10 St Bride Street
London EC4A 4AD
T +44 (0) 7331 2000

techUK.org | @techUK

Synthetic Data

techUK's response to the Financial Conduct Authority's (FCA) call for input on synthetic data to support financial services innovation

Contact: Dani Dhiman, Policy Manager, Data
T 020 7331 2000 | E dani.dhiman@techuk.org

Introduction

As the Financial Conduct Authority (FCA) has rightly identified, sufficient access to personal and non-personal data remains a key challenge for businesses of all sizes and sectors seeking to innovate, particularly SMEs. This is especially the case for the development of new products and services such as digital identity, which deploy artificial intelligence (AI) and tend to depend on large volumes of high-quality data to train algorithms that can deliver fair and ethical outcomes.

Data access and sharing is also vital for businesses to develop digital solutions which serve the public good. For example, in the context of financial services, data drives efforts towards combatting financial fraud and improving financial inclusion. techUK welcomes the FCA's detailed exploration of how synthetic data can help to address the challenge of lack of data access while helping to stimulate innovation and competition in markets, such as financial services. The call for input also raises ethical considerations on the use of synthetic data, as well as the role regulators can play in governing and encouraging its widespread use.

techUK understands that synthetic data refers to data that is generated using algorithms which preserves an original dataset's statistical features while producing entirely new data points[1]. As the Turing Institute states, synthetic data generators enable users to share data, to work with data in safe environments, to fix structural deficiencies in data, to increase the size of data, and to validate machine learning systems by generating adversarial scenarios.

As thinking develops on synthetic data, Government and regulators must recognise the considerable work that still needs to be done to remove existing barriers to unlocking the full value of non-synthetic or 'real data' i.e., data generated by actual events. A key barrier which must be addressed is lack of access to real data, which will also be relevant for generating synthetic data since it depends on observing and replicating properties of real data. There are also many other long-standing data challenges that exist which risk persisting in the context of synthetic data if not addressed e.g., the data skills shortage amongst others which we explore below.

Below, techUK has set out several recommendations which should be taken into consideration as Government, regulators, and bodies such as the Digital Regulation Cooperation Forum (DRCF) develops its thinking in this area. If the FCA considers new initiatives or market interventions related to synthetic data, it should ensure engagement and discussion with a diverse range of stakeholders, including industry.

Recommendations

1. **Step up efforts to address existing barriers to data access and sharing, such as lack of data standards, data quality as well as privacy and security concerns,**
2. **Outline a clear plan for the continued opening up of Government and public sector data sets, with the aim to move toward near real-time reporting of data,**
3. **Ensure that synthetic data is deployed for the public good and its use underpinned by careful privacy considerations**
4. **Leverage the Digital Regulation Cooperation Forum (DRCF) to ensure joined up and consistent approaches to synthetic data,**

[1] Synthetic data generation for finance and economics | The Alan Turing Institute

5. Narrow the data skills gap and combat skills shortages by investing in training, upskilling, and reskilling of the UK's workforce.

1. **Step up efforts to address existing barriers to data access and sharing, such as lack of data standards, data quality as well as privacy and security concerns,**

   Some of the current challenges for accessing and sharing data will not be resolved through the generation of synthetic data and will remain a challenge for businesses seeking to innovate. If these challenges persist, they risk hindering the development and use of synthetic data. For example, without a set of industry-led standards on how data (whether synthetic or real) should be recorded e.g., format, management, tagging etc. it remains a key barrier for businesses to share data sets across the economy and impedes on the consistency of the quality of data being produced.

   Another reason why businesses may not share data is due to the lack of commercial incentive to do so. techUK supports measures that will encourage voluntary data sharing, where appropriate, across all sectors and industries. Government must seek to develop market mechanisms which can be introduced to help deliver trusted avenues for commercial data sharing arrangements that are fair and inclusive. However, it is important that interventions do not drive investment away from good data practices or stifle emerging business models and innovation.

   The CMA and industry's work to transition from Open Banking to Open Finance is a strong example of the success and benefits sector-specific data sharing can unlock and should be extended to other sectors through BEIS' Smart Data Workstream. We urge Government to make swift progress in laying primary legislation to encourage greater industry participation in Smart Data schemes. Government should also begin to consider how Smart Data frameworks can be used to facilitate wider cross-sectoral data sharing.

   Regulators will also need to find a balance between competing priorities of boosting competition in digital markets through increased data sharing while not undermining privacy and data protection standards. For example, techUK has welcomed proposals in the Government's upcoming reform to the data protection regime which could offer businesses access to more personal data for research purposes, without undermining or weakening data protection rights.

   techUK also recognises that synthetic data offers a promising way to address the tension between innovation and privacy, which is another key barrier to data sharing. Other mechanisms such as Privacy Enhancing Technologies (PETs) and a thriving data intermediary ecosystem will also play a key role in managing this tension and we encourage the Government and regulators to drive efforts on all these fronts to better enable the market to share data in ways that support innovation and competition. For example, techUK has welcomed the UK-US partnership on prize challenges focused on advancing PETs.

2. **Outline a clear plan for the continued opening up of Government and public sector data sets, with the aim to move toward near real-time reporting of data,**

   While synthetic data will be significant in providing businesses with more access to datasets, efforts must still be put in to understand existing types of real datasets, such as those held by Government and public services and how they can be best leveraged. This will be key for businesses to innovate, but to also generate their own synthetic datasets, by modeling synthetic datasets from public sector data.

techUK and members have long called for the opening up of key Government datasets which will be vital in spurring the commercial development of products and services. Although the UK was ranked a world leader in Open Data initiatives in 2017, it is disappointing to see the UK take a step backwards in an area where it was once seen a world leader in opening up datasets. techUK believes the UK can regain its leadership in this area, but action is needed now to make sure this happens.

The Government could also consider generating and opening data as synthetic datasets to mitigate concerns or risks around privacy and ethics. For example, the Medicines and Healthcare products Regulatory Agency (MHRA) used primary care data to produce entirely artificial data that did not contain any original data from "real" patients, to reduce risks to patient privacy. These synthetic data sets have helped medical researchers to develop cutting-edge medical technologies, such as medical devices to fight COVID-19 and cardiovascular disease.

When considering which datasets to open up – whether as synthetic or real – Government must consult with industry and organisations to better understand which data sets could unlock the most value and outline a clear plan on how this will be put into action. For example, in the case of digital identity solutions, access to databases which include marriage registry, births and deaths registries, the passport register, and land registry entry would be most useful.

3. Ensure that synthetic data is deployed for the public good and its use underpinned by careful privacy considerations

The use of synthetic data has the potential to support business innovation and bring economic benefits including the deployment of new services such as digital identity which can help to improve efficiency, cut cost, and help combat fraud. For example, businesses using AI-driven digital verification technology to combat financial fraud use synthetic data to accelerate the maturity of algorithms as well as evaluate the overall performance and accuracy of their systems. It can also play an important role in achieving wider social benefits and improve services by enabling wider data-sharing while protecting people's right to privacy.

The FCA's own DataSprint in the summer of 2020 provides a strong example of the development of high-quality synthetic datasets for public good, with aims including preventing fraud and supporting the financial resilience of vulnerable customers. As technologies to support such purposes rely on highly sensitive information, synthetic data provides one of the ways in which they can be developed and deployed without infringing on privacy rights.

It should be a priority for regulators to continue to facilitate collaboration around the creation of synthetic datasets, given how significantly these could benefit the public. techUK also welcomes further exploration of what a more significant role for regulators could look like, including hosting and providing access to synthetic data against a fee. If the service is monetised, a fair relationship needs to be worked out between private providers who contribute with real data samples and the regulator. Regulators should also consider providing some synthetic data for free or at discounted rates if used for purposes that are of particular benefit for the public (such as preventing fraud or protecting vulnerable customers).

Lastly, we agree with the potential risks and limitations of synthetic datasets, including the risk of re-identification and mirroring biases in real datasets. If regulators begin playing a role as providers of synthetic data, one of the benefits should include reassurance that the synthetic data provided fulfils the highest security standards and that bias has been accounted for in the generation process.

4. **Leverage the Digital Regulation Cooperation Forum (DCRF) to ensure joined up and consistent approaches to synthetic data,**

   As the FCA explores the use of synthetic data in the financial services industry, it is vital that key findings and outputs from the consultation are shared with the DRCF to ensure a collaborative approach and culture of knowledge sharing between regulators on this nascent topic. This should include engagement with the Competition and Markets Authority (CMA) who will play a key role in assessing the impact of synthetic data on competition.

   With businesses navigating complex and oftentimes overlapping regulatory regimes, the DRCF plays a vital role in ensuring consistent approaches, shared visions, and cooperation between regulators on topics with mutual interest. Since the uptake synthetic data will cut across multiple sectors, a joined up and consistent approach between regulators will be essential.

   This will be especially true if regulators adopt the role of coordinating, generating and/or hosting synthetic data, which will require further consultation with industry to determine what types of synthetic datasets should be prioritised for generating and sharing, and how access to this data will be facilitated.

5. **Narrow the data skills gap and combat skills shortages by investing in training, upskilling, and reskilling of the UK's workforce**

   If businesses and organisations are to begin generating synthetic data through the deployment of algorithms, digital and data skills will be vital. Synthetic data is generated programmatically and depends on highly skilled computer scientists with expertise in deep and machine learning models. There is already an existing and considerable data skills gap, in part due to businesses, organisations, and Government lacking the technical skills to manage data, as well as the skills to think creatively about data.

   To address this skills shortage, Government must invest in the training, upskilling, and retraining of the workforce. Both industry and Government need to emphasise the development of lifelong learning to prepare the workforce for the technological changes to come. The UK's digital and data skills shortage must be tackled at all levels of education including secondary, higher, further, adult, vocational (such as apprenticeships and T-levels), and upskilling within industry.

   One way the Government can address digital skill shortages to boost growth is by expanding the coverage of the Help to Grow: Digital Scheme, supporting SMEs to invest in digital reskilling through a Digital Skills Tax Credit and continuing to reform the Apprenticeship Levy. The tech industry, working with Government and key stakeholders, has a responsibility to tip the scales so that motivations for learning outweigh any barriers faced.