



WHITEPAPER

# ENSURING SECURITY SAFELY, EFFICIENTLY, & EFFECTIVELY

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

## INTRODUCTION

Innovation in healthcare is nearly entirely dependent on access to data – whether to evaluate new drugs and devices, ensure quality, drive new technology development, or ensure resource allocation in increasingly complex supply chains. However, due to the sensitive nature of health data, decisions are often made on abstracted, synthetic, or pre-anonymized datasets, resulting in increased costs to healthcare systems. Furthermore, these data may contain errors or miss key elements, resulting in flawed insights.

This document focuses on the current state of healthcare data, how the healthcare industry is currently working to integrate data, and how TripleBlind's patented innovations can advance and scale access to protected data, thus reducing costs and offering new opportunities.

Applications of TripleBlind, a solution built with patented innovations on top of decades of verified, peer-reviewed technologies, will streamline secure access to data, allowing scalable, global opportunities to improve the healthcare innovation and delivery landscape while protecting against big data breaches, reducing costs, and enabling the next era of digital health.

## CONTENTS

Introduction	2
The Data Problem in Healthcare	4
Present Day Approaches to Healthcare Data	6
Optimal Solution to the Data Problem	8
The TripleBlind Solution	9
The TripleBlind Product	10
Potential Impact in Healthcare	12
Healthcare Use Cases	13
Conclusion	14

## THE DATA PROBLEM IN HEALTHCARE

While the value of access to data across all silos and segments of healthcare is well recognized, there has been competing recognition between the value of near continuous access to raw data and the need to keep such data private, in order to protect individual patient privacy or the intellectual property rights of healthcare organizations or their industry partners. Various initiatives have been developed over the years to reconcile these opposing concerns – regulatory initiatives to define “de-identification” (HIPAA) or the use of codified ontologies (to allow for data abstraction without the need for raw data access). However, given the increasing complexity of healthcare data and the resulting evolving questions surrounding what constitutes “adequate” de-identification (eg, for genetic data, radiographic data, etc), the realities of accessing such data have only become more complex.

Several industry and governmental regulations have long recognized the value of continuous data access to optimize care delivery and innovation. In 2007, the US Congress directed the FDA to create a post-marketing surveillance system to monitor the safety of medical products, through direct access to the electronic health records (known as the Sentinel Initiative). However, despite the lofty ambitions set forth, the Sentinel network, rather than being a national integrated initiative, became a coalition of the largest insurance and healthcare entities sharing primarily claims data with a smaller percentage being primary health record and lab data. In turn, there is a lack of integration with manufacturer data, vastly limiting the potential accuracy of insights derived out of such registries based off of limited primary data or meta-data.<sup>1</sup>

The key limitations to deriving accurate, real-time, meaningful insights from health data are the fact that present day approaches to de-identifying, abstracting, and/or normalizing data are:

- | hardware dependent
- | only offer partial security
- | have residual risks for reidentification
- | often only function on structured data or specific data types.

The need to account for the dynamic nature of health data and data sharing limitations imposed by regulatory considerations can be overcome by digital privacy enhancing solutions. These solutions allow data to stay resident at its source of origin, permit dynamic transformation for data interoperability, eliminates risk of decryption, is agnostic to the form of the data, and opens up analytic frameworks to allow queries to be enacted in real-time on data as it evolves and expands.



## PRESENT DAY APPROACHES TO HEALTHCARE DATA

The potential for digital health to offer efficient and scalable solutions to improve the healthcare system, innovate new technology, and reduce costs has been touted for over two decades. Paired with recent advances in analytics, including big data, machine learning, and artificial intelligence techniques, the value of data access has become increasingly obvious. With these changes, legal and regulatory requirements have rapidly evolved to keep abreast of the needs for data to be aggregated and accessed securely. Thus, the number of data localization laws has increased from 67 in 2017 to more than 144 in 2021. With these laws, companies are racing to address how data can be encrypted, shared, and analyzed at scale.

Working on data requires three considerations:

- Securing privacy of the data at its point of residence, ensuring external parties cannot meaningfully access the data

- Securing privacy of data as it is transferred between parties to limit likelihood of interception

- Securing privacy of data when being computed upon, to eliminate risk of extracting sensitive information while tracking how the data is used to ensure “intended use”

The evolution of tools and technologies to address these considerations are broadly referred to as “privacy enhancing technologies” (PETs) or “privacy enhancing computation technologies” (PECs).

Modern approaches to securing data for the purposes of enabling analytics often include some combination of these PETs, including, but not limited to:

- manual or tokenization-enabled removal of personal identifiable information (PII) (which limits cross-dataset patient matching)
- encryption of data during transfer with decryption keys made available to the data user (which limits how data use can be tracked, has residual risk of malicious third parties accessing the data, and often carries limitations to scalability with larger, more complex data sets)
- storage in secure enclaves (which is hardware-dependent and requires significant additional investment)

	Degree of Privacy	Ability to Operate at Scale	Types of Data	Speed	Supports Training New AI/ML	Digital Rights Managed	Algorithm Encryption (3 Types)	Compliance (GDPR, HIPAA)	No Masking, Synthetic Data, or Hashing	Not Hardware Dependent	Interoperable with Third Parties	No Accuracy Reduction
TripleBlind	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Homomorphic Encryption	◐	●	◐	●	●	●	●	●	✓	◐	◐	✓
Secure Enclaves	◐	✓	◐	◐	◐	●	◐	●	✓	●	●	✓
Tokenization	◐	◐	●	✓	◐	●	●	✓	●	✓	◐	●
Synthetic Data	◐	◐	●	✓	◐	●	●	●	●	✓	●	●
Differential Privacy	◐	✓	✓	◐	◐	●	●	✓	✓	✓	●	●

● Negligible Value
 ◐ Criteria Mostly Fulfilled  
◐ Criteria Minimally Fulfilled
 ✓ Criteria Fulfilled  
◐ Criteria Somewhat Fulfilled



## OPTIMAL SOLUTION TO THE DATA PROBLEM

The optimal solution to the healthcare “data problem” would:

- | Allow for automated, secure, irreversible obfuscation of PII while eliminating the needs for manual intervention
- | Maintain the computational value of important fields while guaranteeing that individuals cannot be re-identified
- | Ensure computational accuracy and precision
- | Be agnostic to the use of structured or unstructured data (including the use of image, voice, genetic and other data sources)
- | Minimize computational load, thus limiting hardware dependency and maximizing scalability
- | Offer trackability to how, when and for what purpose data is accessed, thus ensuring data owners’ control over how their data is used
- | Protect the IP of complex, proprietary analytic algorithms developed or used by “data users” so they can securely validate and deploy on external parties’ healthcare data



# 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.

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, Otkrit Gupta, Ramesh Raskar  
 Massachusetts Institute of Technology  
 Cambridge, MA 02139, USA  
 {vepakom, abh124}@mit.edu

**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 inevitability 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 large 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 names of which

Learn more about the technologies underpinning our solution.

# THE TRIPLEBLIND PRODUCT

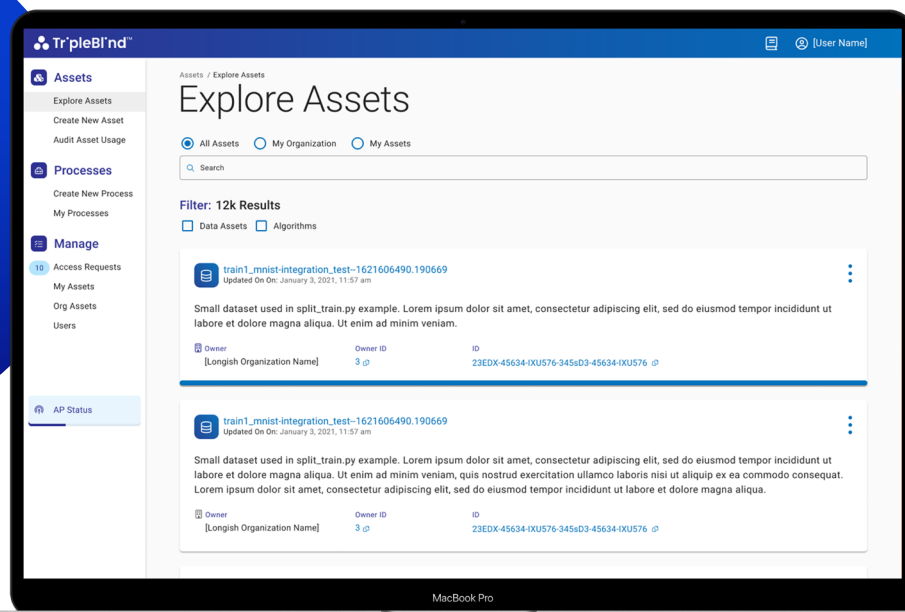
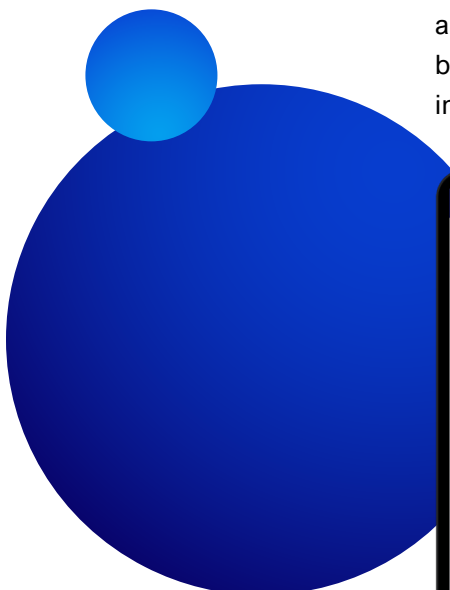
TripleBlind users fall into two categories: asset owners and asset users. An “asset” can describe either a proprietary dataset or algorithm. Users can register their assets through TripleBlind, making them “discoverable” to partners using TripleBlind.

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.



# OUR PRIVATE DATA COLLABORATION SOLUTION

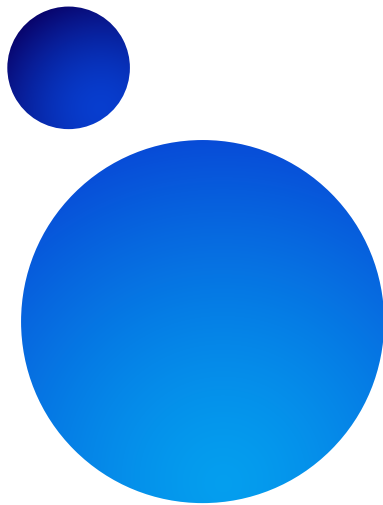
The TripleBlind Solution equips organizations with the tools they need, and importantly can effectively use, to leverage the power of the underlying cryptographic techniques, without requiring any previous knowledge of cryptography.

	<p><b>“SECRET SAUCE”</b> Mathematical techniques and privacy primitives used to run a myriad of processes.</p>		<p><b>BLIND AI TOOLS</b> AI model training and inference, on distributed private datasets.</p>
	<p><b>BLIND DATA &amp; ALGORITHM API</b> Our method of securely connecting and managing processes.</p>		<p><b>BLIND QUERY</b> Tools for learning from protected datasets without exposing private data.</p>
	<p><b>BLIND DATA TOOLS</b> Data tools your teams expect, like pre-processing and EDA, designed around privacy.</p>		<p><b>BLIND ALGORITHM TOOLS</b> Allow easy distribution of models while maintaining full control over your IP.</p>

## 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.

## POTENTIAL IMPACT IN HEALTHCARE



The operations of every facet of healthcare are dependent on high quality, real-time, efficient access to data. When we consider the broad data needs of healthcare, integration of high quality primary data arising from real world experience, diagnostic and monitoring technologies, basic science and clinical trials, and other sources such as pharmacies, insurance carriers, and social media is critical to ensuring that the insights and tools derived from digital health analytics are scalable, broadly applicable, computationally efficiently and have minimal cost implications.

Key value that can arise from implementation of TripleBlind to enable interactions between parties who need access to data for development of new analytics, discovery, or validation include:

- | Safer, more secure collaboration internally (across regulatory boundaries) and externally (between organizations)
- | Reduced friction in accessing data related to concerns surrounding removal of PII and other sensitive aspects
- | Reduced chances of non-compliance via simplified, smoother contracting processes
- | Increased availability of raw data beyond structured data, including genetics, images, and voice data
- | Increased ability of data owners to control how their data is accessed and used by internal and external parties

## HEALTHCARE USE CASES:

The applications of TripleBlind to healthcare are broad, ranging from facilitating the development, validation and deployment of new machine learning algorithms, enabling standard analytics to be done in real-time and securely for the purpose of clinical trial site selection and data safety monitoring globally, and allowing for a more thorough appreciation of regional and global supply chains. Several common use cases are listed below:

### CLINICAL TRIAL OPTIMIZATION

- Rapid site selection
- Reduce CRO / clinical trial costs
- Real-time data assessments

### UNLOCK THE VALUE OF DATA ASSETS

- Need to gather data across multiple parties
- Securely enable third party data usage

### AI ALGORITHM DEVELOPMENT & DEPLOYMENT

- Securely validate and commercialize
- Need for diverse access to data securely

### MULTINATIONAL CORPORATION DATA LITERACY

- Regulatory standards currently limit ability to aggregate institutional data

### PHARMACOVIGILANCE / POST-APPROVAL MONITORING

- There is need for real-time, continuous monitoring for adverse events, outcomes
- Requires secure connections to patient records

## CONCLUSION

As the next era of digital health evolves, there is increasing reliance on secure connectivity between various data owners and data users. The missing piece is how to efficiently and securely facilitate this collaboration in a non hardware-dependent, regulatory-compliant framework. The TripleBlind solution enables enterprises to collaborate around otherwise impossible or difficult to access sensitive data via its proprietary non-decryptable processes. This resultantly gives healthcare parties the ability to more quickly and securely access data to facilitate analytics and discovery.

By giving data owners full control over their own data and removing the need to transfer data, TripleBlind further minimizes risk to data leaks that may arise from dependence on security of data transfer processes or of the third party who is gaining access to the data. Furthermore, by ensuring data residency while enabling analytics, these tools allow companies to interact globally with their own or other's data assets while abiding by various local, national, and regional regulations.

It is critical to note that in this process TripleBlind never hosts or touches the data or the algorithms at any point in the data life cycle. Ultimately, TripleBlind is the fastest, most scalable privacy framework which enables accurate analytics, a high degree of interoperability, and minimization of risks related to data privacy violations.