TripleBlind™

WHITEPAPER

DATA IS ABUNDANT AND UNDER UTILIZED

Unlock the Business Value of Data while
Preserving Privacy and Enforcing Compliance

INTRODUCTION

Privacy concerns, government regulations, and operational complexities prevent an estimated 43ZB of data stored by enterprises today from being accessed and realizing their business value.

This whitepaper delves into the current state of data, explains how enterprises are unable to unlock the full potential of their data, and argues that incumbent data privacy solutions are insufficient. It then presents a solution from TripleBlind, built with patented innovations on top of thirty years worth of verified, peer-reviewed technology to help enterprises work with protected data to improve processes and create new opportunities.

Applications of such technology will improve the accuracy of medical diagnoses, thwart hackers and fraudsters, and prevent the next big data breach.

CONTENTS

CURRENT STATE OF DATA

We live in a digital world where data is abundant but underutilized for a number of reasons.

Enterprises are only able to use small portions of their own data because it is mostly siloed due to data regulations and other barriers. Data residency laws, for example, require data to be kept within the borders of the country in which it is generated. Strict regulations are becoming increasingly common, with well over 100 countries adopting data residency laws in the past 5 years.

The number of localization laws increased from 67 in 2017 to more than 144 in 2021.[1]

Sharing insights derived from Personally Identifiable Information (PII) and Protected Health Information (PHI) with any partner is risky and requires the establishment of trust in both the people, processes, and technology involved.

Due to government regulations and liability risks, businesses are encouraged to safeguard data for the protection of their employees and customers. This often leads to the creation of data siloes, which lock up data out of reach from bad actors, but also render it unusable for legitimate purposes by authorized parties.

The extreme caution driving enterprises towards these strict preventative measures is valid, as hackers have developed sophisticated attacks to compromise state-of-the-art security initiatives.

Furthermore, when two companies, or separate offices of the same company, need to share data, the current norm is for the first party to encrypt and send its data to the counterparty, which decrypts the data for use. At this point, the counterparty has full access to the raw data, and the first party has lost permissions enforcement over how thier data may be used.

This process results in intended and unintended data abuses, privacy violations, regulatory compliance hassles, lengthy contract negotiations, and an abundance of stress and headaches.

[1] REFERENCE: https://itif.org/publications/2021/07/19/how-barriers-cross-border-data-flows-are-spreading-globally-what-they-cost

PRESENT DAY PRIVACY SOLUTIONS AND THEIR LIMITATIONS

While legal agreements and regulations set the stage for safe, private data usage, technology is required to ensure that data can be put to work effectively and compliantly.

There are three main types of data privacy which technology addresses. The first is privacy at-rest, which often includes encryption and security mechanisms to ensure that unauthorized users are never able to access protected data where it lives.

The second type is privacy in-transit, which applies when data is sent from one party to another with protections in place to prevent nefarious parties from intercepting the information.

The third type of privacy gaining considerable attention is privacy in-use, which aims to protect data while it is being computed upon. Privacy in-use innovations will greatly expand opportunities for compliant data usage across multiple parties, while eliminating the need to ever transmit or send raw data over the internet again.

Most present day privacy in-use technology solutions revolve around developments in known concepts such as homomorphic encryption, differential privacy, federated learning, synthetic data, tokenization, and secure enclaves.

Providers of these solutions, which have become known as Privacy Enhancing Technologies (PETs) or Privacy Enhancing Computation technologies (PECs), span the range of startups to major tech firms.

Some of the companies providing homomorphic encryption include Enveil, Baffle, Duality Technologies, PreVeil, Google, IBM, Intel, and Microsoft.

Immuta, Microsoft, and SAP use differential privacy.

Owkin, Nvidia, and Rhino Health provide federated learning.

Vendors that sell synthetic data are Statice, Mostly.ai, Syntho, MDClone, and Tonic.

For tokenization, IBM participates with Cloud Pak for security, along with Informatica with its Data Privacy Management product.

Lastly, Google Cloud, IBM Cloud, Microsoft Azure, Oasis Labs, and Anjuna offer secure enclaves as part of their solution.

Experts on these technologies will attest that each, while promising, have major limitations today:

- Homomorphic encryption is too slow for practical uses, fails to meet the criteria for most regulations, and is only useful for simple computations on tabular datasets.
- Differential privacy reduces data fidelity by adding noise to the data.
- Federated learning places an extremely high computational and storage burden on counterparties and leaves algorithm IP vulnerable.
- Tokenization and masking of data elements reduces the computational value of the dataset by stripping out potentially important fields and does not guarantee that individuals will not be re-identified using the remaining information.
- Secure enclaves require data to be compiled in one place and do not solve privacy challenges from regulations like HIPAA, GDPR, and data localization.
- Synthetic data reduces the accuracy and precision of computations, as it substitutes real data for fake data generated algorithmically from the real data.
- Finally, business agreements, the status quo, are written by expensive lawyers, take too long to negotiate, and rely on goodwill.

THE TRIPLEBLIND SOLUTION

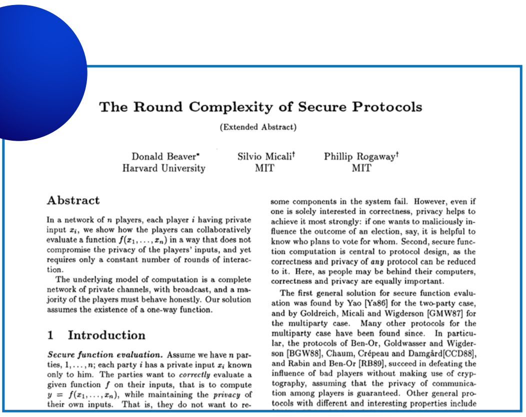
TripleBlind presents a solution built with innovations on top of principles which have been well understood and well documented via substantial peer reviewed papers and commercial use over the past 30 years. The extensions and improvements reduce these well known principles to greater practical use, by drastically improving the scalability, speed, and breadth of use cases involving protecting companies' data and algorithms in-use.

The unique solution removes the risks involved in data sharing by eliminating decryption and movement of raw data, while facilitating privacy-intact computations to occur. TripleBlind's technology enables companies to safely provide and consume sensitive data and algorithms in encrypted space, without compromising privacy, security, or scalability.

TripleBlind's technology helps companies use third party datasets and better leverage first party distributed datasets. Organizations may also make their data available for computation by others. The TripleBlind toolset allows users to easily gain insights from datasets owned by others, without taking possession of any data. Companies can utilize previously inaccessible data to glean new insights, improve the accuracy of machine learning models, and decrease algorithm bias.

New sources of revenue are unlocked through enabling companies to realize the business value of their data and algorithms without losing ownership of the asset or exposing intellectual property.

Data is inherently protected through the technology, so companies can collaborate freely without having to rely on "good faith" adherence to their terms of use.



The Round Complexity of Secure Protocols

(Extended Abstract)

Donald Beaver[*]
Harvard University

Silvio Micali[†]
MIT

Phillip Rogaway[†]
MIT

Abstract

In a network of n players, each player i having private input x_i, we show how the players can collaboratively evaluate a function $f(x_1, \ldots, x_n)$ in a way that does not compromise the privacy of the players' inputs, and yet requires only a constant number of rounds of interaction.

The underlying model of computation is a complete network of private channels, with broadcast, and a majority of the players must behave honestly. Our solution assumes the existence of a one-way function.

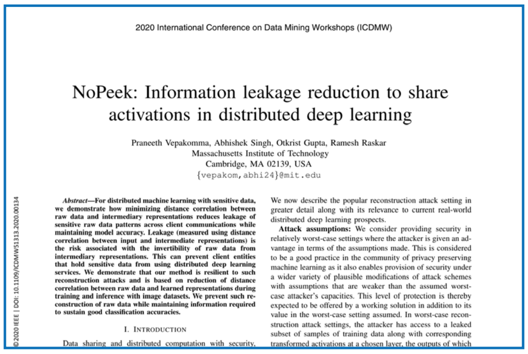
1 Introduction

Secure function evaluation. Assume we have n parties, $1, \ldots, n$; each party i has a private input x_i known only to him. The parties want to *correctly* evaluate a given function f on their inputs, that is to compute $y = f(x_1, \ldots, x_n)$, while maintaining the *privacy* of their own inputs. That is, they do not want to re-

some components in the system fail. However, even if one is solely interested in correctness, privacy helps to achieve it most strongly: if one wants to maliciously influence the outcome of an election, say, it is helpful to know who plans to vote for whom. Second, secure function computation is central to protocol design, as the correctness and privacy of *any* protocol can be reduced to it. Here, as people may be behind their computers, correctness and privacy are equally important.

The first general solution for secure function evaluation was found by Yao [Ya86] for the two-party case, and by Goldreich, Micali and Wigderson [GMW87] for the multiparty case. Many other protocols for the multiparty case have been found since. In particular, the protocols of Ben-Or, Goldwasser and Wigderson [BGW88], Chaum, Crépeau and Damgård[CCD88], and Rabin and Ben-Or [RB89], succeed in defeating the influence of bad players without making use of cryptography, assuming that the privacy of communication among players is guaranteed. Other general protocols with different and interesting properties include

Learn more about the technologies underpinning our solution.



2020 International Conference on Data Mining Workshops (ICDMW)

NoPeek: Information leakage reduction to share activations in distributed deep learning

Praneeth Vepakomma, Abhishek Singh, Otkrist Gupta, Ramesh Raskar
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{vepakom,abhi24}@mit.edu

©2020 IEEE | DOI: 10.1109/ICDMW51313.2020.00134

***Abstract*—For distributed machine learning with sensitive data, we demonstrate how minimizing distance correlation between raw data and intermediary representations reduces leakage of sensitive raw data patterns across client communications while maintaining model accuracy. Leakage (measured using distance correlation between input and intermediate representations) is the risk associated with the invertibility of raw data from intermediary representations. This can prevent client entities that hold sensitive data from using distributed deep learning services. We demonstrate that our method is resilient to such reconstruction attacks and is based on reduction of distance correlation between raw data and learned representations during training and inference with image datasets. We prevent such reconstruction of raw data while maintaining information required to sustain good classification accuracies.**

I. Introduction

Data sharing and distributed computation with security,

We now describe the popular reconstruction attack setting in greater detail along with its relevance to current real-world distributed deep learning prospects.

Attack assumptions: We consider providing security in relatively worst-case settings where the attacker is given an advantage in terms of the assumptions made. This is considered to be a good practice in the community of privacy preserving machine learning as it also enables provision of security under a wider variety of plausible modifications of attack schemes with assumptions that are weaker than the assumed worst-case attacker's capacities. This level of protection is thereby expected to be offered by a working solution in addition to its value in the worst-case setting assumed. In worst-case reconstruction attack settings, the attacker has access to a leaked subset of samples of training data along with corresponding transformed activations at a chosen layer, the outputs of which

Learn more about the technologies underpinning our solution.

THE TRIPLEBLIND PRODUCT

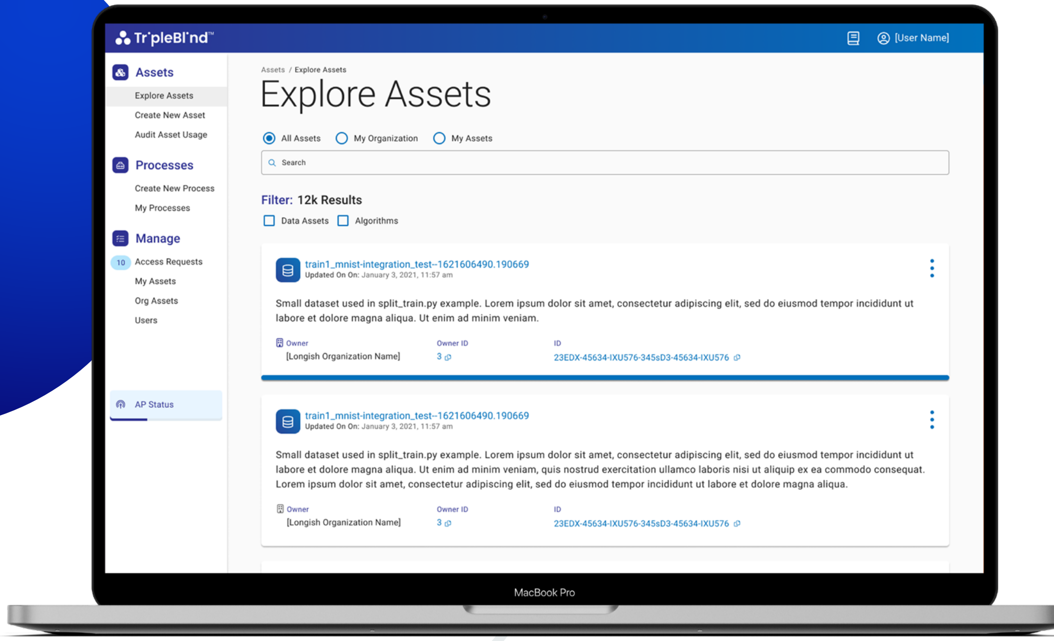
TripleBlind users fall into two categories: asset owners and asset users. An “asset” can describe either a proprietary dataset or algorithm. Users can register their assets through TripleBlind, making them “discoverable” to partners using TripleBlind.

Importantly, the asset is never uploaded to or seen by TripleBlind or any of its systems. With TripleBlind, data and algorithms interact in a peer-to-peer fashion that requires no trust in the third party.

The asset owner retains full and granular control of the dataset’s “discoverability” - whether other researchers can find general information about the dataset and request permission for its use.

Through a sleek user interface and easy-to-use package of APIs, owners can seamlessly and intuitively form agreements with others who wish to privately use their data or algorithms.

All assets remain protected and encrypted throughout the process. They are one-way encrypted at the time of use with software that sits behind the users’ firewalls. Asset owners and users interact in a peer-to-peer secure connection, sharing only one-way encrypted bits, which can never be linked to the raw data or algorithm, if intercepted.



train1_mnist-integration_test--1621606490.190669
Updated On On: January 3, 2021, 11:57 am

Small dataset used in split_train.py example. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam.

Owner	Owner ID	ID
[Longish Organization Name]	3	23EDX-45634-IXU576-345sD3-45634-IXU576

train1_mnist-integration_test--1621606490.190669
Updated On On: January 3, 2021, 11:57 am

Small dataset used in split_train.py example. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Owner	Owner ID	ID
[Longish Organization Name]	3	23EDX-45634-IXU576-345sD3-45634-IXU576

OUR PRIVATE DATA SHARING SOLUTION

The Private Data Sharing Solution provided by TripleBlind equips organizations with the tools they need, and importantly can effectively use, to leverage the power of the underlying cryptographic breakthroughs, without requiring any previous knowledge of cryptography.

BLIND COMPUTE
Mathematical techniques and privacy primitives used to run a myriad of processes.

BLIND AI TOOLS
AI model training and inference, on distributed private datasets.

BLIND VIRTUAL DATA EXCHANGE API
Our method of securely connecting and managing processes.

BLIND QUERY
Tools for learning from protected datasets without exposing private data.

BLIND DATA TOOLS
Data tools your teams expect, like pre-processing and EDA, designed around privacy.

BLIND ALGORITHM TOOLS
Allow easy distribution of models while maintaining full control over your IP.

THE SOLUTION

Utilizes a distributed data model - data is accessed, but never moved or revealed.

Always keeps assets fully one-way encrypted, ensuring third parties cannot access specifics - even while running operations.

Employs innovative methods that create new data and algorithm licensing opportunities previously unavailable.

Lives in the cloud and is based and provider agnostic, but also works on premises.

Supports any mathematical function, including Artificial Intelligence (AI) training and inferences, Machine Learning (ML) processes, and statistical models.

Supports any type of data, including tabular, image, video, voice, and even large genetic datasets.

Provides granular Digital Rights Management and auditability of every transaction, allowing suppliers to control specifically who, when, how often, and for what purpose their assets are used.

Offers Blind Join and other preprocessing measures.

Enables Horizontal Stacking and Vertical Partitioning of distributed data.

Uses an advanced encryption model - the correct cryptography is chosen and used for every task.

Supports one-way algorithm encryption. Most privacy approaches focus on keeping the data safe. However, TripleBlind can also keep the intellectual property of the algorithm safe.

BUSINESS VALUE

Entities can derive more business value with TripleBlind, as the technology offers a number of benefits such as:

- Safer collaboration both internally (across regulatory boundaries) and externally (between different organizations).
- Increased availability of high-quality data - protected datasets can be made available for consumption by other entities without extensive data preparation.
- Better quality analysis and modeling due to increased availability of diverse data, reducing algorithm bias and increasing accuracy.
- Increased opportunities for revenue generation - previously unused data can be made available to other organizations, unlocking tremendous business value.
- Reduced chances of non-compliance with simplified, smoother contracting processes.

BUSINESS USE CASES:

TripleBlind's solution is extremely versatile and can be used in most scenarios that have previously required transmission of data. The following two examples are pulled from the more than two dozen market validated use cases across healthcare, life sciences, financial services, insurance, advertising, energy, manufacturing, and more.

HEALTHCARE

As diagnostic algorithms become increasingly prevalent, access to quality data is imperative to reduce algorithm bias and fine-tune the accuracy of models. Researchers are enlisting TripleBlind's Blind AI Tools to train AI models on data from disparate hospitals around the world, spanning the United States, Europe, and the Middle East, without moving any of the raw data from its sources. This allows the researchers to easily achieve and maintain compliance with privacy laws including HIPAA, GDPR, and strict data residency laws, as source images are obfuscated and cannot be reverse-engineering to re-identify an individual.

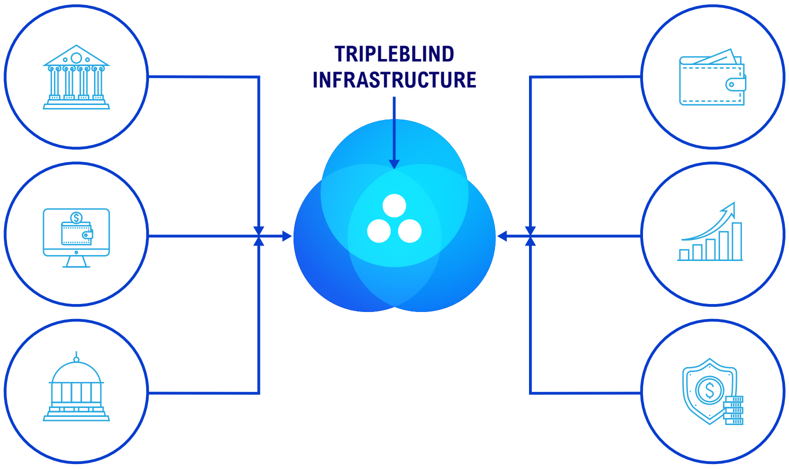
With better models and easier compliance, researchers are free to focus on important initiatives, like building a global data-analysis capability, training on newly available datasets, and validating novel algorithms for use on one-way encrypted data.

NOTE: TripleBlind is not accessing any Mayo Clinic data.

FINANCIAL SERVICES

Sophisticated methods of committing fraud and money laundering are constantly evolving and challenging financial services organizations to be vigilant and adaptive. Today, independent financial institutions are collaborating using TripleBlind's solution to detect and prevent financial crime. Even competitors can agree to use each other's data for mutual benefit: achieving an overall reduction in fraud and obtaining higher levels compliance, without giving a competitive edge to rivals. Institutions can now track suspicious behavior across multiple organizations and geographies without viewing any private or sensitive information.

Allowing Banks, Fintechs, Credit Bureaus, Investors, Insurers, Technology Vendors, Governments, and others to collaborate on challenging issues.



ON A FINAL NOTE

As companies round out their digital transformations, a final missing piece is how they decide to leverage their own data and that of their partners, especially with mounting regulatory restrictions on how that data can be shared.

The TripleBlind solution enables enterprises to collaborate around previously unusable sensitive data via one-way encryption (decryption is impossible). The ability to unlock this data allows for faster and more accurate treatment and diagnosis in healthcare and quicker, more secure safety monitoring in financial services.

TripleBlind keeps data and algorithms private at every stage of a data project and never stores data itself. The technology protects assets from various forms of misuse, including:

- Malicious attempts to gain access to data or algorithms
- Trusted but curious parties viewing raw data
- Unauthorized uses of data resulting from stray raw data being left behind after a job is completed
- Violations of data privacy laws, including GDPR, HIPAA, and data residency

TripleBlind never hosts or touches the data at any point of the data lifecycle.

Ultimately, the tools allow companies to build new revenue streams from underutilized datasets and algorithms, all the while complying with relevant privacy regulations.

TripleBlind is the fastest, most accurate, and scalable privacy framework with the highest interoperability and minimizes risks associated with data privacy violations.