

Synthetische Daten

Anonymisierung als Lösung?

Dr. Zoi Opitz-Talidou, LL.M. CIPM

10. Münchner Datenschutztag

01.12.2022

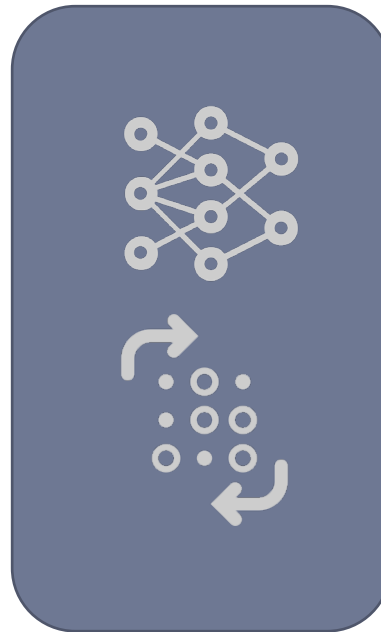
Was sind Synthetische Daten?

Synthetic data is generated by applying a sampling technique to real-world data or by creating simulation scenarios where models and processes interact to create completely new data not directly taken from the real world. **Gartner**



Echte Daten

- Kundendaten
- Transaktionsdaten
- etc.



Generative Adversarial Network

Privacy Tests
Accuracy Tests



Synthetische Daten

- Identische Struktur
- Identische statistische Eigenschaften
- 100% synthetisch, d.h. keine Rückschlüsse auf echte Daten möglich

Synthetische Daten versus anonyme Daten: Was ist der Unterschied?

Synthetische Daten

- künstliche erzeugte Daten
- Nachahmung von Struktur und statistischen Eigenschaften von Originaldaten
- Erlernen von Strukturen/Eigenschaften der Originaldaten und Bau eines Modells, der diese Strukturen/Eigenschaften wiedergibt
- Keine Eins-zu-Eins Beziehung zu den Originaldaten

Anonyme Daten

- Informationen, die sich nicht auf eine bestimmte oder beziehbare natürliche Person beziehen lassen (ErwG 26 DSGVO)
- Das Entfernen von Klardaten wie Name, Adresse von einem Datenbestand reicht nicht aus
- Weitere Schutzmechanismen wie k-Anonymity erforderlich

Wichtig! Für beide Verfahren Sicherstellung der Robustheit: Keine Aufhebung der Anonymisierung durch

- Datenbasierte Isolierbarkeit einer Einzelperson (Singling out)
- Verknüpfbarkeit
- Identifizierung durch Inferenztechniken

Rechtliche Einordnung der Synthetisierung

- **Synthetisierung = Anonymisierung**
 - Technisches Verfahren mit Hilfe dessen neue, anonyme Datensätze mit größtmöglich ähnlichem Informationsgehalt erzeugt werden
 - Re-Identifizierung der betroffenen Personen, deren Daten synthetisiert wurden, ist technisch nahezu unmöglich
- **Synthetisierung/Anonymisierung = Löschung**
 - Gemäß Art. 5 Abs. 1 lit. e) i. V. m. Art. 6 Abs. 1 lit. c) DSGVO sind Daten zu löschen, wenn diese nicht mehr erforderlich sind
- **Synthetisierung/Anonymisierung = Erzeugung eines nicht-personenbezogenen Datenbestandes**
 - Gem. Art. 6 Abs. 1 S. 1 lit. f) bzw. Art. 6 Abs. 4 DSGVO kann eine verantwortliche Stelle Daten synthetisieren, um Verarbeitungen zur Verfolgung von berechtigten Interessen bzw. von kompatiblen Zwecken datenschutzfreundlich durchzuführen

Synthetisierung: Datenschutzrechtliche Must-Haves

- ✓ Für die Rechtmäßigkeit der Anonymisierung ist nach herrschender Meinung eine Rechtsgrundlage erforderlich
 - ✓ Art. 6 Abs. 1 lit. f DS-GVO: Berechtigtes Interesse z. B. mit Hilfe von synthetischen Daten Entwicklung und Testing von IT-Systemen zu erreichen
 - ✓ Art. 6 Abs. 4 DSGVO einen neuen Zweck zu erreichen, der mit dem ursprünglichen Zweck (Vertragsanbahnung/Vertragsdurchführung) kompatibel ist
- ✓ Einhaltung des Grundsatzes der Transparenz erforderlich
- ✓ Implementierung von toMs bei den Systemen, wo Synthetisierungssoftware implementiert

Synthetisierung: Datenschutzrechtliche Must-Haves

- ✓ Bewertung der Notwendigkeit der Durchführung einer DPIA (Schwellwert-Analyse), immer zu bejahen, wenn neuartige Technologie (z. B. AI Methoden) in Einsatz kommt
- ✓ Dokumentation: Planung, Bewertung der Synthetisierung sowie Durchführung einer Interessenabwägung bzw. Kompatibilitätsprüfung
- ✓ Durchführung von Privacy-Tests beim Ergebnis der Synthetisierung (Sicherstellung, dass keine Re-Identifizierung möglich ist)
- ✓ Evaluierung der Robustheit des Verfahrens in regelmäßigen Abständen
- ✓ Festlegung einer für die Businessbereiche angemessene Löschfrist der synthetischen Daten

Synthetic Data Use Cases @ TEF DE

Anwendungsgebiet	Beschreibung	Anwendungsfälle
Citizen Data Sciene	 <ul style="list-style-type: none">➤ Den Fachbereichen (Marketing, Sales, Customer Service, etc.) wird es ermöglicht, einfache Analysen selbstständig auf Einzeldatensatzebene durchzuführen.	 <ul style="list-style-type: none">✓ Angebotsgestaltung✓ Customer Service Optimierung
Performance Tests	 <ul style="list-style-type: none">➤ In der Migration von großen Datenmengen kann mit beliebig vielen synthetischen Datensätzen getestet werden.	 <ul style="list-style-type: none">✓ Infrastruktur-Migration✓ Hardware-Migration✓ Skalierung
Data Science Modeling	 <ul style="list-style-type: none">➤ Typische Probleme (Bias, Schiefe, zu wenige Daten, etc.) beim Data Science Modeling können mit synthetischen Daten umgangen werden.	 <ul style="list-style-type: none">✓ Reduzierung des Bias✓ „Saubere“ Daten



Telefónica