FINANCIAL CONDUCT AUTHORITY

Synthetic data to support financial services innovation

March 2022

How to respond

We are asking for comments on this Call for Input by **22 June 2022**.

You can send them to:

RegTech and Advanced Analytics
Innovation, Data and Technology
Financial Conduct Authority
12 Endeavour Square
London E20 1JN

Telephone:
020 7066 4250

Email:
syntheticdata-callforinput@fca.org.uk

Moving around this document

Use your browser's bookmarks and tools to navigate.

To **search** on a PC use Ctrl+F or Command+F on MACs.

Contents

1 Introduction 4
2 Data access and innovation 6
3 Synthetic data 8
4 The role of the regulator 14
5 Next steps 16

Annex 1
Table of questions in this Call For Input 17

Annex 2
Glossary of terms used in this document 19

Executive Summary

Financial services are increasingly digital. Consumer interactions and engagement with products and services build digital footprints, generating ever increasing amounts of data that can reveal much about our identity and behaviours. It is vital that this data is protected and that consumers' right to privacy is safeguarded.

Simultaneously, this data can also drive valuable innovation. Advances in data and analytics can enable automation, improved decision-making and risk management, and personalisation of services. Ultimately, this has the potential to deliver societal benefits such as greater market efficiency and integrity, financial inclusion, and the prevention of financial crime.

It is widely recognised that artificial intelligence (AI) holds significant potential in the financial services industry. However, the extent of its potential will depend on the wide availability and accessibility of data to innovators who will build the next generation of products and services.

Financial data, while valuable, is highly sensitive, and is subject to data privacy laws that place conditions on sharing this data for innovation and research purposes to protect the privacy of consumers. There are conditions where financial data can be readily shared in accordance with data privacy laws, however third-party providers such as RegTechs and FinTechs will often have to spend a period of time navigating complex due diligence and onboarding processes with an institution to access this data. Whilst data privacy laws are critical to the protection of consumers' privacy rights, the challenges associated with access to financial data, especially for new market entrants, can inhibit the development of new products and services in the market.

We are therefore interested in solutions that will enable greater data sharing for the purposes of competition, without undermining data protection laws that are in place to protect consumers.

'Synthetic' data is a privacy preserving technique that could open up more opportunities for data sharing by generating statistically realistic, but 'artificial' data, that is readily accessible. Synthetic data is already used across other sectors, such as robotics and autonomous vehicles, and there is an early but growing level of exploration within financial services. We have already explored synthetic data through our TechSprints, and more recently the Digital Sandbox pilot, where we have sought to alleviate the data access challenge by making synthetic financial data available to participating firms. Participants in the Digital Sandbox pilot cited synthetic data as the most valuable feature on the platform, while simultaneously calling for development of these data assets to enable more effective testing and product development.

With this Call For Input, we would like to conduct an introductory exploration of market attitudes towards synthetic data, and its potential for opening data sharing between firms, regulators and other public bodies. We want to understand industry views on the potential for synthetic data to support innovation and the requirements to be effective, as well as potential limitations and risks.

1 Introduction

1.1 We see innovation as a vital component of effective competition. As well as providing novel and inventive solutions to meet consumers' needs, innovation enables agile start-ups to enter the market and challenge incumbents, while driving incumbents to compete harder to retain customers. Technological innovation can also reduce operating costs, improve efficiency, and more effectively manage risk. Responsible innovation can lead to outcomes that improve market integrity and ultimately lead to downstream benefits for consumers such as more affordable products and services which more effectively meet their needs.

1.2 We have developed a variety of tools to foster an innovation friendly environment and culture in the UK in line with our competition objective.

1.3 Through initiatives such as the Regulatory Sandbox, TechSprint programme, Direct support, and Advice Unit, we have engaged with and supported hundreds of innovative firms. Running these market-facing services has also given us rich and deep insights into the challenges and opportunities facing innovators. In an ever more digital landscape, we are keen to ensure our services continue to support the needs of an industry moving at pace, and to explore new ways we can support responsible innovation.

1.4 Innovation within financial services is increasingly driven by access to large volumes of high-quality data, with new innovations underpinned by advances in AI and machine learning, cloud computing, big data, and Open Banking infrastructure.

1.5 On the consumer side, as financial services become more digital, our interactions with products and services creates a larger digital footprint and exponentially more data points. Both new market entrants and incumbents delivering large-scale digital transformation programmes are attempting to turn this metadata into business intelligence and unlock new avenues of value for consumers or operational efficiencies. Correspondingly, we have witnessed an increasing demand from the market for regulatory support services that focus on data and data access.

1.6 However, financial data, such as consumer transaction records, account payments, or trading data, is sensitive personal data subject to data protection obligations, as well as often being commercially sensitive. In instances where data is shared between or even within an organisation, it is vital that the appropriate protections and safeguards are in place to protect individuals' privacy. Consequently, the process of sharing data can require several months of complex due diligence and onboarding processes. The difficulty in accessing financial data has created challenges particularly for new markets entrants and third-party providers such as RegTechs.

1.7 Simultaneously, we are witnessing the development of privacy preserving techniques to enable greater data sharing in this industry. 'Synthetic' data is one such technique which we have explored through our TechSprints, and more recently the Digital Sandbox pilot, where we have sought to accelerate the validation and testing of products by making synthetic financial data available to participating firms.

1.8 Synthetic data is not 'real' data created naturally through real-world events, rather it is 'artificial' data, generated using algorithms. The benefit of using synthetic data is that it simulates real data without identifying specific individuals, therefore as long as no real individuals can be identified from the synthetic data, data protection obligations, such as GDPR, do not apply. Synthetic data is created by observing patterns and the statistical properties of real data, and using algorithms to replicate these patterns within the synthetic dataset, aiming to make it a realistic replica of the real data. Whilst the utility and analytical value of the synthetic datasets are dependent on the quality of the model and data used to generate them, these 'artificial' datasets can be shared for a wide range of uses.

1.9 Cutting-edge techniques for synthetic data generation are still being developed, and there is ongoing research into the level of privacy risk that should be attached, i.e., whether it can be possible to de-anonymise a real individual from a synthetic population under certain circumstances. Alternative privacy enhancing techniques, including differential privacy, can be used in such circumstances to further reduce the risk of de-anonymisation, although statistical outliers may continue to pose a distinct privacy risk. However, synthetic data is generally considered to be one of the most promising tools for sharing data where privacy protection is a prerogative. Synthetic data can also be generated at scale, making it a viable tool for training AI in environments with strict data privacy controls.

1.10 Recent advances in technology, research, and computational power have led to improved synthetic data accuracy and stronger privacy preserving guarantees. In turn, there has been an increase in experimentation with synthetic data across industries such as automotive and robotics, healthcare and medicine, logistics, and government national statistical offices for policymaking and research.

1.11 To ensure that our innovation services remain fit-for-purpose in the digital age, we want to better understand different market participants' views on the extent to which synthetic data can expand data access and data sharing opportunities in the market. We are also seeking to evaluate the maturity of synthetic data usage within financial services, and the extent to which both regulated and unregulated firms are currently using it. Finally, we are interested in what industry sees as the role of the regulator, particularly regarding our competition remit.

Next steps

1.12 This Call For Input will be of interest to:

- academics
- incumbents
- start-ups
- RegTechs and FinTechs
- technology firms
- regulators and policy-making bodies

1.13 We welcome discussion and feedback. Please share your views, including responses to our questions by 22 June 2022. A complete list of questions in this Call For Input is available in Annex 1. Responses do not need to answer every question.

2 Data access and innovation

2.1 For many firms, access to financial data is key to developing new and innovative products and solutions.

2.2 The use of artificial intelligence (AI) and machine learning (ML) has the potential to disrupt financial services. AI is often used as a catch-all term for a host of different technologies and approaches, but broadly it can be defined as the theory and development of computer systems able to perform tasks which previously required human intelligence. ML is one of many sub-categories of AI. Unlike traditional 'rules based' algorithms where rules are created from human knowledge and experience, ML is a technique where computer programmes fit a model or recognise patterns from data, without being explicitly programmed and with limited or no human intervention.

2.3 Over recent years, computational power and the volume of data that is collected and processed has grown exponentially. This has led to advances in ML techniques, and for ML models to become an order of magnitude larger and more sophisticated than traditional techniques. As a result, ML models can often make better predictions than traditional models or find patterns in large amounts of data from increasingly diverse sources.

2.4 The application of this technology within businesses is still relatively early, however it has the potential to make financial services more accessible, more efficient, and to improve consumer outcomes. Current data-driven innovations are being deployed in areas such as:

- financial crime and fraud prevention
- customer engagement
- credit scoring
- sales and trading
- insurance pricing and insurance claims management
- asset management and portfolio optimisation

2.5 In response to a joint BoE-FCA survey in 2019, 67% of regulated firms indicated that they are using ML in live production environments. Likewise, industry surveys report similar findings to attitudes towards AI and ML: 81% of C-suite respondents in financial services saw AI as important to their company's future success, and more than half thought it gave them a competitive advantage. However, potential new market entrants who might once have competed with incumbents, or third-parties that would offer services to incumbents, face innovation barriers in accessing large volumes of high-quality data required to develop and implement these types of strategies. In the case of a RegTech for example, it is very difficult to build a new ML based solution without beforehand going through complex due diligence and costly onboarding processes with an institution to access their data.

2.6 The transformation potential of these technologies therefore depends on the wide availability of data to innovators. Accessing data at an individual level is possible through mechanisms such as consent, for example through Open Banking infrastructure, but to truly develop these technologies requires widespread access to large data sets. The widening data gap risks inhibiting competition, where new entrants

and challengers lack the raw material (data) required to develop technology and strategies to compete with incumbents.

2.7 Evidence gathered from our TechSprint programme and Digital Sandbox pilot has further demonstrated that access to data is a missing rung on the ladder for firms in the early-stages of developing new products and services.

2.8 In the first Digital Sandbox cohort, despite limited subject areas that firms could apply for, we received 94 applications within a 4-week application window. Of the 28 successful firms in the cohort, 18 (64%) were developing products/solutions underpinned by machine learning. There were over 850,000 API calls to the synthetic data sets provided, and 92% of participants identified the synthetic data as the most important feature of the Digital Sandbox programme. Interestingly, this is despite the fact that the synthetic data available was only generated to be of sufficient quality to meet a minimum standard needed for a pilot cohort. Further feedback from the participating firms was that higher-quality synthetic data (higher fidelity, referentially linked, and additional data sets) would have even further accelerated development and enabled rapid creation of accurate models.

2.9 Making sensitive data available for innovation or research purposes is a significant challenge. Data relating to customers (such as banking or transactional data) contains personal data and is subject to obligations under GDPR and the UK Data Protection Act.

2.10 We believe that access to data is a key driver of modern innovation, and that synthetic financial data could play a role in supporting innovation and enabling new entrants to develop, test, and demonstrate the value of new solutions.

Q1: **How important do you think access to data is for innovation within financial services? What else do you view as significant barriers to innovation?**

Q2: **Do you agree that it is challenging to access high-quality financial data sets? If so, specifically what challenges do you face? (for example, understanding legal requirements around data access, commercially expensive, or technology infrastructure.)**

3 Synthetic data

3.1 The UK Office of National Statistics defines synthetic data as: '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.'

3.2 Currently, synthetic data is used for a wide range of activities and industries. We are focusing this Call For Input on financial services, so the synthetic data we are discussing is predominantly assumed to be alphanumeric. However, synthetic data can also include unstructured data such as text, speech, and visual media including images and video, and relationships between data points such as hierarchical and network models. A well-known type of synthetic data is so-called 'deep fakes' (a portmanteau of 'deep learning' and 'fake'), which produces realistic looking but computer-generated media. These are created using 'Generative Adversarial Networks' (GANs) which, as discussed below, can similarly be used to extract and model synthetic numerical data.

3.3 In financial services, synthetic data is being used as test data for new products and tools, for model validation, and in AI model training. Many problems of modern AI come down to insufficient data: either that the available datasets are too small, insufficiently labelled, or cannot be accessed without breaching individuals' privacy rights. Furthermore, historical data is often biased and unrepresentative, and algorithms trained with this data will replicate these biases. Synthetic data in principle offers solutions to these problems.

3.4 We see three overarching benefits for synthetic data:

- Data Privacy – when data privacy requirements make collecting, sharing, and accessing real data difficult or with prohibitive timeframes. This is common in fields such as healthcare and finance where data sets frequently contain sensitive personal information.
- Real data is limited or does not exist – where the data required is rare, does not exist in sufficient quantities for training purposes, or it does not yet exist and must be simulated for as yet unencountered conditions. Synthetic data can be used to model realistic but potentially unlikely or uncommon scenarios, for example for risk management and stress testing.
- Cost efficiency – large volumes of training data are needed for training accurate machine learning algorithms. However, it can sometimes be more efficient to generate high volumes of synthetic data than capture and/or label real data. One of the main applications of synthetic data today is in computer vision for autonomous vehicles. The software is trained on synthetically generated images rather than capturing and labelling millions of hours of real-life footage.

Q3: Do you agree with the high-level benefits for synthetic data? Are there any other benefits for synthetic data for your organisation, both now and in the future?

Q4: **Does your organisation currently generate, use, purchase or otherwise process synthetic data? If possible, please explain for what purpose(s).**

Data Privacy and privacy preserving techniques

3.5 The ability to share data among different entities has valuable applications across many different industries and sectors, and the access challenge is commonplace. Consequently, several advanced statistical techniques to generate synthetic data have been developed in an attempt to enable data-sharing while preserving the privacy of data subjects. There is no perfect solution to this utility-privacy trade off as all privacy preserving techniques will result in some loss of data utility compared to real data. However, the goal of these techniques is to maintain as much utility as possible while preserving an individual's privacy.

3.6 An important consideration for financial data is consumer sentiment. A 2013 study[1] revealed that individuals across the US and EU ranked their financial data as the most sensitive data type, above health and genetic data, location data, information about children, telephone call history, internet history, and email, for example.

3.7 The traditional approach to overcoming this data access challenge has been anonymisation, a process of removing personal identifiers, direct and indirect, that could result in an individual being identified. However, there has been a well-documented history of de-anonymisation of anonymised data, undoing the protections it is meant to create.

3.8 In a 2015 paper[2] published in the American Journal Science, researchers were given access to an anonymised dataset of 3 months of credit card transactions of 1.1 million users in 10,000 shops in an OECD country. Despite name, account number, and all other identifiable information stripped out of the datasets, the researchers demonstrated that 90% of individuals could be re-identified by combining the anonymised dataset with limited amounts of outside data – in this case 4 additional 'spatiotemporal points'. In other words, knowing an individual's location on four different days would be enough to identify them in the anonymised dataset, in 90% of cases.

3.9 Requiring additional outside data to de-anonymise a dataset may initially appear to be a significant hurdle; however, the increasing size of the data footprint individuals create through modern digital lifestyles means there is an ever present and expanding privacy risk. There have been a number of real life examples where seemingly fully anonymised datasets released publicly have been de-anonymised to reidentify individuals.

3.10 Pseudonymisation is an alternative privacy preserving technique which, similar to anonymisation, aims to replace or remove personal data that can be attributed to an individual. The technique may involve replacing names or other identifiers that are easily attributed to individuals in order to reduce the links between a dataset and an individual's original identity. Pseudonymisation allows organisations to process personal data in accordance with data protection laws, although pseudonymised

1 Rose et. Al, 2013
2 De Montjoye et. Al, 2015

personal data is still classed as personal data, and remains within the scope of the UK GDPR. Furthermore, as with anonymisation, it is possible to reverse engineer pseudonymous data to reidentify individuals.

3.11 As the limits of anonymisation and pseudonymisation became better understood, a more recent privacy preserving approach known as 'differential privacy' has emerged. A simplified explanation of the techniques used to achieve differential privacy is that they work by adding random noise into a dataset or generator algorithm. Adding this randomness provides a mathematically provable guarantee of privacy protection against attempts to learn private information related to specific individuals for the purposes of reidentification.

3.12 Differential privacy provides stronger protection than anonymisation and pseudonymisation and has also been described as 'one of the most promising anonymisation techniques' by the European Data Protection Board. However, a challenge of this approach is that queries to a differentially private dataset should be limited by a 'privacy budget', simply meaning a limited number of attempts allowed at querying the dataset. The purpose of limiting the number of queries is to prevent unlimited non-trivial queries on a dataset as the whole dataset could be revealed that way. This approach therefore requires the monitoring of the privacy budget over time.

Synthetic data generation approaches

3.13 Synthetic data is a further technique that is fast gaining traction to enable privacy preserving data sharing. As discussed above, **synthetic data can be thought of as 'artificial' data that has been generated to accurately represent real data.** By understanding the statistical relationships in a real dataset, a model can be built which replicates those relationships in an artificially created (fake) dataset, which should not in theory contain any real personal data.

3.14 The concept of synthetic data was first proposed by Rubin (1993), while working to release privacy-preserving versions of US Census Bureau data. Due to exponential increases in computational power and AI research, there are an emerging range of techniques to generate synthetic data with far greater levels of accuracy. Broadly, these can be classified as:

- **Data-driven methods**, which take a sample of real data and use a model to generate new data with the same inherent statistical distribution:
  - Extracting data relationships from a distribution: This approach works by observing real statistical distributions and reproducing them in a fake dataset. It can also be performed using generative models.
  - Deep learning models and GANs: these approaches use neural networks. A neural network is a series of algorithms that attempt to recognise underlying relationships in a dataset through a process that mimics the way the human brain operates. Generative adversarial networks (GANs) are used to improve the accuracy of the synthetic data. A GAN features two algorithms that face off against each other, one known as a 'generator' and one known as a 'discriminator'. The discriminator has been trained to spot fake data, and the generator will iteratively produce new data until the discriminator can no longer tell the difference between the real and the fake data.

- **Process-driven methods**, which generate new data by mimicking the underlying process that formed the real data in the first place:
  - Agent-based modelling: The actions and interactions of individuals (agents) within a dataset are modelled in order to understand the behaviour of the system as a whole and predict its outcomes. Although significantly more challenging than just matching statistical patterns, this approach may lend itself better to use cases where there is a need to model data for scenarios that have not happened in real life, for example stress testing or rare events.

3.15 There is no single best generative technique, and different approaches will be more or less effective depending on the characteristics of the underlying dataset which needs to be synthesised. Furthermore, as with many privacy preserving techniques, synthetic data generation will diminish the utility of the data to some extent, depending on the techniques used.

Q5: **If your organisation generates synthetic data, please describe at a high level the techniques used. Why have you chosen to use this approach?**

Q6: **What do you see as the difficulties and barriers for firms in creating high-utility, privacy secure synthetic data?**

Q7: **Does your organisation engage with privacy enhancing technologies or privacy preserving techniques other than synthetic data? How would you assess the utility and benefits of synthetic data in comparison to other techniques?**

Synthetic financial data for financial services innovation

3.16 The term 'financial data' comprises an incredibly broad spectrum with hundreds of subdomains and alternative data types. As a result there is no single best approach to synthetic data generation – a more pragmatic approach may be to build up a library of models over time, specific to individual use cases.

3.17 The potential for synthetic data to alleviate the data-access challenge for innovation purposes has been discussed above. In addition, access to readily available synthetic data could have the following benefits:

- Data can be made available from third-parties, such as RegTechs and B2B FinTechs, to construct better models and develop new techniques or use computational resources that might be unavailable to incumbent holders of sensitive data.
- These third-parties could pool synthetic data from multiple sources, revealing trends, patterns and insights that are more accurate, or indeed only apparent in the pooled data. This could have valuable applications in detecting and preventing financial crime for example, by facilitating cooperation between multiple organisations that are prevented from efficiently co-operating at a granular data level. Pooling synthetic data however increases the risk of de-anonymisation, and so third parties would need to exercise caution when engaging with this approach.
- Synthetic data and data generation techniques could be either shared or made publicly available, which could potentially provide an important step towards

reproducibility of results. In a future world with more common AI processes, this may be an important compliance mechanism.

3.18 During our Digital Sandbox Sustainability cohort, a number of firms requested synthetic ESG-related data in order to train and develop algorithms designed to identify cases of 'greenwashing'. We would like to understand the range of use cases at this level across financial services, as well as the current maturity and level of synthetic data use by both regulated and unregulated firms within financial services.

Q8: **What do you see as the highest priority use cases that would benefit from synthetic data?**

Q9: **Are the synthetic data use cases you have mentioned significant for early business phases or mature operations/processes within your organisation?**

Q10: **How would your organisation make use of synthetic data if it was available (if at all)?**

Q11: **What synthetic data sets would you find most valuable to have access to? For example, Open Banking, Customer profiles, account to account payments, Credit card transactions, trading data, etc. What challenges would these data sets help your organisation to solve? E.g. AML and fraud detection, ESG, etc. Please be specific.**

Q12: **What requirements would you need for the synthetic data to feasibly meet your use cases? Please be as specific as possible (for example, details on volume, accuracy, referential integrity between sets).**

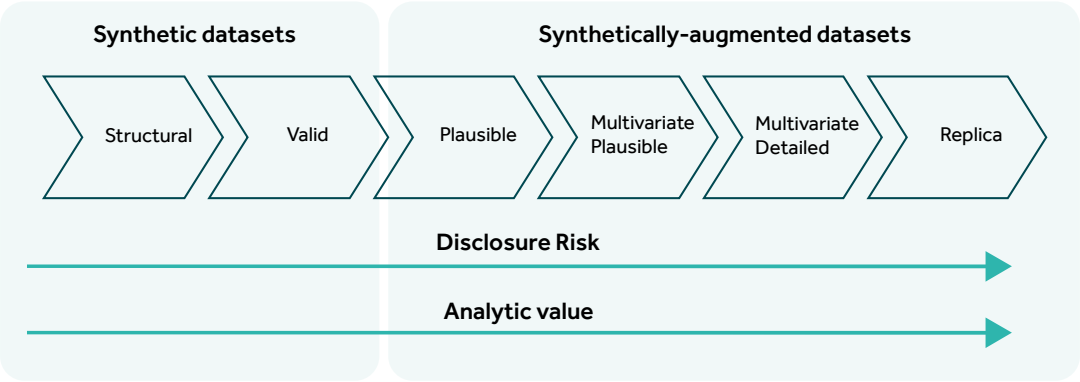
Risks and limitations of synthetic data

3.19 Like the other privacy preserving techniques defined above, the widely recognised limitation of synthetic data is that the more the synthetic data mimics the underlying real data, the greater the risk it may be reverse engineered under certain circumstances to reveal real-world information. There is currently no perfect solution to the data privacy-utility challenge.

3.20 Models for generating synthetic data must often train on real datasets, so while the artificially produced data is fake, the statistical properties are real. In the case of extreme outliers for example, one may expect an accurate synthetic dataset to replicate such outliers, and if the outlier has a combination of unique characteristics, it could be inferred that the synthetic outlier corresponds to a real individual with those unusual characteristics. Furthermore, if real data is limited or biased, there is a risk that this bias will be replicated in the synthetic dataset, unless this is accounted for during the generation process.

3.21 The UK ONS has suggested there may be a need for a spectrum for considering the de-anonymisation risk of synthetic data:

'At one end we define a purely synthetic dataset suitable for code testing with no ambition to replicate underlying data patterns and at the other end of the spectrum sits a dataset created by clever augmentation of the original, which aims to replicate the patterns contained within the original data source. Clearly, the latter carries extremely high disclosure risk and needs to be approached cautiously'.



3.22 Other approaches have also been suggested for managing the privacy risk associated with synthetic data with recent efforts having focused on differentially private synthetic data. There is an emerging number of differential privacy techniques that use carefully crafted random noise within the synthetic data generation process to provide strong privacy guarantees. These techniques guarantee differential privacy, and can be used to statistically quantify the privacy risk of the synthetic data generation process.

3.23 Academics have referred to differentially private synthetic data as a 'best of both worlds' scenario in maximising utility while retaining privacy.

3.24 The processing of real data to generate synthetic datasets presents another key risk. As mentioned, some techniques for synthetic data generation require real data as an input to train the model. Depending on whether the synthetic data generation approach falls under the definition of data 'processing', the real data will be subject to privacy law requirements, and therefore must comply with the UK GDPR and Data Protection Act 2018. A failure to comply with data protection laws will result in an infringement on consumers' privacy rights.

Q13: Do you agree with our assessment of the potential limitations and drawbacks of synthetic data? Are there any others?

4 The role of the regulator

4.1 We're interested in exploring the potential of synthetic data and its relevance to our operational objectives. We believe that opening up novel forms of data sharing could potentially create new opportunities for innovation and a competitive arena for new market participants. Synthetic data could also provide the opportunity to achieve effective data sharing in a way that reduces the risk to consumers' data privacy.

4.2 Synthetic data therefore has the potential to advance all three of our operational objectives:

- Market competition: Larger companies have traditionally held a competitive advantage in data analytics and related fields due to large data harvesting capabilities, whereas smaller or new competitors often lack sufficient quantity and quality data to develop algorithms and data-led strategies. Synthetic data could help democratise data access across the industry by opening access to data assets with incumbents and start-ups alike. Diversifying the range of firms with access to quality data could advance the development of disruptive products and services not currently offered by the market.
- Market integrity: Promoting innovation through effective data sharing could help to accelerate the development of RegTech tools and initiatives, which in turn has the potential to strengthen firms' compliance programmes, helping them to meet their regulatory obligations, and ultimately create safer markets.
- Protecting consumers: Synthetic data reduces many of the data privacy risks associated with real data sharing, as it removes personal data. High-quality synthetic datasets could potentially be used by regulators in the future to evaluate and compare algorithmic decision making by firms, as one tool to safely monitor the deployment of AI by firms and ensure customers are being treated fairly.

4.3 **We would like to conduct an introductory exploration of market attitudes towards synthetic data, and its potential for opening data sharing between firms, regulators and other public bodies.** By examining the market need and harm caused by disproportionate data access across the market, we hope that this Call For Input will advance regulatory knowledge and inform a future direction of travel.

4.4 In particular, we would like to explore the potential for synthetic data to both broaden and enhance collaboration initiatives within the public sector. Like the markets we regulate, cross-regulatory initiatives are increasingly data-driven and regulators and other policy making bodies both nationally and internationally could benefit from the ability to co-operate at a granular data level (for example, in the design and implementation of data standards and frameworks). It is now more important than ever for public sector bodies to work together to overcome the challenges and exploit the opportunities of a new technological age.

4.5 We currently engage with a spectrum of regulators and public bodies, nationally and internationally, to explore shared priorities and challenges. We would like to assess the appetite to expand this approach into data-driven co-operation, for example the co-creation for SupTech (Supervisory Technology) between regulators. Synthetic data could be important to such an approach, as it would provide quality, shareable data

assets to test and develop products and solutions, where previously shareable data assets have not been available.

What role could the FCA play in synthetic data generation/ sharing?

4.6 There are different potential roles that the regulator could perform regarding the further exploration and potential use of synthetic data in financial services. Broadly:

- Data Generator: The regulator collaborates with industry experts and academia to generate synthetic data in-house, to be shared with the industry. The regulator could obtain real data from multiple entities, ensuring a cross-section of industry is sampled and the data are not biased towards a single organisation. Synthetic data could also be shared with organisations holding real data for iterative benchmarking purposes, improving the quality of the data over time.
- Central Host: The regulator provides an independent hosting platform through which synthetic data can be stored, shared and accessed for the purposes of product development and testing.
- Coordinator: The regulator as a co-ordinating body to facilitate data sharing and/or collaboration opportunities for synthetic data generation.

4.7 These roles are not independent, and a regulator could perform one of the above roles or a combination of the three. To perform any of these roles, however, would require involvement from the industry on several fronts.

4.8 We would like to assess the appetite of firms to co-operate with the FCA and other organisations, for example by providing sample real data as an input into the generation process, as well as synthetic data generation expertise to ensure that the synthetic data produced is of optimal quality. Furthermore, industry-wide synthetic data sharing would first and foremost require significant engagement from firms and public organisations. We are therefore seeking to understand the extent to which organisations would be interested in using synthetic data, how often they would use it, and the volume of datasets they would require to realise significant benefits from synthetic data sharing.

Q14: **Do you believe that regulators should play a role in the provision of synthetic data? If so, what do you think the extent of that role should be? (e.g. co-ordination, generation, hosting, etc)**

Q15: **To what extent would you be willing to collaborate with regulators and/or other organisations to generate synthetic data? For example, would you provide real data samples, or benchmark synthetic data against real data sets?**

Q16: **Do you think access to synthetic data should be a public utility for the purposes of innovation and research? Would you pay for access if it was delivered at-cost, or monetised?**

5 Next steps

5.1 We are keen to hear from all market participants about their views and experiences. We are seeking evidence, views and answers to the questions we have asked throughout this Call For Input from all market participants and other interested stakeholders by **22 June 2022**.

5.2 We would also like to hear about any other issues relating to synthetic data not covered in this Call For Input but which you think we should also consider. We look forward to engaging with industry over the course of this work.

5.3 Please submit any responses via the online form, email syntheticdata-callforinput@fca.org.uk or via post to the address set out on page 2 by **22 June 2022**. When providing feedback please include as much detail as possible and provide evidence for your responses where possible to do so.

5.4 Following the receipt of responses we plan to publish a Feedback Statement setting out our analysis, findings and any next steps.

Annex 1
Table of questions in this Call For Input

Q1: How important do you think access to data is for innovation within financial services? What else do you view as significant barriers to innovation?

Q2: Do you agree that it is challenging to access high-quality financial data sets? If so, specifically what challenges do you face? (for example, understanding legal requirements around data access, commercially expensive, or technology infrastructure.)

Q3: Do you agree with the high-level benefits for synthetic data? Are there any other benefits for synthetic data for your organisation, both now and in the future?

Q4: Does your organisation currently generate, use, purchase or otherwise process synthetic data? If possible, please explain for what purpose(s).

Q5: If your organisation generates synthetic data, please describe at a high level the techniques used. Why have you chosen to use this approach?

Q6: What do you see as the difficulties and barriers for firms in creating high-utility, privacy secure synthetic data?

Q7: Does your organisation engage with privacy enhancing technologies or privacy preserving techniques other than synthetic data? How would you assess the utility and benefits of synthetic data in comparison to other techniques?

Q8: What do you see as the highest priority use cases that would benefit from synthetic data?

Q9: Are the synthetic data use cases you have mentioned significant for early business phases or mature operations/processes within your organisation?

Q10: How would your organisation make use of synthetic data if it was available (if at all)?

Q11: **What synthetic data sets would you find most valuable to have access to? For example, Open Banking, Customer profiles, account to account payments, Credit card transactions, trading data, etc. What challenges would these data sets help your organisation to solve? E.g. AML and fraud detection, ESG, etc. Please be specific.**

Q12: **What requirements would you need for the synthetic data to feasibly meet your use cases? Please be as specific as possible (for example, details on volume, accuracy, referential integrity between sets).**

Q13: **Do you agree with our assessment of the potential limitations and drawbacks of synthetic data? Are there any others?**

Q14: **Do you believe that regulators should play a role in the provision of synthetic data? If so, what do you think the extent of that role should be? (e.g. co-ordination, generation, hosting, etc)**

Q15: **To what extent would you be willing to collaborate with regulators and/or other organisations to generate synthetic data? For example, would you provide real data samples, or benchmark synthetic data against real data sets?**

Q16: **Do you think access to synthetic data should be a public utility for the purposes of innovation and research? Would you pay for access if it was delivered at-cost, or monetised?**

Annex 2
Glossary of terms used in this document

This glossary should not be considered an indication of regulatory definitions. The definitions and explanations contained herein are only to clarify references to the associated concepts in the Call For Input.

Term	Description
Innovation	Innovation is the creation of new knowledge and ideas to facilitate new business outcomes, aimed at improving internal business processes and structures and to create market driven products and services. Innovation encompasses both radical and incremental innovation.
De-anonymise	Sometimes referred to as re-identification, de-anonymisation is a data mining technique that attempts to re-identify encrypted or obscured information.
Artificial Intelligence (AI)	Artificial intelligence is a computerised system that exhibits behaviour that is commonly thought of as requiring intelligence.
Machine Learning (ML)	Machine learning is the process in which a computer distils regularities from training data.
Personal data	Information that can be used, either directly or indirectly, to identify an individual.
Incumbent	A firm which is already in an established position in the financial services market. Definition includes organisations such as banks with a large market share, but would not for example, include large technology firms that enter financial services.
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. (Office for National Statistics)
Differential privacy	Differential privacy is a formal mathematical framework for quantifying and managing privacy risks when analysing or releasing statistical data.
Metadata	Metadata provide information about other data, including a description of the data. This includes information that provides context to the data (for example, how it was collected), or the coverage of the data (for example, publication date, description and search keywords). (Office for National Statistics)

Term	Description
GDPR	The General Data Protection Regulation (GDPR) is an EU regulation that controls the processing of personal data and the free movement of such data in the European Union and the European Economic Area and was onshored post-Brexit.
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 information 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.
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.

Pub ref: 007728.1

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