# Anonomatic

# PII Vault

## A Simple Way to Reduce Risks and Costs Associated With Handling PII Data

# CONTENTS

# Anono**matic**

## Overview

To achieve the most robust analytical outcomes, data scientists need access to the richest data possible. Due to increasing regulations and rising risks, having access to sensitive personal data may pose a liability for organizations and the data scientists they work with. Accidental exposure or loss of private records like medical or financial data, can result in millions of dollars in fines and legal actions. This white paper presents a very simple and cost effective solution for circumventing risk while still allowing for access to rich datasets.

## Industries Most at Risk for Data Breaches and PII Theft

In 2019 there were more than 3,800 data breaches affecting U.S. companies, representing an increase of over 50% in less than five years. In 2020, according to a

report prepared by IBM and the Ponemon Institute, the average cost related to stolen or exploited data was $3.6 million. More importantly, the average cost, if PII was exposed, was $150 per record. This figure can climb to $450 per record if medical or financial data is exposed. While the majority of breaches derive from attacks by

# Anonomatic

outside bad players, a significant portion results from "friendly fire" incidences involving employees or partners.

When examining which industries suffer the most breaches and data losses, finance, healthcare, and academia make the top of the list. Though banks and

other financial organizations routinely implement robust cyber security measures, a persistent lack of training among personnel frequently renders them vulnerable to loss and attacks.

Criminals don't just seek to steal data, but to also hold valuable data hostage until a victim pays. Hospitals are increasing as targets of ransomware schemes.

In the education sector, primary and secondary public schools provide soft targets due to their small IT departments and older equipment. On the higher end, colleges and universities have stronger cyber security protocols; however, both lower and higher education organizations are exploited by hackers who inject ransomware into their systems or try to gain access to student and faculty credit data.

Due to the heightened level of risk involved, it's incumbent upon data scientists to use robust PII management when working with any data that includes PII.

## Institutions Most at Risk for Data Theft

1. Financial Institutions

2. Primary and Secondary Public Schools

3. Colleges and Universities

4. Hospitals and Medical Facilities

# A Simple, Cost-effective Approach to PII Management

Historically the options for protecting PII have been few. While limiting or completely removing access to data containing PII is [one way](#) to reduce costs and risks, it's not a scenario that's feasible for every situation.

Another option is to implement a large and complex platform solution which controls and manages all data access, however, these types of platforms are cumbersome and expensive to maintain.

A new and simpler solution is to preserve PII in the original source data, and temporarily replace it with a unique identifier so that it can be safely transported and handled for research. For true flexibility, such a process should also support data matching between single and disparate datasets, while keeping PII completely protected.

[Anonomatic](#) offers this exact service for organizations, thereby relieving them of the burden and the expense associated with creating extensive processes, workflows and audits to ensure compliance while handling PII. The portability and ease-of-use of Anonomatic's Web Services also permits independent data scientists to gain access to more sources of data that would otherwise be closed off to them.

# AaaS: PII Anonymization as a Service

Anonomatic has developed its **PII Vault**, delivered as a web-based service, so organizations of every size, and across every industry, can safely work with rich data at a vastly reduced cost. Using patented technology, PII Vault provides secure, automated **PII anonymization** using a stand-alone, self-serve API.

The beauty of the solution lies in its simplicity: there is no software to install or interface to learn, all functions

# Anonomatic

are performed via Web Service calls which clients may easily implement regardless of their internal tech stack.

## How Anonomatic's PII Vault Anonymizes Data

It's common for data warehouses to store PII alongside Fact data (financial records, medical records, buying habits, location, etc.) It is also true that PII is almost never included in analysis. When PII is found stored with Fact data in data warehouses it's either part of the record and all data moved to the warehouse, or it's used to combine data from different sources.

Regardless of why PII may be included, when PII is stored or shared, a significant cost in time, money and resources is accrued due to measures required to protect it. The core feature of the PII Vault is its process for separating PII from Fact data so that Fact data can be more efficiently handled.

> **The beauty of Anonomatic's PII Vault lies in its simplicity: no software to install or interface to learn, all functions are performed via Web Service calls.**

Privacy compliance is accomplished through a process called **Poly-Anonymization™**. Poly-Anonymization involves taking any personal identifying pieces of information (name, gender, address, social security number, etc.) and swapping it out for an anonymous value (**Poly-ID**). This value is unique, inconsistent, unpredictable, and not hashed.

The process to replace PII with a Poly-ID is straightforward and follows the following flow:

# Anonomatic

1. Data is extracted from the source system, complete with the PII and all desired Fact data.

2. The PII which needs to be protected is packaged, per profile, and sent to the PII Vault Web Service, either *GetPolyId* or *GetPolyIdBulk*.

3. A Poly-ID is returned for each individual profile and this value then replaces all of the PII values in the source data.

4. The resulting data has all the Fact data desired for analysis but none of the PII which would otherwise make sharing it risky.

After anonymization has been completed, clients are able to share the resulting data either internally or externally minus the usual risk that would be associated with such activity.

The unique security features of Poly-IDs are covered in more detail in the section of this paper titled *How Poly-Anonymous IDs Protect PII.*
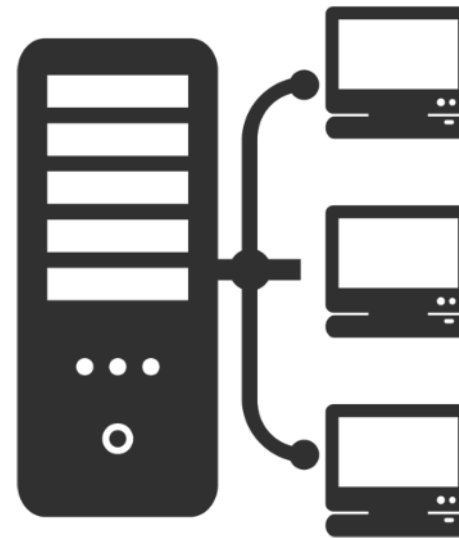
## How to Combine Anonymous Datasets with PII Vault

Robust insights are more attainable when data is processed from multiple or disparate sources. Combining data from multiple sources usually means including PII with Fact data. The risk of unintentional PII exposure is heightened the more that data is combined from multiple sources. It can be particularly difficult for a

# Anonomatic

researcher to convince disparate agencies or organizations to share datasets for combining, and attaining necessary clearance to proceed with research can take several months.

Anonomatic's PII Vault overcomes these obstacles with its unique ability to **A)** Poly-Anonymize data at the source, and **B)** enable the matching of Poly-Anonymized data. Using this service, data scientists, researchers and analysts (Data Processors) are able to match Fact data from disparate data sources without PII ever being shared with the Data Processor.

**Below is an illustration of how the process works:**

Once a Data Processor has received multiple Poly-Anonymized datasets, they have safe Fact data against which they may perform their analysis, AI (artificial intelligence), or ML (machine learning) assisted processes. However, as each Poly-ID value is different for every anonymized individual from each data source, the Data Processor is unable to merge the data from different sources at the individual level.

To be able to combine multiple, anonymized datasets, the Data Processor must first be **authorized** within Anonimatic's PII Vault by each of the sources whose data is to be combined.

Once approval is received, a Data Processor may call the PII Vault, using two additional Web Services, to receive a **Matching Table**.

The Matching Table provides the Data Processor with all of the details they need to combine the various anonymous records from each data set. The Matching Table contains:

1. **A Poly-ID which represents an individual.**

2. **A second Poly-ID which represents the same individual but from a different data source.**

3. **The method by which those profiles were linked (this list is controlled in the PII Vault portal by the Data Processor).**

# Anonomatic

4. The confidence level of the match (as stipulated by the data source for each matching method).

The Matching Table will have a record for the same two Poly-IDs for every matching algorithm (and confidence level) through which they matched. This provides the Data Processor with complete control over when data is to be combined.

## Case Example

Below is an actual case scenario in which PII Vault, Poly-Anonymization, and Anonymized Data Matching played a key role:
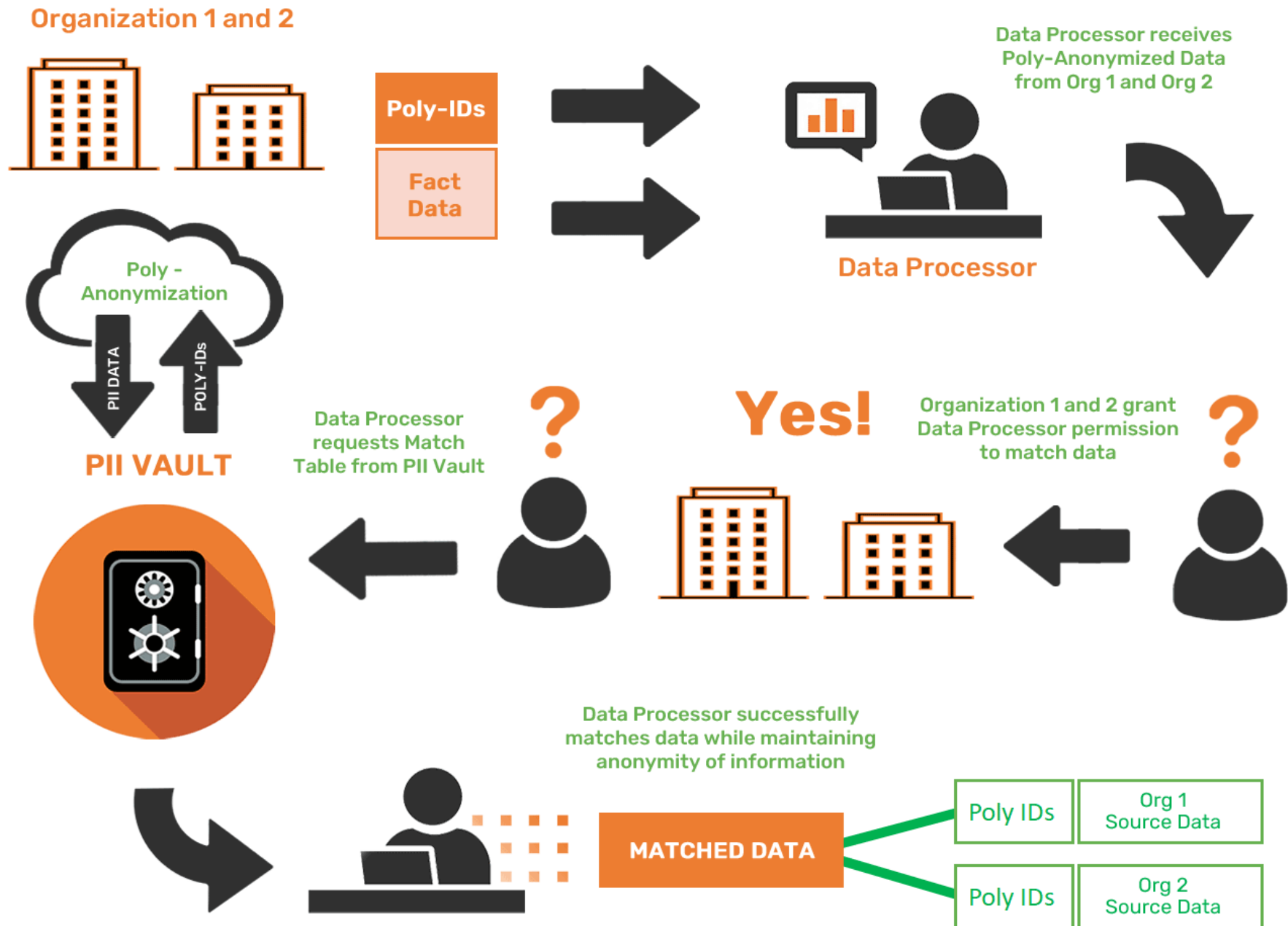
*A non-profit overseeing all healthcare services provided to 600K+ students for the Los Angeles Unified School District wanted to join academic data with detailed student medical services records. Their goal was to identify services which provided the greatest impact on student performance.*

1. First, Anonomatic's PII Vault and Poly-Anonymization feature were used to anonymize their data—replacing the PII with Poly-IDs.

2. Next, the organizations sent the resulting anonymous data to the Data Processor.

3. Upon receiving the data, the Data Processor was able to request a Matching Table from the PII Vault which compared records between the two disparate datasets, and combine the records of the same individual across multiple datasets.

4. Finally, the Data Processor was able to mine the combined data for insights that were impossible to obtain from any of the data sources individually.

Using the service, the non-profit was able to combine data from:

• 12+ medical service providers

• 300K+ individual medical encounters

• 450M+ academic records from LAUSD

# Anonymizing and Matching Data Safely via PII Vault

Organization 1 and 2

Poly-IDs

Fact Data

Data Processor receives Poly-Anonymized Data from Org 1 and Org 2

Data Processor

Poly - Anonymization

PII DATA

POLY-IDs

PII VAULT

Data Processor requests Match Table from PII Vault

?

Yes!

Organization 1 and 2 grant Data Processor permission to match data

?

Data Processor successfully matches data while maintaining anonymity of information

MATCHED DATA

Poly IDs | Org 1 Source Data

Poly IDs | Org 2 Source Data

# Anonomatic

Below are other key benefits clients receive by using PII Vault and Poly-anonymization:

- **The service offers unlimited scalability since it's 100% cloud-based, which means no project is too big to handle.**

- **It streamlines the data sharing process by reducing cost and risk of PII exposure.**

- **It provides secure storage for "at rest" data.**

- **Source data warehouses can easily re-match anonymous data once Data Processors have completed their work.**

- **No worries about Anonomatic having unauthorized access to any sensitive information, even we are unable to read at rest data.**

## How Poly-Anonymous IDs Protect PII

Core to the PII Vault's functionality are Poly-Anonymous IDs. One definition of Poly-Anonymous IDs is that they are multi-value, non-identifying identifiers. They are identifiers because their value represents the person, or other entity, that was used to generate PII. They are non-identifying because no Poly-ID value should ever exist in more than one database and no one who is able to view the value of a Poly-Id can ever use that value to link the PII and Fact data.

When the PII Vault receives new PII it will generate an **Internal Poly-ID** and store the PII with this internal identifying value. It will then calculate an **External Poly-ID** value and return that value to the sender. The existence of a dual-value Poly-ID protects the identity of an individual or entity in the unlikely event that a determined hacker was ever able to breach both the analytics database containing Fact data and the PII Vault.

# Anonomatic

If the values of the anonymous Id were the same in both systems a hacker would be able to easily re-match all of the Fact data with its PII. However, with disparate values that are different in these two systems, there is no way for a bad actor to reconnect the records.

It's also important to note that the Internal Poly-ID is never distributed or exposed in any way. There are no screens, pages, APIs, reports or other PII Vault outputs which contain this value.

## Anonomatic's Security Features

Anonomatic was founded on the principle that data analysis needs to follow ethical and legal best practices for protecting PII. We utilize a blend of standard security features and patent-pending technologies to ensure our clients can rest easy and focus on completing their work instead of worrying about the safety of using our service.

**We're compliant with data protection law.** Our API is developed to meet or exceed GDRP, CCPA and SOC 2 requirements and assists our users with remaining compliant as they fulfill their research or other tasks. This means that no matter where our clients are located in the world, their work will meet legal regulations for anonymization and data minimization.

**We use patent-pending Sundering technology.** Sundering is a process we developed that provides all the benefits of encryption without the limitations. It

# Anonomatic

prevents the data from being read at rest but also allows us to search, and join on data values, without having to decrypt entire database tables.

Sundering works by parsing data records into components and obfuscating those elements so they cannot be recombined without the presence of unique, multi-part (Symmetric-like) Sunder Keys.

When a PII Vault account is created, the account holder is provided with their half of a Sunder Key, a value we never store. This half of the Sunder Key is provided to us when a Web Service call is made, allowing us to combine it with our half of the Sunder Key so we can work on the data. Anemometric never stores both halves of a key; this means the data is secure at rest. Not even our users can read the data held in the PII Vault.

**Client privacy is built-in.** We exist to perform one function and one function only: solve your PII privacy needs. We store no other data beside PII and internal values. No internal key stored with us has a value that exists outside of our PII Vault. Lastly, with our system,

there is no external way to identify what data belongs to an Anonomatic client.

## How to Start Using PII Anonymization Today

The team at Anonomatic is always happy to work with organizations and independent data scientists who seek to reduce the risk and cost of working with PII data. To learn more about how our unique API tool will fit into your next research project, contact our team today.