Privacy-Enhancing Technologies
PRIZE CHALLENGES

Financial Crime Prevention Technical Brief

Transforming Financial Crime Prevention through Federated Learning with End-to-End Privacy

Background

Federated learning (FL), or more generally collaborative learning, shows huge promise for machine learning applications derived from sensitive data. FL enables training on distributed datasets without raw data being shared amongst the participating parties.

The goal of this prize challenge is to mature federated learning approaches and build trust in adoption by accelerating the development of efficient privacy-preserving federated learning solutions that leverage a combination of input and output privacy techniques to:

- Drive innovation in the technological development and application of novel privacy-enhancing technologies
- Deliver strong privacy guarantees against a set of common threats and privacy attacks
- Develop technology that is capable of generating effective models for identifying anomalous financial transactions when applied to real world data.

This challenge use case is focused on enhancing cross-organization, cross-border data access to support efforts to combat fraud, money laundering and other financial crime. Participants are asked to develop innovative, privacy-preserving FL solutions to enable the detection of potentially anomalous transactions, utilizing synthetic transaction data created by SWIFT, the global financial messaging provider, and synthetic account-related data representative of data held by banks. “Anomalous transactions” covers a range of payments that vary significantly from the norms seen in the dataset, and thus may be indicative of fraud, money laundering, or other financial crime. For example, a transaction that is of an unexpected amount or currency, uses unusual corridors (senders/receivers), has unusual timestamps, or contains other unusual fields. The training datasets are labeled with anomalies, and therefore participants do not need a detailed understanding of financial crime issues.

For the purposes of this challenge, a privacy-preserving solution is defined as one which ensures that sensitive attributes in the datasets remain confidential to the respective data owners across the machine learning lifecycle. This requires that access to the raw data is protected (input privacy), and that sensitive information cannot be reverse-engineered during model training or inference (output privacy).

This is a high-impact and exciting use case for novel privacy-enhancing technologies. There are currently challenging trade-offs between enabling sufficient access to data to build tools for effectively detecting illegal financial activity, and limiting the identifiability of innocent individuals represented in the data, as well as ensuring their personal information is kept confidential. The scale of the problem is vast: the UN estimates that US$800-2000bn is laundered each year, representing 2-5% of global GDP.

Though novel innovation for this use case alone could achieve significant real-world impact, the challenge is designed to incentivize development of privacy technologies that can be applied to other use cases where data is distributed across multiple organizations or jurisdictions, both in financial services and elsewhere. The best solutions will deliver meaningful innovation towards deployable solutions in this space, with consideration of how to evidence the privacy guarantees offered to data owners and regulators, but also have the potential to generalize to other situations.

FL produces a global model that aggregates local models obtained from distributed parties. As data from each participating party does not need to be shared, the approach provides a baseline level of privacy. However, privacy vulnerabilities exist across the FL lifecycle. For example, as the global federated model is trained, the parameters related to the local models could be used to learn about the sensitive information contained in the training data of each client. Similarly, the released global model could also be used to infer sensitive information about the training datasets. Protecting privacy across the FL pipeline requires a combination of privacy-enhancing technologies and techniques that can be deployed efficiently and effectively to preserve privacy while still producing ML models with high accuracy and utility. The core of this challenge is to develop FL approaches that provide such end-to-end privacy, in accordance with the privacy threat profile detailed in the *Privacy Threat Profile* section of this document.

Structure

The challenge is split into three phases:

- **Phase 1: White Paper** (also referred to as a Concept Paper, with the two used interchangeably). You will develop a technical white paper that describes your proposed approach
- **Phase 2: Solution development.** You will build and develop the solution proposed in your white paper
- **Phase 3: Red Teaming.** The top solutions will be tested by competing red teams.

Further details about the phases are provided below, and on the challenge website.

A range of support and opportunities will be provided to participants during the challenge:

- Funding (see separate UK and US details on the challenge website)
- Opportunities to engage with data protection and financial regulators, and public sector organizations operating in the financial crime space, namely:
    - UK Information Commissioner's Office (ICO)
    - UK Financial Conduct Authority (FCA)
    - UK National Economic Crime Centre (NECC)
    - US Financial Crimes Enforcement Network (FinCEN).
- Opportunities to engage with financial institutions
- Technical support and guidance from the SWIFT Innovation Team, including workshops during Phase 1 of the challenge detailing the use case and how to work with the provided datasets.

The organizers plan to offer opportunities to showcase the best solutions in front of a global audience at the second Summit for Democracy, to be convened by President Joe Biden, in the first half of 2023.

The Challenge

Objective

The objective of the challenge is to ***develop a privacy-preserving federated learning (PPFL) solution*** that is capable of training an anomaly detection model on the datasets, while providing a demonstrable level of privacy against the defined threat profile.

This PPFL solution should aim to:

- Provide robust privacy protection for the collaborating parties
- Minimize loss of accuracy in the model, as compared to a centralized model
- Minimize additional computational resources (including CPU, memory, communication), as compared to a centralized model.

In addition to this, the evaluation process will reward participants who:

- Display a high degree of innovation

- Demonstrate how their solution (or parts of it) could be applied or generalized to other use cases
- Effectively prove or demonstrate the privacy guarantees offered by their solution, in a form that is comprehensible to data owners and regulators
- Consider how their solution, or a future version of it, could be applied in a production environment.

Datasets

Organizers will provide two datasets to participants via a secure method:

1. **Dataset 1:** A synthetic dataset representing transaction data created by SWIFT, the global provider of secure financial messaging services
2. **Dataset 2:** Synthetic customer / account metadata flags representative of data held by banks

There are approximately 4 million rows across the two datasets.

Note: The challenges are based on synthetic data to minimize the security burden placed on participants during the development phase; of course, the intent of the challenge is that privacy solutions are developed that would be appropriate for use on real datasets with demonstrable privacy guarantees. However, participants must adhere to a data use agreement for the synthetic data (see Annex A).

Dataset 1: synthetic transaction data created by SWIFT

Challenge participants will be provided a synthetic dataset representing transaction data created by SWIFT, the global provider of secure financial messaging services. Each row in this dataset is an individual transaction, representing a payment from one sending bank to one receiving bank. The dataset will:

- Contain data elements as defined in the ISO20022 pacs.008 / MT103 message format
- Comprise transactions between fictitious originators and beneficiaries, sender and receiving banks, payment corridor, monetary amount and timestamps.

Expertise in financial crime or ISO20022 messaging is not an expected prerequisite for entering the challenge, and the assessment process will not focus on detailed understanding of the use case itself. However, participants unfamiliar with this space may find it helpful to consult a general introduction to ISO20022, e.g., https://www.swift.com/campaign/iso-20022/iso-20022-dummies. Participants may also find the ISO20022 message definitions informative.

The dataset will reflect a snapshot of transactions sent by an ordering customer or institution to credit a beneficiary customer or institution. The dataset will cover roughly a month's worth of transactions involving 50 institutions.

The synthetic data is not generated based on any real traffic and will not contain any statistical properties of the real SWIFT transaction data (SWIFT will apply normal and uniform distributions).

The dataset will contain the fields described in Table 1 below.

Table 1: synthetic transaction data

Field	Description
MessageId	Globally unique identifier within this dataset for individual transactions
UETR	Unique End-to-end Transaction Reference - a 36-character string enabling traceability of all individual transactions associated with a single end-to-end transaction
TransactionReference	Unique identifier for an individual transaction
Timestamp	Time at which the individual transaction was initiated
Sender	Institution (bank) initiating/ordering the individual transaction
Receiver	Institution (bank) receiving the individual transaction
OrderingName	Name for the originating ordering entity
OrderingAccount	Account identifier for the originating ordering entity (individual or organisation) for the end-to-end transaction
OrderingStreet	Street address for the originating ordering entity
OrderingCountryCityZip	Remaining address details for the originating ordering entity

BeneficiaryName	Name for the final beneficiary entity
BeneficiaryAccount	Account identifier for the final beneficiary entity (individual or organisation) for end-to-end transaction
BeneficiaryStreet	Street address for the final beneficiary entity
BeneficiaryCountryCityZip	Remaining address details for the final beneficiary entity
SettlementDate	Date the individual transaction was settled
SettlementAmount	Value of the transaction net of fees/transfer charges/forex
InstructedCurrency	Currency of the individual transaction as instructed to be paid by the Sender
InstructedAmount	Value of the individual transaction as instructed to be paid by the Sender
Label	Boolean indicator of whether the transaction is anomalous or not. This is the target variable for the prediction task.

Each row in this dataset is an individual transaction, representing a payment from a sender bank to a receiver bank. An end-to-end transaction is a transaction from an originating ordering entity (a.k.a. ultimate debtor) to a final beneficiary entity (a.k.a. ultimate creditor) and may involve one or more individual transactions.

The end-to-end transaction is one individual transaction in the case where the originating orderer's bank sends payment directly to the final beneficiary's bank. However, it may be the case where the payment is not directly sent, but is instead routed through one or more intermediary banks. In such a case, there are multiple individual transactions belonging to the single end-to-end transaction, with each individual transaction representing a bank-to-bank payment. Each end-to-end transaction is uniquely identified by the UETR field. In the case of a sequence of multiple individual transactions for one end-to-end transaction, all individual transactions share a value for UETR, and the Sender and Receiver banks form a chain from the originating ordering bank through one or more intermediary banks to the final beneficiary bank.

Because each end-to-end transaction is defined by one originating orderer and one final beneficiary, this means the Ordering* columns for the orderer and Beneficiary* columns for the beneficiary have been included in this dataset in a denormalized fashion—the values are duplicated across all the individual transactions (rows) belonging to the same end-to-end transaction. Additionally, this means that the OrderingAccount and BeneficiaryAccount in a given row may not necessarily belong to the bank in that row's Sender and the bank in that row's Receiver, respectively. The correct way to associate an OrderingAccount to the correct bank is to identify the Sender bank in the originating (first) individual transaction in that end-to-end transaction, and the correct way to associate a BeneficiaryAccount to the correct bank is to identify the Receiver bank in the final (last) individual transaction in that end-to-end transaction.

MessageId	UETR	Sender	Receiver	OrderingAccount	BeneficiaryAccount	...
...	...	...	...	...	...	...
10	00012345-...	A	B	111	222	...
11	00012345-...	B	C	111	222	
12	00012345-...	C	D	111	222	
...	...	...	...	...	...	...

Illustrative example showing how to associate the originating orderer and final beneficiary information with the correct banks for one end-to-end transaction made up of three individual transactions. The orderer and beneficiary account information is duplicated across all rows in this group, and the sender and receiver banks form a chain. The bank and account information of the originating orderer is highlighted in blue, and the bank and account information for the final beneficiary is highlighted in yellow.

Dataset 2: Synthetic account data held by banks

Participants will also be provided access to synthetic account-related data representative of data held by banks. This dataset will contain account-level information including flags signaling whether the account is valid, suspended, etc.

The bank data provided during Phase 1 is generated using the same process described for the SWIFT synthetic dataset, with higher concentrations in certain currency flows to reflect regional characteristics. During Phase 2, participants will be provided with new synthetic bank data generated using a modified process that provides higher quality statistical properties. This dataset will have an identical structure and schema to that provided in Phase 1, but is intended to be of higher quality, potentially enabling models of higher accuracy to be developed.

The dataset will contain the fields described in Table 2 below.

Table 2: Synthetic account-related data

Field	Description
Bank	Identifier for the bank
Account	Identifier for the account
Name	Name of the account
Street	Street address associated with the account
CountyCityZip	The remaining address details associated with the account
Flags	Enumerated data type indicating potential issues or special features that have been associated with an account. Flag definitions are: ● 00 - No flags ● 01 - Account closed ● 03 - Account recently opened ● 04 - Name mismatch ● 05 - Account under monitoring ● 06 - Account suspended ● 07 - Account frozen ● 08 - Non-transaction account ● 09 - Beneficiary deceased ● 10 - Invalid company ID ● 11 - Invalid individual ID

Data can be linked using the Account field in the bank data and the OrderingAccount or BeneficiaryAccount in the SWIFT transaction data. Please see the previous section for details on how to identify which bank an OrderingAccount or BeneficiaryAccount should be linked to.

Note that bank nodes will not have access to data on the SWIFT node and vice-versa—a case of vertical data partitioning. It is up to you to determine how to exchange this information in a secure and private way.

Also note that the flags may not be representative of real-world practices. For example, in the real world, banks may use different flags and may interpret or weight them differently based on appetite for risk.

Evaluation datasets

The datasets being provided are intended for local development use in both Phase 1 and Phase 2. The transaction dataset has been split in time – the bulk of the dataset is the training set, and the final week of the dataset is a test set. The prediction task, as detailed in a later section, is to predict a confidence score for each individual transaction in the test set as to whether it is an anomalous transaction. The ground truth is provided for both the training and set sets in the development dataset.

In Phase 2, a separate and held-out dataset will be used for solution evaluation. You can expect the Phase 2 evaluation dataset to be of the same size and schema as the development dataset. The statistical distributions of the evaluation dataset will generally be similar to the development dataset, but some aspects may be changed that should be learnable by your model. In Phase 2, you will submit code for your solution to a code execution environment. The code execution runtime will run cold-start federated training on the new dataset's training split and then run inference to generate predictions for the new dataset's test split. Your solution's performance will be measured by evaluating its predictions against the ground truth for the new dataset's test split.

Challenge scenario: developing privacy-preserving anomaly detection models

The analytical objective is to **train a model that enables SWIFT to identify anomalous transactions**. In the context of this challenge, this is a classification model to be trained on provided training data with ground truth labels. In real world deployments, such transactions might be subject to additional verification actions or flagged for further investigation, dependent on context.

A number of banks are working with SWIFT to collaboratively train such a model. The parties are working jointly to do this, and can take a common approach to technical design, infrastructure etc., but are **not able to enable access to each other's raw data.** In the real world there are a number of barriers that might prevent this; banks are subject to a variety of privacy, competition and financial industry regulations, may be operating in different jurisdictions, and have legitimate commercial and ethical reasons for not sharing customer data with competitors.

The key task of this challenge is to design a privacy solution so that SWIFT can safely train and deploy such a model without compromising the privacy requirements (more details on the requirements, and an associated threat model, are described below).

For the purposes of the challenge, participants should demonstrate their solution by training two models:

- M_C = a centralized model trained on datasets 1 and 2 in a non-privacy-preserving way
- M_{PF} = a privacy-preserving federated learning model trained using their privacy solution.

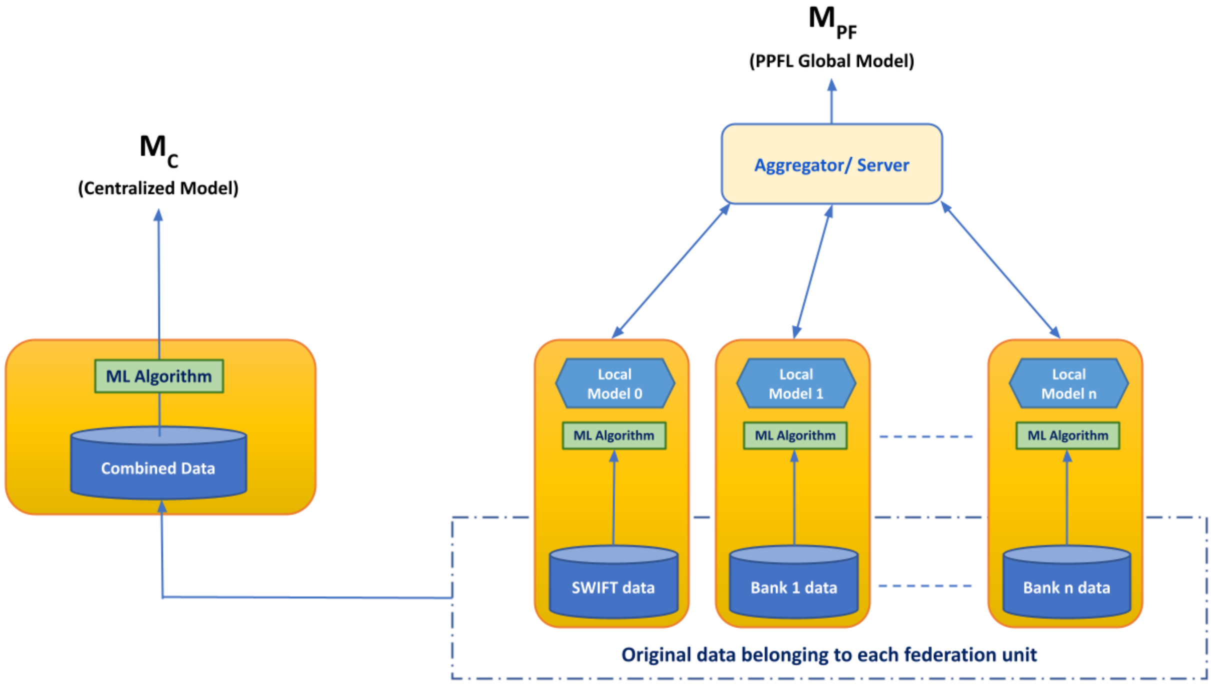

Figure 1. Simple illustration of a centralized ML model and a privacy-preserving federated learning model

SWIFT will provide participants with sample Python code for training a centralized anomaly detection model (M_c) on the ISO20022 training data. This code snippet will use the datasets provided as input and train a simple anomaly detection model using the XGBoost classifier. Participants are permitted to use this baseline code as the basis of their PPFL solution, or take an entirely different analytical approach.

In either case, the core of the evaluation will be assessing the comparison between a centralized model M_C (trained without consideration of privacy), and an alternative model M_{PF} that combines a federated learning approach with innovative privacy-preserving techniques.

In the real world, SWIFT may wish to train a model collaboratively with a number of banks, in order to increase the volume and variety of data being used to train the model. Participants should therefore aim to develop scalable solutions that enable additional nodes to be integrated into the federated network whilst incurring an acceptable additional performance overhead.

The federated learning scenario thus consists of one node hosting the synthetic SWIFT dataset, and N nodes hosting bank data. We will evaluate solutions for values of N between 1 and 10, in order to assess how well solutions scale as more banks are added to the network. This approach will be further detailed in an *Evaluation Methodology* document to be supplied to participants at the start of Phase 2. During solution development, participants have full autonomy over how they partition the bank dataset in order to understand the scalability of their solutions.

Details of the evaluation criteria are given below which, at a high level, consider:

- The ability of the solution to deliver (and evidence) relevant privacy properties
- The accuracy of model M_{PF} compared to M_C
- The performance/computational cost of training M_{PF} compared to M_C
- The scalability, usability, and adaptability of the solution.

It is important to note that the **accuracy and performance measurements are comparative**; the challenge is designed to reward strong privacy solutions which minimize accuracy loss and can be run with acceptable compute, memory, storage and communication costs. Privacy solutions which can support more effective machine learning approaches are encouraged (and will likely score higher in some areas), but the overall accuracy of the centralized model M_C is not a key factor in scoring.

Participants are free to determine the set of privacy technologies they use, with the exception of specialized or bespoke hardware. This exclusion is to ensure a fair baseline for Phase 2 evaluation. Solutions will be evaluated in a common technical environment, with each solution running on identical (virtualized) hardware, with access to the same compute, memory, storage and network infrastructure. We are therefore unable to provide access to specialist hardware such as secure enclaves/trusted execution environments in this challenge. However, the challenge organizers retain a deep interest in hardware-based privacy technologies, and encourage researchers or companies working in this space to engage with us to explore how we could collaborate in future to advance research and adoption of such technologies.

There are no restrictions on the software that challenge participants use in their solutions. We anticipate that proposals may leverage various de-identification techniques, differential privacy, cryptographic techniques, or combinations thereof. But this is not a prescriptive list, and we highly encourage submissions that propose novel technological approaches, or innovative application of existing technologies.

Privacy Threat Profile

Overview

Participants will design and develop end-to-end solutions that preserve privacy across a range of possible threats and attack scenarios, through all stages of the machine learning model lifecycle. Participants should therefore carefully consider the overall privacy of their solution, focusing on the protection of sensitive information held by all parties involved in the federated learning scenario. The solutions designed and developed by participants will include comprehensive measures to address the threat profile described below. These measures will provide an appropriate degree of resilience to a wide range of potential attacks defined within the threat profile.

Scope of sensitive data

Participants' solutions must prevent the unintended disclosure of a) sensitive information in the synthetic SWIFT transaction dataset; and b) sensitive information in the bank dataset, to any other party, including other insider stakeholders (for example, SWIFT and other financial institutions) and outsiders.

The following information in the dataset should be treated as sensitive:

- For the synthetic SWIFT dataset: All personally-identifiable information about the originating orderer (a.k.a ultimate debtor) and final beneficiary (a.k.a) ultimate creditor parties, for example personal details like names and addresses, and group membership information. This includes but is not limited to the raw private data about the orderer/beneficiary stored directly in the Account number, name, and address fields, and the Transaction identifiers and Timestamps.
- For the synthetic bank dataset: All personally identifiable information about parties involved in the transactions, for example personal details like names

and addresses, and group membership information. This includes but is not limited to the raw private / business data reflected in Account numbers/Names/Addresses and Flags fields.

Lifecycle

Participants will consider risks across the entire lifecycle of a solution including, in particular, the following stages:

- Training
    - Raw training data should be protected appropriately during training
    - Sensitive information in the training data should not be left vulnerable to reverse-engineering from the local model weight updates.
- Prediction/inference
    - Sensitive information in the training data should not be left vulnerable to reverse-engineering from the global model. The privacy solution should aim to ensure that those with access to the global model cannot infer sensitive information in the training data for the lifetime of the model's production deployment.

Actors and intention

Participants will consider threat models that range from honest-but-curious[1] to malicious (aggregators and participating clients) and propose solutions accordingly. While participating organizations can be trusted, such threat models help capture a broad spectrum of possible risks, such as the outsourcing of computation to the untrusted cloud; and, in the event trusted private cloud infrastructure is used, the remaining possibility that malicious external actors could compromise part of that infrastructure (for example one or multiple banks), leading to a potential reduction in the trustworthiness of components within the system.

Privacy attack types

Any vulnerabilities that could lead to the unintended exposure of private information could fundamentally undermine the solution as a whole. Participants will therefore consider a range of known possible privacy attacks, and any new ones relevant to the privacy techniques employed or to this specific use case. Participants will primarily be expected to consider inference vulnerabilities and attacks, including the risks of membership inference and attribute inference. In the white paper,

[1] An honest-but-curious party is a legitimate party in the federated learning scenario who will not attempt to deviate from the defined protocol but will attempt to learn all possible information from data received legitimately from other parties.

participants should clearly indicate the threat models considered and any other assumptions.

Participants will be expected to address the risks associated with the considered threat model through the design and implementation of technical mitigations in their solutions, and to explain in their white paper how solutions will mitigate against these. Participants will be expected to consider whether technical innovations introduced in their proposed solution may introduce novel privacy vulnerabilities and to clearly articulate potential privacy attacks and mitigations. Throughout both the white paper and solution development phases, participants should also take into account established privacy and security vulnerabilities and attacks, and corresponding best practice mitigations.

Challenge Phases & Evaluation

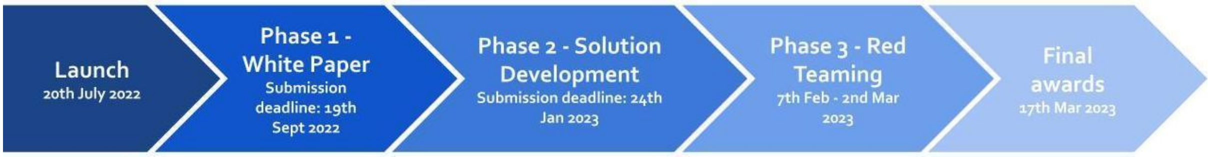

Figure 2. Timeline for the challenge phases

Phase 1: White Paper

In Phase 1, participants are asked to produce a technical white paper setting out their proposed solution, and provide information to support initial decisions about funding eligibility and Phase 1 prize awards. It is expected that some initial implementation/prototyping activity may be already underway when participants submit white papers, but code does not need to be submitted at this point.

White papers should be a **maximum of 10 pages** (excluding references), with size 11 font.

In addition to the white paper, participants need to respond to a number of other questions to establish their eligibility to participate in the prize challenge. The details of these, and the process for submitting them, are different for UK and US participants - please see information on the challenge website.

The technical white paper should:

- Clearly describe the technical approaches and sketch out proof of privacy guarantees based on the threat model considered, including:
    - The design of any algorithms, protocols, etc. utilized

    - Formal or informal arguments of how the solution will provide privacy guarantees.

- Clearly list any additional privacy issues specific to the technological approaches used
- Justify initial enhancement or novelty compared to the state-of-the-art
- Articulate:
    - Expected efficiency and scalability of the privacy solution
    - Expected trade-offs between privacy and accuracy/utility
    - How the explainability of model outputs may be impacted by your privacy solution.
- Describe how the solution will cater to the types of data provided to participants, and articulate what additional work may be needed to generalize the solution to other types of data.

Evaluators will score the white papers against the weighted criteria outlined in the table below. Note that the different criteria are not fully independent of one another. For example, solutions will likely need to carefully consider trade-offs between privacy and accuracy, accuracy and efficiency, etc. Participants should take the weightings of the criteria into account when considering these trade-offs. Importantly however, proposals must demonstrate how acceptable levels of both privacy and accuracy will be achieved – one must not be completely traded off for the other (a fully privacy-preserving but totally inaccurate model is not of use to anyone). Proposals that do not sufficiently demonstrate how both privacy and accuracy will be achieved (as determined by an independent expert assessor) will not be eligible to score points in the remaining criteria.

Topic	Specific Criteria	Weighting (/100)
Technical Understanding	Does the white paper demonstrate an understanding of the technical challenges that need to be overcome to deliver the solution?	10
Privacy	Has the white paper considered an appropriate range of potential privacy attacks, and how the solution will mitigate those?	25

Accuracy	Is it credible that the proposed solution could deliver a useful level of model accuracy?	10
Efficiency and Scalability	Is it credible that the proposed solution can be run within a reasonable amount of computational resource (e.g., CPU, memory, storage, communication), when compared to a centralized approach for the same machine learning technique? Does the white paper propose an approach to scalability that is sufficiently convincing from a technical standpoint to justify further consideration, and reasonably likely to perform to an adequate standard when implemented? Solution scalability will be evaluated primarily for a) the number of connected banks/financial institutions involved, b) volume of transactions, and c) volume of customer data held by banks	15
Adaptability	Is the proposed solution potentially adaptable to different use cases and/or different machine learning techniques?	5
Feasibility	Is it likely that the solution can be meaningfully prototyped within the timeframe of the challenge?	10
Innovation	Does the white paper propose a solution with the potential to improve on the state of the art in privacy-enhancing technology? Does the white paper demonstrate an understanding of any existing solutions	20

	or approaches and how their solution improves on or differs from those?	
Usability and Explainability	Does the proposed solution show that it can be easily deployed and used in the real world, and provide a means to preserve any explainability of model outputs?	5

Phase 2: Solution Development

In Phase 2, participants are asked to develop working prototypes of their solutions. These solutions are expected to be capable of being used to train a model against the evaluation dataset, with measurement of relevant performance and accuracy metrics. However, we are not expecting fully productionized solutions; for example, participants will be able to actively support deployment on any testing platforms, and any test runs against evaluation data.

The following section describes how we plan to evaluate solutions. Further details, including technical details for submission, will be supplied in an *Evaluation Methodology* document to be supplied to participants at the start of Phase 2. Though our intent is that the details below will remain unchanged, organizers reserve the right to make changes to specific evaluation criteria or weightings if we consider that this is necessary to fairly and efficiently evaluate the full range of solutions proposed, or for other reasons.

Evaluation

Participants will submit the following for evaluation:

1. Centralized model M_C:

- Code for training a centralized model M_C
- Documentation for how to train and make inferences from the centralized model, including a list of any dependencies (e.g., a requirements.txt)

2. Privacy-preserving federated learning model M_{PF}:

- Code for training the PPFL model M_{PF}
- Documentation for how to train and make inferences from the PPFL model (e.g., a requirements.txt).

3. Key metrics:

- Self-reported privacy, accuracy, and efficiency metrics for the two models (these will not be used to evaluate solutions, but help to flag potential issues if assessors obtain very different metric values).

Submitted solutions will be deployed and evaluated on a technical infrastructure hosted by the challenge organizers. This infrastructure will provide a common environment for the testing, evaluation, and benchmarking of solutions in accordance with the Phase 2 evaluation criteria below.

Participants will be provided with the following in the *Evaluation Methodology* document to be supplied at the start of Phase 2:

- Hardware specifications (CPU/GPU, memory, data storage etc.) for the runtime environment that will be used to benchmark and evaluate solutions
- Software specifications for the runtime environment, including any fundamental requirements for code to successfully execute
- Upper limits for a) execution time, and b) resource requirements (CPU/GPU, memory, data storage etc.) for code execution during the model training stage
- Example code submissions.

An independent assessor will take the following steps to perform the evaluation, using a sequestered training and test dataset that participants have not had access to:

1. Train the centralized model M_C on a single node, with synthetic SWIFT data and synthetic bank data stored on local disk
   - Measure efficiency metrics during training
   - Measure accuracy metrics using test data on the resultant model.
2. Train the PPFL model M_{PF} across N+1 nodes, where one node is the server/aggregator and the remaining nodes represent the banks. Predetermined values of N between 1 and 10 will be used to evaluate the scalability of the solutions
   - Measure efficiency metrics during training
   - Measure accuracy metrics using test data on the resultant model
   - Measure privacy metrics during training (to assess risk of leakage from local model updates and any other exchanged information) and inference (to assess risk of leakage from the resultant global model).

3. Qualitative assessments will be made of the solution's adaptability, usability, explainability, and level of technical innovation.

Details of the specific criteria and how they will contribute towards overall scoring are given in the table below. We will provide more specific information on how these will be measured practically in the *Evaluation Methodology* document to be provided prior to the start of Phase 2.

As with the Phase 1 evaluation, the different criteria are not fully independent of one another, and the outcomes of trade-off considerations made in the white paper should be reflected in the developed solution. Importantly, solutions must meet a minimum threshold of privacy and accuracy (which will be quantitatively measured) to be eligible to score points in the remaining criteria.

Topic	Factors	Weighting (/100)
Privacy	Information leakage possible from the PPFL model M_{PF} during training and inference, for a fixed level of model accuracy[2] Ability to clearly evidence privacy guarantees offered by solution in a form accessible to a regulator and/or data owner audience	35
Accuracy	Absolute accuracy of the PPFL model M_{PF} developed (e.g., F1 score) Comparative accuracy of PPFL model compared with a centralized model, for a fixed amount of information leakage	20

[2] For example: If differential privacy is employed to protect output privacy, and F_1 score is an appropriate accuracy metric, what is the smallest value for the privacy budget ε that can be configured to achieve an F_1 score that is a fixed amount less than the F_1 score of the centralized model M_C.

Efficiency and Scalability	Time to train PPFL model M_{PF} vs comparison with the centralized model M_C Network overhead of model training Memory (and other temporary storage) overhead of model training Ability to demonstrate scalability of the overall approach taken for additional nodes	20
Adaptability	Range of different use cases that the solution could potentially be applied to, beyond the scope of the current challenge	5
Usability and Explainability	Level of effort to translate the solution into one that could be successfully deployed in a real-world environment Extent and ease of which privacy parameters can be tuned Ability to demonstrate that the solution implementation preserves any explainability of model outputs.	10
Innovation	Demonstrated advancement in the state-of-the-art of privacy	10

	technology, informed by above-described accuracy, privacy and efficiency factors	

Partitions

For local development, you are provided a full, unpartitioned dataset. In Phase 2 evaluation, evaluation will occur with predetermined partitioning along institutional boundaries. The SWIFT data will always belong to a single federation unit that represents the SWIFT Data Store and only has access to the SWIFT data. Banks will be split up among federation units such that one bank's account data entirely belongs within one partition. In cases where there are fewer bank partitions than the number of banks, each bank partition may contain data from more than one bank.

Any partitioning of the data that you might perform in your local development experiments should take this into account. Your solution should be able to handle any number of bank partitions, and in Phase 2, we may evaluate your solution with a number of bank partitions between 1 and 10.

Prediction Target and Evaluation Metric

The target variable for the modeling task is a confidence score (between 0.0 and 1.0) for whether each individual transaction is anomalous. As discussed previously, anomalous is not precisely defined and should be learned by your model via supervised learning on provided training data.

The evaluation metric will be Area Under the Precision–Recall Curve (AUPRC), also known as average precision (AP), PR-AUC, or AUCPR. This is a commonly used metric for binary classification that summarizes model performance across all operating thresholds. This metric rewards models which can consistently assign anomalous transactions with a higher confidence score than negative non-anomalous transactions. AUPRC is computed as follows:

$$\text{AUPRC} = \sum_{n} (R_n - R_{n-1}) P_n$$

where P_n and R_n are the precision and recall, respectively, when thresholding at the nth individual transaction sorted in order of increasing recall.

Phase 3: Red Teaming/Testing

In Phase 3, red teams will plan and launch privacy attacks against the highest-scoring solutions developed in Phase 2. Solutions will be re-evaluated based on the outcomes of the red teaming attacks, and each solution will be assigned a final score which will be used to determine the allocation of prize awards. The criteria outlined in Phase 2 will be used for this re-evaluation, taking into account the impact of the red team attacks on the solutions. Most notably, it is expected that privacy scores will change according to how resilient the solution was to the red teams' privacy attacks.

Success of red team attacks will be assessed by a panel of judges using the criteria below, in order to evaluate the empirical results reported, the approaches taken and the severity of the flaws red teams are able to exploit. Details of the specific criteria and how they will contribute towards overall scoring are given in the table below.

Each red team will be assigned multiple solutions to test, with each solution therefore being tested by multiple red teams. Individual red teams will be scored, in part, by comparing the outcomes of attacks carried out against the same solutions by different red teams. Additionally, an individual red team's attacks against the solutions it was assigned to attack will be assessed for consistency, difficulty, novelty, rigor, and practicality.

Further details on red teaming and recruitment of red teams will be provided to participants in Autumn/Fall 2022.

Topic	Factors	Weighting (/100)
Effectiveness	How completely does the attack break the privacy claims made by the target solution? (e.g., what portion of user data is revealed, and how accurately is it reconstructed)?	40
Applicability / Threat Model	How realistic is the attack? How hard would it be to apply in a practical deployment?	30
Generality	Is the attack specific to the target solution, or does it generalize to other solutions?	20
Innovation	How significantly does the attack improve on the state-of-the-art?	10

Annex A: conditions of data use

To participate in Phase 2 of the challenge, participants are required to accept and comply with a lightweight data use agreement for one or both of the synthetic datasets. The datasets do not contain personal data, but their use is restricted to the purposes of this challenge.

The process for accepting the data use agreement, and securely downloading the data, is different for UK and US participants. Please consult the UK and US challenge briefing materials on the challenge website for further details.

Item	Description
Public communications	The Participants shall not directly or indirectly cause or permit (a) the oral or written release of any public statement referring to SWIFT's involvement in the Challenge and contribution of the datasets or (b) any use of SWIFT's name or trademarks, without SWIFT's prior written consent. Notwithstanding the foregoing, the Participants may acknowledge the existence and nature of SWIFT's involvement in the Challenge when required by applicable laws and regulations.
Synthetic datasets and other materials	SWIFT provides the synthetic cross-border payment transaction dataset and synthetic bank transaction datasets to support the Challenges. The datasets provided by SWIFT are fully synthetic. The Participants are allowed to use any transaction datasets provided by SWIFT only for the purpose of the Challenges. The use of the datasets provided by SWIFT for any other purposes is strictly prohibited. For Participants that would like extra support in AI/ML model development, SWIFT can provide sample Python code for training a centralized anomaly detection model on the ISO20022 training data. This code snippet will use the dataset provided as input and train a simple anomaly detection model using the XGBoost Classifier.
IP rights	**General conditions:** All IP rights in the synthetic datasets, sample code on AI/machine learning models, the methodology, code, materials, data, and any other materials provided by SWIFT shall remain vested in SWIFT exclusively. Except as otherwise provided herein, the Participants may only use the materials described in Annex A for the purposes of the Challenges. **Synthetic datasets** SWIFT grants the Participants of the Challenges a free, non-transferable, non-exclusive license to use these synthetic datasets only for the purposes of the Challenges. **Other materials** SWIFT grants the Participants a free perpetual, non-exclusive, non-transferable license to use such sample code for an AI/ML model provided by SWIFT for the Challenges for any purposes, including without limitation the rights to use, copy, modify, merge,

	publish, distribute, sublicense, and/or sell copies of the software developed based on such materials. Participants must agree to acknowledge SWIFT's IP rights for the sample code to the extent it is embedded into their own software, models, or solutions. **Disclaimer of Warranty** SWIFT's datasets and other materials are provided "as is", without warranty of any kind, express or implied, including but not limited to the warranties of merchantability, and fitness for a particular purpose. In no event shall SWIFT be liable for any claim, damages or other liability, whether in an action of contract, tort or otherwise, arising from, out of, or in connection with such materials or the use or other dealings in such materials. The Participants will own the IP rights upon creation by the Participants on any of their modifications, enhancements and upgrades of the sample code for AI/ML model provided by SWIFT for the Challenges. The Participants must agree to acknowledge SWIFT's IP rights in case they embed the code into their own software, models, or solutions. **Residuals** The Challenges do not limit in any way SWIFT's right to independently develop or acquire similar, or competing machine learning/AI models, PETs, products, processes or services. The Participants acknowledge that SWIFT or its affiliates may currently, or in the future, develop information internally or receiving information from third parties, that is substantially similar to the solutions developed by Participants during the Challenges. For the avoidance of doubt, SWIFT may use general knowledge acquired during the Challenges and other residual information for any purposes including without limitation use in development, manufacture, promotion, sale and maintenance of its products and services. Residual information means any information that is retained in the unaided memories of SWIFT or its affiliates' employees, consultants, or contractors who receive access to the Challenges information.
Liability and warranties	SWIFT will not bear any liabilities or offer any warranties related to its participation in the Challenges and the provisioning of the synthetic cross-border transaction datasets and other materials.
Use of winning solution(s)	SWIFT will have the right, following the Challenges, to enter into discussions with the Participants of its choice to explore the feasibility of adopting the winning solution(s) as part of its own products and services, at SWIFT's entire discretion. Collaboration with the winning Participant(s) is not part of the present terms and conditions, and will remain subject to separate bilateral agreements with the winner(s) of the Challenges, at SWIFT's discretion. Participants shall be under no obligation to enter into any outside agreements with SWIFT based on their participation in the Challenges.
Confidentiality of synthetic datasets	The distribution of any SWIFT datasets to any third party requires SWIFT's prior written approval. Public distribution or sharing of the SWIFT's datasets is strictly prohibited.
Security	The Challenges' Participants shall apply adequate technical and organizational security measures to protect the synthetic datasets provided by SWIFT and prevent any data leakage.

	Participants will, among others: ● Keep the data protected and accessible only to the people working on the Challenges. Data may not be distributed to any third parties; ● Always encrypt data in transit; ● Delete data after the Challenges are completed.
Personal data protection	Personal data processed during communication with SWIFT (if any) shall be processed and stored according to the applicable personal data protection legislation. The synthetic datasets provided by SWIFT do not include any personal data.

Annex B: Version History

Version/Date	Changes
V1.0 (20th July)	Initial version at launch
V1.1 (17th Aug 2022)	Various improvements to technical content to better align with the final version of the data sets as released to participants, specifically: ● In the “Dataset 1: synthetic transaction data created by SWIFT” subsection, adding a sentence to make clear what each row in the dataset represents: “Each row in this dataset is an individual transaction, representing a payment from one sending bank to one receiving bank.” ● Updates to the dataset details provided in “Table 1: synthetic transaction data”: ○ Introduction of a MessageId field - a globally unique identifier within this dataset for individual transactions ○ All field names converted to CamelCase (e.g. Transaction-reference → TransactionReference) ○ Modifications to the Beneficiary* and Ordering* field names to more accurately reflect the the nature of the underlying data ○ Settlement-Date/Currency/Amount split into two fields SettlementDate and SettlementAmount ○ Instructed-Currency/Amount split into two fields InstructedCurrency and InstructedAmount ○ Descriptions of the fields modified for clarity and consistency ● Content has been added immediately after “Table 1: synthetic transaction data” describing how end-to-end transactions are represented in the dataset. ● Updates to the dataset details provided in “Table 2: Synthetic account-related data”: ○ Bank field added ○ Account Identifier field renamed Account

	○ Entity Name field renamed Name ○ Entity’s Street Address field renamed Street ○ Entity’s County-City-Zip field renamed CountryCityZip ○ Flag definitions have been specified ● A clarifying sentence added on the bank flags: “Note that the flags may not be representative of real-world practices. For example, in the real world, banks may use different flags and may interpret or weight them differently based on appetite for risk.” ● New “Partitions” subsection added, detailing how the data is partitioned ● “Evaluation Datasets” section has been expanded to describe more clearly the different datasets that will be used during development and evaluation. ● New “Prediction Target and Evaluation Metric” subsection added, detailing how the Area Under the Precision–Recall Curve (AUPRC) will be used to assess model accuracy.