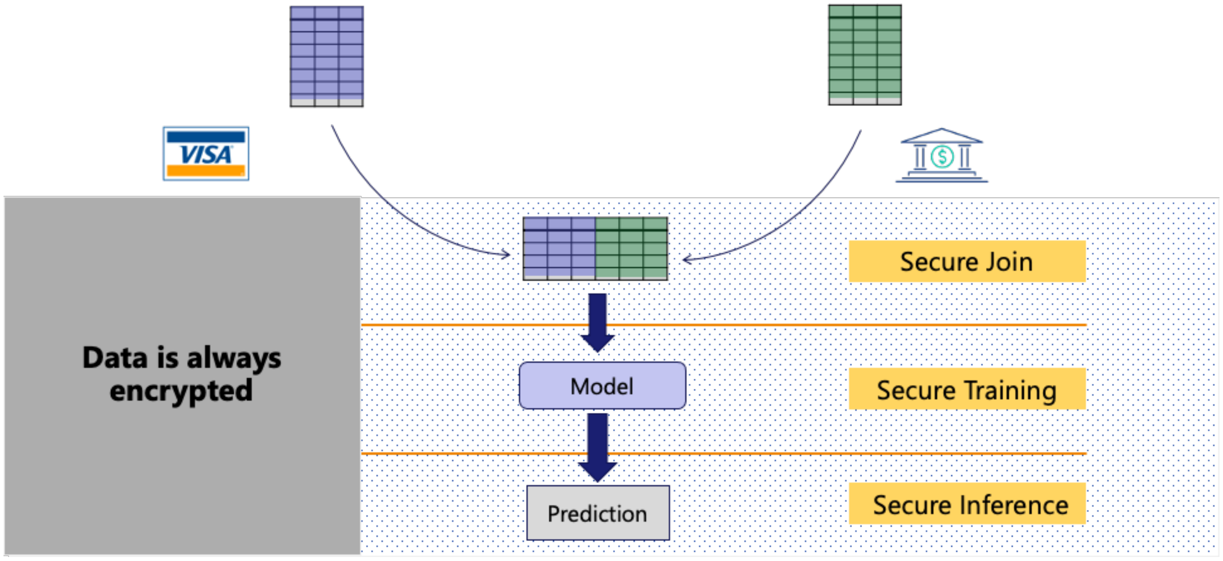
Secure Collaborative Machine Learning

Traditional cryptography has played a fundamental role in securing the payments ecosystems for the past few decades. New advances in cryptography allow us to further strengthen our protections and enable new capabilities while maintaining high levels of data privacy. Powerful techniques such as multi-party computation and homomorphic encryption, enable computations over encrypted data. Our research in these areas is facilitating flows which allow us to perform complex and collaborative data analytics without every exposing any of our, or our partner's, data.



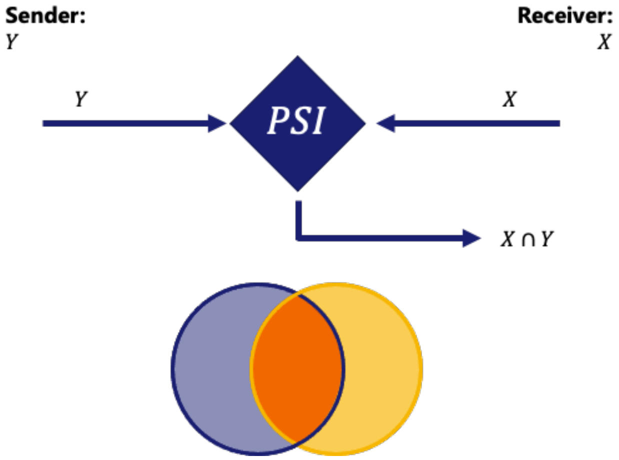
The ability to combine Visa's data with those of our partners has enormous potential to provide new insights and improve areas such as fraud detection. This begins with our pioneering work in Private Set Intersection to enable the joining of such datasets in a secure and efficient manner. Building on this we are developing techniques to efficiently perform machine learning in a fully privacy-preserving manner where both training and inference are executed over encrypted data. Combining these techniques creates a pipeline which ensures the confidentiality of data throughout the entire analytics process.

Secure Multiparty Computation (MPC) is a powerful collection of techniques which we use to help us achieve these complex tasks within the encrypted domain. So, what is the MPC model? In this context, each party locally encrypts their own data and holds the decryption key. These keys are chosen in a specific way that allows the parties to combine their encrypted data and perform encrypted computation. Decryption of the result can only be performed with the consent and participation of a quorum of the parties, e.g., all of them. The utility of this is twofold, first it provides greater data privacy. Secondly it guarantees that the data can only be used for the intended computation and can't be leaked or misused by one of the parties.

Data collaboration:

The first step in our secure data processing pipeline is for the parties providing their data to perform an encrypted join to combine associated data attributes. This process is achieved using an MPC technology called Private Set Intersection (PSI). Conceptually, this phase takes as input each party's data and generates an encrypted database table of these joined attributes. Here we are assuming the parties have some common set of identifiers that the datasets can be joined on. The internal workings of this phase can be quite involved and utilizes advanced encryption techniques.

Although the research topic of PSI has existed for a long time, dating back to the 1980s, only recently has advancements in encryption technology enabled practical deployments. For example, our recent publication [1] on PSI allows two datasets of one million items each to be joined in 0.2 seconds and 200MB of communication over the cost of performing the same computation on plaintext data. Compared to prior work this represents an order of magnitude improvement in performance.



Once the data is joined and encrypted, it can be used in subsequent encrypted computations to derive useful insights. For example, Visa and a partner bank could join their customer data to obtain a more holistic (and encrypted) picture of their customers. This is turn would enable Visa and the partner bank to make more accurate fraud predictions in case the customer's payment credentials are compromised in some way. Importantly, this is achieved while also respecting the customers' privacy since their actual data remains fully encrypted.

Privacy-preserving Machine learning:

This ability to perform fraud predictions on the encrypted user data is more generally known as privacy-preserving machine learning. Like machine learning itself, this can be divided into a training phase where a predictive model is generated based on existing data, and an inference phase where the model is used to predict some outcome of interest, e.g., whether a pending transaction is fraudulent. Privacy-preserving machine learning aims to enable both phases to operate on encrypted data.

Our secure data processing pipeline aims to enable both of these flows. During training our secret join/private set intersection techniques can allow the parties to join their respective datasets and then feed these into an encrypted machine learning framework to an encrypted model. All of this is achieved without any party being able to decrypt the data. Visa Research has developed several of the state of art techniques for performing encrypted machine learning in an efficient manner [2,3,4]. Once an encrypted model is produced it can either by used to make encrypted predictions on new data or the parties can agree to decrypt the model and use it in the traditional way.

While this technology offers great promise and has had several early successes. The main challenge in widespread adoption is the efficiency of the encryption schemes and the ease of use by machine learning practitioners. Today when preforming machine learning techniques in the encrypted domain result this results in large computational overheads. To address this, new techniques need to be developed an area which we are actively pursuing. Secondly, usability remains a big factor with the task of performing machine learning tasks in the encrypted domain as quite challenging to non-cryptographers. Developing tools which reduce this barrier to entry is also an active area of interest.

[1] *Blazing Fast PSI from Improved OKVS and Subfield VOLE.* Peter Rindal and Srinivasan Raghuraman.
[2] *SecureML: A System for Scalable Privacy-Preserving Machine Learning*. Payman Mohassel and Yupeng Zhang
[3] *ABY3: A Mixed Protocol Framework for Machine Learning.* Payman Mohassel and Peter Rindal
[4] *Practical Privacy-Preserving K-means Clustering.* Payman Mohassel, Mike Rosulek, and Ni Trieu