日本総研
The Japan Research Institute, Limited

プライバシー保護合成データの概説と動向

2023年6月13日

株式会社日本総合研究所
先端技術ラボ

＜本件に関するお問い合わせ＞森　毅(mori.takeshi@jri.co.jp)

改訂履歴

版	改訂内容	改訂日
1.0	初版	2023年6月13日
1.1	• 体裁の軽微な修正 • P.20 仮名加工情報の第三者提供可否についての記載を修正	2023年12月28日

はじめに

- レポートの内容：　プライバシー強化技術としての合成データに焦点を当て、概要と事例をまとめたもの
- 想定する読者：　プライバシー強化技術に関心を持つ、合成データの活用を検討しているビジネスパーソン
- 前提とする知識：大半の内容において前提とする知識はない（一部でAIに関する知識が必要）

エグゼクティブ・サマリ

背景・ レポートの構成	• AI・データ利活用の促進と並行してデータ保護規制の動きも拡大している。多くの組織では、保有するデータに含まれるプライバシー情報を保護した状態でデータ利活用を進める方法を模索しており、プライバシー強化技術は解決策の一つである。 • **合成データ(Synthetic Data)は実在するデータと同じ構造で異なる値を持つデータ**であり、実在する個人のデータを直接用いないことで、データ主体*1のプライバシーを保護できる**プライバシー強化技術としても注目されている**。 • 本レポートではプライバシー保護を目的とする合成データ(プライバシー保護合成データ)に焦点を当て、技術の概説と事例をまとめた。また、活用に向けて現状と課題について整理し、今後の展望および活用を検討する組織に対する推奨事項を考察・提言した。
技術概説	• 近年提案されている合成データの生成手法の多くは、元データの統計的特徴を保持する手法であり、**プライバシーを保護した状態で分析に利用できる**といった利点がある。 • 一方、元データの統計的な特徴を保持していることから、**合成データから元データの情報が推測される等の脅威を抱えている**。脅威への対策として、差分プライバシー等により**プライバシー保護を強化した合成データ(Privacy-Preserved Synthetic Data)を生成する手法も提案**されている。
活用動向	• 合成データのユースケースは、組織間・組織内でのデータ蓄積・共有、外部のデータ分析者の活用、データ販売による収益化等多岐にわたる。 • 主に国外において合成データを活用したデータ分析の事例が増えている。COVID-19患者の分析に用いる事例や、外部企業へのデータ連携における社内承認フローを簡易化・高速化する事例等が挙げられる。
課題・提言	• **プライバシー強化技術として合成データが認知されていないこと、合成データの安全性に関する評価基準が明確に定まっていないことなどから、社会全体への普及には時間がかかる**と考えられる。 • 合成データの活用を検討する組織に向けた推奨事項として、**①技術やツールの調査・評価**、**②コンプライアンス・法律等の複合的な観点による評価**、**③データガバナンス・プライバシー影響評価体制の整備等による社会受容性の確保**が挙げられる。

*1 GDPR(EU一般データ保護規則)内では「data subject(データ主体)」は「識別された、もしくは識別可能な自然人」として定義されている。本レポートでは個人データの該当者を表す単語として使用する。

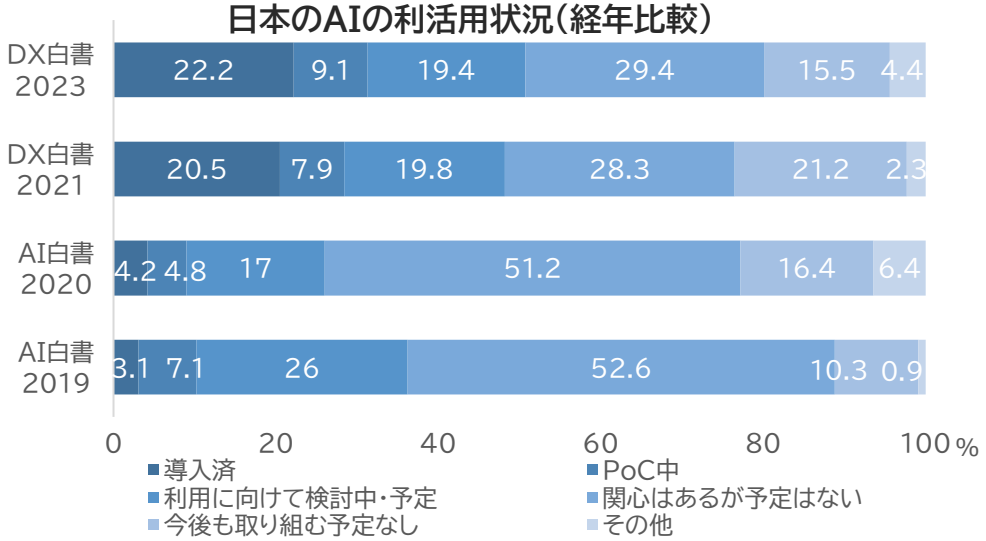
日本総研 The Japan Research Institute, Limited

1.1 データ利活用の進展と課題

- デジタル化によるデータ量の増加とAIの進化によって、データ利活用への期待が高まっている。
- データ利活用の推進と並行してデータ保護規制の動きも拡大しており、保有するデータに含まれるプライバシー情報を保護する方法が重要になる。

AI社会実装の進展

- 米IBMの調査*1によれば、2022年時点において、世界のAI導入率は35%に達し、加えて、42%がAIの導入を検討している。
- またDX白書2023（IPA発行）*2によれば、日本企業のAI導入率は着実に増加しており、22.2%に達する（下図）。



IPA「DX白書2023」 図表1-34、IPA「DX白書2021」*3 図表42-48を基に日本総研が作成

*1「IBM Japan Newsroom - ニュースリリース『IBM、「世界のAI導入状況 2022年（日本語版）」を発表』」, 2022-07-12, https://jp.newsroom.ibm.com/2022-07-12-AI-Adoption-Index-2022,（アクセス日 2023/04/28）
*2「DX白書2023」, 2023-03-16, https://www.ipa.go.jp/publish/wp-dx/gmcbt8000000botk-att/000108041.pdf,（アクセス日 2023/04/28）
*3「DX白書2021」, 2021-12-01, https://www.ipa.go.jp/publish/wp-dx/qv6pgp0000000txx-att/000093706.pdf,（アクセス日 2023/04/28）

データ保護規制の強化

- 2021年には中国、2022年にはタイで個人情報保護法が施行されており、世界的にデータ保護規制のための法整備が進んでいる。
- Gartnerは2023 年末までに世界中の企業の80%以上が、少なくとも 1 つの、プライバシーに焦点を当てたデータ保護規制に直面すると予測している。*4

*4「Gartner Says Digital Ethics is at the Peak of Inflated Expectations in the 2021 Gartner Hype Cycle for Privacy」, https://www.gartner.com/en/newsroom/press-releases/2021-09-30-gartner-says-digital-ethics-is-at-the-peak-of-inflate（アクセス日 2023/04/28）

EUのデータ保護規制によって日系IT企業が行政処分を受けた例*5

- 2022年11月に、NTTデータのスペイン子会社が、GDPR*6違反により6万4000ユーロの制裁金を科された。
- NTTデータスペインの顧客である保険会社が、顧客情報の漏洩問題を起こしたことに対し、顧客管理システムを提供していたNTTデータスペイン側にも過失があったとするもの。

*5「NTTデータの海外子会社がGDPR違反で制裁金、ついに日系IT企業が摘発対象に」. 日経クロステック. 2022-12-23, https://xtech.nikkei.com/atcl/nxt/column/18/00989/122100105/（アクセス日 2023/04/28）
*6 General Data Protection Regulation、EU一般データ保護規則

日本総研
The Japan Research Institute, Limited

1. 2 プライバシー強化技術

- データ保護規制の強化に伴い、プライバシー保護を実現するための技術（プライバシー強化技術*1）が注目されている。
  - 実現するプライバシー原則の要件や、技術を適用する方法・タイミングに応じて複数の技術が提案・利用されている。

プライバシー原則とプライバシー強化技術

- プライバシー保護規制のベースには、OECD *2やISO *3が定めるプライバシー原則がある。
  - 利用目的を定める「目的明確化」や、目的に沿ったデータへのアクセスを最小限にする「最小化」等
- プライバシー原則を実現・強化するための技術として、プライバシー強化技術が注目されている
- プライバシー強化技術全体の概説は、日本総研発行のレポート「プライバシー強化技術の概説と動向」を参照。

プライバシー強化技術の主な要素技術	
認証・アクセス制御	連合学習
デジタル署名	ゼロ知識証明
ブロックチェーン	匿名化
秘密計算	暗号化
差分プライバシー	合成データ

*1 Privacy Enhancing Technologies、PETsとも呼ばれる。近年プライバシーテックという呼称も登場
*2 Organization for Economic Co-operation and Development（経済協力開発機構）の略
*3 International Organization for Standardization（国際標準化機構）の略

プライバシー強化技術に関するレポート

2023年2月には国連から、2023年3月にはOECDからプライバシー強化技術に関するレポートが発行されている

図出典
（左）「2023_UN PET Guide.pdf」, https://unstats.un.org/bigdata/task-teams/privacy/guide/2023_UN%20PET%20Guide.pdf （アクセス日 2023/04/28）
（右）「Emerging privacy enhancing technologies: Maturity, opportunities and challenges」, https://www.oecd-ilibrary.org/deliver/bf121be4-en.pdf?itemId=%2Fcontent%2Fpaper%2Fbf121be4-en&mimeType=pdf（アクセス日 2023/04/28）

日本総研 The Japan Research Institute, Limited

1. 3 プライバシー強化技術：合成データ（Synthetic Data）

- 合成データ（Synthetic Data）は、実在するデータと同じ構造で異なる値を持つデータの総称である。
- 実在するデータを直接用いないことでデータ主体のプライバシーを保護し、組織のデータ利活用を促進させるプライバシー強化技術の一つとして、近年注目されている。
- 本レポートでは、テーブル形式のパーソナルデータに対するプライバシー保護を目的として利用する合成データを対象にし、技術の概説・活用動向をまとめる。

合成データ

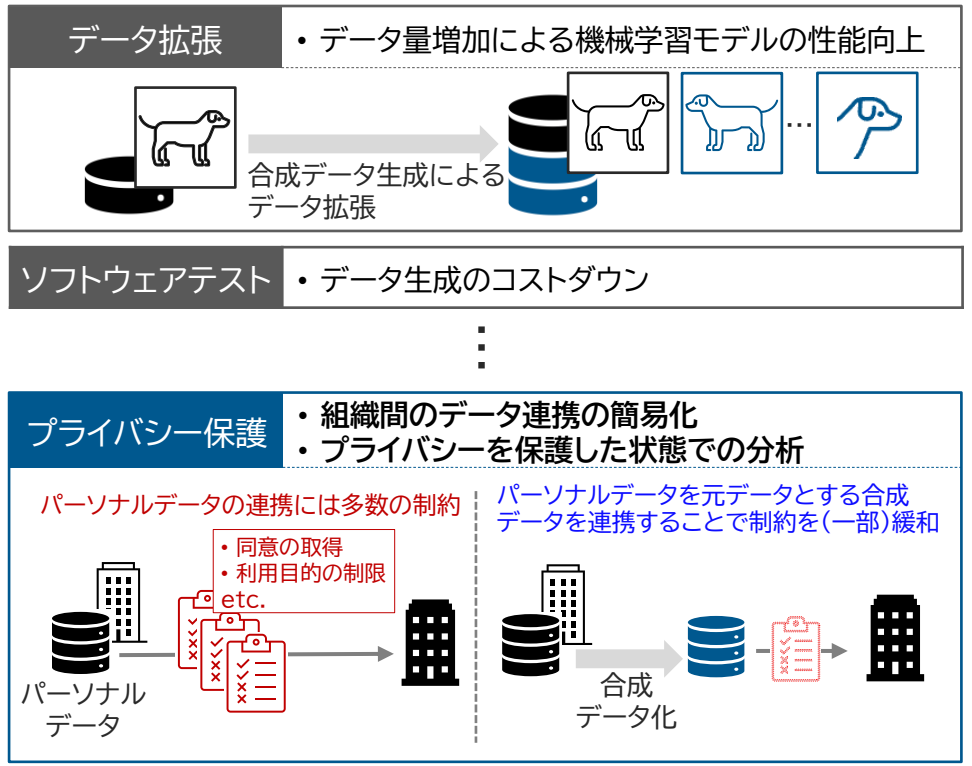
- 合成データ（Synthetic Data）は、実在するデータと同じ構造で異なる値を持つデータの総称である。
- 合成データの例
  - データ拡張*1によって得られるデータ
  - 画像生成AI（Stable Diffusion・Midjourney等）により生成されるデータ
  - Deepfakeによって生成されるデータ
  - デジタルツイン*2から得られるデータ

プライバシー強化技術としての合成データ

- 合成データはプライバシー強化技術の一つとして注目されている
  - 合成データを用いることで、実在するパーソナルデータへのアクセスを最小限に留める「最小化」が実現・強化される
- プライバシー強化技術として合成データを利用する場合、元データ*3に含まれるプライバシー情報を保護するための技術的な工夫がなされることが多い（詳細 2章）

*1 保有するデータを元に加工等によって新しいデータを生成し、データのサイズを拡張する手法のこと。Data Augmentationとも呼ばれる
*2 現実空間をサイバー空間上に再現したもの。もの作りの試作やシミュレーションによるデータ収集に活用可能
*3 合成データを生成するための元となるデータを本レポートでは「元データ」と呼ぶ

合成データの活用先



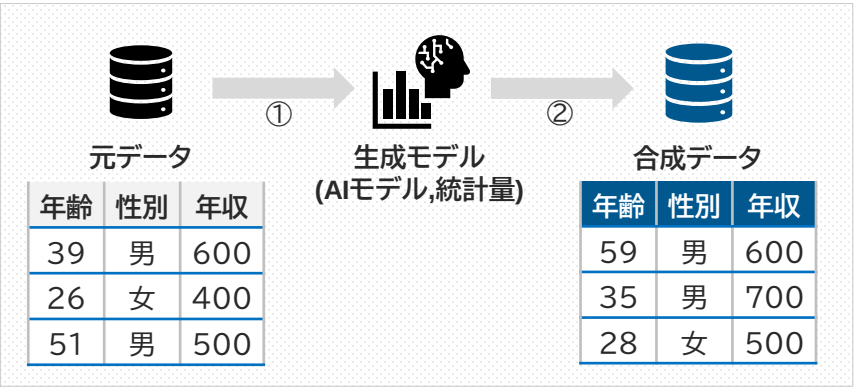
日本総研 The Japan Research Institute, Limited

2.1 合成データの生成

- 合成データは一般に、生成モデルが学習した元データの特徴を保持したデータを生成することで得られる。
  - 生成モデルによる生成手法の多くは統計的な性質を保持した合成データを生成出来るため、分析用途に利用可能。
- ベースとなる生成モデルに応じて複数の合成データの生成手法が研究・提案されている。

生成モデルを用いた合成データの生成

① 生成モデルが元データの統計的な特徴を学習
（元データとの1対1の関係は保存されない）
② 生成モデルが学習した特徴を保持したデータを生成



年齢	性別	年収
39	男	600
26	女	400
51	男	500

年齢	性別	年収
59	男	600
35	男	700
28	女	500

生成モデルの種類

生成モデル	概説	手法例
① 統計値ベース	周辺確率分布やクロス集計表を復元する合成データを生成する	• 統計値を使用した手法[*1]
② Baysian Network	元データから確率モデルを構築し、それを元にデータを生成する	• Baysian Networkを用いた手法[*2]
③ Copula	統計的な依存関係を学習したモデルを元にデータを生成する	• Gaussian Copula • Copula Flow
④ GAN	生成モデルと識別モデルを競い合わせ、本物に近いデータを生成する	• TGAN • CTGAN
⑤ Diffusion Model	元データにノイズを加える過程の逆過程により、ノイズからデータを生成する	• TabDDPM

合成データの歴史

- 物理学における未観測データの推定や、音声の合成といった領域で用いられてきた。
- プライバシー保護を目的とした合成データの利用は1993年に提案されている。[*3]
  - 米国の国勢調査データの機密保持のために、マスキングではなく合成マイクロデータの公開を提案した。
- 2000年以降は機械学習・深層学習の発展と共に合成データの生成方法が高度化している。
  - 2023年にはLLM(大規模言語モデル)を用いて、テーブルデータの合成データを生成する手法が提案されている。[*4]

*1 岡田 莉奈 他,統計値を用いたプライバシ保護擬似データ生成手法, コンピュータセキュリティシンポジウム, 2017
*2 Jim Young et al., Using Bayesian Networks to Create Synthetic Data, Journal of Official Statistics, Vol. 25, No. 4, 2009, pp. 549–567
*3 Donald B. Rubin, Discussion: Statistical Disclosure Limitation, Journal of Official Statistics, Vol. 9, No. 2, 1993, pp. 461-468
*4 Vadim Borisov et al., Language Models are Realistic Tabular Data Generators, arXiv preprint arXiv: 2210.06280, 2023

日本総研 The Japan Research Institute, Limited

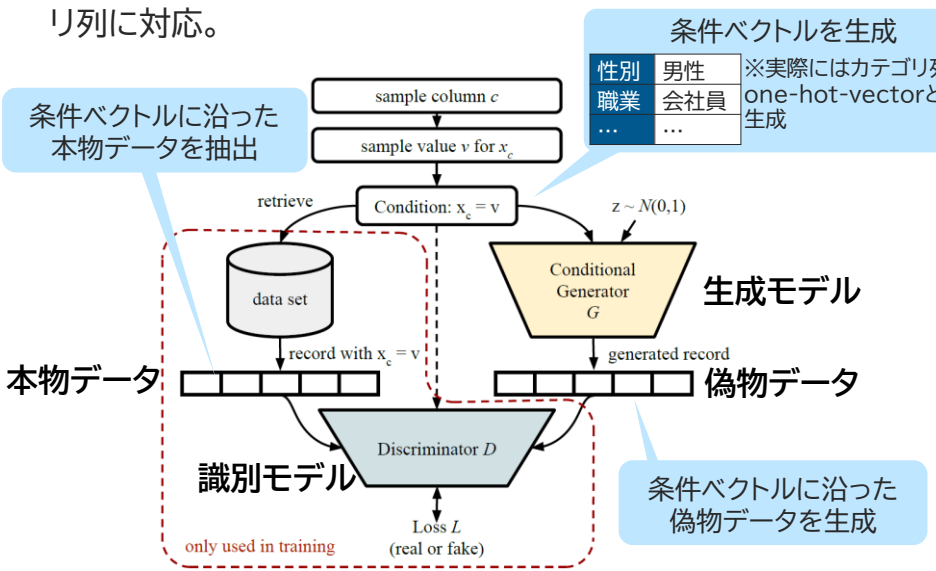
（参考）代表的な合成データ生成モデル：CTGAN

- 合成データの生成モデルの代表的な例として、CTGANが挙げられる。
- GAN*1ベースのテーブルデータ生成モデルであり、テーブルデータ特有の課題に対応した生成が可能。

テーブルデータ特有の課題	解決策
• アンバランスなカテゴリ列	① 条件ベクトル (condition vector)
• 非ガウス性 • 多峰性	② モード固有の正規化 (mode-specific normalization)

① 条件ベクトル

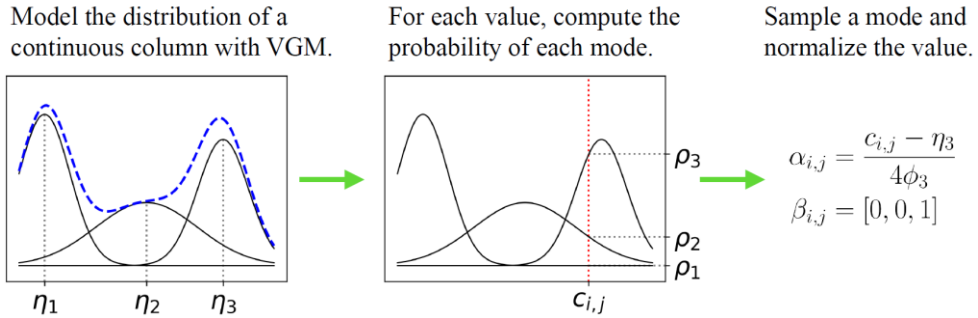
- 特定のカテゴリ列の条件を表現する条件ベクトルを生成し、その条件に沿った本物データと偽物データ同士で学習を行う。
- 条件ベクトルの生成確率を調整することで、アンバランスなカテゴリ列に対応。



Vadim Borisov et al., Deep Neural Networks and Tabular Data: A Survey, arXiv preprint arXiv:2110.01889v2, 2021 説明のため一部文章を追記

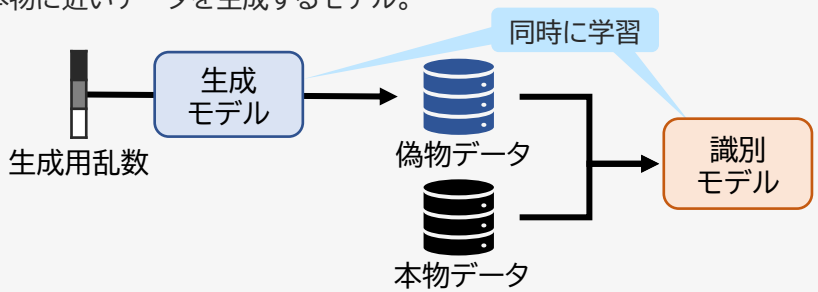
② モード固有の正規化

- テーブルデータは画像データと異なり、正規分布に従わず（非ガウス性）、各列が多峰性を持つという特徴がある。
- 多峰性を持つ確率密度関数を複数の正規分布で近似することで、よりリアルなテーブルデータの生成を可能にしている。



Lei Xu et al., Modeling Tabular Data using Conditional GAN, 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), 2019

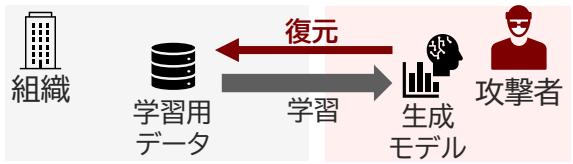
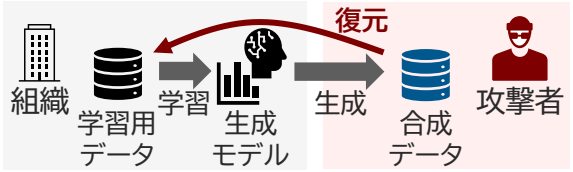
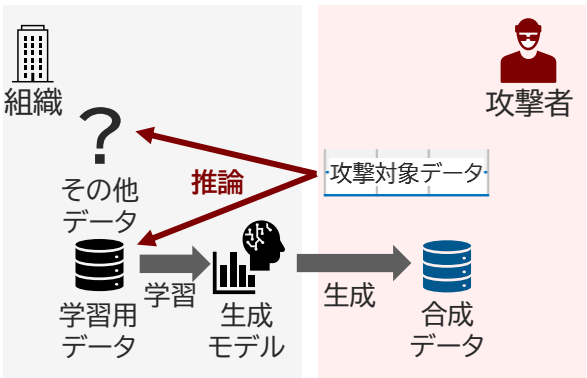
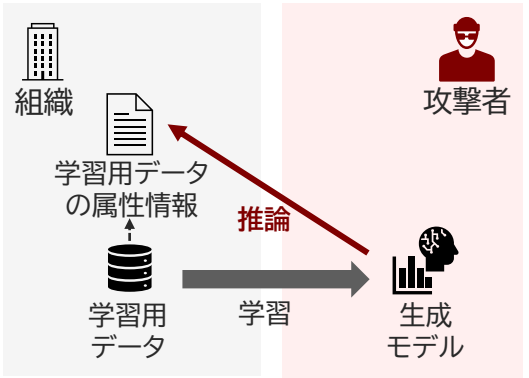
*1 GAN（敵対的生成ネットワーク）
本物に近い偽物のデータを生成する生成モデルと、本物データと偽物データを高精度で分類する識別モデルを、競わせながら同時に学習することで、より本物に近いデータを生成するモデル。



日本総研
The Japan Research Institute, Limited

2. 2 合成データに対する攻撃

- 合成データや生成モデルから、元の学習データの情報が漏れるというリスクがある。
- 攻撃の目標、攻撃者が保有する事前情報によって様々な攻撃手法が提案されている。
- プライバシー保護を目的とする合成データの利用には、これらのリスクを考慮する必要がある。

攻撃手法	Model Inversion Attack*2/ Reconstruction Attack（モデル反転攻撃/再構成攻撃）	Membership Inference Attack（メンバシップ推論攻撃）	Property Inference Attack（属性推論攻撃）
目標	• 学習用のデータを直接復元 • 学習データ内の属性情報を取得	• 攻撃対象のデータが学習データに含まれるかどうかを推論	• 学習データの属性・特徴を推論（あるカテゴリの割合等）
事前情報	• 生成モデル • 合成データ	• 合成データ・攻撃対象データ • 生成モデル・攻撃対象データ	• 生成モデル • 合成データ
攻撃イメージ（一例）	事前情報：生成モデル  事前情報：合成データ 	事前情報：合成データ・攻撃対象データ 	事前情報：生成モデル 

※ 機械学習モデル全般に対する攻撃手法も含む

*2「Algorithms that remember: model inversion attacks and data protection law | Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences」, 2018-10-15, https://royalsocietypublishing.org/doi/10.1098/rsta.2018.0083

日本総研 The Japan Research Institute, Limited

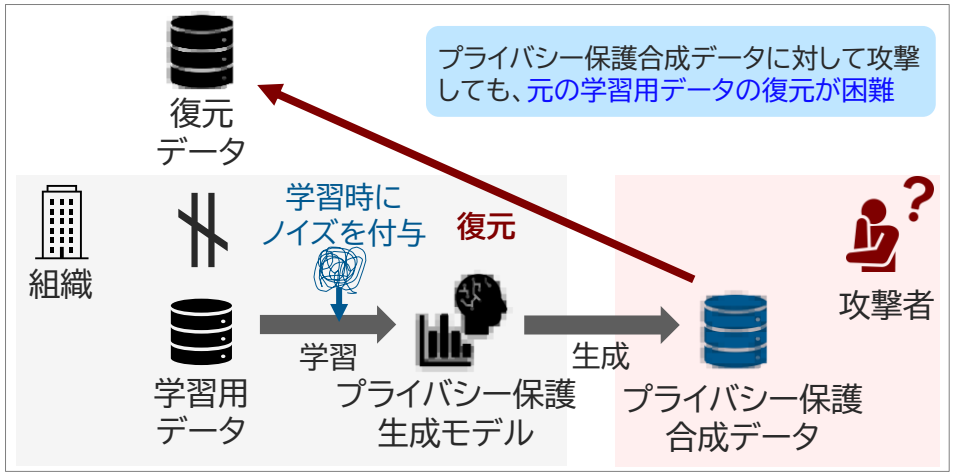
2.3 プライバシー保護合成データ(Privacy-Preserving Synthetic Data)

- 合成データおよび生成モデルに対する攻撃が困難な、生成モデルの学習手法やデータ生成手法が研究されている。
  - ノイズの付加等により、差分プライバシーと呼ばれる安全性指標を満たす手法が主流。
- 攻撃を困難にすることで、元のデータに対するプライバシー保護をより強化している。

プライバシー保護合成データの生成例

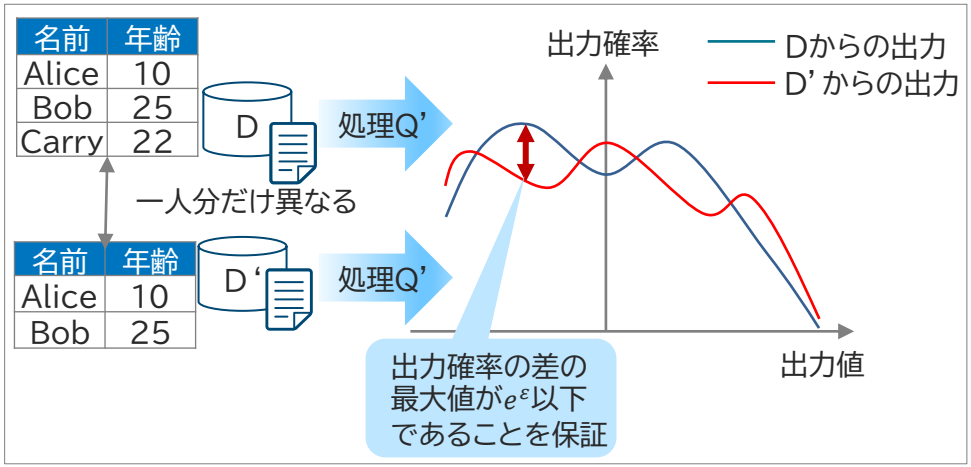
- 生成フロー
  - ① 生成モデルの学習の際に**ノイズを付与**
  - ② ノイズを付与した生成モデルから合成データ(プライバシー保護合成データ)を生成
- ノイズによって生成モデル/合成データに対する攻撃が困難。

例:プライバシー保護合成データへのModel Inversion Attack



差分プライバシー(Differential Privacy, DP)

- 個人ごとのデータを記録し、各々に含まれるデータが一人分だけ異なる(隣接する)2つのデータベースをDとD'、データベースに対する処理をQ'とする。なお、処理Q'には確率的な要素が含まれるものとする。
- 2つのデータベースに対する処理Q'の結果がある値になる確率の対数差を考える。この対数差が、たかだかεに抑えられるとき、処理Q'はε-差分プライバシーを満たすという*1。



*1 厳密には任意の隣接するD,D'の組に対して以下の式を満たす場合、Q'はε-DPを満たすという。Sは処理Q'の出力空間の任意の部分集合である。

$$\Pr[Q'(D)\in S] \leq e^{\varepsilon} \cdot \Pr[Q'(D')\in S]$$

日本総研
The Japan Research Institute, Limited

2.4 プライバシー保護合成データの生成手法

- 各生成モデルに対して、差分プライバシーを適用した手法が提案されている。
- 深層学習モデルの学習時にDP-SGDを用いる手法などがある。

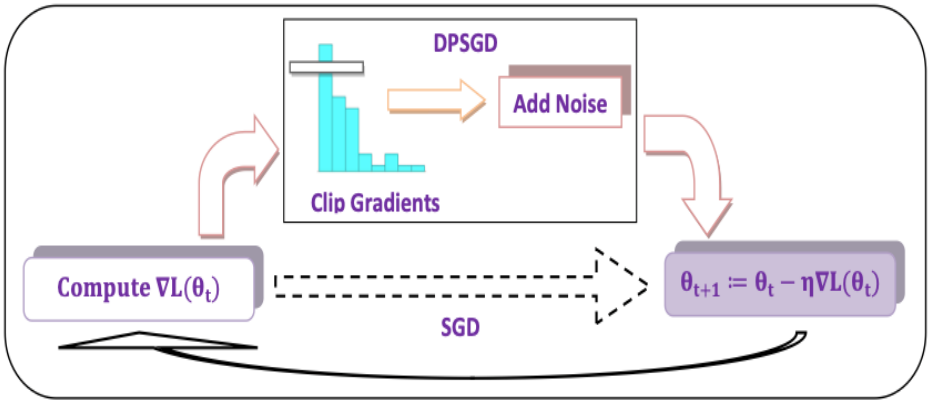
差分プライバシー適用手法の種類

- Diffusion Modelは近年発展したモデルの為、テーブルデータに対する差分プライバシーを適用した手法は調査の限り見当たらない。(2023年4月現在)
- しかし、Diffusion Modelベースの画像生成モデルに対して差分プライバシーを適用した手法が提案されており、テーブルデータへの拡張手法も今後登場すると見込まれる。

生成モデル	手法例	差分プライバシー適用手法
① 統計値ベース	• 統計値を使用した手法	• Private-PGM • MST
② Baysian Network	• Baysian Networkを用いた手法	• Privbayes
③ Copula	• Gaussian Copula • Copula Flow	• DPCopula
④ GAN	• TGAN • CTGAN	• DPCTGAN • PATECTGAN
⑤ Diffusion Model	• TabDDPM	

DP-SGD

- ニューラルネットワークにおけるパラメータの更新方法である、SGD(Stochastic Gradient Descent:確率的勾配降下法)に対して差分プライバシーを保証。合成法則により学習全体の差分プライバシーを保証している。
- 工夫点として、SGDにおける更新量の上限に制約を設ける(クリップする)ことで大域的感度*1を抑え、差分プライバシーを保証するためのノイズを低減する。



(図出典)M. A. Rahman他, Membership inference attack against differentially private deep learning model. Transactions on Data Privacy 11, 2018 Figure 2

*1「データベース中の情報が1つ変わった際に処理の出力が最大でどれだけ変化するか」という指標

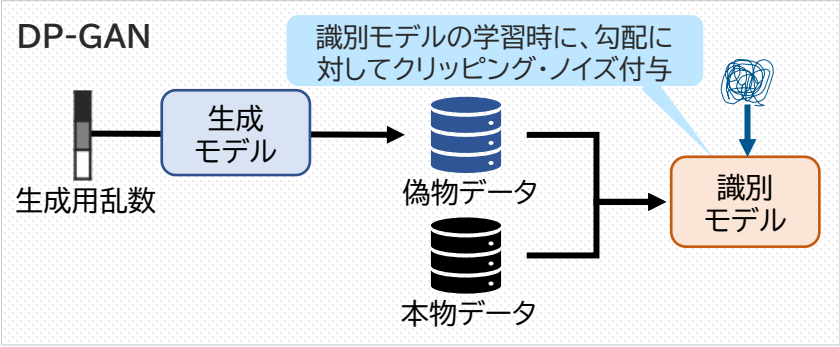
日本総研 The Japan Research Institute, Limited

（参考）代表的なプライバシー保護合成データ生成モデル

- 差分プライバシーを保証する方法は手法により異なる。

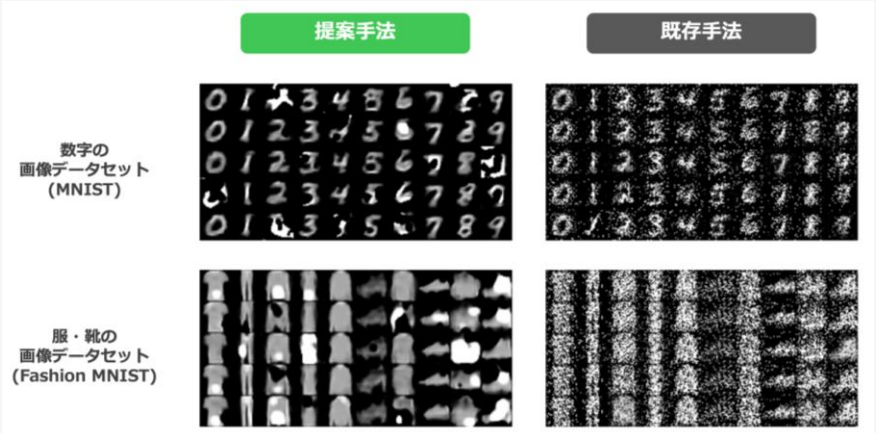
手法例：DP-CTGAN

識別モデルの学習時にDP-SGDを用いることで、GANにおいて差分プライバシーを保証するDP-GANをCTGANに適用した手法。



Mei Ling Fang et.al., DP-CTGAN: Differentially Private Medical Data Generation using CTGANs, https://ml-research.github.io/papers/fang2022dpctgan.pdf, Fig. 1を基に日本総研が作成

LINE社は差分プライバシーを保証した高品質なデータを生成する手法を提案、世界トップレベルの国際会議「ICLR 2022」にて論文が採択されている。



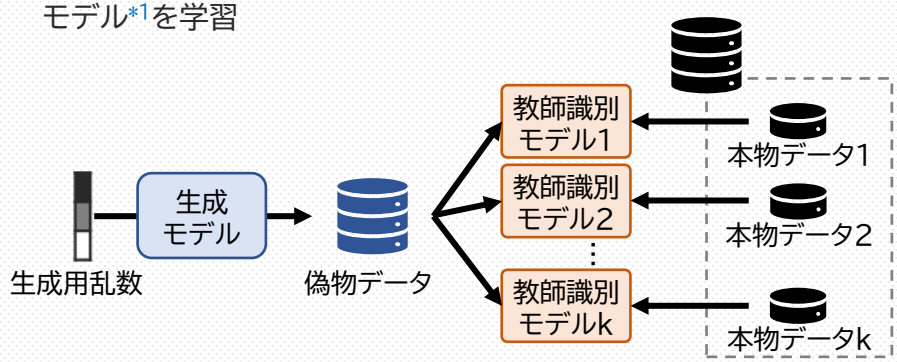
「LINE、深層学習における世界トップレベルの国際学会「ICLR 2022」にて論文採択 | ニュース | LINE株式会社」, https://linecorp.com/ja/pr/news/ja/2022/4112（アクセス日 2023/04/28）

手法例：PATE-CTGAN

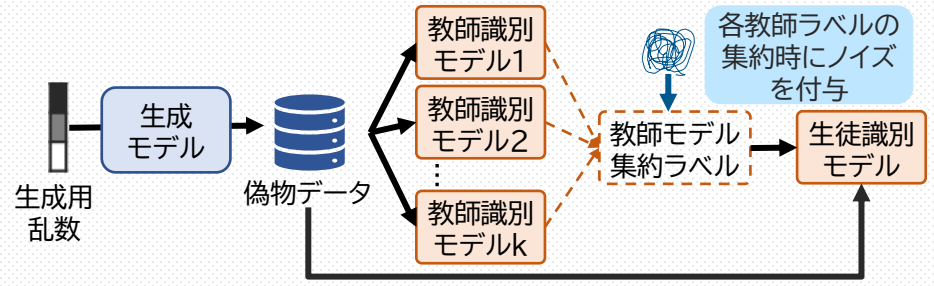
差分プライバシーを保証するGANの学習手法の一つである、PATE-GANをCTGANに適用した手法。

PATE-GAN

1. 分割した各本物データと、生成した偽物データを用いて、各教師識別モデル*1を学習



2. 生成した偽物データと、全教師モデルの出力を集約してノイズを加えた値を用いて、生徒識別モデル*2と生成モデルを学習



James Jordon,et.al., PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees, ICLR 2019, 2019, Figure 3.4 を基に著者作成

*1 あるモデルを学習する際の基になるモデル
*2 教師モデルを基に学習されるモデル

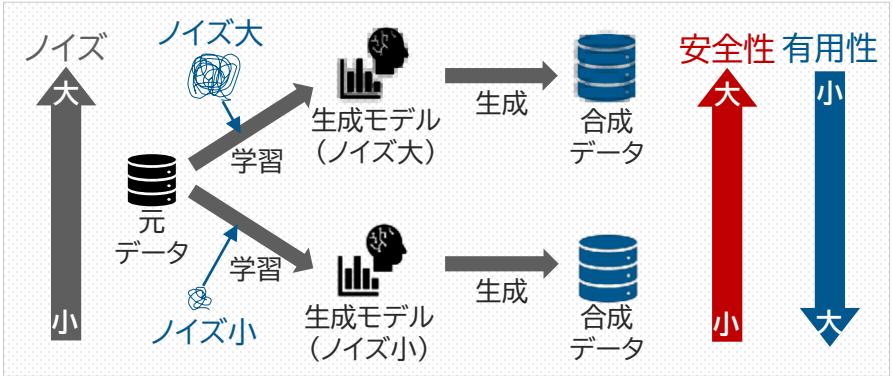
日本総研 The Japan Research Institute, Limited

2. 5 合成データの評価指標

- プライバシー保護を目的とした合成データの代表的な評価指標として、「有用性」・「安全性」が挙げられる。
- 合成データを活用する際には、目的に沿って上記指標の要件を定めることが望ましい。

有用性と安全性のトレードオフ

- 一般に大きいノイズを加えると高い安全性が実現できる。
  一方、大きいノイズを加えると有用性が下がる。
- ノイズを加える箇所や方法は様々であり、有用性と安全性を共に高める手法が研究されている。(前頁LINE社の例)



	有用性	安全性
概要	合成データの統計的な特徴が元データとどの程度類似しているか	合成データがプライバシー上のリスクをどの程度抱えているか

評価指標	• 元データと合成データの統計的な類似度
	  • 相関行列
	  • クロス集計表
	• 合成データで学習した機械学習モデルの性能
		• 合成データに対する各種攻撃の性能
		• 各合成データサンプルと元データの類似度
		  • NNDR
		  • DCR

評価イメージ	相関行列の類似度 元データ → 集計 → 元データ相関行列 合成データ → 集計 → 合成データ相関行列 差分計算 → 相関行列の差 相関行列の差が小さいほど元データと類似している 【相関行列】 相関係数を元に複数の変数間の関係を表した行列 身長 体重 BMI 身長 1.0 0.7 0.4 体重 0.7 1.0 0.6 BMI 0.4 0.6 1.0	NNDRによる評価 元データ 合成データ → NNDRの計算 データID / NNDR Syn_1 0.20 Syn_2 0.90 集計 NNDR集計値 最小値 0.17 平均値 0.85 NNDRが大きいほど元データが推測されにくい 【NNDR】 ある点に対する、最近傍点と次に近い点との距離の比 ●:元データ ●:合成データ --:距離 2番目に近い点 最近傍点 3 2 4 6 $NNDR = 2/3 = 0.67$

※評価指標には「合成データが元のデータに含まれる多様な性質をどの程度カバーしているか」を評価する「多様性」といった評価指標も存在する

日本総研
The Japan Research Institute, Limited

3.1 合成データのユースケース

- 合成データのユースケースは以下の通り。
- 法規制によるデータ蓄積･活用の制約に対する解決策としての活用が期待される。

合成データのユースケース

- ①～⑤がプライバシー保護を主な目的とした合成データのユースケース

	項目	内容
①	組織間･組織内でのデータ共有	セキュリティやコンプライアンスのリスクから組織外･組織内のデータ共有が困難になっている。合成データによってデータの共有が容易になる可能性がある。
②	クラウドデータ連携	クラウドインフラの利用にもセキュリティやコンプライアンスのリスクが伴う。機密データの合成データを活用することでクラウドへの移行が容易になる可能性がある。
③	データ販売による収益化	合成データを第三者に販売することで、収益化が期待できる。
④	データ保持	社内のデータ保持ポリシー等の規則によって、企業が個人のデータを保存できる期間も制限される。合成データによって規則を遵守したデータ保持が期待できる。
⑤	外部のデータ分析者の活用	生データの代わりに合成データを活用することで、組織外にあるデータの分析リソースを活用することができる。
⑥	データ拡張による機械学習モデルの性能向上	金融の不正検知や製造業の不良検知など稀な事象を予測する機械学習モデルは、データサイズが小さいと精度が低くなる。合成データを用いてデータサイズを増やすことでモデルの性能を向上させる。
⑦	シミュレーションによるデータ生成	データの収集はコストがかかる。シミュレーションによりデータを生成することでデータ収集のコスト･時間を削減する。
⑧	テストデータ	ソフトウェアテストや品質保証用のテストデータに合成データを用いることで、テスト時間の短縮による開発の高速化や開発時の柔軟性の向上に繋がる。

（参考）
「Top 20 Synthetic Data Use Cases & Applications in 2023」, https://research.aimultiple.com/synthetic-data-use-cases/
「Synthetic data use cases - MOSTLY AI」, https://mostly.ai/all-synthetic-data-use-cases
「Synthetic data for privacy compliance - Statice」, https://www.statice.ai/ （アクセス日 2023/04/28）

日本総研
The Japan Research Institute, Limited

3. 2 合成データの活用事例

- 合成データの活用事例が複数登場しており、実用化の段階に移りつつある。
- センシティブな情報を保有する金融・医療領域に多く活用事例が存在する。

領域	ユースケース	事例	企業・組織
金融	組織間・組織内でのデータ共有	• 銀行クライアントのアプリケーションを構築する際に必要な顧客データとして合成データを使用	Accenture
		• 金融サービスにおけるAIの研究開発を促進するため、リアルな合成データセットを生成するための研究とアルゴリズム開発を実施	J.P.モルガン
		• データのプライバシーリスクの評価にかかる時間を最大で3カ月短縮 • 合成データを使用して機械学習モデルを学習させ、有用性の評価で元データの97%を達成	Provinzial
	組織間・組織内でのデータ共有/テストデータ	• 連携先のFintech企業のツールテストに合成データを使用 • 実データでは連携までに最大6カ月かかる承認プロセスを3日に短縮	NationWide
	外部のデータ分析者の活用	• 人工的に作成した金融データを使用して外部コンペを開催	金融データ活用推進協会（FDUA）
通信	データ販売による収益化	• 人流に関する合成データを生成し、分析プラットフォームを第三者向けに提供	GEOTRA
医療	外部のデータ分析者の活用	• マウスのゲノム分析に合成データセットを使用した研究事例 • 遺伝子研究における合成データセットの有用性について主張	Gretel.ai
		• COVID-19に関する分析に合成データを使用した研究事例 • 実データの分析結果とほぼ同じ結果が得られた	MDClone
		• ヘルスケアデータの合成データ交換プラットフォームを公開 • 新たな知見の発見・製品の高度化が目的	Humana
その他	組織間・組織内でのデータ共有	• 退役軍人へのサービス提供に向けた分析に合成データを使用 • 品質管理やIRB（倫理委員会）のレビュー無しにデータ提供が可能に	米退役軍人保健局（VHA）

（参考）青字は次頁以降に詳細解説

日本総研 The Japan Research Institute, Limited

3. 3. 1 個別事例：組織間・組織内でのデータ共有

- 組織間・組織内におけるデータ共有に、合成データを用いる事例が存在する。
- 実データの代わりに合成データを用いることで承認プロセスを簡易化し、承認に必要な時間を短縮することが可能。

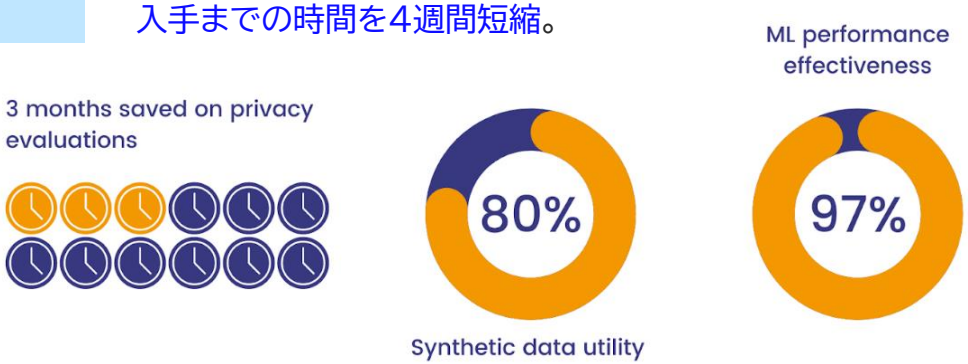
組織内でのデータ共有：Provinzial社の事例

課題

- プロジェクト開始時に全てのデータ分析のユースケースを予見できないため、利用目的が制限される。
- データの利活用方法の検討や漏洩リスクの評価には、数週間から数カ月かかることもある。

成果

- データのプライバシーリスクの評価にかかる時間を最大で3カ月短縮。
- 合成データを使用して機械学習モデルを学習させ、有用性の評価で元データの97%を達成。
- 社内のデータ共有ワークフローを調整することなく、データ入手までの時間を4週間短縮。



（参考・図出典）「Case study: Provinzial runs predictive analytics on synthetic insurance data, achieves 97% model performance efficacy. | Statice」, https://www.statice.ai/case-study/provinzial-predictive-analytics-synthetic-insurance-data（アクセス日 2023/04/28）

組織間でのデータ共有：NationWide社の事例

背景・課題

- Fintech企業が提供しているサービスとの連携を検討している。
- 連携先のツールテストにあたりデータが必要だが、データのアクセス許可が課題。
  - 第三者に渡すためのデータの準備や承認プロセスを通すために最大6ヶ月かかることもある。
  - 大幅に編集・マスキングされたデータしか連携されないため、可能なテストが制限される。

手法・成果

- トランザクションデータに含まれる統計情報と複雑な関係を学習した、元の情報を含まない合成データセットを作成。
- 第三者へのデータ準備や承認プロセスに必要な時間を最大6カ月から3日に短縮。

「Nationwide unlocks rapid innovation with synthetic data - Hazy」, https://hazy.com/resources/nationwide-rapid-innovation（アクセス日 2023/04/28）

日本総研
The Japan Research Institute, Limited

3. 3. 2 個別事例：国内における事例

- 国内においても合成データを活用する事例が存在する。
- 人流に関するデータを合成することで、従来のメッシュデータでは分析できないシミュレーションを可能にする事例がある。

データ販売の機会創出：GEOTRA社の事例

課題

- 一般的な人流データはメッシュ化・集計化がされているため、分析の用途が限られる。

手法・成果

- 人流データを合成し、合成データを用いた分析プラットフォームを提供
- GEOTRAは合成データ技術等の技術を用いることで、集計されていないトリップデータを提供可能。

ID	性別	年代	出発時間	到着時刻	移動目的	移動手段	始点(経度)	始点(緯度)	・・・
034	男性	30代	7:12	8:00	通勤	車	139.11	36.44	・・・
111	女性	40代	7:14	8:58	通勤	鉄道	139.11	36.44	・・・
006	女性	60代	7:31	7:54	買い物	徒歩	139.11	36.44	・・・
239	男性	20代	7:33	8:33	通学	鉄道	139.11	36.44	・・・
099	男性	50代	8:00	8:45	出勤	鉄道	139.11	36.44	・・・
542	女性	20代	8:10	8:30	食事	徒歩	139.11	36.44	・・・
090	男性	30代	8:16	8:40	通院	車	139.11	36.44	・・・
034	男性	30代	8:00	8:25	食事	徒歩	139.29	34.32	・・・
・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・

非集計トリップデータのイメージ

（図出典）「株式会社GEOTRA（ジオトラ）」, https://www.geotra.jp/service （アクセス日 2023/04/28）

外部のデータ分析者の活用：金融データ活用推進協会の事例

事例詳細

- MUFG、みずほFG、三井住友信託銀行、SBIホールディングス、SBI新生銀行が共催
- 延べ1,658名が参加
  - コンペティションサイト「SIGNATE」金融分野のコンペにおける歴代1位の参加者数
- コンペティション参加者である外部のデータ分析者から、データに関する新しい知見を獲得可能。

「第1回 金融データ活用チャレンジ | SIGNATE - Data Science Competition」, https://signate.jp/competitions/841（アクセス日 2023/04/28）

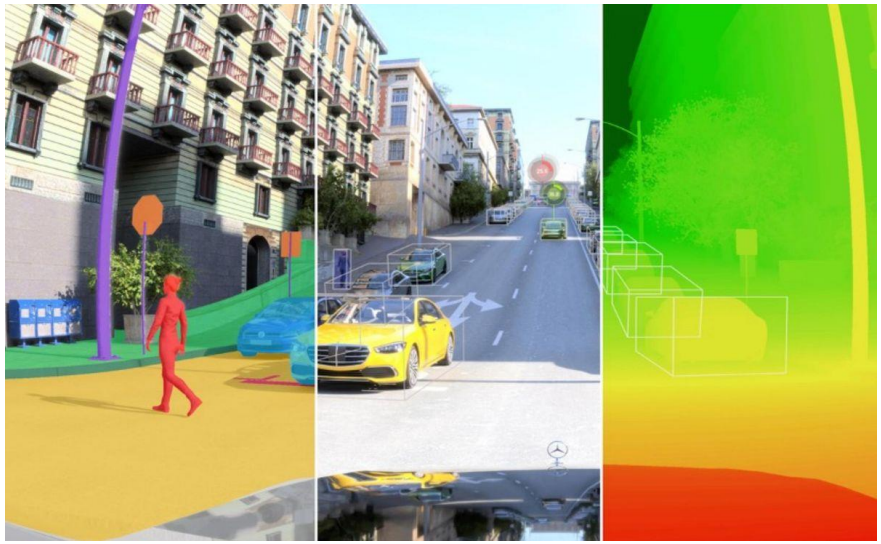
日本総研 The Japan Research Institute, Limited

（参考）データ拡張を目的とする合成データの活用事例

- データ拡張による機械学習モデルの性能向上のため、合成データを活用する事例も存在する。
- 実環境では発生しえない希なデータを人工的に作り出し、訓練データとして活用することで機械学習モデルの性能向上させる取り組みが進展（特に画像認識）。

画像認識：NVIDIA社の事例

- 画像認識（画像分類・物体検知等）のAIモデルの開発に合成データを活用する事例が増えている。
- 自動運転で活用されるAIモデルの開発に使用される事例がある（下図）。実際に取得したデータでは再現できない環境（例：雪が降っている）を再現して、モデルの学習データとして活用している。



（図出典）https://spectrum.ieee.org/synthetic-data-ai

金融分野：アメリカンエキスプレス社の事例

- アメリカン・エキスプレスは合成データを使用し、不正検知モデルの性能を向上させている。
- 実環境で取得しにくい詐欺パターンについて、データを合成的に作成して学習データに加えている。

「Fake It to Make It: Companies Beef Up AI Models With Synthetic Data」, https://www.wsj.com/articles/fake-it-to-make-it-companies-beef-up-ai-models-with-synthetic-data-11627032601（アクセス日 2023/04/28）

通信分野：Amazon社の事例

- Alexaの新しい言語の対応に合成データを使用。
- 新しい言語における幾つかの標準的な発話をテンプレートとし、テンプレートを組み合わせることで新しいデータを生成。
- ユーザーとのやり取り無しにNLU（Natural Language Understanding, 自然言語理解）システムを学習可能。

「Tools for generating synthetic data helped bootstrap Alexa's new-language releases - Amazon Science」, https://www.amazon.science/blog/tools-for-generating-synthetic-data-helped-bootstrap-alexas-new-language-releases（アクセス日 2023/04/28）

日本総研
The Japan Research Institute, Limited

3. 4 合成データを提供するベンダー

- 合成データ技術を提供しているベンダーは複数存在している。
- 海外ベンダーの多くは公式HPに法規制の対応について記載している。

主な合成データベンダー

分類	開発組織（製品名）	所在地	生成手法	法規制対応	安全性指標
国内	NTTテクノクロス（tasokarena）	日本	統計値	非公開	照合可能性, k-匿名性
海外	Gretel	アメリカ	Timeseries DGAN	GDPR*1, HIPAA*2	差分プライバシー, 過剰適合防止, 類似性フィルター, 外れ値フィルター
	MDClone	イスラエル	カーネル密度推定	HIPAA	差分プライバシー
	MOSTLY AI	オーストリア	ニューラルネットワーク（詳細不明）	GDPR, HIPAA, CCPA*3	完全一致率, 最近レコード距離, 最近傍距離量
	Octpize	フランス	k近傍法	GDPR	識別, 連結, 属性推定
	Statice	ドイツ	GAN+VAE	GDPR, HIPAA	差分プライバシー, 識別, 連結, 属性推定
	Ydata	アメリカ	GAN	GDPR, CCPA, HIPAA, PIPEDA*4	差分プライバシー

以下を基に日本総研が作成
- 「[New] List of synthetic data vendors— 2022 | by Elise Devaux | Medium」, https://elise-deux.medium.com/new-list-of-synthetic-data-vendors-2022-f06dbe91784, (アクセス日 2023/04/28)
- 月間誌『統計』2022年8月号 特集:「プライバシー保護技術の新展開」

*1 GDPR：General Data Protection Regulation、日本語訳「EU一般データ保護規則」
*2 HIPPA：Health Insurance Portability and Accountability Act、日本語訳「米国医療保険の携行性と責任に関する法律」
*3 CPAA：California Consumer Privacy Act、日本語訳「カリフォルニア州消費者プライバシー法」
*4 PIPEDA：Personal Information Protection and Electronic Documents Act「カナダ個人情報保護および電子文書法」

日本総研 The Japan Research Institute, Limited

4.1 データ流通における合成データ

- 国内のパーソナルデータの活用方法として、統計情報・匿名加工情報[*1]・仮名加工情報[*2]等が挙げられる。
  - データ流通における障壁（第三者提供のための同意取得、利用目的の制限等）の緩和が可能。
  - 一方、粒度が荒くなるため有用性が喪失する等の課題もある。
- 合成データは有用性を維持したまま、幅広い分析用途に活用できるデータとして期待されている。

データ流通における各情報の比較

	合成データ	統計情報	匿名加工情報	仮名加工情報
定義	実在するデータと同じ構造で異なる値を持つデータ	複数人の情報から共通要素に係る項目を抽出し、同じ分類ごとに集計等して得られる情報	特定の個人を識別することができないように個人情報を加工して得られる個人に関する情報	他の情報と照らし合わさない限り特定の個人を識別できないように加工（削除）した個人に関する情報
個人情報かどうか	定まっていない	個人情報に該当しない	個人情報に該当しない	個人情報に該当する（加工元が個人情報の場合）
第三者提供	合成データが個人情報の場合同意が必要となる	同意無しに可能	同意無しに可能（公表が必要）	委託、事業継承、共同利用の場合、法令に基づく場合を除き不可
粒度	○ 元のデータの粒度を保持	× 個人に対する情報は保持しない	△ 一般化等の加工により粗くなる	○ 元のデータの粒度を保持
真実性[*3]	× 実データとは異なる値を持つデータが生成される	△ 粒度は粗いものの、実データを元に作成	△ 加工により実データとは一部異なる	○ 削除情報以外は実データのまま

*1 匿名加工情報：2017年施行
*2 仮名加工情報：2022年施行
*3 真実性：実データの情報を保持しているかどうか

日本総研 The Japan Research Institute, Limited

4. 2 合成データの活用に向けた課題

- 合成データの活用にあたり、満たすべき安全性の評価基準については国内外ともに定められていない。
- 事例が増えることで、合成データに対する法的な見解や活用時のガイドラインが公表されることが期待される。

個人情報保護法における合成データの扱い

- 「個人情報を元に生成された合成データが個人情報に該当するかどうか」についての見解は様々である。
  - 生成モデルが統計情報として見なせる場合、モデルから生成されたデータは個人情報ではないという意見もある。
  - また、機械学習モデルをベースにした生成モデルについても、学習済みのパラメータが個人と対応しない場合は個人情報に該当しない。*1
  - 一方、生成モデルの学習方法によっては、モデルから生成されたデータから元の個人が容易に識別される恐れがある。
- 非個人情報として扱うために満たすべき安全性の評価指標や基準についても定められていない。（2023年4月現在）

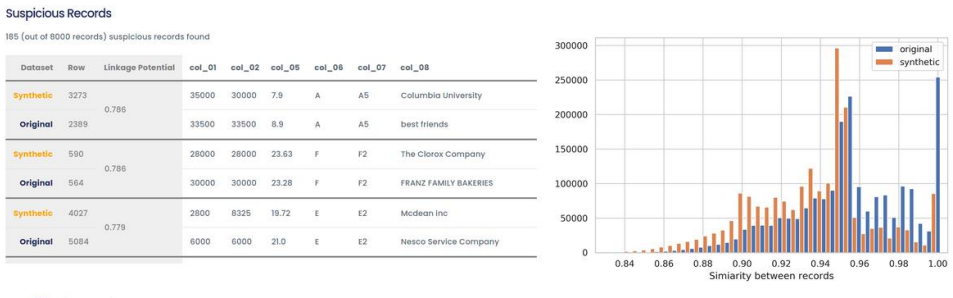
*1「複数人の個人情報を機械学習の学習用データセットとして用いて生成した学習済みパラメータは、個人情報に当たりますか。｜個人情報保護委員会」, https://www.ppc.go.jp/all_faq_index/faq1-q1-8/（アクセス日 2023/04/28）

CNIL（フランス共和国データ保護機関）がGDPRに準拠していると指定したデータについて、再構築攻撃に脆弱であるという報告をした例がある。

「Microsoft Word - Differential Privacy Amicus Brief - FINAL(3205841.1)」, https://www.brennancenter.org/sites/default/files/2021-04/Amicus%20Brief_dataprivacyexperts_%202021-04-23.pdf（アクセス日 2023/04/28）

GDPRに準拠した合成データ:Statice社の事例

1. 個人に対応していない統計的な分布を学習し、そこから合成データを生成しており、元データと合成データは1対1対応しない。
2. 元データの複雑なパターンをモデルが学習する（元データと同じデータを生成する）ことの対策として、差分プライバシー等の技術を適用する。
3. 1,2によって生成したデータも完璧なプライバシー保護は保証していない（ある程度の実用性を保ちながら完璧なプライバシー保護は保証できない）。GDPRでは、再識別の残存リスクを評価することが求められる。現状、この評価方法は各組織に委ねられている。



元データと合成データの類似度によるリスク評価

「How to manage re-identification risks with synthetic data - Statice」, https://www.statice.ai/post/how-manage-reidentification-risks-personal-data-synthetic-data（アクセス日 2023/04/28）

日本総研
The Japan Research Institute, Limited

5.1 現状と課題・今後の展望

- 技術面では高度なモデルを用いた生成手法や、有用性・安全性を高める手法が研究されている。これらの研究は引き続き進展すると見込まれる。
- 合成データを活用する事例は増加傾向にあり、実用化の段階に移りつつある。一方、法律面の観点等から社会に広く普及するには時間がかかる。
- 医療・行政に関する社会の解決といった、個人情報の活用が受容されやすいと想定されるユースケースから活用の進展が見込まれる。

	現状・課題	今後の展望
技術	• Diffusion Modelや大規模言語モデル等、他領域のアーキテクチャを用いたテーブルデータ合成手法が提案されている。	• 合成データの生成手法についても、深層学習の発展と共に高度化していくと見込まれる。
	• 安全性を高めるための処置（例:ノイズの付加による差分プライバシーの適用）によって合成データの有用性が下がる。	• 有用性と安全性を共に高く保つ合成データの生成手法の研究が引き続き進む。
活用・法規制	• 国内外において実証実験や実務への適用に取り組む組織も登場している。 • しかし、合成データの活用に向けた法整備は進んでおらず、社会全体への普及には時間がかかる。	• 個人情報の活用に当たっては、個人への情報開示や同意の取得など、慎重な取り扱いが必須である。その上で、個人情報の活用が受容されやすいと想定される、医療や行政に関する社会課題の解決といったユースケースから活用が進展すると考える。 • 社会実装が増え、認知度が高まることで、合成データ活用におけるガイドラインが策定される等法整備が進む。

日本総研
The Japan Research Institute, Limited

5. 2 合成データの活用に向けた推奨事項

- 合成データの活用に向けた推奨事項は以下の通り。

1. 技術やツールの調査・評価	• 合成データを活用するため、技術やツールの調査・技術評価を実施する。 • 安全・高品質な合成データの生成のために、幅広い分野の動向を注視し、合成データ生成手法への活用可能性を検討することが重要。
2. 法律面等の複合的な観点による評価	• 合成データの活用を検討する際には、技術的な観点だけでなく、コンプライアンスや法律といった複合的な観点での評価を実施する。 • 必要に応じて各分野の外部専門家の意見を取り入れながら検討する。
3. 透明性の保持等による社会受容性の確保	• 合成データを含む新技術を用いた分析を実施する際には、社会受容性を高めることも重要である。 • 具体的な方策として、データガバナンスの整備・プライバシー影響調査（PIA）*1を実施し、データの取り扱いや分析によって生じるリスクについて透明性を保つことが挙げられる。

*1「プライバシー影響調査（PIA）とは、個人情報及びプライバシーに係るリスク分析、評価、対応検討を行う手法」のこと。経済産業省「 DX 時代における 企業のプライバシーガバナンスガイドブック ver1.2 」