Synthetische Generierung von tabellarischen Testdaten mit Deep Learning

Studiengang: MAS Data Science

Mittels Machbarkeitsstudie wurde geprüft, ob es mit aktuell verfügbaren Verfahren möglich ist, realistische Testdaten aufgrund von Originaldaten vollsynthetisch zu erzeugen. Für einzelne Tabellen kann diese Frage mit ja beantwortet werden. Bei relationalen Datenbanken ist dies aktuell noch nicht möglich. Das Haupthindernis ist die Modellierung der logischen Zusammenhänge des Datenschemas sowie der Applikation.

Ausgangslage

Die Auftraggeberin Bedag entwickelt und betreibt zahlreiche Applikationen für öffentliche Verwaltungen. Diese Applikationen enthalten sensitive wie auch personenbezogene Daten, welche speziell schützenswert sind, da sie Rückschlüsse auf reale Personen zulassen. Um diese Daten vor einem allfälligen Missbrauch zu schützen, sollten die Daten einem möglichst kleinen Personenkreis zugänglich sein.

Zielsetzung

Da die Durchführung von Softwaretests nur mit realistischen Testdaten möglich ist und nur in der Produktion mit Originaldaten gearbeitet werden sollte, soll ein Verfahren zum Erzeugen von gültigen Testdaten entwickelt werden, ohne dass Rückschlüsse auf reale Personen möglich sind. Für die Bearbeitung dieser Aufgabenstellung wurde ein Literaturstudium in Kombination mit einem Proof of Concept durchgeführt, wobei aufgrund der aktuellen Erkenntnisse der Forschung geprüft wurde, ob qualitativ hochwertige Testdaten synthetisch erzeugt werden können. Da die Auftraggeberin davon ausging, dass synthetische Daten eine erhöhte Vertraulichkeit bieten, sollten die Entwicklungs-, Demo- und Testumgebungen im Falle eines positiven Ergebnisses künftig mit synthetischen Testdaten getestet werden.

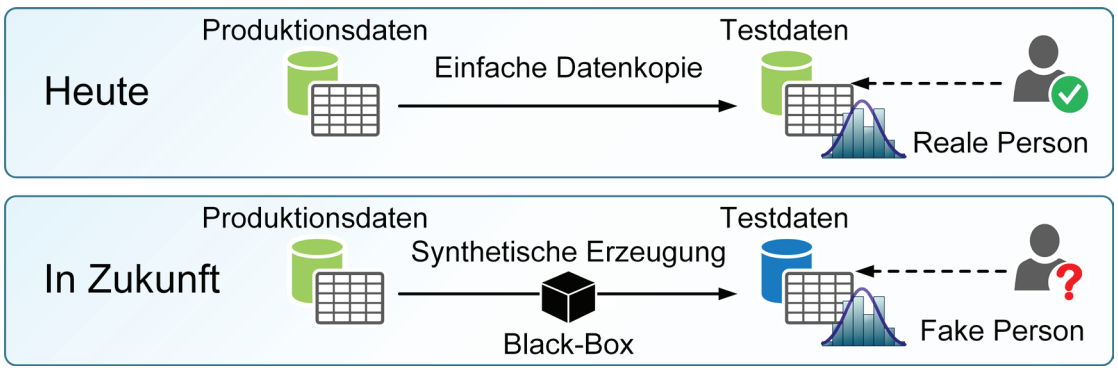
Die Vertraulichkeit von synthetischen Daten

Die Untersuchung des aktuellen Stands der Forschung hat gezeigt, dass die Vertraulichkeit von sensitiven als auch personenbezogenen Informationen in den Originaldaten durch die synthetische Erzeugung von Implikaten gewährleistet werden kann. Dabei bietet der vollsynthetische Ansatz noch mehr Vertraulichkeit als der teilsynthetische, welcher nur ausgewählte Spalten der Originaldaten synthetisiert. Basierend auf dieser Erkenntnis wurde im Rahmen der Arbeit der vollsynthetische Ansatz tiefer untersucht.

Ergebnisse und Ausblick

Die theoretischen sowie praktischen Erkenntnisse zeigen, dass das Erzeugen von vollsynthetischen Testdaten im Falle einer einzelnen Tabelle bereits heute möglich ist. Obwohl es einen ersten Ansatz für die vollsynthetische Modellierung von relationalen Datenbanken gibt, weist das experimentell getestete Modell noch grössere Mängel auf. Dabei ist insbesondere die Modellierung der logischen Zusammenhänge wie Beschränkungen des Datenschemas oder die Nachbildung der Applikationslogik nicht gelöst. Folglich sollte der teilsynthetische Ansatz weiterverfolgt werden, da die Vertraulichkeit für dieses Vorhaben wohl ausreichend wäre und aufgrund der vorliegenden Erkenntnisse dieser Arbeit einfacher zu realisieren ist.

Adrian Bronner
adrian.bronner@bedag.ch



Schematische Darstellung der Zielsetzung