect could be counterproductive as the data distributions do not align with real world distributions.

Note that similar challenges arise in high-fidelity synthetic data; it is often unclear what conclusions one can draw from synthetic data. The level of data fidelity required to obtain similar results with synthetic data and real data is data set and problem dependent.

A data set that produces near identical results to real data with a specific synthetic data generation technique does not imply the synthetic data technique will give similar levels of accuracy on other data sets. A synthetic data generator that models each feature of a data set independently may give accurate results on problems where single variables in isolation are strong indicators of fraud but may give poor results where multi-dimensional relationships are key to fraud detection.

Lessons learned

Lower fidelity synthetic data sets can provide significant value where fidelity is not essential for the use case. For the PETs prize data, significant community insights into the strengths and weaknesses of different approaches to privacy enhancing technologies were obtained and shared within the community.

Transparency on the fidelity of the data is key to ensuring appropriate use and assessing the reliability of any conclusions being drawn from that data.

Use Case 2: Data sharing to increase efficiency and effectiveness of anomaly detection controls for fraud and anti-money laundering

Problem statement

The effectiveness of fraud and anti-money laundering (AML) systems is limited by the breadth of data available within their design and implementation. Organised fraud and AML operations often involve a network of accounts involving active or passive (unwitting) actors spanning multiple financial institutions. Each institution may have different demographic distributions and (perhaps thus) different transaction distributions.

As such, models that benefit from access to transaction data insights from multiple financial institutions have the potential to be more effective than solutions focused on transactions within a single bank.

Synthetic data methodology

This use case is based on research into synthetic data generation for development of machine learning anomaly detection models carried out at a network level using synthetic data derived from cross border transaction patterns across thousands of financial institutions and over 200 countries. This project was a collaboration between a payment network and a national research centre.

Such analytics can be used to create detailed synthetic models of both regular (legitimate) and illicit (money laundering or fraudulent) payment patterns including patterns that may not be observable from the lens of a single institution. A separate notable example with similar objectives to the project described here is Project Aurora.

For accurate machine learning research, a high-fidelity data set was needed for which advanced techniques were necessary to generate the synthetic data. To ensure utility, the synthetic data needed to preserve specific statistical characteristics of the original data, including:

- The diversity of behaviours observed across the different entities within payments. This can range from entities that transact a single time to entities that transact many times a day.
- The temporal aspects of payments. Payments from a specific entity are typically not uniform but occur in bursts or are clustered around specific periods of a day, week or month.
- Entity relationships. Payments can involve multiple counterparties and intermediaries. Accurate modelling of network relationships and topologies across entities was essential for accuracy.
- Preservation of multivariate relationships between transaction variables (e.g. currencies, banks, countries, amounts).
- Preservation of sequences of interactions. Payments are not independent, but instead show sequences where outgoing payments depend on incoming payments that were received by the same entity at an earlier time.

The technologies experimented with the aim to achieve the above goals were refined over time, using a self-built set of benchmarking tools that compared the statistical properties of real data with their synthetic equivalent.

It was observed that the preservation of temporal patterns within the network data was essential for the effectiveness of fraud solutions and typical approaches for synthetic data such as generative adversarial networks (GANs) and diffusion models currently lack such capabilities.

The chosen solution consisted of two parts:

- An *encoder* that captures the detailed statistical properties of the real data.
- A *decoder* that takes the encoded data as input and generates synthetic data as output.

For the encoder the most effective technique discovered was Temporal Graph Networks (TGNS). TGNs were originally developed for use in social media analytics due to their ability to learn complex systems.

For the decoder, originally a Recurrent Neural Network (RNN) was used but this was found to be ineffective in preserving the temporal patterns within the data. It was discovered that by utilizing a Transformer approach (utilising a variant of the Transformer models within Large Language Models) the temporal statistical characteristics of the original data were preserved.

For the specific use case, the TGN when combined with Transformers provided significantly higher fidelity and utility than alternatives while also sufficiently meeting privacy requirements. The applicability of these techniques for a broader set of use cases is however currently unknown and subject of further research.

Evaluating privacy, utility, and fidelity

As mentioned above, benchmarking tools were created comparing performance of real data with synthetic data as an aid to measuring utility and fidelity. In these measures, the real and synthetic datasets are compared to produce metrics such as period lengths of temporal self-similarity, distributions of graph centralities and clustering statistics, and distributions of temporal motifs.

For measuring data privacy especially, privacy associated with aggregated statistics within synthetic data, metrics were derived including Jensen-Shannon distance that compared distance between real and synthetic equivalents. Given the commercial interests of the financial institutions involved, single-payment privacy was inadequate, and some of these measures focused on ensuring a distance between synthetic and real data sets at an aggregate level. This avoids revealing potential sensitive market insights such as market shares, country footprints and/or currency portfolio make-ups.

It is intended that these types of privacy metrics can be used within a synthetic data governance framework where different uses (e.g. internal or external) would require different thresholds. However, whilst this approach was taken for this use case, we note that distance-based privacy metrics alone are known to be ineffective at protecting all

privacy concerns for synthetic data. As highlighted by recent research, membership inference attacks can be a valuable approach to evaluate privacy and identify where data may not be adequately protected against from a privacy perspective.

Trip hazards and risk of poor practice

Synthetic data can play an important role in supporting sharing data with privacy protection. However, there is a risk of insufficient protection through differential privacy could lead to exposure of sensitive information. This is particularly acute when dealing with aggregated data, which might contain unique or identifiable patterns.

Therefore, it is important to determine the level of privacy at priory with consideration of the protection provided. Once a synthetic data set is released the privacy has been fixed, and it is not possible to independently release further synthetic data from the same source data without impacting the overall privacy budget, i.e., two synthetic data releases generated using the same original data reveal more information than the data set either on their own.

Also, when releasing synthetic datasets, it is important to label when the data was generated, the generator type, the data source and the validity of the data. That is, those circumstances in which the data is useful. Synthetic data is in general best released for specific purposes.

Implementing advanced synthetic data generation techniques, such as TNG combined with Transformer models, requires deep expertise and precision. Errors in the process could compromise the synthetic data's relevance and reliability.

The success heavily depends on accurately replicating the complex statistical characteristics of original transaction data, including the diversity of entity behaviours, temporal payment patterns, entity relationships, and multivariate transaction variables. Any inaccuracies in these areas could lead to ineffective or misleading anomaly detection.

Lessons learned

Achieving a balanced approach in synthetic data generation is crucial for maintaining its utility for anomaly detection while protecting sensitive features in the data.

Cross-disciplinary collaborations, as demonstrated between a payment network and a national research centre, are vital for addressing complex challenges such as financial crime, combining domain knowledge with advanced data science techniques.

Engaging in discussions with the owner or provider of the origin/source data is important to comprehend the required level of protection, i.e., the privacy budget, alongside grasping the fidelity needs and key use cases from the data users to gauge utility accurately.

The necessity for continuous refinement and validation of synthetic data generation methodologies is underscored, with benchmarking against real data serving as a vital tool for assessing and amplifying its fidelity and utility, thereby ensuring the effectiveness of developed machine learning models.

Furthermore, establishing robust synthetic data governance frameworks, inclusive of privacy metrics and considerations capable of adeptly managing privacy nuances, is essential, to safeguard data integrity and confidentiality, while optimising its utility.

Regulatory considerations for internal and external data sharing

As explored across the two use cases in this section, elements around data privacy are important for internal and external data sharing. No (synthetic) data can contain or imply the identity or personal details of individuals, legal entities or financial institutions. Even single words with free-text can be problematic if that word is associated with a well-known company. It should not be possible to infer the identity of an actor within a transaction by cross referencing the transaction with other public or private data. Examples include: the amount or time of transaction can act as a signature if sufficiently unique.

As detailed in use case two, the system should not reveal commercially sensitive information even within aggregated statistics and distance-based privacy metrics alone are not always effective at guarding all privacy concerns. For example, by analysing the statistical characteristics of the synthetic data, information may be inferred on market share distributions or activity totals within a specific business area. Whether such inferences are commercially sensitive can be use case specific. *Model stealing attacks* within encoder/decoder models were identified as a specific risk that required a risk assessment from third party domain experts.

Note: this box provides an overview of the important regulatory considerations relating to use case one and two above from SDEG members. These considerations are not a comprehensive list and should not be taken as indication of a policy position or guidance.

Chapter 7

Considerations of creating synthetic data

7.1 Across the themes and use cases outlined above, there are a number of lessons learned and important points for practitioners to consider when creating synthetic data. Effectively generating synthetic data requires careful consideration of different technical, regulatory and ethical elements. It is important that when the relative merits of using synthetic data are benchmarked, where possible, against alternative techniques and approaches. The points below summarise considerations of creating synthetic data as derived from the report. They do not represent formal guidance or recommendations relating to synthetic data creation.

7.2 From the use cases above, there are several considerations relating to synthetic data creation:

1. **Regulatory/legal input:**
   There are regulatory considerations associated with synthetic data creation and usage, which vary according to the methodology, type of data and region. For examples, the use case may include personal data which requires a lawful basis for processing; this would further entail, data retention policies to be introduced as well as, adherence to data protection principles set out in the UK GDPR (and/or other international data protection laws where appropriate).

2. **Dependency on real world data:**
   The quality and accuracy of real-world data can impact the synthetic data being generated, understanding how the synthetic data is generated is an important factor to assessing the appropriateness for each use case. Where data is dynamic, careful monitoring and updates are needed to keep them up to date.

3. **Methodology:**
   It is still an open question regarding what generation methods and combinations of real-life and synthetic data are optimal for model accuracy, ethically desirable and aligned with regulatory frameworks. Where possible, it is important to test multiple generation methods and stress test what combination of real and synthetic data leads to better evaluation metrics. The selection of appropriate technologies and models is also important depending on the use case (e.g., different models might be appropriate for generating, text, tabular data, timeseries images etc. or depending on the distribution of data, or the amount of available real data).

4. **Deployment of data:**
   The accepted degree of privacy, utility and fidelity of synthetic data is likely to be influenced by the intended use of the data, for example, whether it will be used in public domains or within private system testing. In the former, privacy is usually prioritised whereas utility may take precedence in private use cases.

5. **Combination of Privacy Enhancing Technologies (PETs):**
Synthetic data is one technique that can be used to protect data privacy and is often used alongside other PETs such as differential privacy, secure multi-party computation and federated learning. It is important to be able to assess where other PETs may enhance the synthetic data or be better suited to use instead of synthetic data, especially where there may be concerns with data leakage. Where possible, comparing synthetic data to an appropriate baseline can help to understand whether synthetic data is the best approach. This can include the use of alternative privacy enhancing or traditional anonymisation techniques, experts' knowledge of the statistical properties or simulated data to critically evaluate the strength of synthetic data.

6. **Biases in generated data:**
Biases and inaccuracies in real world data can be replicated in synthetic datasets. Data and model testing and assessments are needed once the data have been generated to help address any biases that are present.

7. **Validating the data:**
Understanding the privacy, utility and fidelity aspects of the synthetic data set is very important. Acceptable levels of privacy, utility and fidelity of the data should be considered on a case-by-case basis and according to the deployment of the data. For example, privacy is more likely to be the priority element where synthetic data is shared externally than in use cases where synthetic data is created for internal usage and no personal identifiable information is used.

8. **Technical and Domain expertise:**
A strong level of data science/engineering know how and close collaboration with subject matter experts is needed to balance the variety of methodological choices whilst tailoring synthetic data to the use case in question.

9. **Resource and feedback:**
Generating and using synthetic data is an iterative process which takes resource, including personnel time. The process can be streamlined by first understanding which stakeholders inside an organisation should be included in the generation and/or evaluation phases; these may include legal teams, compliance professionals and executives. It is important to consider whether an organisation has the appropriate talent and skills required to generate and use synthetic data. Gathering feedback from fellow data scientists, developers and users following deployment can help to improve synthetic data sets.

Chapter 8

Conclusion

8.1 Synthetic data presents a valuable resource for a variety of financial services use cases, including understanding consumer behaviours, enhancing financial operations and as a tool to develop collaborative solutions to important societal issues. In the examples explored throughout this report, synthetic data plays a crucial role across the data life cycle, suggesting diverse applications to help address broader public challenges in financial services. In these use cases, synthetic data emerges as a tool with the potential to help augment data quality and quantity, mitigate biases, and facilitate the testing of systems and models whilst protecting privacy concerns, whilst still subject to the fundamental validation trade-off. Moreover, synthetic data can contribute to overcoming organisational boundaries by fostering collaboration and advancing interdisciplinary efforts to address societal challenges.

8.2 The quality of machine learning and artificial intelligence models is intricately linked to the quality of their data inputs. Synthetic data emerges as a valuable technique among various machine learning approaches, to enhance the performance of data-driven machine learning models. In practice, synthetic data could contribute to a range of applications, from fraud detection and reject inference to the generation of synthetic identities and customer patterns for testing open-banking solutions.

8.3 From the use cases explored in this report, there are several important considerations for practitioners to think about when creating synthetic data. These are centred around ensuring clarity on regulatory and legal aspects which may vary depending on the approach and the use case. Additionally, thorough evaluation of the methodology, including whether synthetic data is the best approach, whether other privacy enhancing technologies alone or in combination with synthetic data may improve model quality, should be considered.

8.4 Validating synthetic data is an ongoing challenge for practitioners, and existing applications indicate various approaches are currently taken to evaluate aspects such as privacy, utility and fidelity, and often vary according to the specific use case and where the data is being deployed. The use cases in the report reflect the importance of needing a high level of technical and domain expertise within the generation process and, the value of gaining feedback from users and other data scientists on the synthetic data to make continuous improvements.

8.5 Organisations need to be thoughtful of ethical considerations and intentional about the use of synthetic data beyond technical development. To effectively generate synthetic data, firms are likely to benefit from introducing internal governance processes that support the responsible and legal use of synthetic data. This includes being able to communicate effectively with technical and non-technical stakeholders to understand the risks and build appropriate governance procedures alongside internal guidelines.

8.6 The integration of synthetic data within the financial services landscape represents a forward-looking approach to harnessing the power of data. As organisations navigate this evolving landscape, strategic governance, developed with ethical considerations of the points above, will help to realise the full potential of synthetic data.

Chapter 9

Next steps

9.1 Synthetic data has a role to play in addressing societal challenges requiring collaboration, for example, between organisations, regulators, law enforcement, academics and other subject matter experts. The next phase of the SDEG will look to further understand where synthetic data is used to drive beneficial innovation in financial services, and how stakeholders can collaborate to overcome potential barriers.

9.2 Synthetic data presents a potential solution to data scarcity and quality issues by bringing different stakeholders together to generate novel data sets. However, there are still open questions for practitioners to consider including when it is ethically permissible to use synthetic data. Through the SDEG, we hope to enable beneficial innovation and contribute to the safe and responsible development, usage, and deployment of synthetic data in financial services.

9.3 To engage with us on synthetic data related matters, discuss the report or contribute to the collaboration framework, please reach out to SyntheticDataEG@fca.org.uk.

Appendix 1

Group members and acknowledgements

1. The use cases in this report are authored by the SDEG members. With thanks to their contributions, insights and expertise that have shaped this report.

2. The FCA launched an expression of interest for the Synthetic Data Expert Group in February 2023 and the members were appointed in March. The SDEG is a sub-group of the IAG and operates within the IAG Terms of Reference and is chaired by the FCA.

Members of the Synthetic Data Expert Group:

- Alexandra Ebert, MostlyAI
- Caroline Louveaux, Mastercard
- Carsten Maple, University of Warwick
- David Buckley, Responsible Technology Adoption Unit
- David Tracy, Smart Data Foundry
- Elena Strbac, Standard Chartered
- Giulia Fanti, Carnegie Mellon University
- Ismini Psychoula, Ofcom
- Janet Bastiman, Napier
- June Brawner, The Royal Society
- Lee Gregory, Barclays
- Luk Arbuckle, Privacy Analytics
- Lukasz Szpruch, Alan Turing Institute
- Marilena Karanika, Experian
- Michael Meehan, Howso
- Nick Clark, Cambridge Regulatory Innovation Hub
- Oxana Samko, HSBC
- Paul Comerford, Information Commissioner's Office (ICO)
- Robin Glover, Swift
- Tom Fiddian, Innovate UK

FCA Contributions

3. Emelie Bratt, Henrike Mueller, Leo Gosland, Matt Lowe, Pavle Avramovic, Rebecca Humphry, Simran Singh and Yu Bian.

Acknowledgements

4. With thanks to the BIS Innovation Hub Nordic Centre Team and Imperial College London who provided feedback on the report.

Appendix 2

FCA synthetic data journey

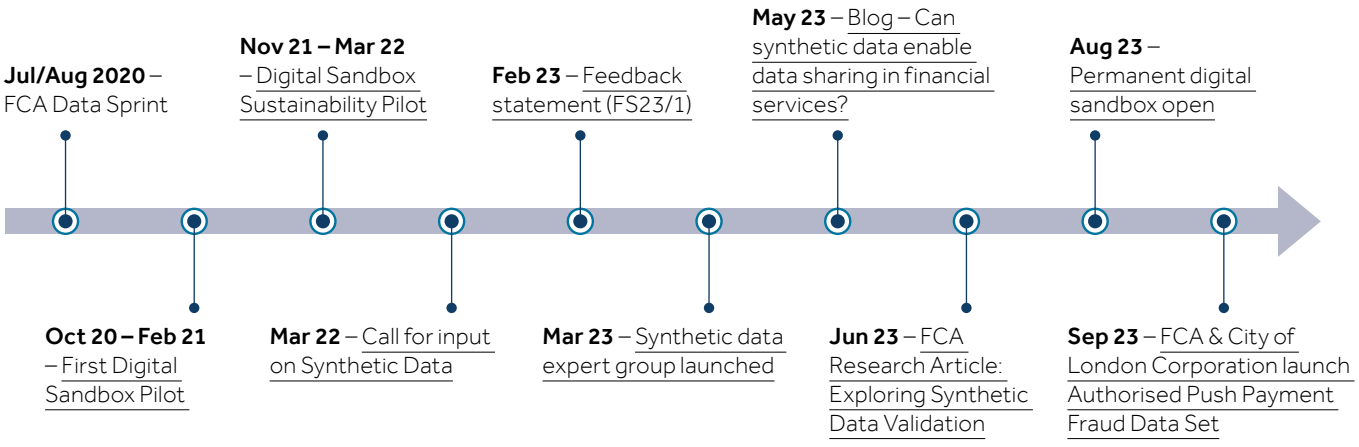
1. In advancing its commitment to innovation, the FCA has undertaken a number of synthetic data related initiatives, commencing with a DataSprint in June 2020. During the DataSprint multi-disciplinary teams collaborated in the digital sandbox to create synthetic datasets mimicking real-world financial scenarios to help innovators develop new solutions. Focusing on banking, lending, blacklists, and telecommunications, the sprints aimed to combat fraud, support financial resilience, aid vulnerable consumers, and enhance access to finance for SMEs.

2. Building on this foundation, the FCA, in partnership with the City of London Corporation, launched two digital sandbox pilots. These initiatives granted participants access to synthetic data assets, an API marketplace, coding environments, and expert mentors.

   i. The first pilot, involving 28 organisations, underscored the value of synthetic data, emphasising the need for more referentially linked datasets and finer granularity. Subsequently, the Kalifa review recommended a permanent digital sandbox, a testament to the initial pilot's success.
   ii. The second pilot, from November 2021 to March 2022, focused on solving regulatory challenges related to Environmental, Social and Governance (ESG) data and disclosure. The synthetic data generation journey for the second pilot aimed to deepen the FCA's understanding of data generation requirements, process and methodologies.

3. Building on the evaluation from the two pilots, the FCA officially launched the permanent digital sandbox in August 2023, followed by the release of Authorised Push Payment Fraud Synthetic data in September 2023. This comprehensive offering invites applications from innovators, data providers, and mentors. Throughout this support lifecycle, the FCA actively collects and evaluates opportunities and challenges associated with synthetic data, informing our understanding of its impact to the UK financial service industry, consumers and markets.

4. These initiatives help the FCA to grow technical understanding on existing barriers to synthetic data preventing innovation and positively shape digital markets by working with industry to tackle societal issues.



Appendix 3

Glossary

1. The terms below are not regulatory definitions, they included solely for the purpose of providing explanations and clarification in reference to the terms used in this document.

Word	Description
Agent-based modelling	A simulation modelling technique to analyse a system by its individual agents and associated interactions (Bonabeau, 2002).
Data Protection Act	The Data Protection Act 2018 is the UK's implementation of the General Data Protection Regulation (UK GDPR). It controls how personal data is used by organisations, businesses or the government.
Digital Sandbox cohort	The Digital Sandbox cohort is an 11-week initiative hosted by the FCA and the City of London Corporation, designed to stimulate and foster the development of innovative products and solutions within financial services. Participants are given access to data, mentors and collaboration platforms to prototype and test their proof of concepts, with the aim of reducing time to market.
Fidelity	Refers to measures that directly compare the synthetic dataset with the real dataset i.e. the statistical similarity of the synthetic dataset to the input real data (ICO).
General Adversarial Network	A type of machine learning model where two neural networks engage in a competitive process employing deep learning techniques to enhance the precision of their predictions.
Privacy	Measures the risk that specific individuals (or other sensitive data) can be re-identified from the synthetic dataset.
Personal Identifiable Information (PII)	Any representation of information that relates to an identified or identifiable individual, either directly or indirectly.
Synthetic data	Microdata records created to improve data utility while preventing disclosure of confidential respondent information. Synthetic data is created by statistically modelling original data and then using those models to generate new data values that reproduce the original data's statistical properties. (ONS)
TechSprint	The FCA TechSprints are events that bring together participants from across and outside financial services to develop technology based ideas or proof of concepts to address specific industry challenges. The events usually last between 2-5 days, and help us to shine a light on issues and expand the discussion and awareness of potential solutions.
Utility	A synthetic dataset's 'usefulness' for a given task or set of tasks, for example for training AI or Machine Learning models (ICO).

Sources: Office for National Statistics (ONS), Information Commissioner's Office (ICO), Alan Turing Institute, Bonabeau (2002)

Appendix 4

References

Ashwin et al. (2016), Motifs in Temporal Networks

Bank for International Settlements (2023), Project Aurora

Bonabeau, E. (2002), Agent-based modeling: Methods and techniques for simulating human systems

Duddu et al. (2022), Quantifying Privacy Leakage in Graph Embedding

El Emam et al. (2020), Practical Synthetic Data Generation

FCA (2022), Authorised Push Payment Fraud TechSprint

FCA (2022), Supporting innovation in ESG data and disclosures – the digital sandbox

FCA (2021), Supporting innovation in financial services: the digital sandbox pilot

FCA (2023), FS23/1 - Feedback Statement on Synthetic Data Call for Input

Financial Action Task Force (2022), Partnering in the Fight Against Financial Crime: Data Protection, Technology and Private Sector Information Sharing

Ganev, G. and Cristofaro, E.D. (2023), On the Inadequacy of Similarity-based Privacy Metrics: Reconstruction Attacks against "Truly Anonymous Synthetic Data''

Herbold, S. and Haar, T. (2022), Smoke testing for machine learning: simple tests to discover severe bugs

IBM (2024), What are recurrent neural networks?

ICO (2023), Final Guidance on Privacy-Enhancing Technologies (PETs)

Jordan, J. et al. (2022), Synthetic Data - what, why and how?

National Institute of Standards and Technology (2024), SDNist: Synthetic Data Report Tool

Potluru, V.K. (2023), Synthetic Data Applications in Finance

Royal Society (2023), Privacy Enhancing Technologies

HM Treasury (2021), Kalifa Review of Fintech

Rossi et al. (202), Temporal Graph Networks for Deep Learning on Dynamic Graphs

Vaswani et al. (2017), Attention Is All You Need

All our publications are available to download from www.fca.org.uk.

Request an alternative format

Please complete this form if you require this content in an alternative format.

Or call 020 7066 6087

Sign up for our **news and publications alerts**

Pub ref: 1-008184

12 Endeavour Square London E20 1JN
Telephone: +44 (0)20 7066 1000
Website: www.fca.org.uk