CITY OF LONDON

THE GLOBAL CITY

Unlocking AI Skills in Financial and Professional Services: Synthetic Data Marketplaces

2025 Edition

Executive Summary

This report explores the potential of synthetic data to address the growing demand for artificial intelligence (AI) and machine learning (ML) skills in the Financial and Professional Services (FPS) sector. AI is expected to have significant impact on the sector, contributing an estimated £35 billion and 50% productivity increase over the next five years. Despite this, there is a notable shortage of professionals with the necessary AI and ML skills, exacerbated by the rapid pace of technological advancements.

To bridge this skills gap, the report proposes the establishment of a synthetic data lake through the Financial Conduct Authority's (FCA) 'supercharged' Digital Sandbox. This initiative aims to enable course providers, such as apprenticeship programmes, business schools, and other skills providers to create hands-on AI training courses tailored to financial services roles. The synthetic data, which mimics real-world data without the privacy and compliance issues, offers a promising solution for experiential learning and upskilling.

The report outlines the technical feasibility of generating and maintaining high-quality synthetic datasets, the security and compliance advantages, and the integration with the FCA's existing Digital Sandbox and AI lab platform. We discuss the financial feasibility, including initial investment, ongoing maintenance costs, and potential revenue models to achieve self-sufficiency. Additionally, the report addresses regulatory feasibility, aligning with UK GDPR and FCA regulations, and the role of the regulator in hosting and overseeing the synthetic data environment.

The initiative aligns with the UK government's AI policy, particularly the AI Opportunities Action Plan, and aims to support the UK's ambition to be an 'AI superpower' by upskilling thousands of professionals in AI. There is also clear alignment with more recent policies to use the immigration system to fund domestic skills. The return on investment is clear: the proposed Immigration Skills Charge increase of 32% will raise roughly £773 million.

For less than £1 million investment, the government could seed an AI training programme marketplace in Financial Services to address the AI skills gap, train domestic talent for high-paying jobs in a sector with the highest average tax contributions and create a self-sustaining marketplace within 5 years.

The report concludes that the synthetic data lake initiative is highly feasible and represents a prudent investment in the infrastructure for innovation, ensuring the UK remains at the forefront of financial innovation and responsible AI deployment.

Benefits of Centralised Synthetic Data Solution

Cost
A centralised synthetic data solution distributes costs across the industry, making it more accessible for all firms, particularly SMEs.

AI Policy
The initiative supports the UK's ambition to be an 'AI superpower' by upskilling professionals in AI.

Compliance
Synthetic data mitigates the risks associated with using real financial data for training purposes.

Learning
Diverse, extensive datasets without the privacy and compliance issues of real data, significantly enhancing hands-on learning opportunities.

Scalability
Easily scaled to meet growing demand across the sector, providing datasets to content creators, business schools, and firms specialising in AI skills training.

Standardisation
A common dataset could lead to more standardised training across the industry, potentially improving overall skill levels.

Key considerations

Data Privacy and Compliance:
Ensure that synthetic data adheres to privacy regulations and does not compromise sensitive information.

Technical Feasibility:
Assess the capability to generate and maintain high-quality synthetic datasets that accurately mimic real-world data.

Scalability and Latency:
Address the challenges of scaling synthetic data solutions and managing latency concerns.

Cost and Investment:
Evaluate the financial feasibility, including initial investment, ongoing maintenance costs, and potential revenue models.

Industry Collaboration:
Foster collaboration between various departments and industry stakeholders to enhance the quality and applicability of synthetic data.

Regulatory Alignment:
Ensure alignment with UK GDPR, FCA regulations, and other relevant legal frameworks.

Security and Reliability:
Mitigate privacy and security risks while ensuring the reliability and fidelity of synthetic data.

Government and Policy Support:
Aligns with the UK government's ambition to be an 'AI superpower'.

Background and Context

The demand for AI and ML skills in the FPS sector is growing rapidly. This surge is driven by the sector's ongoing digital transformation and the adoption of AI technologies to enhance efficiency, decision-making, and customer experiences.[1][2] AI is predicted to add significant value to the FPS sector, with estimates suggesting it could contribute as much as £35 billion and 50% productivity over the next five years.[3]

Current Skill Shortages and the Need for Upskilling to Maintain Competitive Advantage

Despite the high demand, there is a notable shortage of professionals with the necessary AI and ML skills. The Financial Services Skills Commission's 2024 Skills Report showed that AI and ML was the largest skills gap in FS.[4] This skills gap is exacerbated by the rapid pace of technological advancements, which outstrip the current supply of qualified talent. The traditional talent pipeline cannot address the demand fast enough – 80% of 2030's workforce is already employed. Firms must upskill, and they know it.

A recent IBM study found that over the next three years, 87% of business leaders expect at least 25% of their workforce will be required to reskill in response to generative AI and automation, with 45% predicting that over half will need to reskill.[5] To maintain a competitive edge, companies must invest in upskilling their workforce, focusing on both technical and transferrable skills.[6] But finding cost-effective ways to do so in a highly regulated environment is challenging, especially when those new technical skills require hands-on learning experience manipulating data in algorithms.

Experiential Learning

Studies show that experiential learning, learning by doing, has a positive effect for children not only in math, science, problem solving and critical thinking, but it also improves their motivation, engagement and sense of agency.[7][8] Such learning also has proven positive impacts on mental wellbeing for university learners.[9] These impacts are not unique to traditional educational structures but apply to adults upskilling too.[10] Despite very welcomed recommendations in the recent AI Opportunities Action Plan[11] to support AI education (e.g. recommendations 17 and 19 which seek to expand pathways into AI and encourage lifelong learning), Government have to date focussed more on the traditional talent pipeline rather than on upskilling/reskilling solutions for helping today's workforce transition to tomorrow's work place. Government should look to support experiential learning programmes which offer those with existing technical skills the ability to get hands-on learning experience using data and testing algorithms.

Synthetic Data

Synthetic data, which is artificially generated data that mimics real-world data, offers a promising solution for experiential learning and upskilling. It allows for the creation of diverse and extensive datasets without the privacy and compliance issues associated with real data. Such data can often offer improved data quality, scalability, ease of use and reduced bias. This can significantly enhance hands-on learning opportunities and help bridge the skills gap.

FPS firms often interact with this type of data in sandboxed environments, an innovation championed by the UK and recently furthered by the Financial Conduct Authority's (FCA) commitment to their AI lab. However, sandbox environments are usually confined to the individual firm's data, and stakeholder engagement shows that costs and data richness can prove prohibitive for smaller firms.

But there are clear opportunities to resolve these. Participants in the FCA's earlier Digital Sandbox pilot cited synthetic data as 'the most valuable feature' and called for expanding these data assets to support more effective testing and learning.[12]

Summary

The report summarises our research into a policy initiative to establish a synthetic data lake through the Financial Conduct Authority's (FCA) 'supercharged' Digital Sandbox, enabling course providers (e.g. apprenticeship programmes like Multiverse and business schools) to create hands-on AI training courses tailored to financial services roles. It is informed by two roundtables and over 40 firm interviews to discuss barriers, identify trends and offer possible solutions. The approach outlined in this report addresses several key challenges we heard raised consistently from across the sector: risk appetite and compliance concerns, cost, depth and richness of data (especially for new market entrants), and accessibility.

In alignment with the UK's ambition to be an 'AI superpower,' this initiative would leverage regulatory innovation to upskill thousands of professionals on AI, supporting both competition in financial services and the broader AI adoption goals of HM Treasury and the Department for Science, Innovation and Technology (DSIT).

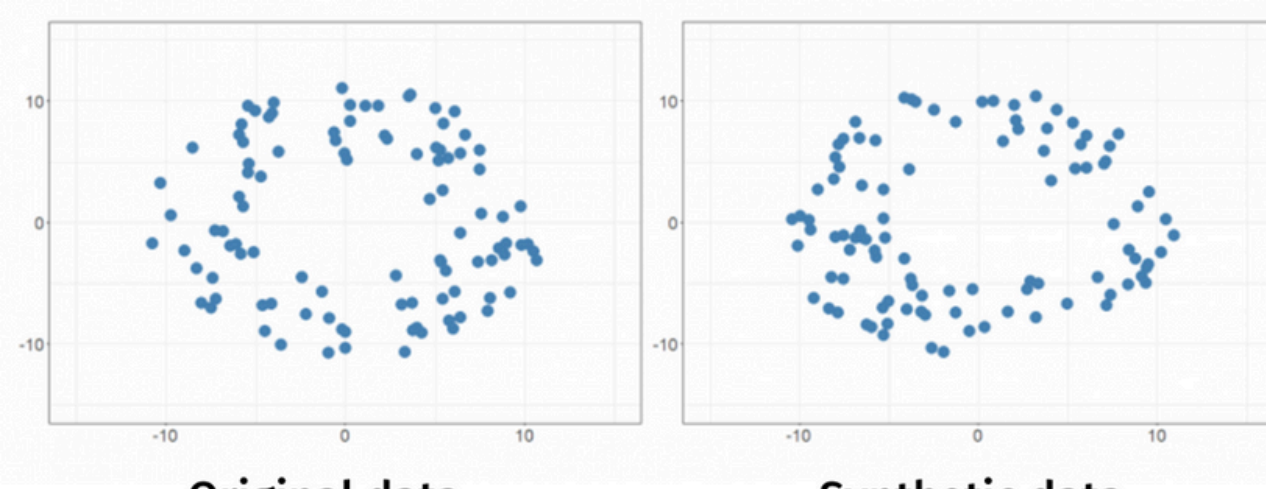
Technical Feasibility

Synthetic Data Generation & Maintenance

Generating high-quality synthetic datasets for financial services is technically viable given recent advances but requires careful design. Synthetic data uses algorithms (such as statistical models or GANs) to create 'artificial' data that mirror the patterns of real financial data without exposing actual personal information.[13] This approach has already been piloted by regulators and banks for use cases like fraud detection and anti-money laundering. For example, the FCA and CoLC recently launched a synthetic dataset mimicking real banking transactions to combat APP fraud.[14] Likewise, the Alan Turing Institute is working with industry partners, e.g. HSBC, on novel techniques to generate realistic financial data.[15] These efforts indicate that it is feasible to produce rich datasets representing transactions, customer profiles, market data, etc., suitable for training AI models in a sandbox environment. . Maintaining the data lake will involve periodically updating and expanding the datasets to reflect new trends (such as emerging fraud tactics or market conditions) and ensure they remain realistic.

Industry collaboration can help here. The FCA's Synthetic Data Expert Group has explored co-creation models where firms contribute sample data and expertise to improve synthetic data quality.[16]

However, it must be acknowledged that synthetic data generation is an evolving field; current techniques still face challenges in perfectly preserving data utility and complexity. Ongoing validation and refinement will be needed to ensure the synthetic data remains a high-quality substitute for real data over time.[17] Stakeholders also noted concerns not only about how good the quality of the data was, but also latency issues in scalability. Even with synthetic data, the five V's of Big Data apply: Velocity, Veracity, Volume, Variety and Value. Starting small, with the aim of having a select few course providers addressing a single theme and testing within small cohorts would allow for a better understanding of the infrastructure needed.



The synthetic data retains the structure of the original data but they are not the same

Source: Synthetic data: Unlocking the power of data and skills for machine learning - Data in government (blog.gov.uk)

Security, Compliance and Reliability

Using synthetic data significantly mitigates privacy and security risks compared to using real customer data. By design, synthetic datasets do not contain actual personal identifiers and instead replicate the statistical structure of real data. This greatly reduces the risk of sensitive information leakage, enabling open use in training environments without violating customer confidentiality.

In the context of the FCA's sandbox, all datasets would be vetted to be GDPR-compliant and anonymised by design. That said, there are important reliability and ethics considerations. Course participants must trust that the synthetic data is representative of real-world scenarios. If the data is too 'clean' or fails to capture edge cases, AI models trained on it might not perform well in production. Ensuring fidelity (that synthetic data reflects real data distributions) while maintaining privacy is a known trade-off.[18]

There is also a small risk that poorly generated synthetic data could inadvertently encode biases or be reverse-engineered to infer something about the source data

. To address this, the initiative would need to implement robust validation processes. The FCA, ICO, and Turing Institute have noted that developing validation metrics for synthetic data's utility and privacy is critical for adoption. Techniques like differential privacy, deep learning generators, and continuous testing of AI models on synthetic vs. real sample data could be employed to gauge reliability.

In summary, from a security perspective synthetic data is highly advantageous (since it avoids exposing real data), and with proper controls and iterative improvements, it can be made reliable enough for training purposes. The sandbox's secure cloud infrastructure will further ensure that even the synthetic datasets are protected from unauthorised access or tampering. And the FCA name carries significant reassurance with firms that would contract with skills providers. Knowing the underlying course data is vetted and housed within the FCA's Digital Sandbox is critical to addressing risk aversion.

Integration with FCA's Supercharged Sandbox

A core strength of this initiative is building on the FCA's established Digital Sandbox platform. The FCA has already made the sandbox permanent (as of 2023) with a rich library of over 300 datasets (synthetic, public, or anonymised) and more than 1,000 APIs available for testing innovative solutions. The recently launched 'supercharged' enhancements, e.g. greater computing power, enriched datasets, and increased AI testing capabilities[19], align perfectly with the needs of an AI training data lake.

Course creators would be able to plug into this environment rather than starting from scratch. Practically, the integration could work as follows: the synthetic datasets (e.g. a large set of fictional transaction records, customer account data, loan applications, etc.) would be hosted in the sandbox's cloud environment. Training providers (once onboarded to the sandbox) can access these datasets through secure APIs or download portals and leverage the sandbox's computing resources to run machine learning experiments.

For instance, an apprenticeship course on fraud analytics could give learners access to a synthetic transactions dataset via the sandbox API and task them with building an AI model to detect suspicious patterns. The sandbox also offers development environments and collaboration tools (as noted in its features), which course participants could use to write and test code in a safe setting.

Integration is made easier by the fact that the sandbox already supports external innovators: any firm (or potentially educational institution) that meets eligibility criteria can apply for access and, once accepted, use the data and tools for a defined project period.

The initiative might establish a specialised 'Education/Skills' cohort within the sandbox, streamlining access for approved training providers and their student cohorts. By leveraging the FCA's infrastructure, the technical overhead for course creators is lowered; they do not need to procure separate secure data storage or worry about compliance vetting of data, as these are provided by the sandbox. The 'supercharged' upgrades (more compute and richer data) will further support running AI models at scale, allowing, for example, the training of complex models or simulation of high-volume transaction streams for real-time AI exercises.

Technically integrating the synthetic data lake into the FCA's sandbox is feasible given the existing platform capabilities. The key will be coordinating user access and perhaps customising some sandbox features for educational use (such as resetting environments for each cohort, providing user-friendly documentation of the datasets, etc.), but these are manageable within the sandbox's flexible framework. That all course content is driven by an underlying dataset vetted and housed by the regulator gives credibility to the content providers and critically addresses risk, or data compliance concerns, from the sector.

Financial Feasibility

Initial Investment

Launching this initiative will require upfront funding to create the synthetic data lake and pilot the first cohort of AI courses. The major cost components at the start include data generation and infrastructure setup and course development subsidies. Generating high-fidelity synthetic financial datasets is a specialised task, resources may be needed to procure or develop data generation tools and to involve domain experts (possibly via partnerships with organisations like the Alan Turing Institute or fintech vendors experienced in synthetic data). This could involve contracting data scientists to produce the datasets and validate them, as well as computing costs for running generation algorithms on large real datasets.

Additionally, supporting course creators in designing experiential AI modules will be important to ensure the content aligns with both industry needs and the available data.

This might take the form of grants or co-development contracts for a few pilot programmes (for example, funding an apprenticeship provider like Multiverse to create a 'Financial Services AI Analyst' course using the sandbox, or a business school to develop a fintech AI case study curriculum). The initial cohort of learners may also be subsidised, or instance, offering free or low-cost enrolment to a certain number of participants from various firms to demonstrate the value.

While detailed cost estimates require further scoping, a pilot could be launched with a modest investment spread across data engineering, platform enhancement, and educational content creation. Notably, this leverages existing infrastructure (the FCA Sandbox), which reduces the need to build a platform from scratch. Even in the FCA's recent collaboration with the City of London to release a synthetic fraud dataset, the approach was to build on the permanent sandbox and invite participants through an existing process, suggesting that incremental investments can piggyback on prior ones.

We anticipate that HM Treasury and DSIT could provide seed funding or innovation budget allocations for this initiative, given it sits at the intersection of financial innovation and AI skills development, both of which are policy priorities.

A preliminary estimate of the upstart costs is outlined below in Table 1. This estimate suggests a total initial investment in the range of £585,000 to £620,000+, acknowledging that this is a preliminary figure that requires further detailed analysis and a comprehensive implementation plan to refine. This estimate does not include ongoing operational costs for the maintenance and further development of the platform and the synthetic datasets.

Table 1: Estimated Upstart Costs

Cost Category	**Description**	**Estimated Range (GBP)**	**Notes**
Infrastructure Enhancements	Upgrades to FCA Digital Sandbox for larger data lake and platform support	£150,000 - £200,000	Dependent on existing capacity and required features.
Synthetic Data Tools & Platforms	Procurement or development of advanced data generation capabilities	£100,000 - £150,000+	Varies based on complexity, licensing models, and whether off-the-shelf or custom solutions are
Data Storage	Initial storage for the synthetic data lake	£15,000 - £20,000	Dependent on the volume and type of synthetic data.
Content Creator Support	Grants, training, and technical assistance	£70,000 - £100,000+	Scale will depend on the number of content creators engaged and the level of support provided.
Personnel Costs (Initial Setup)	Project management, technical expertise, stakeholder engagement	£100,000 - £150,000+	Includes salaries and benefits for the initial phase of setting up the initiative.
Total Estimated Upstart Costs	**Sum of above**	**£585,000 - £620,000+**	This is a preliminary estimate and requires further detailed analysis. Excludes ongoing operational

Ongoing Maintenance & Expansion Costs

Beyond the pilot phase, maintaining the synthetic data lake and scaling up course offerings will incur continuing costs. Data maintenance will require periodic updates to the synthetic datasets: adding new data fields, generating fresh data to prevent models from overfitting to one static dataset, and creating new datasets for additional use cases (e.g. insurance claims, trading data, etc. as the course catalogue expands). This implies retaining a small technical team (data engineers and analysts) to manage the data lake. The FCA's innovation division may host this team, possibly in partnership with academic or industry secondees.

Infrastructure costs for hosting data and providing compute will scale with the number of users and complexity of exercises. The Digital Sandbox is cloud-based; thus, expenses will include cloud storage for large datasets and computational expenses when course participants run AI model training or queries.

However, since the sandbox is already an ongoing service, incremental costs for additional users/data might be moderate, especially if the platform's planned enhancements are implemented (greater computing power likely comes with an established budget).

That said, if the programme grows significantly (imagine dozens of training providers and thousands of learners using it), the platform might need further investment in server capacity and support staff. It will be important to monitor usage and performance during the pilot to project future capacity needs. As this should be viewed as an initial investment in creating a marketplace, the FCA could potentially pass operational costs through to content creators, who themselves would set prices.

Another cost factor is user support and administration: handling onboarding of course providers, providing documentation and maybe training for instructors on how to use the sandbox tools, and ensuring any technical issues are resolved.

his likely necessitates a support function, potentially an extension of the FCA's Innovation Hub team. In financial terms, maintenance could be managed through a combination of budgeted regulatory expenditure and cost-sharing with industry (discussed below). Each additional dataset or feature should be weighed for its value, e.g. developing a new synthetic dataset might be justified if multiple courses demand it or if it addresses a high-value skill area (like ESG data or blockchain transaction data).
Additionally, supporting course creators in designing experiential AI modules will be important to ensure the content aligns with both industry needs and the available data.

Revenue Models and Phased Funding

While initially grant-funded or subsidised, the initiative can explore revenue models to move toward financial sustainability. One potential model is a tiered access fee: training providers or firms that want to use the synthetic data environment for internal upskilling could pay a subscription or license fee. For example, a large bank that enrols its employees in these AI courses might contribute via a membership fee to the sandbox. Since the FCA is involved, a light-touch regulatory fee mechanism could be considered – perhaps an innovation levy or an extension of the FCA's fee structure to cover use of the educational sandbox (similar to how firms pay fees that fund the FCA's operations). However, care would be needed to ensure fees are not a barrier for smaller firms or non-profits wanting to improve skills.

A phased approach is likely optimal: Phase 1 (pilot and initial cohorts) would be free or heavily subsidised to encourage uptake and demonstrate success, i.e. proof of concept. Phase 2, once value is proven, could introduce partial cost recovery.

For instance, course providers begin paying a nominal fee for continued access, or learners pay a course fee (possibly offset by employer sponsorship or the Apprenticeship Levy/Growth and Skills Levy funding). Using the Levy would align Skills England's sector champion approach with the Industrial Strategy's recognition of Financial Services.

Course charges can also be phased in: initially, courses developed under this initiative could be offered at low cost, but as they become part of standard training offerings, providers might charge tuition. Some providers (like professional training companies) might build their own commercial courses around the synthetic data. The initiative might then charge those providers for access to the data environment. Additionally, there is the potential for public-private co-funding. If certain datasets are particularly valuable - say a synthetic Small and Medium-sized Enterprise (SME) lending dataset that fintech startups also want - those startups or a consortium of banks might sponsor the expansion of that dataset, sharing the cost.

The AI Sector Champion for financial services, as recommended in the AI Opportunities Action Plan, could help rally industry contributions, highlighting that this skills initiative serves the sector's collective interest.20

Path to Self-Sufficiency

Over time, e.g. 3-5 years, the goal would be for the data lake and course ecosystem to become largely self-sustaining. In a mature state, the initiative could be managed by a dedicated entity or within the FCA Innovation division, funded by a mix of participant fees and ongoing support from regulators or government if needed.

One vision is that once a critical mass of courses exists, employers and individuals see clear value in them and willingly pay, creating a market for experiential AI training in finance. If, for example, dozens of firms are sending employees through an AI-in-finance apprenticeship that utilises the synthetic data, their fees (via the Growth and Skills Levy or training budgets) flow to the course provider, who in turn can pay for continued data access.

Another possible revenue stream is certification or accreditation. The initiative could partner with industry bodies (like the Chartered Institute for Securities and Investment or CISI for finance professionals) to create an AI skill certification backed by these courses, making it prestigious and motivating firms to invest in it. The City Corporation has success in working with these bodies to do just that, having pushed the Ethical AI Initiative through the 695th Lord Mayor.21 The role of government can gradually shift from funder to facilitator: initial pump-priming ensures the resources exist, and later, governance can ensure fairness and regulatory compliance, while the day-to-day costs are covered by user-driven funding.

We note that the FCA's interest in synthetic data stems partly from its competition mandate. By lowering barriers to innovation, the market benefits. In that spirit, even if some public funding remains necessary, the economic benefits (more AI-skilled workers, more competitive fintech services, etc.) likely justify it.

Nonetheless, careful planning is required so that as the initiative scales, it does not simply become a permanent cost centre. A break-even analysis should be done after the pilot: projecting how many course providers and learners would be needed, paying what fees, to cover operating costs. The phased introduction of fees must be handled such that it balances financial viability with accessibility, ensuring SMEs and a diverse range of participants can still benefit.

With initial government support and a well-designed fee structure, the initiative can transition from reliance on public funds toward a sustainable model supported by the financial services industry it benefits. It would be a world first and support the aims of the Industrial Strategy, Skills England and the AI Opportunity Actions Plan.

Regulatory Feedback

Synthetic Data under UK GDPR and FCA Regulations

From a legal standpoint, using synthetic data for training purposes can be designed to comply with data protection requirements. UK GDPR and the Data Protection Act place strict limits on processing personal data, especially for purposes like training or product development. Synthetic data offers a way to sidestep these issues because properly generated synthetic datasets do not contain personal data about real individuals. If the synthetic generation is done correctly, the resulting data is fully anonymised (in fact, it's fictitious) meaning GDPR would not apply to the synthetic dataset itself, as no 'identifiable natural person' can be linked to it.

The FCA has explicitly framed synthetic data as a 'privacy-preserving technique' that can enable data sharing without undermining data protection laws.[20] However, regulators will insist on rigorous safeguards to ensure that synthetic data truly lives up to this standard. Key considerations include: the process of generating synthetic data often involves using real data as training input.

That initial step must have a lawful basis, e.g. the real data could come from firms that have consent or another legal basis to share it for this innovation purpose, potentially under regulatory oversight.

Additionally, one must ensure no residual risk of re-identification. The Information Commissioner's Office (ICO) has been actively examining synthetic data and other Privacy-Enhancing Technologies, and they stress the importance of techniques that prevent someone from reverse-engineering synthetic data to recreate original personal records. In practice, proper validation and possibly attestation will be needed: course creators and users may need reassurance (or contractual commitments) that the synthetic datasets provided have passed privacy tests. This could involve metrics that the FCA/ICO have been exploring, for example, measuring how closely synthetic records might match any real record which should be very low to avoid identification.

On the FCA regulatory side, there is no rule prohibiting the use of synthetic data for training or development. In fact, the FCA has been encouraging it as a solution to data access problems.

"Training is increasingly moving into the legal teams - what was about curiosity is now more about data access, legal questions, etc. Sandboxing can help to keep the curiosity space open."
– Skills provider

The FCA's own Business Plan 2022/23 referenced experimenting with synthetic data for financial crime controls.[21] As host of the data lake, the FCA would likely set terms of use to ensure it's used responsibly, e.g. not to be merged with any real personal data in ways that could re-identify individuals, not to be used to train models that will then be secretly applied to personal data without proper testing, etc.

If managed under the FCA umbrella, with ICO input, the synthetic data approach is legally sound and exemplifies compliance by design; it enables learning from data without processing actual customer information, aligning with both the spirit and letter of privacy laws.

Role of the Regulator in Hosting and Oversight

A pivotal aspect of this initiative is the FCA, and potentially other regulators, acting as a host for the synthetic data environment. This represents a progressive approach to regulation, moving beyond setting rules to actively facilitating safe innovation and skills development.

The FCA has signalled receptiveness to this role and indeed appears to be embracing it through its AI Lab.[22]

By hosting the data lake on its Digital Sandbox, the FCA ensures a level of governance and trust that would be hard to replicate in a purely private setting. Regulators are seen as neutral and focused on the public interest, so firms may be more willing to contribute data or participate, knowing the platform is overseen for fairness and compliance. That said, this role must be executed within the FCA's statutory remit. Fortunately, fostering innovation and competition is squarely within the FCA's objectives. The Financial Services and Markets Act gives the FCA a competition objective, and it views innovation as a key to achieving that.

Providing a sandbox with rich data to upskill industry participants can be seen as an extension of its existing innovation services (which already include sandboxes and tech sprints).

The FCA will need to ensure operational integrity of the data lake, for example, establishing clear usage terms, security measures, and perhaps an oversight committee (which could include representatives from DSIT, industry, and academia) to guide the initiative.

Additionally, other regulators might be involved: if datasets pertain to specific sub-sectors, regulators like the Bank of England (for prudential aspects) or the Information Commissioner's Office (for data privacy guidance) should be consulted or included. Closer work regulators on AI skills would be well suited to the Digital Regulatory Cooperation Forum (DRCF) and could form part of future business plans.

The regulator-as-host model also raises the question of long-term ownership and liability. The FCA would likely include disclaimers that while they host the data, they do not warrant that using it guarantees any particular outcomes, and users use it at their own risk (particularly to avoid any responsibility if a model trained on synthetic data fails in the real world).

These legal nuances can be managed via participation agreements. Importantly, the security of the sandbox environment will remain paramount. The FCA's Innovation Hub will ensure that the data lake is isolated and that any external integrations (APIs, etc.) do not introduce vulnerabilities.

As the platform grows, the FCA may need to balance its role as an enabler with its core supervisory functions: for instance, ensuring that broad access to data does not inadvertently give someone an unauthorised advantage or that all participants are legitimate. The existing eligibility criteria for sandbox access, which assess the bona fides of applicants and the public interest of their projects, would naturally extend to course providers as well.

Having the FCA (supported by other regulators) host the synthetic data sandbox is not only feasible but advantageous, provided governance structures are put in place. It aligns with the FCA's innovative track record (e.g. running TechSprints and the Digital Sandbox) and demonstrates a commitment to proactive support of the industry's evolution.

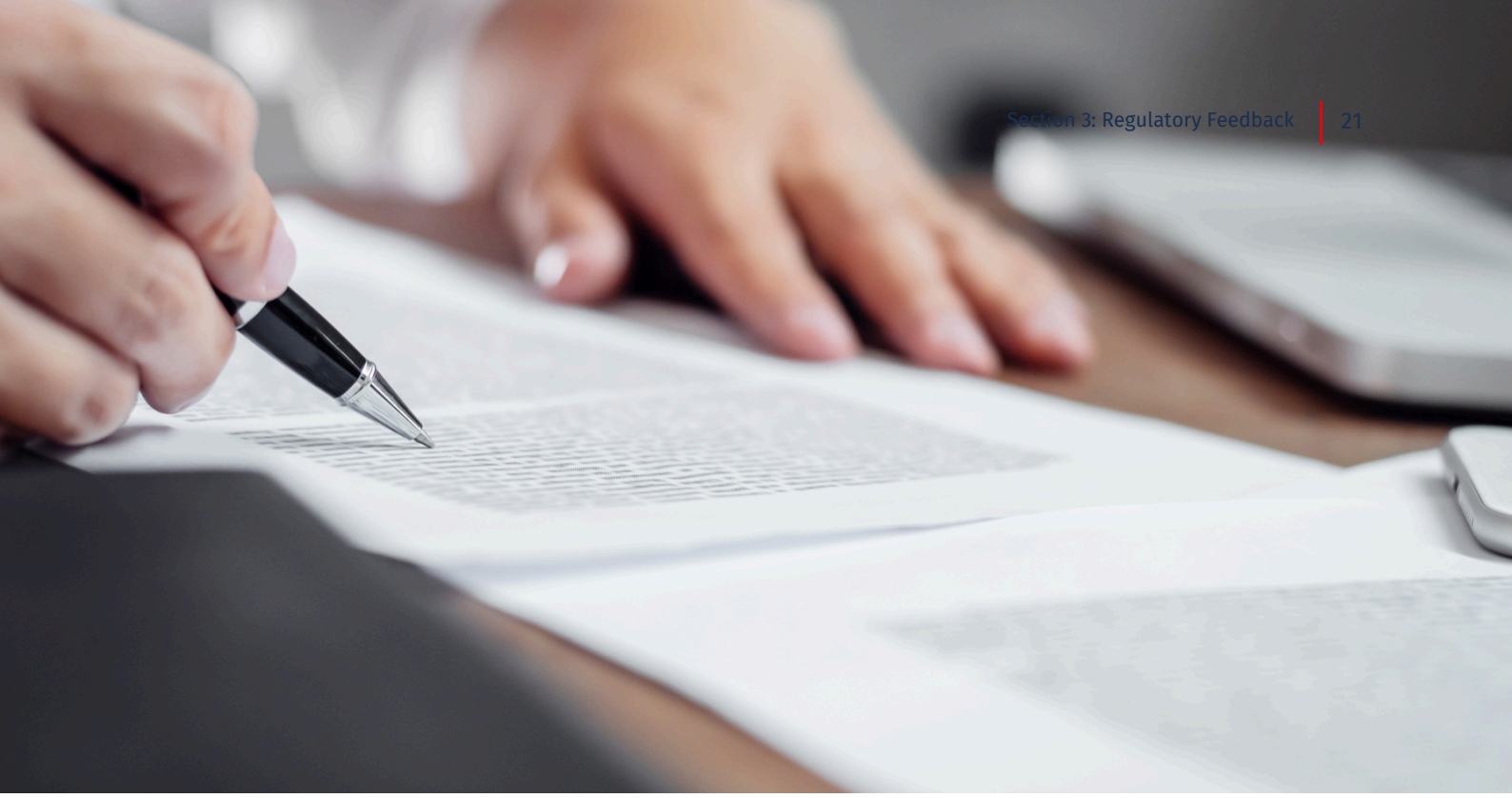
Alignment with UK Government AI Policy (AI Opportunities Act)

This initiative strongly aligns with the direction of UK government AI strategy and recent policy recommendations. The AI Opportunities Action Plan (2025), led by Matt Clifford, emphasises 'ramping up AI adoption across the UK to boost economic growth' and highlights the need for both improved data access and widespread skills development.[23] One key recommendation is that government and industry should 'explore use of synthetic data generation techniques to construct privacy-preserving versions of highly sensitive datasets' to unlock innovation. Our proposed synthetic data lake directly responds to this call by creating shareable, realistic financial data for training AI, without compromising privacy.

The Action Plan also calls for training tens of thousands of additional AI-skilled individuals in the coming years to meet demand. Recommendation 17 specifically suggests expanding alternative pathways like apprenticeships and employer-led training for AI, not solely relying on traditional degrees.

By working with apprenticeship providers, business schools and other skills providers to deliver experiential AI learning, this initiative embodies that approach, creating a practical route for upskilling current professionals and new entrants in AI skills relevant to finance. Moreover, the plan recommends (Recommendation 48) sector-specific interventions and the appointment of AI Sector Champions in industries such as financial services to drive adoption and address sectoral needs.

The synthetic data sandbox for financial services could be a flagship initiative under such a sector plan, demonstrating how government, regulators, and industry can collaborate to overcome key barriers. The UK government's vision is to make the country not just an AI research hub but also a leader in AI application across the economy; finance and professional services are crucial parts of the economy where the UK has a competitive edge. By improving AI proficiency in this workforce, the initiative would help maintain the UK's global lead in fintech and financial innovation.



It also complements other government efforts like Skills Bootcamps in AI and Data, the AI Scholarship programmes, and the emphasis on lifelong learning for digital skills. HM Treasury is likely to view this favourably as it strengthens the financial sector's resilience and innovation (a tech-savvy financial sector is better equipped to combat fraud, improve customer service, and develop new products). DSIT, which champions the National AI Strategy, will see this as a concrete example of turning strategy into action, combining data availability, regulatory support, and skills training to seize AI opportunities. Because this is a collaborative model (regulator with industry and educators), it could serve as a template for other sectors (e.g. synthetic health data for medical AI training, etc.), fitting into the broader AI ecosystem the UK is building.

The initiative is not only feasible within the current regulatory and policy framework, but it actively advances several high-priority policy goals expressed in the UK's AI plan and related government reports. It represents a proactive step in ensuring the UK's financial services workforce can safely and effectively harness AI, echoing the aims of the AI Opportunities Act to 'fairly and effectively seize the opportunities of AI' for economic growth and job creation.

Potential Blockers

Data Quality and Scope

Ensuring the synthetic datasets are sufficiently realistic, up to date and comprehensive is challenging. If the data lacks fidelity or important edge cases, the training value diminishes. Ongoing collaboration with industry data experts is needed to avoid a scenario where courses rely on oversimplified data. We must also prioritise which financial datasets to synthesise first; focusing on high-impact areas (e.g. fraud transactions, loan default data) will maximise usefulness. In speaking to industry, there were two areas that emerged as best potential themes: applied AI for FS data analytics or fraud detection. Views were dependent and varied on several factors including the individual's aims and the firm's position and skillsets. Though many echoed a warning: the pilot of this would likely have significant interest and uptake, if the data were poor quality, word would spread and we may not get a second chance.

Technical
Expertise and Resources

Developing and maintaining synthetic data at scale requires specialised skills that may be in short supply. The initiative must secure the right talent or partnerships. There is a risk that under-resourcing this could lead to delays or subpar data. Additionally, providing enough computing resources in the sandbox for potentially hundreds of simultaneous learners (if the programme grows) will require careful capacity planning.

Industry Adoption and Engagement

The success of the programme hinges on course providers and financial firms embracing it. We need to verify that apprenticeship programmes, business schools and other training providers are eager to develop content around the synthetic data lake and that employers will recognise and value the resulting skills. If industry uptake is slower than expected, the financial self-sufficiency model could falter. Early engagement with firms (possibly via the AI Sector Champion for finance) can mitigate this, ensuring the courses meet real employer needs and that firms might contribute data or funding.

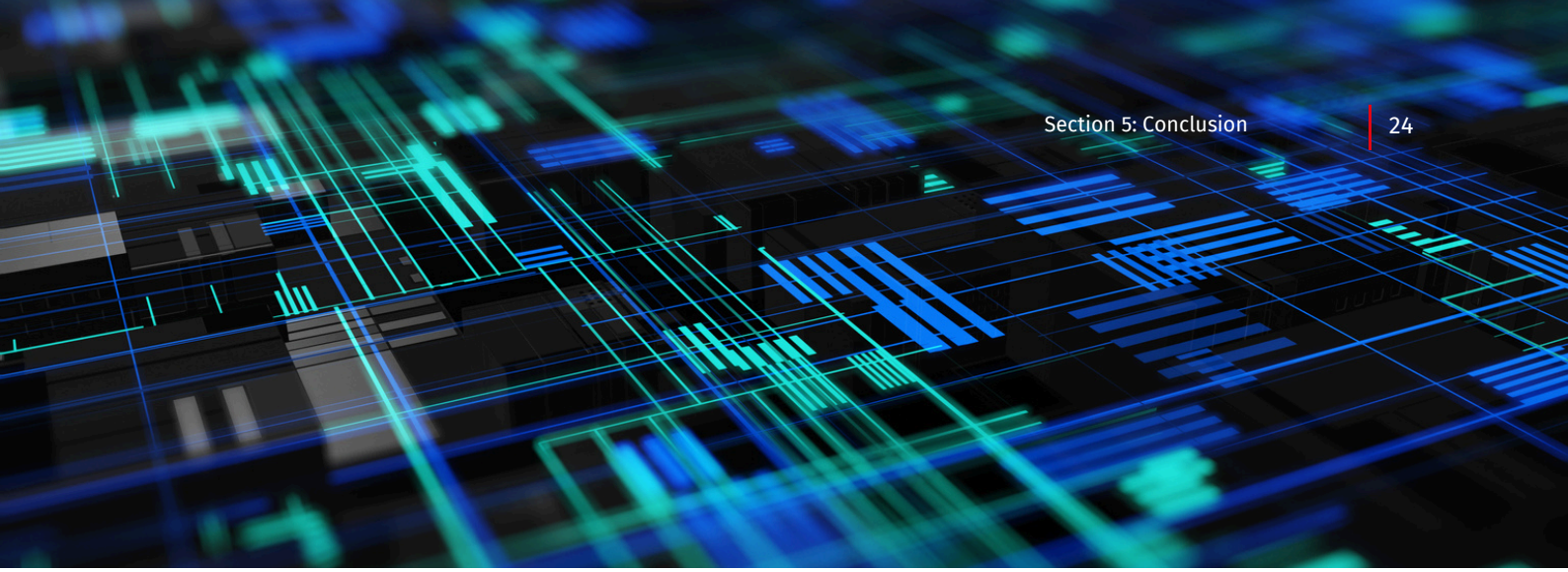
Legal/Compliance Uncertainties

While synthetic data largely circumvents personal data issues, it's a relatively new area in regulation. Guidance from the ICO on synthetic data practices is still evolving. We must remain vigilant to any changes in how synthetic data is treated under law. For instance, might future regulations require certification of anonymity or impose standards on synthetic data generation? Additionally, intellectual property rights in the generated datasets (who 'owns' the synthetic data) should be clarified, especially if industry partners contribute to its creation.

courses on some widely available platforms were 'good but still generic…this has always been a problem. What you learn from a book is good, but you need real-world experience to make it make sense.'
- Asset Management firm

Platform Governance and Competition Considerations

If the FCA is hosting a central data resource, we must ensure this does not inadvertently favour certain providers or become a bottleneck. Governance should be transparent, possibly involving multiple stakeholders, to oversee dataset updates and access rules. Also, as the sandbox becomes more widely used for training, we should watch out for any antitrust issues (unlikely, but for completeness: e.g., if access were restricted in a way that benefits some commercial trainers over others). Keeping the platform open to all qualified course creators will be important.
Each of these issues can be managed with proactive planning and stakeholder consultation. They should be examined in a follow-up discussion to develop risk mitigation strategies and ensure the initiative's robust implementation.



Conclusion

In conclusion, the proposal to create a synthetic data lake via the FCA's supercharged Digital Sandbox to bolster AI skills in financial services is highly feasible and aligns well with the UK's innovation and skills agenda. Technically, the ingredients are in place, synthetic data technology is maturing and the FCA's sandbox provides a ready platform, though attention must be given to data fidelity and secure integration.

Financially, while upfront investment is needed, the initiative can leverage existing infrastructure and move toward a cost-sharing model as industry buy-in grows, potentially achieving self-sufficiency as the value is demonstrated. Regulatorily, it fits within current legal frameworks and exemplifies forward-thinking regulation, using synthetic data to reconcile the twin goals of data privacy and open innovation. It directly supports the government's AI adoption plans by addressing the data access barrier and accelerating human capital development in AI. By training professionals on realistic financial scenarios in a safe environment, the programme will help financial and professional services firms unlock AI-driven improvements, from better fraud prevention to more personalised customer services, thereby driving competitiveness and growth.

For HM Treasury, DSIT, and the financial regulators, this initiative represents a prudent investment in the infrastructure for innovation: much like physical sandboxes allow fintech startups to experiment under oversight, this data sandbox for skills will allow our workforce to experiment and learn, underpinned by regulatory confidence. With careful execution and collaborative governance, the synthetic data lake initiative can become a cornerstone of the UK's efforts to build an AI-ready financial sector workforce, ensuring that the UK remains at the forefront of both financial innovation and responsible AI deployment.

References

Research cited throughout the report is document below:

1. Financial Services Skills Commission. (2024). Future Skills Report 2024
2. UK Finance. (2023). The impact of AI in financial services
3. Financial & Professional Services: The Future of AI and the Workforce
4. Financial Services Skills Commission. (2024). Future Skills Report 2024.f
5. IBM. (2023). Female Leadership and AI Report – EMEA.
6. World Economic Forum. Why employers need to embrace upskilling for their workforce (2023)
7. University College London (UCL). Experiential learning benefits students' progress in science and general academic skills.
8. British Educational Research Association (BERA). The effect of experiential learning on academic achievement of children aged 4–14.
9. International Review of Psychiatry. Experiential learning spaces and student wellbeing.
10. Harvard Business Publishing. Experiential Learning.
11. UK Government. AI Opportunities Action Plan
12. Financial Conduct Authority (FCA). (2022). Synthetic data to support financial services innovation: Call for Input
13. Thomas, R., Probyn, E. & Horton, S. Synthetic data and financial services: the FCA's Call for Input.
14. City of London Corporation and FCA launch project to tackle authorised push payment scams
15. Exploring Synthetic Data Validation – privacy, utility and fidelity | FCA
16. Thomas, R., Probyn, E. & Horton, S. Synthetic data and financial services: the FCA's Call for Input.
17. Ibid
18. Financial Conduct Authority. Exploring Synthetic Data Validation – Privacy, Utility and Fidelity
19. Financial Conduct Authority. Supercharged Sandbox Participation Pack
20. Financial Conduct Authority. (2022). Business Plan 2022/23.
21. Financial Conduct Authority. (2022). Business Plan 2022/23.
22. Financial Conduct Authority. AI Lab
23. UK Government. AI Opportunities Action Plan

About The City of London Corporation:

The City of London Corporation is the governing body of the Square Mile dedicated to a vibrant and thriving City, supporting a diverse and sustainable London within a globally successful UK.

We aim to: •

- Contribute to a flourishing society
- Support a thriving economy
- Shape outstanding environments.

By strengthening the connections, capacity and character of the City, London and the UK for the benefit of people who live, work and visit here.

www.cityoflondon.gov.uk

About the Global City campaign:

The Global City campaign is The City of London Corporation's overarching initiative to promote the UK as a world-leading international financial centre. It showcases the UK as a great place for financial and professional services firms to invest, locate and grow.

Theglobalcity.uk