TripleBlind™

WHITEPAPER

TECHNOLOGY OVERVIEW

Learn how TripleBlind applies privacy-enhancing technologies to deliver protections for data in use.

CONTENTS

INTRODUCTION

TripleBlind provides the most complete and scalable solution for privacy-enhancing computation (PEC), enabling enterprises to remotely compute on disparate and distributed datasets without moving or physically aggregating the data.

Normally, when an organization needs to operationalize siloed sensitive data for analysis or machine learning model training, the data is transmitted over the internet. Data is typically moved in an encrypted state from its storage location to its use location. The data user then decrypts the data, which creates a copy or duplicate of it, reducing the data provider's control over how often and for what purposes the dataset can be used. The data user also assumes additional liability as a steward of the data, taking on the responsibility to ensure that no personally identifiable information (PII) is leaked or misused. With the emergence of strict data privacy laws, the costs of enforcing compliance and penalties for infractions further disincentivize firms from participating in the growing global data and analytics ecosystem.

TripleBlind's solution mitigates the risks of operationalizing data by providing capabilities for protecting data in use. **The approach ensures that raw data never moves from its storage location, whether on-premises or on a cloud server, at any stage of the data usage lifecycle yet remains fully usable.** The software-only solution is data scientist- and developer-friendly, providing a full suite of **Application Programming Interface (API)** calls and a web interface. TripleBlind allows users to compute on data as they normally would, without having to "see", copy, or store any data. The technology gives data providers complete **Digital Rights Management (DRM)** over how their data is used on a granular, per-use level. Data is **One-Way Encrypted** within the data provider's firewall, meaning no raw data is ever transmitted over the internet for computation. These protections allow data providers to easily make data available for computation without assuming additional risk or giving up digital rights.

Application Programming Interface (API) – a set of functions or software interface connecting the protocols, functions, or data of one piece of software for use by another.

Digital Rights Management (DRM) – the use of technology to enforce proper governance and management of digital assets, including datasets and algorithms, with adherence to legal agreements and regulatory considerations.

One-Way Encryption – an irreversible, single-use data transformation which allows one-way encrypted data to be used for a defined purpose, without generating a decryption key. Because the one-way encryption is irrverrsible and unintelligible outside of the context of the specified data operation, it can flow freely between counterparties.

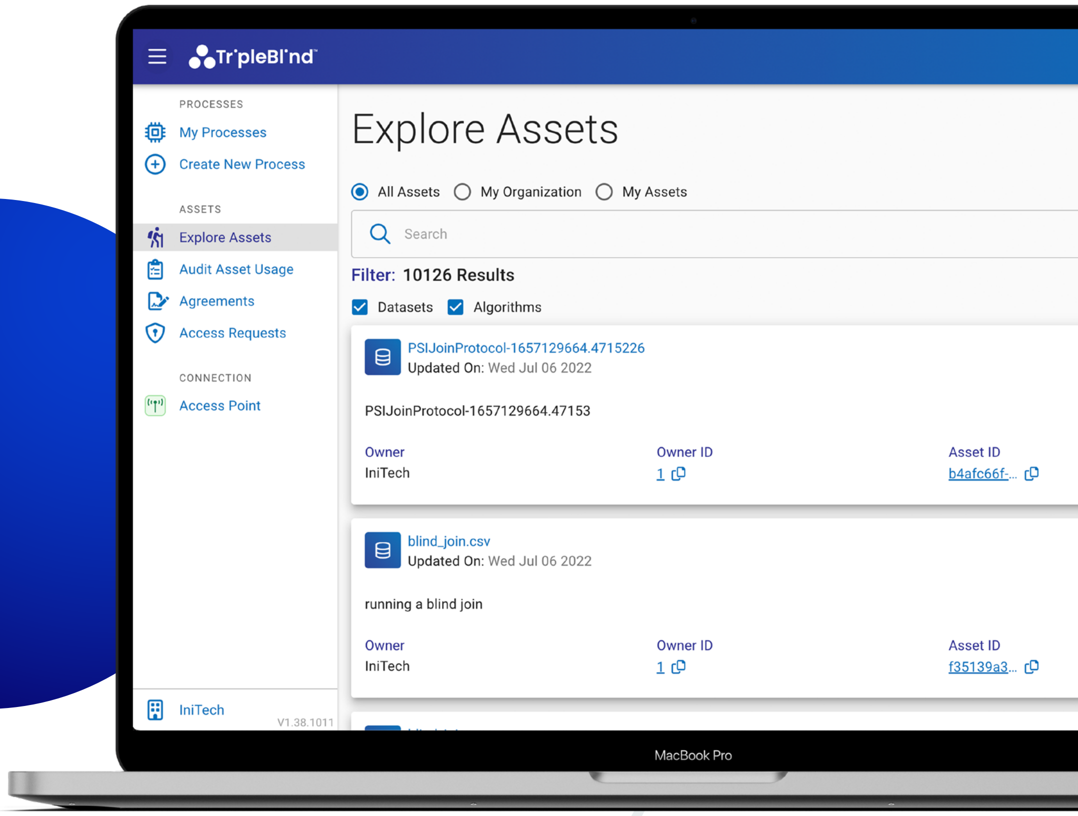
USING THE TRIPLEBLIND SOLUTION

TripleBlind built its solution with user experience in mind. The simple and intuitive web interface allows users to “register” and explore data and algorithm assets, manage processes, set permissions controls, and generate Exploratory Data Analysis (EDA) reports.

The web interface is accessible from any computer or mobile device with an internet connection.

Data providers primarily interact with the solution through the web interface, registering datasets and managing permissions. Data users, on the other hand, also leverage a Software Developer Kit (SDK) containing APIs that resemble popular Python data science libraries like TensorFlow and Scikit-Learn, with support for R. These tools combine to enable privacy-enhanced computations, including the training and inference of machine learning and AI models across distributed datasets, linear and logistic regressions, queries, joins, order statistics, and other computations.

The solution is designed to allow data practitioners to work as they normally would, but with the additional privacy and security benefits of never handing or viewing raw data.



Explore Assets

● All Assets ○ My Organization ○ My Assets

Search

Filter: 10126 Results

☑ Datasets ☑ Algorithms

PSIJoinProtocol-1657129664.4715226
Updated On: Wed Jul 06 2022

PSIJoinProtocol-1657129664.47153

Owner	Owner ID	Asset ID
IniTech	1	b4afc66f-...

blind_join.csv
Updated On: Wed Jul 06 2022

running a blind join

Owner	Owner ID	Asset ID
IniTech	1	f35139a3...

PRIVACY BY DESIGN

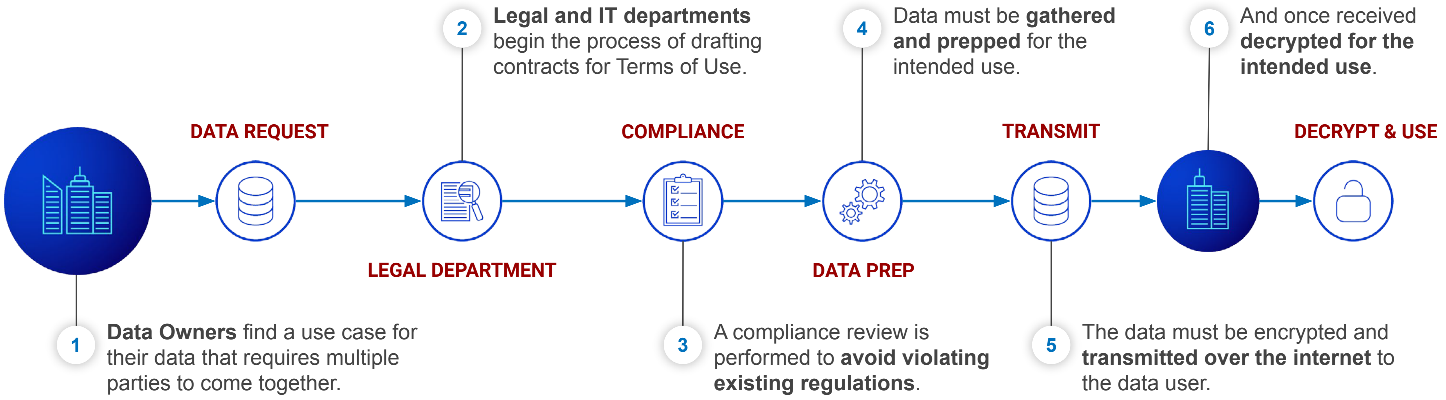
The solution is built on the principle that data should not be required to move onto a third-party environment to be leveraged for insights by authorized parties. Moving data or exposing its raw contents, even to trusted third parties, results in privacy risks and abuses. To achieve a safer alternative, TripleBlind developed a complete, scalable software-based solution that enables any type of computation, including machine learning training and inference as well as more straightforward analytics, to occur on all types of data such as tabular, image, video, voice, and genetic data while the data remains secured behind its firewall. Privacy and regulatory compliance benchmarks are more easily achieved when data users do not need to "see" raw data to work with it. In fact, TripleBlind holds third-party legal opinions stating the solution enables compliance with Europe's **General Data Protection Regulation (GDPR)** and the **Health Insurance Portability and Accountability Act (HIPAA)** in the United States.

General Data Protection Regulation (GDPR) – passed by the European Union (EU), this strict privacy and security law imposes obligations onto all organizations that collect data related to people in the EU and administers steep fines for violations.

Health Insurance Portability and Accountability Act (HIPAA) – a federal US law addressing the use and disclosure of individuals' "protected health information" (PHI) by "covered entities" while also defining individuals' rights to understand and control how their health information is used.

THE TRADITIONAL SUPPLY CHAIN FOR DATA

Supply chain vulnerabilities exist in all industries and processes, but they can be especially harmful in the data science and analytics industry. When sensitive data is compromised and privacy is breached, the cost to organizations and individuals can be much more than monetary. Reputational, or even personal, damage can result, especially when the data contain sensitive personally identifiable information (PII) related to health, lifestyle, and finances. Because the cost of security incidents is so high, the traditional supply chain for data involves a cumbersome process. Once the proper legal and compliance agreements are negotiated, sensitive data elements must be removed from the data asset. Then, the file is encrypted and sent to the other party with a decryption key. The recipient decrypts the file and uses the data, hopefully, for the approved purpose.


2
Legal and IT departments begin the process of drafting contracts for Terms of Use.
4
Data must be **gathered and prepped** for the intended use.
6
And once received **decrypted for the intended use**.
DATA REQUEST
COMPLIANCE
TRANSMIT
DECRYPT & USE
LEGAL DEPARTMENT
DATA PREP
1
Data Owners find a use case for their data that requires multiple parties to come together.
3
A compliance review is performed to **avoid violating existing regulations**.
5
The data must be encrypted and **transmitted over the internet** to the data user.

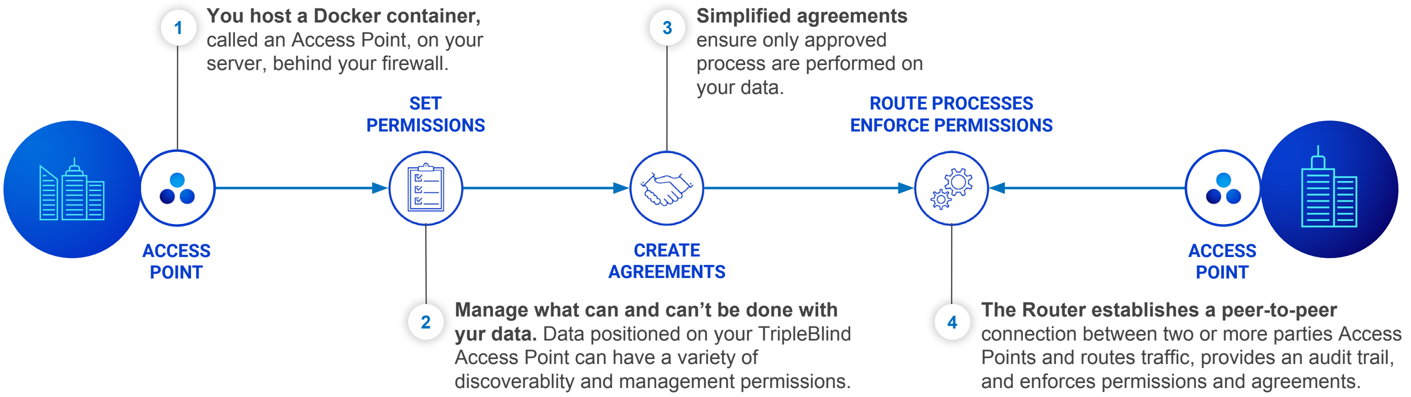
The Traditional Supply Chain for Data

A BETTER SUPPLY CHAIN FOR DATA

TripleBlind enables a better supply chain for data that focuses on privacy and risk minimization afforded by applying peer-reviewed advances in cryptography and privacy-preserving machine learning to minimize raw data exposure while maintaining its utility. One-way encryption techniques obfuscate data at its source without generating a decryption key. Distributed computation methods allow the raw data to stay behind its firewall throughout the entire process, never existing on a counterparty's server, while also providing the option of fully protecting the intellectual property (IP) of the data user's algorithm. This privacy-preserving supply chain for data minimizes risk for all parties: the risk to individuals that their private data is exposed, the risk to data providers that they lose control over their data when they allow it to be used, and the risk to data users that they will be in regulatory noncompliance or assume additional liability from using, or misusing, raw data.

Docker – a software platform that allows developers to deliver software in packages called containers.

Users host the software needed to perform these processes in a **Docker** container provided by TripleBlind. This container is called an "Access Point". Because the Access Point lives on the user's server and all data manipulation occurs on it, neither TripleBlind nor the counterparty ever stores or touches raw data. Instead, a TripleBlind-hosted "Router" establishes a peer-to-peer connection between two or more parties' Access Points. This Router is not a trusted third party, as it never sees, stores, copies, or manipulates data. Algorithms, queries, and models interact only with one-way encrypted data, with optional algorithm protections available.



1 **You host a Docker container,** called an Access Point, on your server, behind your firewall.

2 **Manage what can and can't be done with yur data.** Data positioned on your TripleBlind Access Point can have a variety of discoverablity and management permissions.

3 **Simplified agreements** ensure only approved process are performed on your data.

4 **The Router establishes a peer-to-peer** connection between two or more parties Access Points and routes traffic, provides an audit trail, and enforces permissions and agreements.

A Better Supply Chain for Data

PRIVACY PRIMITIVES

Privacy Primitives – the application of mathematical and cryptographic techniques to computational tasks for the purpose of protecting data and algorithms in-use.

To provide a complete and scalable solution, TripleBlind uses several cutting-edge **Privacy Primitives** that apply the appropriate protections to each data operation. These include privacy-preserving techniques for performing remote computations and inferences, training machine learning models, generating summary statistics reports, and privately joining two or more datasets. These primitives include several patented “secret sauce” advancements TripleBlind has made on existing approaches, including Blind SMPC, Blind Learning, and Blind Stats.

Data and analytics processes are complex and subject to different considerations, regulatory landscapes, and procedures. As such, they require varying levels and types of privacy that no one-size-fits-all approach can meet. By offering several privacy primitives, TripleBlind ensures that the correct approach is applied in the appropriate scenario.

BLIND SMPC

Secure Multiparty Computation (SMPC) – a cryptographic protocol allowing multiple parties to jointly compute an output (answer) without sharing their individual inputs.

Blind SMPC, the innovation which makes **secure multi-party computation (SMPC)** practical, performant, and scalable, is TripleBlind’s most private and secure offering. It serves as the backbone of much of the solution, as it enables data and algorithms to interact without moving either one from its place of storage.

Consider a scenario where a data provider with a dataset D and a data user with an algorithm A exist. The data provider splits its dataset D into two parts, D_1 and D_2, which sum to D but individually appear completely random, giving D_1 to the data user. Likewise, the data user splits its algorithm A into random parts A_1 and A_2, which sum to A. Using SMPC, the data provider and data user cooperate to compute the result of running algorithm A on dataset D in such a way that neither A, nor D, nor any intermediate results of the computation are revealed to either party. Instead, all such values are split between the two parties so that the values seen by each party individually look entirely random. Users can expand this approach to include multiple data providers. Still, the concept of splitting datasets and algorithms, and maintaining that split across all intermediate values of the computation, remains the same.

TripleBlind has iterated on the traditional approach to SMPC (cumbersome and inflexible) to produce Blind SMPC – a performant and scalable solution for a wide range of applications.

BLIND LEARNING

Blind Learning – TripleBlind's patented solution for distributed, privacy-first, regulatory-compliant machine learning at scale. For more information, visit our website.

Federated Learning – an approach to machine learning that involves training an algorithm separately at multiple decentralized data sources and then averaging the resulting models, without moving the data.

Blind Learning is an innovative alternative to **Federated Learning** for decentralized machine learning model training. Training machine learning models on distributed datasets is challenging, especially where privacy laws keep sensitive datasets locked in place.

A technique known as Federated Learning seeks to address this problem, but there are some notable limitations. Federated Learning is a technique of training a machine learning model on distributed datasets while the data remains in place. It involves sending the untrained model to each data provider, who then trains the model and sends it back to the model provider. The model provider then averages the trained models in an attempt to account for their differences.

The approach is promising, but it is neither complete nor scalable. Training machine learning models requires significant computational and storage resources – resources data providers often do not have readily available.

Additionally, the Federated Learning process happens sequentially, which means that all parties must be "online" and perfectly coordinated throughout the lengthy training process.

Finally, Federated Learning provides no protections to the model provider. Should the model provider have developed a proprietary model, each data provider can fully inspect the model's architecture.

Blind Learning, part of the TripleBlind solution, solves these problems by enabling model providers to train their models on data that stays in place without sending the entire model to the data providers. Instead, the model training is split, and only the front portion is trained at the data provider's server. The data is simply input into the first section of the model, behind the data provider's firewall, while enforcing full DRM. The section, made up of just a few layers of the full model, is trained to a specified "split". At the split, the training is paused and finished at the model provider's side.

Blind Learning is "asynchronous", meaning that not all parties must be online simultaneously to complete the training.

BLIND JOIN

It is often beneficial for two organizations holding separate information on the same individuals to combine their data to achieve a more holistic view of their customers, patients, or workforce. To glean more robust, complete insights and actionable knowledge, these organizations may want to analyze the overlap of their datasets. Blind Join enables an organization to augment a dataset using SQL-like methods to select rows with equal or similar (fuzzy match) values in other datasets, then extract other values on the same row, without revealing any non-matched data in those tables. Data providers exhibit complete control over who can perform a Blind Join using their data and which columns can be returned.

BLIND STATS

Almost all data operations begin with some degree of data discovery, which involves gathering a baseline understanding of the datasets available, their **Metadata**, and their basic statistical properties. These prerequisite activities can be challenging when data is stored in multiple locations and cannot be moved, shared, or aggregated for various regulatory and operational reasons.

Metadata – basic information which describes the contents of structured or unstructured datasets, but which does not include the data itself. Metadata about an image, for example, may include information on the type of image, image creator, date created, date labeled, and file size, but would not include the actual image.

TripleBlind has developed a set of privacy-preserving functions called Blind Stats, which enable descriptive statistics, including information like mean, median, range, skewness, standard deviations, population size, and more, to be calculated on multiple disparate data silos without moving, aggregating, or sharing the datasets. Innovative techniques and applications of mathematics make it possible to calculate **Order Statistics** across multiple datasets spread geographically or organizationally without pulling together or sorting the data. This concept and its practical application fill gaps in remote data processing by eliminating the requirement of physically aggregating datasets or handling raw data.

Order Statistics – a class of statistics that requires the ordering of the values. For example, taking the mean of a column would require the values in that column to be arranged in ascending or descending order for the middle, or central-most, value to be identified.

TripleBlind's Blind Stats primitives empower researchers to ask and answer critical qualifying questions about datasets, even when their population is scattered across data silos in different departments, organizations, or countries.

FEDERATED MODE

Also supported and offered through the TripleBlind solution is the option for users to perform computations in “Federated Mode”. Usually, when data from multiple sources needs processing, it is aggregated in a centralized location where all computation occurs. However, in a “federated” setup, the algorithm is brought to and processed on the server or device where the dataset is securely stored. While Federated Mode offers fewer protections for the algorithm provider, it is a viable option for protecting data in use when the intellectual property of the algorithm is not a concern.

Federated Mode can take several forms depending on the use case. For example, a data provider may want to train an AI model to generate insights from their datasets but does not have the in-house resources or staff to build the model. The data provider will then work with a counterparty, which provides an untrained model. The model is trained on the data provider’s server using their data. The data provider owns the resulting trained model, but it is never seen by the model provider, as it now contains valuable intellectual property.

In another case, a third party may want to run simple statistical analyses on the data provider’s asset. Since there is nothing proprietary about the analysis or the algorithm, the data user can perform the computations on the data provider’s server without concern to either party.

To summarize, Federated Mode can be a practical option when protecting the intellectual property of the algorithm is not a concern, as it allows the data to remain securely in place throughout the analysis.

ARCHITECTURE

IDENTITY ACCESS MANAGEMENT: AUTHENTICATION AND AUTHORIZATION

When operating on the enterprise scale, it is vital to account for the fact that each organization involved in a collaborative data process with external parties brings a variety of individual actors who play uniquely defined roles. Some individuals require full access to the organization's use of TripleBlind's tools, while others will need limited access and reduced functionality.

FINE-GRAINED PERMISSIONS

Each organization can set fine-grained permissions over how their users can interact with the tools. Permissions can be set, for example, so that only specific individuals can explore the organization's datasets or that different datasets are discoverable by other teams within the organization.

Though each organization typically has just one Access Point, each user will have their own account, controlled by an organization-level administrator.

OKTA INTEGRATIONS

Additionally, to make the authentication experience easier, TripleBlind has integrated with Okta for enterprise management so that organizations may project their existing management infrastructures on top of the TripleBlind tools. This integration enables more accessible and faster authentication without adding levels of complexity.

APPLICATION SECURITY

Application security is critical to any application interacting with sensitive information or dealing with intellectual property protection. Application security is at the center of TripleBlind, using state-of-the-art tools like Contrast and Snyk to scan, test, and solve code vulnerabilities. Supply chains are incredibly important to protect in any industry, but in the space of privacy and data science, it can be the difference between operating smoothly and being the victim of a data breach and privacy violation.

TripleBlind's software undergoes regular third-party penetration testing to stay ahead of potential vulnerabilities. Additionally, internal testing occurs at every stage of the development lifecycle because delivering a product that is both performant and delivers state-of-the-art protections for its users is integral to the company's mission.

RISK MODEL

TripleBlind's position on risk is that sharing, copying, or otherwise transferring data such that multiple copies of the raw data can exist at one time in different locations introduces unnecessary and often unacceptable risk factors. The mere existence of a decryption key, regardless of where it lives or who possesses it, puts data at risk of being exposed or accessed by unauthorized parties. Gaining the maximum utility from a dataset should not come at the cost of increased risk or decreased privacy, and improvements to cryptography allow for data to be processed in a one-way encrypted format, which is impossible to decrypt.

The out-of-the-box settings for the TripleBlind solution are configured to ensure that at no point during data processing are datasets shared or their contents revealed.

API FLOW

This document has provided an overview of the solutions' technologies and architecture. This section will shift towards explaining how these systems interact by describing the API flow a user would experience when interacting with TripleBlind's solution.

COMPONENTS INVOLVED

Any process carried out using TripleBlind consists of the same components: the TripleBlind owned-and-operated Router, which establishes the peer-to-peer connections between data providers and data users; Access Points, or **Dockerized Containers** of software, hosted by each counterparty; a User Interface (UI) to assist with the management of processes, and a **Software Development Kit (SDK)**, made accessible to each developer at each counterparty.

Dockerized Container – a lightweight software package including all required elements for running an application: code, runtime, system tools, system libraries, and settings.

Software Development Kit (SDK) – a set of software tools and programs which allow developers to create applications from an existing architecture or platform.

API FLOW

Consider a simplified case in which there are two parties, **Company A** (a data provider) and **Company B** (an algorithm provider). **Company B** would like to run its proprietary algorithm on a dataset held by **Company A**, but **Company A** wants to keep its data secure behind its firewall; they do not wish to send it to **Company B** or any other partner. The two companies agree to use TripleBlind to facilitate the computation without sharing their assets.

First, each company will use the web interface to register their dataset or algorithm asset with the TripleBlind Router. Registering an asset does not upload it to the Router; it simply makes the asset discoverable to the appropriate parties who need to access it.

A user at **Company B** can then search for **Company A**'s dataset in the web interface. Using the SDK, the user makes an API call requesting to run its algorithm on **Company A**'s dataset. The computation will not execute until the request is approved. A user with the proper credentials at **Company A** reviews the request, which includes information about the type of analysis to be performed and the type of result to be returned, before approving or declining the request.

If the request is approved, the Router facilitates a secure peer-to-peer connection between the two companies' Access Points. One-way encryption is applied to the dataset while it still sits behind

Company A's firewall. **Company B**'s algorithm is similarly and compatibly one-way encrypted. The peer-to-peer connection and the application of privacy-preserving computation techniques enable the two companies to combine their one-way encrypted assets to produce a privacy-intact analysis without ever sharing their raw assets with one another.

The calculation result is returned only to the requestor, which in this case is **Company B**.

The diagram below depicts the API flow in simplified form, assuming that **Company A** and **Company B** have already registered their assets with the TripleBlind Router. It is important to note that once the TripleBlind Router establishes the connection between the two Access Points, it steps aside and allows the parties to work peer-to-peer. No raw data ever flows through the Router, nor are any calculations ever carried out on the Router.

Diagram Key

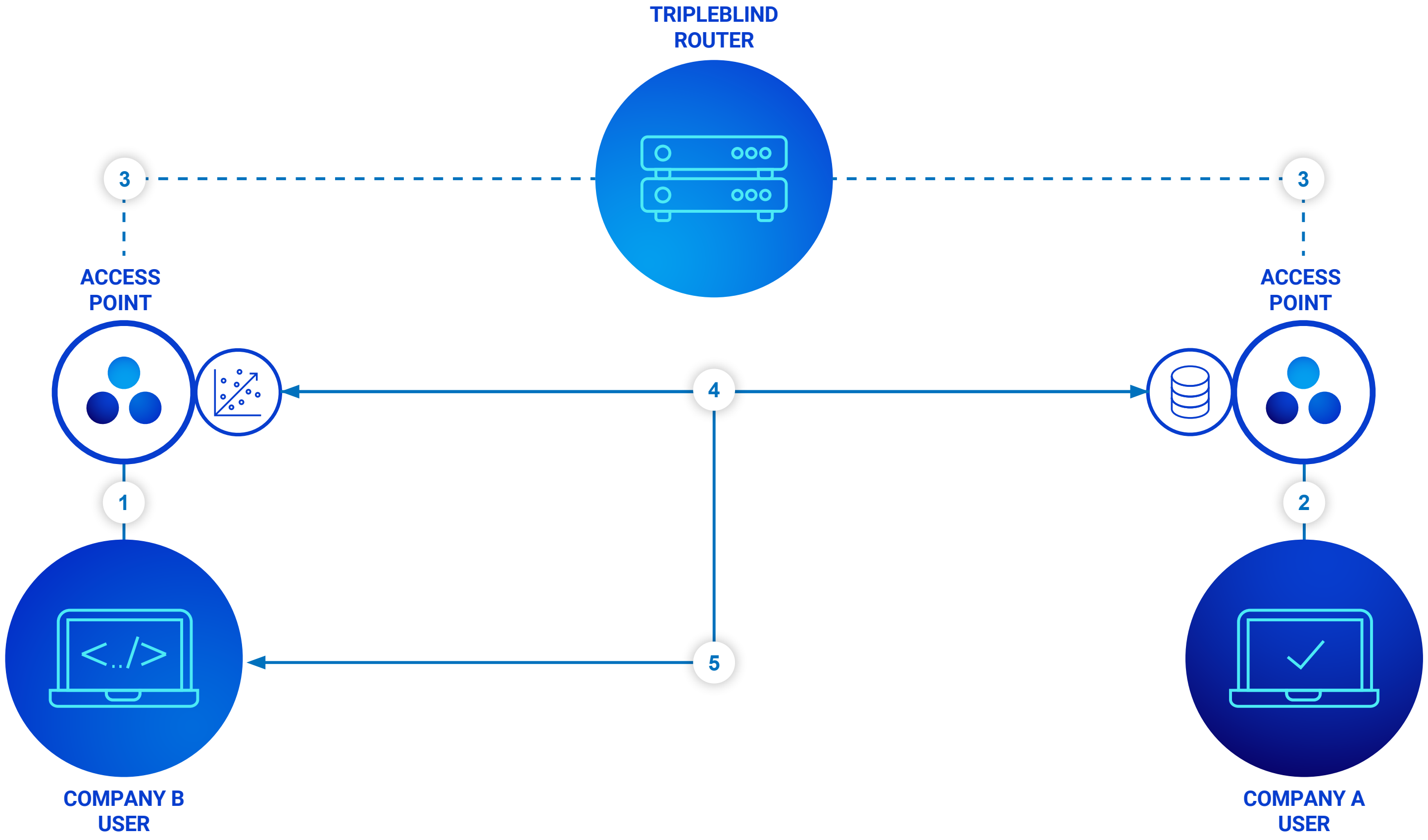
1. API Call

2. Approve Request

3. Establish Peer-to-Peer Connection

4. Privacy-Preserving Computation

5. Result Returned



BLIND LEARNING EXAMPLE

The following practical example walks through the progression of steps involved in training an AI model with Blind Learning.

Consider a scenario in which two high-quality data providers exist, and a separate data user would like to train a proprietary machine learning model, specifically a neural network, using those disparate datasets. In this case, neither data provider is comfortable sending their proprietary data to the data user due to the risks of misuse or data leakage. The three parties are also geographically separated and subject to different data laws, some of which do not legally allow data to move from its country of residence. Additionally, a traditional federated learning approach is not possible in this scenario, as neither of the data providers has the resources available to train machine learning models in-house, and the data user, or model provider, does not want to expose the full architecture of their proprietary model. In this scenario, the model provider will initiate a Blind Learning workflow to achieve the desired result of a trained model while accounting for the privacy and practical hurdles associated with training a model on disparate protected datasets.

Blind Learning makes this seemingly impossible task easy. Here is how:

Each data provider "registers" their dataset with TripleBlind and sets the discoverability such that the data user can find the dataset by name in the web user interface. Notably, the datasets are not uploaded to or stored on any TripleBlind server. "Registering" the dataset simply creates a pointer to a real dataset, which remains only stored at its original location.

The data user logs into the web user interface and searches for the datasets. Clicking on one of the datasets will show a "data profile" of metadata that closely resembles an Exploratory Data Analysis (EDA) dashboard. Without revealing any specifics about the dataset, the tool enables the data user to get a feel for it and determine if it is a good fit for their project.

The data user will also be able to look at some "mock data," which allows them to see column names and information about how different data elements are stored without ever viewing raw data. This mock data is an essential tool for a crucial task in data science called preprocessing, during which the scientist manipulates data so the model can process it.

Now, the data user simply adds a few lines of code to the top of their Python script for the model training to help the file find and use the two datasets. Then, the data user can build out their machine learning model using their preferred methods with TripleBlind's APIs, which closely resemble popular data science tools like Scikit-Learn and TensorFlow. This approach of using the tools that look, feel, and perform how professionals are accustomed to is critically important, as it makes the underlying technology more accessible to a broader range of users. In other words, even novices in the machine learning space can adopt a privacy-preserving approach to training machine learning models.

When the data user tries to run the Python program, an "Access Request" is sent to each data provider. This request appears in the user interface when each data provider logs into their account. The requests must be approved before the training of the AI algorithm begins. Approvals may be automated with prior approval from all parties, but the permissions are set to manual by default.

Once each data provider has approved the request, the machine learning model training begins. Neither the datasets nor the entire model is moved in this process. Rather, the first portion of the model is sent to each data provider's server, where the data is fed into the machine learning model. The first few "**layers**" are trained before a split occurs. The split is a pre-designated position in the architecture of the machine learning model which the data user has chosen. The split exists so that the data providers do not see the entire model nor take on the computational and resource burden of training the entire model. At the split, training stops, and the "**activations**" of the layer are obfuscated and sent back to the server of the data user, where training completes.

Layers – the neurons in a neural network are organized in layers. The number of layers depends on the type of neural network and the design choices of the model architect, but every neural network will have three types of layers: an input layer (where the data is input), hidden layers (where the data is processed), and an output layer (where the result is output).

Activations – refers to the result of a calculation or transformation, known as an activation function, which is applied to the input of an individual neuron, or node, within a layer of a neural network.

The result is a fully trained, highly accurate machine learning model which gets returned to the data user.

This distributed approach ensured that the data providers never sent or revealed their raw datasets to the data user, and the data user never revealed the intellectual property involved in the architecture of their machine learning model. It also ensured that no party was tasked with the computational burden of training the entire model, and the data providers did not need to have machine learning researchers or data scientists on hand.

CONCLUSION

TripleBlind is providing the most complete and scalable solution addressing modern requirements for privacy-enhancing computation. As each major industry undergoes its own digital transformation, many aspects of business activities transition from the physical paradigm to the virtual. In this new landscape, software-based solutions to data privacy are required to optimize the interoperability of data products, maintain flexibility, and maximize the utility of data without introducing costly or time-consuming trade-offs.

TripleBlind's solution provides the underlying technology, the cryptography and mathematical techniques that enable privacy-preserving computation, as well as the tools needed to use that technology. The end-user can perform a variety of powerful tasks, from joining columns of separate datasets to performing statistical analyses on remote datasets to machine learning model training across multiple parties, all in a privacy-preserving way.

Enterprises in every industry can benefit from an increased focus on privacy; it builds trust with customers and protects infrastructure from fraud and the harmful effects of negligence. Legally-enforced privacy is also rapidly becoming the global norm. Each year, governments at all levels are passing increasingly strict privacy laws to protect their constituents. Enterprises must proactively address privacy or react piece-meal to each regulation that affects their business, not knowing if the next regulation will break their business model. TripleBlind provides the technology to allow businesses to take the proactive, future-proof approach.