Sind vom Zukunftspotenzial synthetischer Daten überzeugt (v.l): Mostly AI-Gründer Michael Platzer und Geschäftsführer Tobias Hann

Synthetische Daten – DSGVO-konformes Allheilmittel?

Daten sind für Unternehmen, Forschung und Wissenschaft von unschätzbarem Wert. Je detaillierter, desto besser lautet die Prämisse. Datenschutzexperten schlagen regelmäßig Alarm; das Wiener Start-up Mostly AI bietet hier einen Ausweg. VON MICHAELA SCHELLNER

Handelsunternehmen, Banken und Handy-Anbieter, aber auch Versicherungen, Gesundheitsdienstleister und die öffentliche Hand – sie alle sammeln jeden Tag unzählige Kundendaten. Die daraus generierten Erkenntnisse werden verwendet, um beispielsweise das Sortiment im Supermarkt an die vorhandenen Bedürfnisse anzupassen, maßgeschneiderte Angebote zu entwickeln oder die eigenen Prozesse zu optimieren. Herausfordernd wird es aber dann, wenn sich externe Partner wie zum Beispiel Wissenschaft und Forschung, andere Unternehmen oder sogar verschiedene Mitglieder eines gemeinsamen Kundenclubs für die Daten des jeweils anderen interessieren. Die Weitergabe an Dritte ist bekanntlich streng verboten und in der seit Mai 2018 geltenden Datenschutzgrundverordnung (DSGVO) klar geregelt.

Diese Einschränkung erweist sich jedoch als Hemmschuh in vielen Bereichen, wie Thomas Reutterer, Professor für Marketing und Kundenanalyse an der Wirtschaftsuniversität Wien erklärt: „Man denke etwa an die Standort- und Bewegungsdaten, die Telekommunikationsanbieter sammeln. Verkehrs- oder Stadtplaner würden sich sehr über personenbezogene bzw. individuelle Daten freuen, um etwa geplante Projekte oder Verkehrsströme optimieren zu können. Ein anderes, ganz aktuelles Beispiel ist die Eindämmung der Corona-Pandemie, wo solche Bewegungsdaten auch eine große Rolle für wissenschaftliche und politische Entscheidungen spielen können."

Anonymisierungsverfahren stoßen an Grenzen

Damit Dritte mit nicht selbst erhobenen Daten arbeiten dürfen, müssen diese so weit anonymisiert werden, dass keine Rückschlüsse mehr auf einzelne Personen möglich sind. Genau das ist aber dann besonders schwierig, wenn ein gesammelter Datensatz sehr komplex ist, das heißt, viele verschiedene statistische Parameter enthält. Hier stoßen traditionelle zeit- und kostenintensive Anonymisierungsverfahren, sprich: die Abänderung des Originaldatensatzes durch zum Beispiel Hinzufügen von Zufallswerten oder die Streichung bestimmter Variablen, immer häufiger an ihre Grenzen.

Lösungsansatz: Künstlich generierte Daten

„Wenn ich von einem Supermarkt-Konsumenten 200 statistische Variablen sammle, etwa zu Geschlecht, Alter, Warenkorbzusammensetzung, Einkaufszeitpunkt, Zahlungsmethode und hier noch weiter in die Tiefe gehe, dann müsste ich am Ende wirklich vieles davon löschen, um die Anonymität der jeweiligen Person zu gewährleisten. Die Daten, die dann übrigbleiben, sind für Datenwissenschaftler aber nicht sinnvoll nutzbar. Damit lassen sich nämlich weder Analysen erstellen, noch eine Software programmieren oder testen", erklärt Tobias Hann, Ge-

Foto: Mostly AI

schäftsführer von Mostly AI. Das Wiener Daten-Start-up will genau hier Abhilfe schaffen und erzeugt deshalb seit seiner Gründung im Jahr 2017 aus Standort-, Handelskunden- oder Banktransaktionsdaten mithilfe von künstlicher Intelligenz (KI) sogenannte synthetische Daten. Diese enthalten zu 99 Prozent die statistischen Eigenschaften des Original-Datensatzes, aber keine individuellen Merkmale. Damit sind sie – so Hann – vollkommen anonym, jedoch in ihrer Aussagekraft nicht beeinträchtigt. Und sie fallen nicht mehr unter den Anwendungsbereich der DSGVO. Ein Argument, mit dem auch die Bedenken von Datenschützern ausgeräumt werden sollen.

Technische Exzellenz garantiert Datenschutz

„Unsere Plattform, die aus einem bestehenden Datensatz eine synthetische Daten-Kopie erzeugt, wird direkt bei unseren Kunden vor Ort installiert und ist kinderleicht zu bedienen. Ein realer Datensatz wird eingegeben, ein synthetischer kommt heraus. Die Software hält außerdem keinerlei Rücksprache mit uns und verlässt somit auch nicht die geschützte IT-Infrastruktur." Den Einwand, dass die zum Einsatz kommende künstliche Intelligenz den Original-Datensatz einfach auswendig lernen könne und damit die Anonymität in der Kopie nicht vorhanden sei, kann Hann nachvollziehen: „Hier spielt die technische Exzellenz bei der Daten-Synthese eine entscheidende Rolle. Unsere Anwendung verhindert, dass zu viel gelernt wird, sorgt aber dafür, dass man möglichst nah an den Echt-Daten ist." Mostly AI beschäftige sich seit fünf Jahren mit diesen Prozessen; regelmäßige Test zum Schutz der Privatsphäre seien selbstverständlich. „Zahlreiche Zertifizierungen sowie rechtliche Gutachten bestätigen unsere Professionalität", so Hann. Welche Kosten für Unternehmen für die Nutzung der Software anfallen, sagt der Geschäftsführer nicht; sie habe aber ihren Preis.

Anwendungsfall Betrugserkennung

Als die Gründer Michael Platzer, Klaudius Kalcher und Roland Boubela vor fünf Jahren mit Mostly AI starteten, waren sie zwar vom Potenzial der Nutzung von KI zur Generierung von strukturierten Geschäftsdaten überzeugt – wie das in der Praxis funktionieren kann und welche Prozesse dafür nötig sind, wusste damals allerdings niemand. „Es gab zu diesem Zeitpunkt keine Forschung und keine Mitbewerber, sondern nur eine Idee, die aus den Erfahrungen, die Michael, Klaudius und Roland in ihren Berufen als Datenwissenschaftler gemacht haben, entstanden ist. Jetzt, fünf Jahre später, sehen wir, dass die Ursprungsvision Früchte trägt und das Thema für immer mehr Unternehmen an Relevanz gewinnt", so Hann. Neben der Anonymisierung hätten synthetische Daten aber noch weitere Vorteile. „Sie können modifiziert und bearbeitet werden, um KI-Initiativen zu beschleunigen. Und sie ermöglichen es Unternehmen, ihre Datensätze zu erweitern." Einen Anwendungsbereich sieht der Geschäftsführer hier etwa in der Betrugserkennung, wo Finanzdienstleistern oft nur wenige Daten zur Verfügung stehen. Um beispielsweise illegale Geldwäscheaktivitäten zu erkennen und Algorithmen diesbezüglich entsprechend zu trainieren, braucht es eine angemessene Datenbasis, die mithilfe der Daten-Synthese erzeugt werden kann. Auch im Gesundheitsbereich, etwa bei Patientendaten und deren grenzüberschreitender Nutzung zum Beispiel für die Impfstoff- oder Medikamentenentwicklung sei das Potenzial riesig.

Thomas Reutterer, Professor für Marketing und Kundenanalyse an der Wirtschaftsuniversität Wien: „Innovationsführer verschiedener Branchen setzen bereits darauf, weitere werden folgen."

Aktuell weltweit 20 Mitbewerber

Mittlerweile hat sich Mostly AI zum Weltmarktführer in der Erzeugung von KI-generierten synthetischen Daten gemausert und konkurriert global mit rund 20 Mitbewerbern. Die Mitarbeiterzahl ist auf 35 gestiegen; zu den Kunden zählen große Banken (z. B. Erste Group) und Versicherungen in Europa und Nordamerika sowie Telekommunikationsanbieter. Um weiter zu wachsen, sammelte das Start-up kürzlich 25 Millionen US-Dollar von der britischen Risikokapitalgesellschaft Molten Ventures, den Investor-Unternehmen Earlybird und 42CAP sowie Citi Ventures ein. Mit dem Kapital will man den amerikanischen Markt stärker bearbeiten und die Anzahl der Mitarbeiter verdoppeln.

Auch für Universitätsprofessor Reutterer führt mittelfristig kein Weg an der Nutzung synthetischer Daten vorbei. „Innovationsführer verschiedener Branchen setzen bereits darauf, weitere werden folgen." Ähnliches prognostiziert das auf IT-Entwicklungen spezialisierte Marktforschungsinstitut Gartner. Diesem zufolge werden bis 2024 60 Prozent der Daten, die für die Entwicklung von KI- und Analyseprojekten zum Einsatz kommen, synthetisch generiert sein.

Mostly AI-Anwendungsbeispiel

Um Muster in der Mitarbeiterfluktuation zu erkennen, wollte ein großes Telekommunikationsunternehmen die Daten von über 90.000 Mitarbeitern sammeln und analysieren. Die manuelle Anonymisierung für jedes Analyseprojekt dauerte im Durchschnitt sechs Wochen. Mithilfe synthetischer Kopien konnten die repräsentativen Daten sofort und projektübergreifend genutzt werden. So erkannte das Analyseteam rasch Muster, die zur Abwanderung von Mitarbeitern führten, identifizierte die am meisten gefährdeten Mitarbeiter und entwickelte Maßnahmen zur Bindung von Talenten. Mit einer Senkung der Fluktuationsrate um nur 0,5 Prozent ließen sich Kosteneinsparungen in zweistelliger Millionenhöhe erzielen. Zudem wirke sich eine solche positiv auf die Produktivität des Teams aus.

Foto: WU Wien