clearbox AI

humans && machines

SYNTHETIC DATA FOR PRIVACY PRESERVATION

A synthetic data sandbox

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

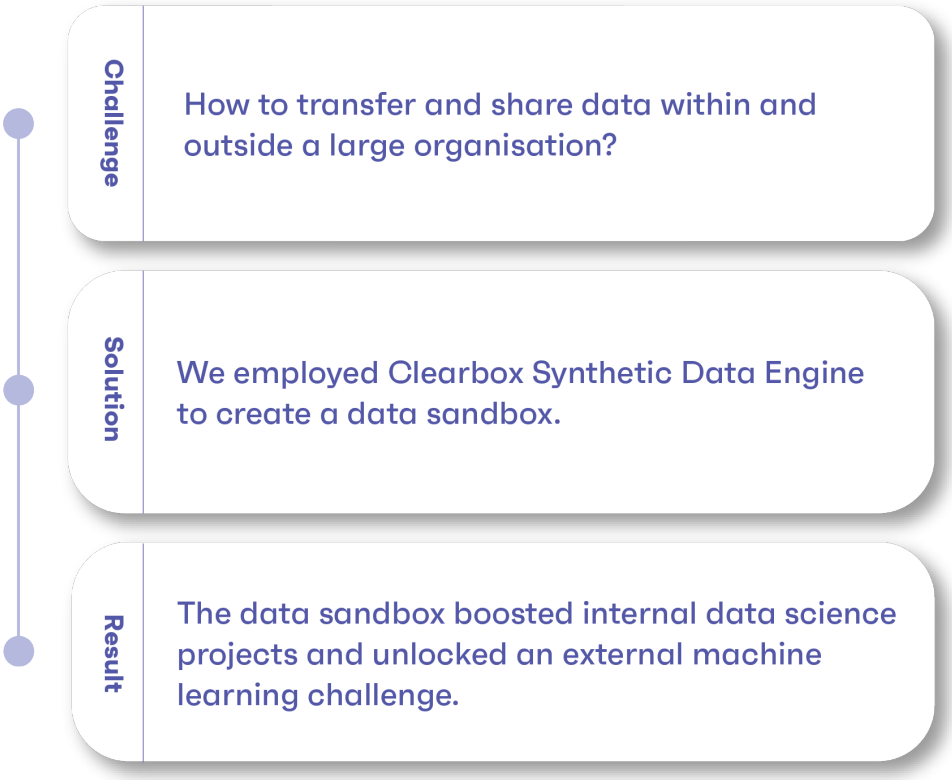
VAT ID: (IT)12161430017

clearbox.ai

Introduction

Moving data outside and within organisations can be **a daunting task**. The same can be said when re-using data for a **different purpose** from the one specified when collecting data. These kinds of **data governance issues** preclude or slow down opportunities that imply data sharing, such as open innovation challenges, cross-organisational analytics and such.

We used our Synthetic Data Engine to help an energy company **create an open data sandbox** ready to be deployed for data challenges. This use case shows that we can effectively use synthetic data to share and move data with **low re-identification risks**.



Challenge

Data regulation can hinder sharing and re-usage, often blocking and slowing down business innovation processes, especially when this **data is sensitive or strategic** for the organisation. The most common approach to **moving data** while complying with regulation is applying **standard anonymisation techniques** such as masking, perturbation, and shuffling.

The problem with this approach is that a successful anonymisation process **might destroy valuable information** from the original data, making it unusable for analytics and data science. We carried out the challenge addressed by this use case in collaboration with a large energy utility affected by the issues mentioned above. Different business units, for example, could not access each other datasets, potentially preventing actionable insights.

Solution

We used our Synthetic Data Engine to ingest and synthesise datasets from different sources across the organisation, **populating a centralised registry with synthetic data**. Each registry entity is associated with a **report** that quantifies the information lost from the original dataset during the generation process.

The same report contains an **evaluation of the re-identification risks** related to the synthetic dataset. The ingestion and generation processes use a generative model that identifies outliers and clean anomalies. This operation helps to avoid re-identification and prepare data science and analytics.

Result

Creating a **data sandbox** allowed the organisation to **boost cross-organisational innovation**. The first result was successfully handling a machine learning challenge based on a dataset that initially contained Personal Identifiable Information. The second result was a better communication of data insights across the marketing and the production unit within the same organisation, and unlocking a data sandbox allowed collaboration with relevant internal and external partners with **facilitated and safe data sharing**.

humans && machines

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

VAT ID: (IT)12161430017

clearbox.ai