Data anonymizer

Tom Rochette <tom.rochette@coreteks.org>

February 16, 2020 — c5ba03ea

1 Problem

I want my clients to share with me confidential data without revealing what the exact values are so that I can train machine learning models on this data.

2 Solution

I wrote a simple python package that uses pandas and scikit-learn to apply some simple transforms to the data. Some transforms that are applied to the dataset can change the distribution of the data, changing its statistical properties, while others preserve them but simply rescale the domain.

Given an anonymizer dataset using this tool, it is possible to do a preliminary data audit and possibly train machine learning models on the data to give a quick idea to clients whether their data looks promising or not without actually revealing the true numbers (except if desired).

The main concern with this approach is that most clients are not technical, and thus having them anonymize their data is generally not easy, if not impossible. Thus it means that such a tool is currently not applicable in the desired context.

3 Reference

- https://github.com/tomzx/data-anonymizer