clearbox AI

humans && machines

SYNTHETIC DATA FOR PRIVACY PRESERVATION

Overcome Data Retention periods with Synthetic Data

C.so Castelfidardo,30/a
10129, Torino (Italy)
info@clearbox.ai

VAT ID: (IT)12161430017

clearbox.ai

Introduction

95%
of statistical information preserved

100%
GDPR-compliance

1.2M
synthetic data records generated

Data retention is an issue companies must carefully address when processing and storing large amounts of personal data. Companies must delete all personal information older than a timespan specified when requesting consent from data subjects. The usual approach is to delete, month after month, those data, losing precious statistical information. During this use case, a financial institution used our Enterprise Solution to create synthetic copies of datasets **that were about to reach the retention limit**. Storing synthetic - anonymous by design - data allowed them to preserve information valuable for analytics and AI projects while complying with the newest data protection regulations.

Challenge

Personal data is subjected to strong data retention constraints, which means that it can be internally kept for a limited amount of time, after which it must be deleted.

Solution

Our client used Clearbox AI's Enterprise Solution to generate high-quality synthetic and anonymized versions of the data that was going to be deleted.

Result

The synthetic datasets generated to replace the deleted data can be stored indefinitely. The synthetic datasets allowed to augment the amount of historical data used for internal analytics and machine learning projects.

Challenge

The General Data Protection Regulation states that *'personal data processed for any purpose or purposes shall not be kept for longer than is necessary for that purpose or those purposes'*. It means that companies collecting personal information must specify a data retention time.

After this time (typically a couple of years), **companies must delete personal data**. Failing to do so will result in hefty fines. Since a data expiration time means fewer data to fuel analytics and AI projects, our client needed a solution to safely preserve the statistical information contained within data beyond its retention time.

Solution

The client installed our Enterprise Solution within their existing infrastructure to ingest and synthesize datasets from different sources across the organization, **populating a centralized registry with synthetic and anonymous copies of datasets**.

Each registry entity is associated with a **report** that quantifies the information maintained from the original dataset during the generation process and all the privacy metrics needed to perform a DPIA. The client finally integrated the GDPR-compliant synthetic data into their internal AI processes.

Result

The financial institution was able to store data beyond its retention period by using its synthetic version. The synthetic datasets generated to replace the original sensitive source have been tested in terms of utility concerning internal business processes, the most important being training machine learning models.

The models trained on the synthetic data achieved **95% of the KPIs** characterising the models trained on the original data.

humans && machines

C.so Castelfidardo,30/a
10129, Torino (Italy)
info@clearbox.ai

VAT ID: (IT)12161430017

clearbox.ai