筑波大学
University of Tsukuba

NTT DATA
Trusted Global Innovator

セキュアなデータ流通と活用のための技術評価

本報告書は、国立大学法人筑波大学と株式会社NTTデータの共同研究
「セキュアなデータ流通・活用のための技術評価」（2021年10月～2022年3月）
の成果をまとめたものです。

株式会社NTTデータ
〒135-6033 東京都江東区豊洲3-3-3 豊洲センタービル
Tel: 03-5546-8051 Fax: 03-5546-2405
https://www.nttdata.com/jp/ja/

目次 contents

1.はじめに

機械学習技術の発展やデジタル化の普及に伴い、データが持つ価値は一層高まっています。さらに今後は企業や業界を超えたデータ活用が増えることが予想されます。一方、そのような垣根を超えたデータ活用ではデータ漏洩のリスクが増大します。また、個人に紐づくデータの場合は、プライバシー保護の観点からより慎重な取り扱いが求められます。そこで本レポートでは、機微なデータを保護しつつ、その価値を最大限発揮するために、企業や業界の壁を超えてデータを統合・分析するための技術について解説します。これにより、技術担当者がデータ保護技術の特徴を理解し、選択できることを目的としています。

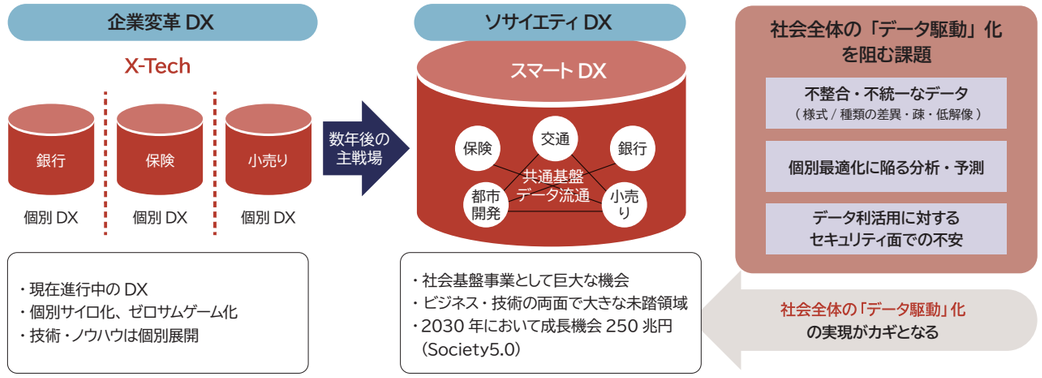
本報告書は、国立大学法人筑波大学と株式会社NTTデータの共同研究「セキュアなデータ流通・活用のための技術評価」(2021年10月～2022年3月)の成果をまとめたものです。

2.データ保護技術について

■データ保護技術が必要となる背景

データ活用を含むDX(デジタルトランスフォーメーション)において、個別企業・個別業界におけるDXから社会広域・複数業界にまたがるDX(ソサイエティDX)へ適用範囲の拡大が見込まれます。その中で個人のデータを含む機微なデータの利活用に対するセキュリティ面での対応が必要となります(図1)。

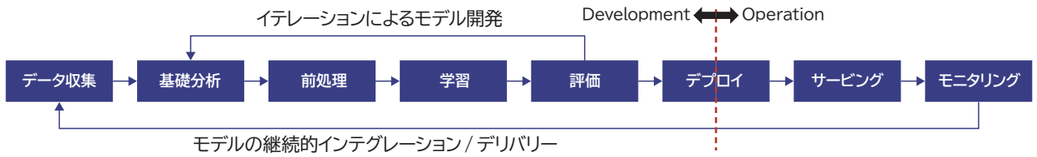
図1 クロスドメインでのデータ活用の課題



また、データ流通およびデータ分析のプロセス(図2)において、データ漏洩や不正利用への対策は必須です。しかし、このプロセスにおいて、データの収集・保管時は暗号化処理をしていても、分析のために生データ(Raw data)になったタイミングを狙われて、データが漏洩してしまうケースがあります。取り扱う生データが個人情報や企業秘密情報の場合、不正利用につながる恐れもあります。

特に、業界や企業を跨ぐクロスドメインのユースケースでは、他組織に生データを流通・共有するという条件は、データ活用の阻害要因となっています。(図2)

図2 データ分析プロセスの例



また、別の背景として、世界的に広がっているプライバシー保護規制が挙げられます。EUにおけるGDPR(一般データ保護規則)や日本における個人情報保護法など、各国の法制度に適切に対応しつつ、データを有効に活用するための技術が期待されています。(表1)

表1 主な個人情報保護関連の規制

(2022 年 4 月現在)

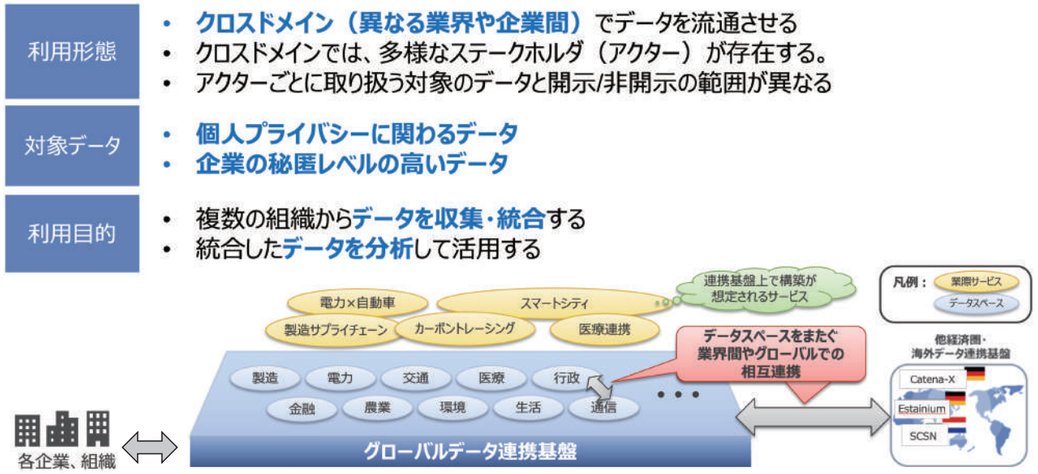
EU	GDPR(一般データ保護規則)	2018 年適用開始
米国	カリフォルニア州消費者プライバシー法 カリフォルニア州プライバシー権利法 バージニア州消費者データ保護法	2020 年施行 2023 年施行予定 2021 年成立
日本	個人情報保護法	2020 年改正・2022 年全面施行
中国	サイバーセキュリティ法 個人情報保護法	2017 年施行 2021 年施行

※他にも世界各国でプライバシーに関する法規制の動向あり　参考:個人情報保護委員会「個人情報保護法 令和 2 年改正及び令和 3 年改正案について」 https://www.meti.go.jp/shingikai/sankoshin/shomu_ryutsu/bio/kojin_iden/life_science/pdf/001_03_02.pdf

■クロスドメインでのデータ連携・活用システム

クロスドメイン環境で秘匿性が高いデータを統合し、そのデータを分析するためのシステム(図3)を構築する場合、「何のデータに対し、誰からどの範囲で守るべきか」を明確にする必要があります。

図3 クロスドメインでのデータ流通・活用システム



■データセットの統合

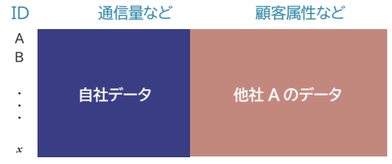
クロスドメインでのデータセットの統合例を示します(図4)。統合の方法は大きく分けて次の2種類あります。

・同じIDをキーとして、水平方向に列(特徴)を増やす(横結合/水平統合)

・同じ列（特徴）で、垂直方向にデータ量を増やす(縦結合/垂直統合)

図4 クロスドメインでのデータセット統合例

① 同一 ID の「特徴」を増やしたい（横結合 / 水平統合）



・自社の顧客情報と、代理店が保有する顧客属性情報を結合
顧客ごとの解約の予兆検知やアップセル確度を予測し、
解約阻止・販促キャンペーンに活用する（通信量 × 顧客属性 etc.）

・自社の法人顧客の取引情報と、当該法人の他社のサービス利用状況を結合
当該法人の取引リスクを予測する（通話 × 電力 etc.）

② 別 ID の「量」を増やしたい（縦結合 / 垂直統合）

・自社の顧客情報と、マーケット層が異なる他の事業者の顧客情報を結合
未知の顧客に対する何らかの予測を高精度に行いたい

■流通対象データの分類

(1)オリジナルデータ(生データ)

暗号化や加工をされていない、もともと持つ情報をそのまま保持したデータをオリジナルデータと規定します。例えば、機微な情報を含むなど外部との共有を前提としないデータをオリジナルデータとするケースがあります。本レポートでは、一般的な”生データ”の用語も同じ意味で使います。

オリジナルデータは、情報の正確性や追跡性の面で後述の他のデータ分類よりもメリットがあります。

機微なデータでも生データのまま分析するケースも数多くあります。ただし、個人情報を扱う場合は、個人情報保護法に則って適切な管理・運用が求められます。

(2)暗号データ

オリジナルデータを暗号化することで、第三者による情報閲覧を防ぎます。

ただし、暗号キー管理や復号後の生データの管理など、データフローに沿って管理と運用を設計する必要があります。さらに、暗号化によるシステム面・性能面でのコストの考慮も必要です。

近年では、データを暗号化したまま計算処理ができる技術や、一部のデータ項目だけを暗号化できる技術も広まりつつあります。

(3)統計データ

個人のデータを統計情報としてまとめ、個々の識別ができない形で可視化し分析するケースがあります。

統計解析は、個々のデータを保護する目的もありますが、統計により有効な結果を導き出すことを主目的にしているケースがほとんどです。ビジネスの場でのデータの統計化や可視化の実績は数多くあります。

統計データを分析のデータソースとして見ると、個々のデータは秘匿化されている反面、情報粒度が粗いことや他データとの結合が難しい場合があることには注意が必要です。

(4)匿名加工データ

匿名化の一般的な定義は、個人の情報を加工して、個人の識別や情報の復元をできないようにすることです。

個人情報保護法の改正により、匿名加工のルールや情報の管理などについての規定が追加されました。ここでは、ガイドラインに沿って匿名加工することで、本人の同意なく第三者提供できることが定められています。(*1)

デメリットとしては、匿名化により情報粒度が落ちることで、分析の精度が望めない場合があります。個人特定のリスクを抑え、かつオリジナルデータの特徴を多く有するような、バランスを保った加工が求められます。

(5)学習モデル

発想を転換して、上記のようなデータではなく、機械学習のモデルやそのパラメータを流通させる分散学習の手法が登場してきています。

データ自体を外部に出す必要がないため、データ流通の観点での漏洩リスクはありません。(ただし、学習モデルからオリジナルデータを推測できないことが前提です。)

表2 流通対象データの分類

データの分類	メリット	考慮点	対応テクノロジーの例
オリジナルデータ（生データ）	情報の完全性、追跡性	情報漏洩リスクへの対策 個人情報の適切な管理	（可視化、統計、機械学習など多数の利用実績あり）
暗号データ	暗号化による秘匿性強化	暗号化による処理コスト増 暗号キーの運用管理	秘密計算
統計データ	統計化による秘匿性強化 多数のビジネス実績	データソースとしての情報粒度 他データとの結合	（統計、可視化ツールなど多数の実績あり）
匿名加工データ	匿名化による秘匿性強化 法的に認められた第三者提供可能な加工 多数のビジネス実績	匿名化による情報損失 結合キーの欠落 匿名性と有用性のバランス	匿名加工技術
学習モデル	モデルパラメータのみ流通 データ自体は非開示	特定のモデリング手法のみ対応	データコラボレーション /Federated Learning

■ロールの分類

「データを誰に開示し、誰に対して秘匿化するか」を明確にするため、 ロールを規定して取扱うデータ対象と開示範囲を定める必要があります。

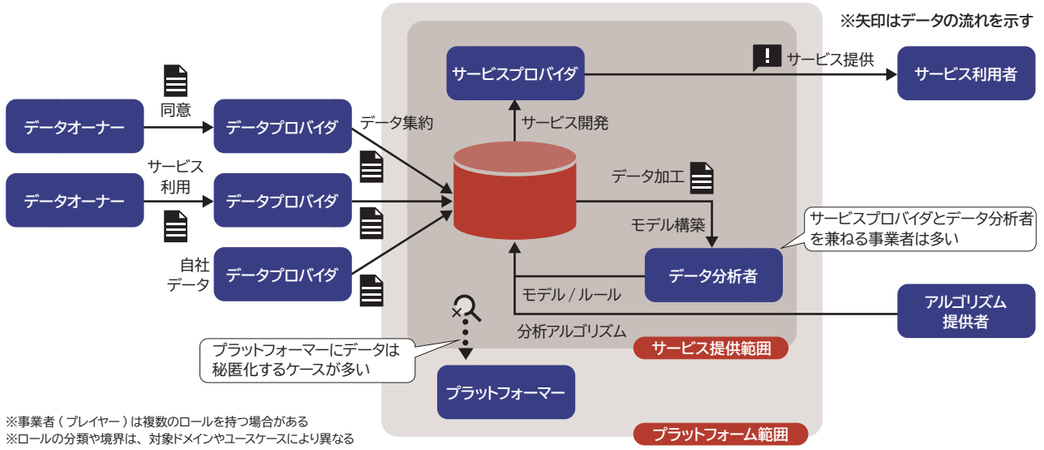
ロールの分類例(表3)と関係図(図5)を示します。

表3 ロールの分類例

ロール	説明	保有データ例
データオーナー	データ発生元、情報の所有主体（原則として発生元がオーナー）	個人発出のパーソナルデータ/ IoT、防犯カメラなどのデータ
データプロバイダ	データオーナーの同意を得てデータを管理	企業の顧客データ 医療機関のレセプトデータ
サービスプロバイダ	データ管理・運用にかかわる事業運営者	情報銀行や PDS に預けられたデータ 認定匿名加工医療情報作成事業者のデータ
プラットフォーマー	クラウドなどの基盤運営事業者	クラウド内のデータ
データ分析者	データ活用（分析）をする企業 データサイエンティスト	分析のためのインプットデータ モデル/ルール
アルゴリズム提供者	サービス提供者やデータ分析者が利用するアルゴリズムを所有・提供する事業者	データ加工・分析アルゴリズム
サービス利用者	データ活用の結果をサービスとして享受する企業・個人	

(*1)個人情報保護委員会 匿名加工情報制度について　https://www.ppc.go.jp/personalinfo/tokumeikakouInfo/

図5 ロールの基本関係図（例）



※事業者（プレイヤー）は複数のロールを持つ場合がある
※ロールの分類や境界は、対象ドメインやユースケースにより異なる

■システム全体におけるデータ保護技術の位置付け

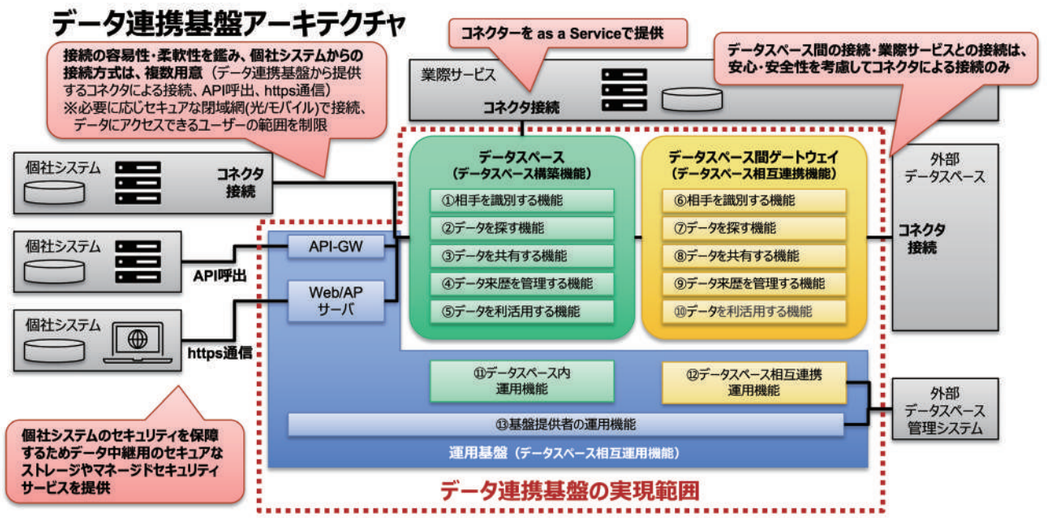
機微なデータを守るために、何かひとつの技術を導入すればよいということはなく、システム全体でセキュリティ関連技術を組み合わせて、総合的に担保することが必要です。

例として、NTTグループで策定したデータ連携基盤アーキテクチャを示します（図6）。

このデータ連携基盤では、相互運用性、セキュリティの担保、データ主権の保護、使いやすさを重視し、必要な機能の整理と基本アーキテクチャを策定しています。

データ保護技術は、「データを共有する機能」、「データを利活用する機能」の中でデータを秘匿化する一機能になりますが、システムセキュリティとしては、認証、アクセス制御、コネクタ/通信におけるセキュリティ管理、基盤運用監視など様々な機能が必要となります。

図6 データ連携基盤アーキテクチャの例



引用元：
組織や業界を横断した安全なデータ流通を実現する グローバルデータ連携基盤のアーキテクチャ構想
https://www.nttdata.com/jp/ja/-/media/nttdatajapan/files/news/information/2022/053100/053100-01.pdf

column

【コラム】データ保護技術の適用検討時の考慮事項

データ保護技術は、「データを隠した状態で統計処理や機械学習ができる」というメリットがありますが、実際に利用検討を進めると、必要性・適用条件・コストなどの壁に当たることがあります。適用検討時に考慮すべき事項を列挙します。

表4 データ保護技術の適用検討時の考慮事項

合目的性	・対象のデータは本当に秘匿する必要があるデータか。 ・目的から考えた場合、他の既存技術やその組合せで実現できるものではないか
セキュリティ	・「データ保護技術を導入するとデータとアルゴリズムを守ることができる」ではない ・認証、アクセス制御、通信暗号化など従来セキュリティと運用も必要
機能実現性	・秘匿化の前に、統計処理や機械学習適用の結果が有効なユースケースになっているか ・データを秘匿化したまま処理すると、オリジナルデータの確認・検証はできない。運用時のユースケースとしてこの点が問題にならないか
ステークホルダ	・ステークホルダが少数で特定される場合、従来の契約やNDAで対応できる範囲ではないか ・各ステークホルダにビジネスメリットはあるか。メリットがないと大切な自社データを提供してくれない
コスト	・想定効果に対してコストをかけられるか

■データ保護技術の分類

データ保護技術は、データを秘匿化した状態で「活用」することが目的です。活用のための手法は主に「秘密計算」、「匿名化」、「プライバシー保護ML」に分類されます。

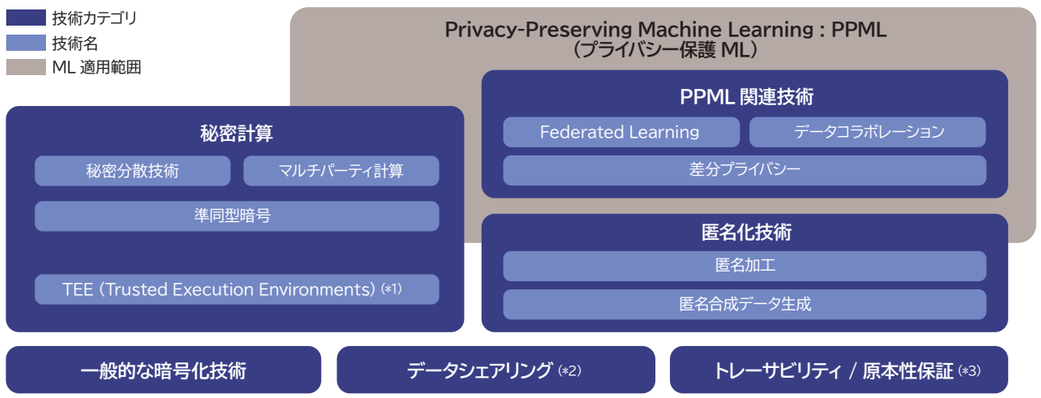
・「秘密計算」では、最初にデータを暗号化し、復号する権利を有する人以外からは中身を見られない状態に変換します。暗号化した状態のままで統計処理などの演算を実施し、計算結果も暗号化された状態で得られます。暗号化された結果についても、復号する権利を有する人にしか取り出すことができません。

・「匿名化」では、生データから個人を特定できる情報を取り除き、加工して個人の特定を困難にします。また、生データから同じ特性を持つ架空のデータを生成することによって個人の情報を保護する技術もあります。

・「プライバシー保護ML」は、個人の情報を秘匿化したままで機械学習を行うための技術です。生データは分散された端末やサーバ、データベースにあり、このデータは他の人やサーバには送られません。端末内で一定の変換を実施し、機械学習のための有用性を失わない、かつ生データは復元できないような形にしてからサーバなどへ集約され、機械学習を実施します。

対象となる具体的な技術について、カテゴリ(図7)と一覧(表5)を示します。

図7 データ保護技術カテゴリマップ



(*1) TEE は、HW の保護領域内でデータを展開するため、秘密計算に分類されない場合もあるが、ここでは広義 (外部から見て暗号化された状態で計算処理) で分類した
(*2) データの秘匿性を担保しつつ、共有を支援するための技術群 : データカタログ、ID 管理、同意管理、アクセスコントロール、共有プロトコルなど
(*3) データ流通において、追跡可能性や原本保証性を確保する技術 : デジタル署名、ブロックチェーンなど

表5 データ保護技術一覧

カテゴリ	技術	技術概要
秘密計算	マルチパーティー計算 (Multi-party Computation)	・元データを暗号化したままデータ統合と統計処理 ・マルチパーティ計算 (MPC) では、データを暗号化、分割して格納する秘密分散の技術をベースとしている
	準同型暗号 (Homomorphic encryption)	・鍵を用いて元データを暗号化したままデータ統合と統計処理 ・復号鍵を持っている人のみ、処理結果の暗号データを復号できる仕組み
	TEE (Trusted Execution Environments)	・セキュアなプログラム実行空間 (Trusted Execution Environment) を持ち、その空間内で復号化、計算処理、再暗号化を実施
匿名化技術	匿名加工	・個人識別や復元ができないように加工する技術
	匿名合成データ生成 (Synthetic Data)	・元データと統計的に類似した架空の個人データを生成する技術
Privacy-Preserving Machine Learning (プライバシー保護 ML 技術)	Federated Learning	・データではなくモデル情報を共有して学習する技術
	データコラボレーション	・データを次元削減等の変換で秘匿化し、統合して機械学習に応用する技術
	差分プライバシー (Differential privacy)	・ノイズを確率的に付与することでデータに紐づく個人特定を困難にする手法

以下の節では、この三つの技術について詳しく見ていきます。

3.個別技術紹介

■秘密計算

【秘密計算概要】

ここでの秘密計算は、(1)データ提供者が持つデータを何らかの形で暗号化し、(2)その状態のまま計算を実施、(3)計算結果も暗号化された状態で受け取る。(4)その後データ利用者が復号して計算結果を得る一連のプロセスを指します(図8)。以下では、秘密計算で主流となっている三つの技術を紹介します。一つはマルチパーティ計算と呼ばれる秘密計算、もう一つは準同型暗号を用いた秘密計算、最後はTEEと呼ばれる専用ハードウェアを用いた秘密計算です。

ソフトウェアレベルでの秘匿化については以下のような性質があります。

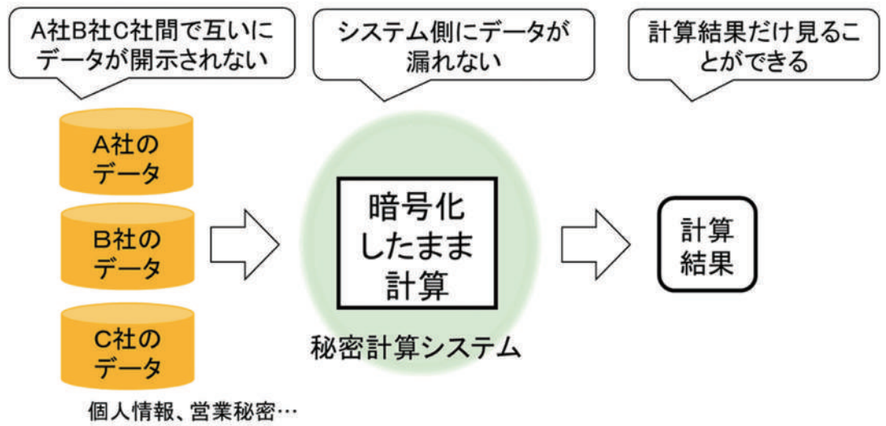
・データ提供者以外(データ分析者、他のデータ提供者など)は生データを参照できない

・解像度を落とさずにデータ共有が可能なため、生データと同一精度の解析が可能

・暗号化状態で行える分析の種類には制限がある

・暗号化状態での分析では、生データを用いた場合よりも処理時間がかかる

図8 秘密計算の原理



秘密計算の原理　https://www.rd.ntt/sil/project/sc/NTT-himitsu-keisan.pdf

【マルチパーティ計算】

マルチパーティ計算(MPC)では、

1.データを分散させる。個々の分散データだけを見ても生データは復元できない。

2.個々の分散データは別々のサーバ上で計算され分散計算結果を得る。

3.分散計算の結果を集めることで、本来の計算結果を得る。

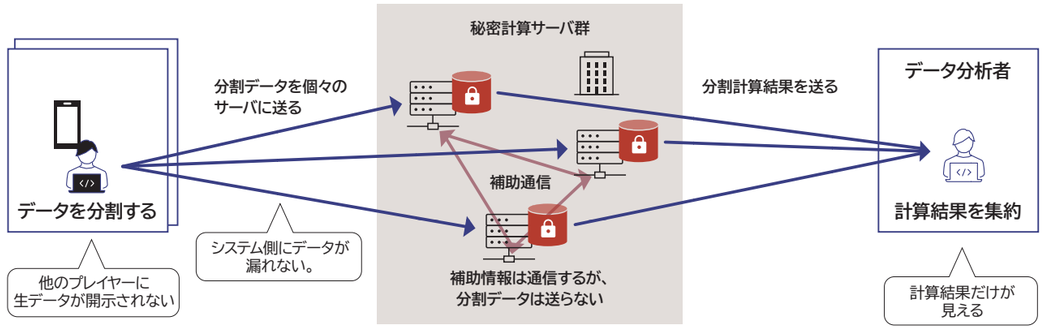
という流れで計算が行われます(図9)。ここで、

・サーバ間で補助情報を通信するが、データ自体は各サーバに分散されたまま。

・各サーバ上のデータ(分散データと補助情報)だけでは分割前のオリジナルデータ(生データ)は復元できない

という特徴があります。データの分散には秘密分散という技術が使われます。

図9 マルチパーティ計算の流れ



暗号化せずに計算を行う場合に比べると、計算方式が変わる分や計算サーバ群同士で通信を行うなどで計算時間が余分にかかりますが、現実的な時間内で計算が可能です。また、計算サーバ群とその内部の通信環境などを準備する必要がありますが、それらがすでに用意されたクラウドサービスも提供され始めています。

秘密計算クラウドサービス例:「析秘(せきひ)」 https://www.ntt.com/about-us/press-releases/news/article/2021/0819.html

Case study

【事例】作物ビッグデータを秘密計算(MPC)で活用

秘密計算を用いて複数組織のデータを秘匿化したまま集めて解析することにより、自組織のデータは外部に出さないまま、一組織だけのデータ分析では見ることのできなかった情報を共有することができます。

NTTと農研機構が秘密計算技術による作物ビッグデータ活用の共同研究を開始
～安全かつ高度な作物データ利活用の活性化した世界を目指して～

複数の組織が保有する作物データに対して、秘密計算技術を用いた解析により、
農業研究開発の効率化に有益な情報を提供

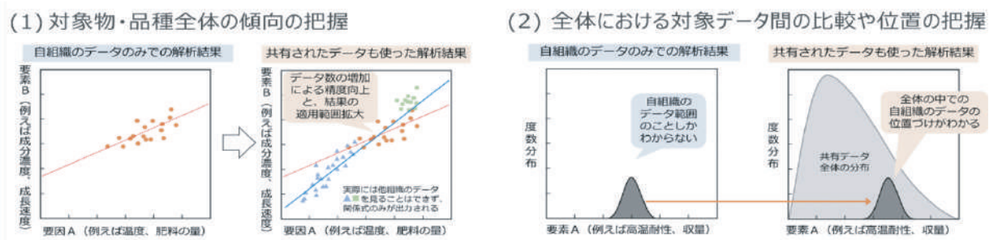


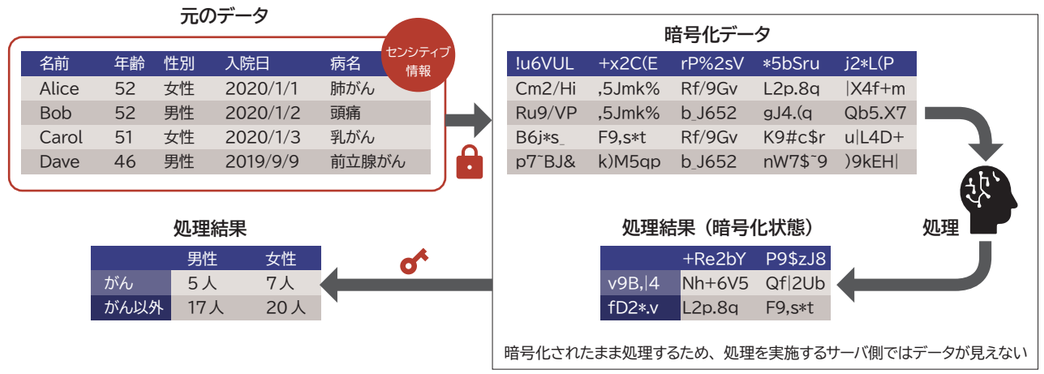
図10 作物ビッグデータでの秘密計算の活用

NTT ニュースリリース https://group.ntt/jp/newsrelease/2020/10/28/201028a.html

【準同型暗号】

準同型暗号とは公開鍵暗号の一種で、暗号化した状態のまま四則演算の一部(加算もしくは乗算)が可能なものです。元データを公開鍵で暗号化してからサーバへ送り、サーバで計算を実施、暗号化された状態の計算結果を受け取って、秘密鍵を用いて復号します(図11)。計算サーバ上には暗号化された情報しか存在せず、集計や統計処理などの演算を行う計算サーバにはデータの中身を見られることなく計算結果を得ることが可能になります。暗号化するプレイヤーと復号するプレイヤーを分けることもできますが、秘密鍵を持っているプレイヤーは元データを暗号化したものも復号できてしまうため、鍵の管理とデータの流通経路も重要となります。

図11 準同型暗号技術の特徴



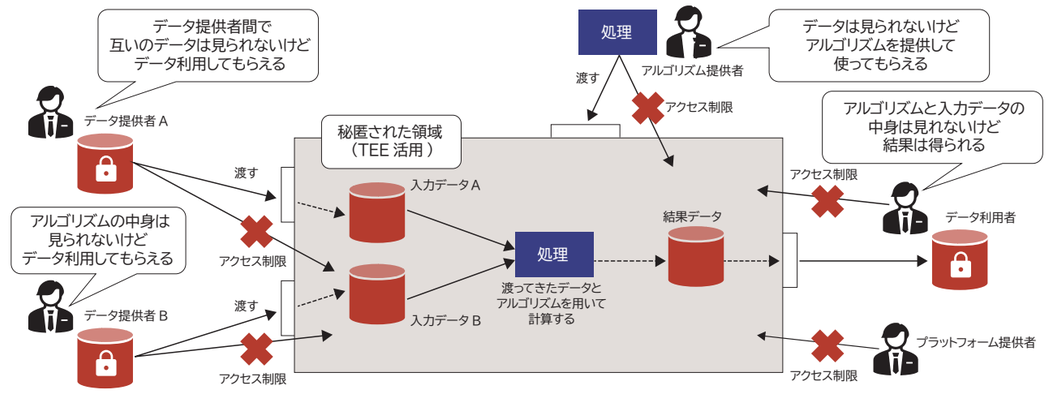
暗号化せずに計算を行う場合に比べると、まだ暗号化状態での計算には大きな時間がかかることが多い状況です。最近ではある程度許容可能な計算速度となり実用化されてきていますが、リアルタイムでの応答処理にはまだ不足と言えるでしょう。準同型暗号演算専用のハードウェアも開発されており、今後さらなる高速化が実現される見込みもあります。

【TEE】

TEE(Trusted Execution Environment)は、ハードウェアベースで保護・秘匿されたセキュアなプログラム実行空間です。データやデータ処理のためのアルゴリズムは所有者が暗号化してからTEE領域へ投入、保護領域内で復号、計算処理、再暗号化を実施します(図12)。これにより外部とは暗号化されたデータのみをやり取りし、生データはセキュアなメモリ空間内のみに存在する状態で計算が行われます。一般的なTEEではデータ提供者とアルゴリズム提供者は一致していることが多いですが、これを複数のデータ提供者、アルゴリズム提供者へ拡張することも可能です。

例:NTTデータサンドボックス技術
https://www.rd.ntt/infrastructure/0001.html

図12 TEE の概要



保護領域内のデータへは誰もアクセスできないため、データ所有者、アルゴリズム提供者、データ利用者は自分の持つデータ以外は見ることができません。計算処理は元データに対して実施するため、一般的な計算処理を実行可能です。演算自体は復号したデータに対して実行するため、暗号化せずに計算を行う場合と比べても実行速度が大きく不利になることはありません。ただし、TEEを提供する特殊なハードウェアを用意する必要があります。また、ハードウェア内では一時的に元データに戻っているため、サイドチャネル攻撃など物理現象を観測する攻撃に対しては元データの情報が攻撃にさらされることになります。

■匿名化

【匿名加工】

匿名加工とは、データ内に含まれる個人を特定できる情報を個人の識別や情報の復元をできないように削除・加工することです。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、また当該個人情報を復元できないようにした情報を匿名加工情報と呼びます。個人情報の保護に関する法律施行規則第19条では、以下の加工基準1号から5号のすべての措置を行う必要がある、とあります。

・[1号] 特定の個人を識別することができる記述の削除(又は置換え)
・[2号] 個人識別符号の削除(又は置換え)
・[3号] 連結符号の削除(又は置換え)
・[4号] 特異な記述の削除(又は置換え)
・[5号] 個人情報データベース等の性質を踏まえたその他の措置

個人情報を含む情報をそのまま第三者が活用することはできず、これらを満たすようにデータを加工し、匿名加工情報にする必要があります。具体的には以下の匿名加工技法を組み合わせて匿名化を実施しますが(表6)、その際には匿名性を持たせつつ有用性を失わないように加工する必要があります。

表6 代表的な匿名加工技法

分類	技法	解説	例
削除	項目削除	個人情報の項目列を削除する	年齢のデータを全ての個人情報から削除
	レコード削除	個人情報のレコードを削除する	特定の年齢に該当する個人のレコードを全て削除
	セル削除	個人情報の特定のセルを削除する	特定の個人に含まれる年齢の値を削除
加工	一般化	上位概念もしくは数値に置き換える	購買履歴データで「きゅうり」を「野菜」に置き換える
	ミクロアグリゲーション	個人情報をグループ化した後、グループの代表的な記述等に置き換える	個別商品の単価ではなく、商品カテゴリーの平均単価に置き換え
	トップ（ボトム）コーディング	数値に対して、特に大きい、または小さい数値をまとめる	年齢データで、90 歳以上で個人特定可能性があるデータを「90 歳以上」にまとめる
	丸め（ラウンディング）	数値に対して、四捨五入等して得られた数値に置き換える	年齢を年代に変換
抽出	レコード一部抽出	個人情報の一部のレコードを（確率的に）抽出する	サンプリングでの抽出
	一部項目抽出	個人情報の項目の一部を抽出する	購買履歴に該当する項目の一部を抽出
かく乱	データ交換(スワッピング)	構成する個人情報相互に含まれる記述等を（確率的に）入れ替える	異なる地域の属性（市区町村など）を持ったレコード同士を統計量が同じになるように入れ替え
	ノイズ（誤差）付加	一定の分布に従った乱数的な数値等を付加することにより、他の任意の数値等へと置き換える	日付データをルールに従って変更
	疑似データ生成	人工的な合成データを作成し、これを個人情報データベース等に含ませる	シミュレーションやトレーニングのため、元データの特徴を把握してダミーデータを追加

匿名加工技法は1種類だけ実施すれば良いということはほぼありません。1種類だけの適用では個人を特定できる情報が残ってしまうか、有用性が失われてしまう可能性があります。有用性と匿名性のバランスをとるためには、いくつかの加工技法を試し、加工後のデータを見ながら、そのバランスを確認する必要があります。そのためのソフトウェアも存在します。

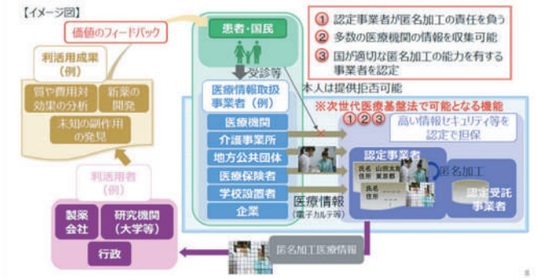
Case study

【事例】医療ビッグデータの匿名化

医療データは個人情報のなかでも特に厳しく管理される必要があるデータですが、一方で多数の症例データに基づいた研究も非常に重要となっています。電子カルテなどの情報を匿名化した匿名加工医療情報を共有することで医療研究を進めることができます。

電子カルテなど医療ビッグデータ活用によるがん患者臨床アウトカム評価の研究開始
～初の次世代医療基盤法に基づく匿名加工医療情報提供に向け契約を締結～

医療ビッグデータを活用した研究を進め、患者さんにより良い医療を提供することを目指す。
次世代医療基盤法：匿名加工された医療情報を安心して円滑に利活用する仕組み。



【次世代医療基盤法の全体像】

①認定匿名加工医療情報作成事業者の整備
医療情報の管理や利活用のための匿名化を適正かつ確実に行うことができる者を認定する。

②認定事業者tに対する医療情報の提供許可
認定事業者は収集情報を匿名加工し、医療分野の研究開発の用に供する。

図13 医療ビッグデータ活用のための匿名加工

イメージ図の出典：内閣官房ホームページ　https://www.mhlw.go.jp/content/10601000/000406831.pdf

【匿名合成】

匿名合成データ生成は、生データの統計的な特徴を学習したモデルを作成し、そのモデルを用いて生データに類似した架空のデータを生成する技術です。特定される個人がそもそも存在しないデータのため自由に使うことができ、さらにデータ数を増やすことも可能です。(図14)

図14 匿名合成データ生成

生データから統計的性質を学習した GAN などのモデルを作成し、 そのモデルを用いて完全に新しいデータを生成する技術

生データ

センシティブ情報

名前	年齢	性別	入院日	病名
Alice	52	女性	2020/1/1	肺がん
Bob	52	男性	2020/1/2	頭痛
Carol	51	女性	2020/1/3	乳がん
Dave	46	男性	2019/9/9	前立腺がん

モデル（GAN など）

合成データ

年齢	性別	入院日	病名
54	男性	2020/3/15	肺がん
50	男性	2020/3/2	頭痛
49	女性	2020/2/3	乳がん
48	男性	2019/10/3	肺がん
44	女性	2019/9/7	頭痛

誰のデータでもない擬似的データ
（統計量・分布等は元データに類似）

参考：従来の匿名化データ

名前	年齢	性別	入院日	病名
Alice	50 代	男性	2020/3/1	肺がん
Bob	50 代	女性	2020/3/2	頭痛
Carol	50 代	女性	2020/3/4	乳がん
Dave	40 代	男性	2019/7/9	***

誰のデータか特定困難

- 合成データは誰のデータでもないため、対応する個人特定のリスクがない
- ランダムに作られたデータではなく、生データの特徴や統計的性質は保持
- 従来の匿名化と異なり、自由にデータ数を調整可能
- モデルは生データを直接持たないため、モデル時点で共有することも可能

主な注意点としては、生データに多数含まれそうなデータほど生成されやすいので、元データが外れ値を含んでいても再現されない場合があります。また、名前やインデックスのようなキーになる情報を持たないため、複数のモデルで生成されたデータを名寄せして横結合することはできません。
作成したモデル自体にも生データは含まれないためモデルを共有することも可能ですが、この場合はメンバシップ推定などの攻撃によって生データが復元されないようにするなどの対策も別途必要です。

■PPML（Privacy Preserving Machine Learning）

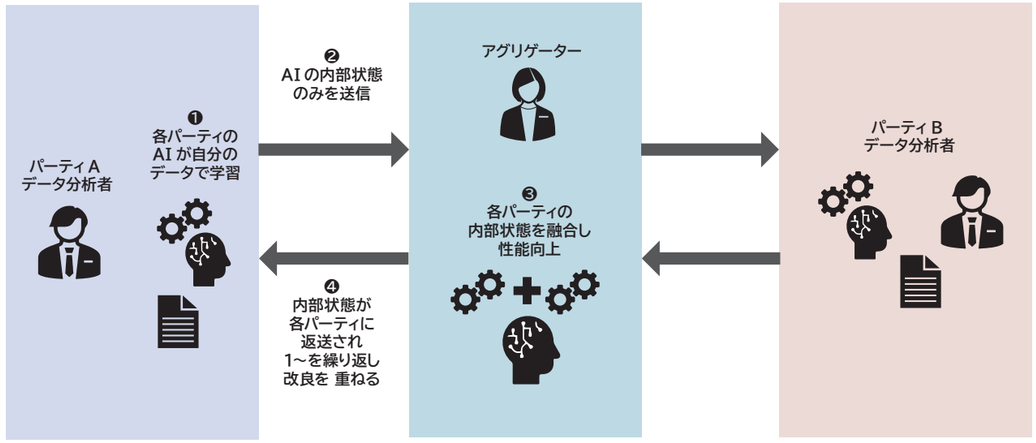
機械学習ではたくさんのデータを集めて学習を行いますが、個人の情報を含む生データを集約することはそれだけで流出のリスクや管理コストを生みます。PPML（Privacy Preserving Machine Learning：プライバシー保護機械学習）は、生データを共有せずAIの内部状態（モデルパラメータ）や、機械学習の過程で生じる途中状態を共有することで個人の情報を流通させないまま機械学習を行います。ここでは、PPMLとして良く知られたFederated Learningと、筑波大学が中心となって開発したデータコラボレーションについて紹介します。

【Federated Learning】

Federated Learningでは、生データを持つデータ分析者（パーティ）が複数いる状態を想定します。各パーティは自分の生データを持っていますが他のパーティとは共有しません。

1.各パーティは自分の生データである程度学習を進めます。
2.各パーティから学習モデルの内部状態のみアグリゲーターに集められます。
3.アグリゲーターは各パーティから来たモデルの内部状態を融合させ、性能の高い学習モデルを作成します。
4.アグリゲーターから融合後の内部状態が各パーティに返送され、各パーティはそれと自分の生データで学習を続けます。
5.上の1.～4.を繰り返してモデルの改良を続けていきます。

図15 Federated Learning 概要



モデルの内部状態を共有するため、生データ自体は各パーティから外へ出ることはありません。各パーティはスマートフォンなど不特定多数のデバイスでも動作させることが可能です。現時点では機械学習モデルとしてはニューラルネットワークを用いたものが主流で、それ以外の機械学習モデルについての利用は限定的です。

Federated Learningを発展させた技術として、ブロックチェーンを利用してパラメータの共有・更新が実装されたSwarm Learning(*2)やモデルパラメータではなくモデルの中間ノードを共有するSplit Learning(*3)などがあります。

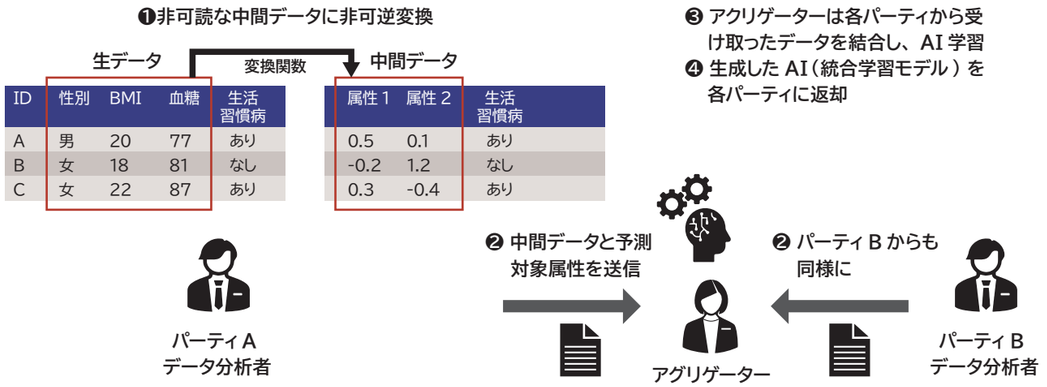
(*2) 参考：Warnat-Herresthal, S. et al. Swarm Learning for decentralized and confidential clinical machine learning. Nature 594, 265-270 (2021).
(*3) 参考：Gupta et al., Distributed learning of deep neural network over multiple agents, Journal of Network and Computer Applications, Vol 116, 2018, pp.1-8.

【データコラボレーション】

データコラボレーションでも生データを持つデータ分析者（パーティ）が複数いる状態を想定します。各パーティは自分の生データを持っていますが他のパーティとは共有しません。

1.各パーティは自分の生データを非可読な中間データへ非可逆変換します。
2.各パーティから中間データをアグリゲーターへ集めます。
3.アグリゲーターは各パーティから来た中間データを統合し学習を実行します。
4.アグリゲーターから生成した統合学習モデルが各パーティに返送されます。

図16 データコラボレーションの概要



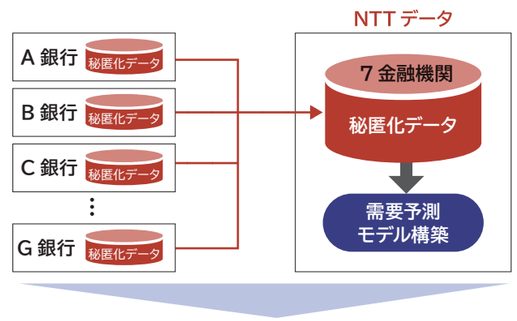
Case study

【事例】データコラボレーションによる金融機関データ活用

事態をまたぐ複数金融機関のデータをコラボレーション技術によって統合し、精度の高い需要予測モデルを構築しています

データコラボレーション技術を用いて業態をまたぐ 7 金融機関のデータを秘匿化した状態で統合し、精度の高い需要予測モデルを構築。（2022/1/11 NTT データニュースリリース ）

概要	複数金融機関からデータを集め、顧客の資金需要を予測するモデルを構築
対象	業態を跨ぐ 7 金融機関
狙い	同一業界データのデータ集約による精度向上
利用データ	融資実績・預金・属性
予測対象	法人顧客の翌四半期における融資需要



結果 正解サンプルが少ない高額融資など 複数機関のデータのメリットを確認。

図17 データコラボレーション事例

https://www.nttdata.com/jp/ja/news/services_info/2022/011100

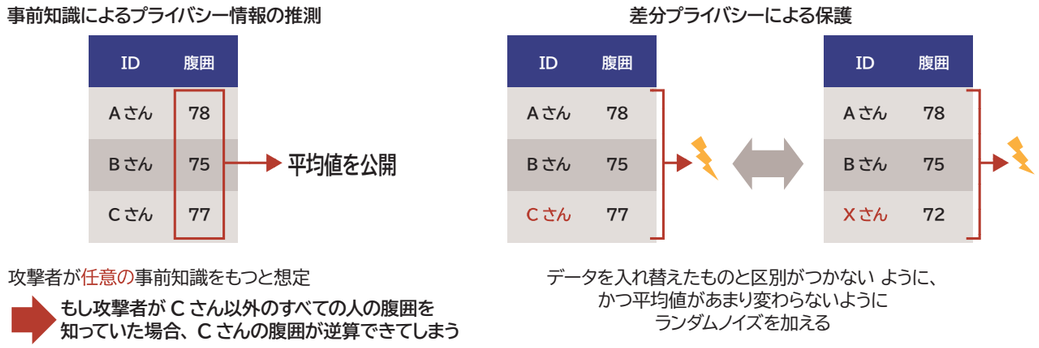
中間データへの変換は次元圧縮などの数値計算であるため、計算コストは大きくありません。また、中間データの形式は比較的自由に選択できるため、アグリゲーター側で作成したいモデルに合わせることが可能です。 Federated Learningでは学習が完了するまで頻繁に内部状態の通信を行いますが、データコラボレーションでは中間データの送信とモデルの返送の一往復ですみます。そのため通信コストが低いだけでなく、システム構築も容易になります。Federated Learning、データコラボレーション、どちらもモデルの情報が盗まれた場合にはメンバシップ推定などモデルからその構築に使われたデータを盗む攻撃にさらされる可能性があります。これらへの対策は別途必要です。

【差分プライバシー】

差分プライバシーはPPML自体ではありませんが、機械学習の分野で発展してきたプライバシー保護技術の一つです。機械学習で利用する際の有用性を保ちつつデータにノイズを乗せて個人の特定ができないようにする手法です。

具体的には、データに対する統計処理などの処理時に確率的なアルゴリズムを用いてランダムノイズを加えてプライバシーを保護します。このとき、ある個人のデータを他のデータと入れ替えたデータセットを考えます。統計処理結果の出力を見ても、この出力がどちらのデータセットから出てきたものなのか判別ができない、かつ平均値などの統計値が変わらないようにランダムなノイズを加えるのが差分プライバシーの基本的な考え方です。

図18 差分プライバシーの概要



4.まとめ

■データ保護技術の現状課題と対応

データ保護技術適用にあたり、現状の課題や制約を把握し、下記の観点で事前検討が必要となります。

・機能実現性

データ保護のための各技術は、実行できる処理や対象のデータ形式(テーブルデータ、非構造データなど)が異なり、それぞれ得意不得意があります。例えば、秘密計算ではテーブルデータに対する統計演算での活用例が多く、Federated Learningでは深層学習の評価実験が多くを占めています。

実施したい処理と対象データを明確にした上で、PoCなどで機能的な実現性を検証する必要があります。

・パフォーマンス

生データでの処理と比較して原理的に処理時間は長くなります。さらに、Federated Learningでは、分散環境での通信コストに課題があります。

最適な通信環境を整備し、実行時間、応答時間が許容範囲かを確認することが必要です。

・セキュリティ

システム化の観点から、データ保護技術以外にも、認証、ID管理、アクセス制御、鍵管理、通信暗号化など、実際の運用を見据えた従来型のセキュリティ設計も必要です。

また、プライバシー保護設計の観点から、保護の仕組みの直感的な理解が難しく、プライバシー強度に対する共通認識形成が課題となる場合があります。一般的にプライバシー強度と有用性はトレードオフの関係にあるため、プライバシー保護対象の共通理解を醸成し、タスクに応じてバランスの取れた設計が求められます。

・法規制対応

技術適用だけで個人情報保護法などに対応している保証はありません。現状の個人情報保護法では、暗号化などによる秘匿化だけでは個人情報として取扱いが求められます(*4)。

匿名化や他の復元不可能なデータへの加工など、ユースケースに応じて個別検討が必要です。

・コスト対効果

データ保護技術による処理は、生データのみを扱う場合よりも、システム化コスト、パフォーマンスコスト、運用コストがかかります。さらに、コストに見合ったAI性能向上が実現できるかも見極める必要があります。

「AI性能向上に対する効果の定量評価」と「技術導入の想定コスト」を早い段階で比較することが推奨されます。また、既存の機能の組合せでの代替策やNDA締結などリーズナブルな運用を検討した方がよい場合もあります。

(*4) 個人情報保護委員会「個人情報の保護に関する法律についてのガイドライン(通則編)」 2-1 個人情報(法第2条第1項関係)より、
『「個人に関する情報」とは、氏名、住所、性別、生年月日、(中略)であり、暗号化等によって秘匿化されているかどうかを問わない。』

■本ホワイトペーパーのまとめ

・データ保護技術は、クロスドメインの環境で、データを秘匿化したまま統合し、分析・活用するユースケースで有効
・機能実現性、パフォーマンス、セキュリティ、法規制対応、コスト対効果などの観点で事前検討が必要
・PoCなどでのフィージビリティ確認を推奨
・各技術の特徴と留意事項(表7)を踏まえた上で、技術を選択する

表7 各技術の特徴と留意事項

カテゴリ	技術	特徴	留意事項
秘密計算	マルチパーティー計算	サーバ分散・暗号化状態で統計処理、機械学習	サーバの分散運用、パフォーマンス
	準同型暗号	暗号化状態でデータ統合と統計処理、鍵による管理	パフォーマンス、暗号鍵の運用管理
	TEE	HW で保護された領域での秘匿データの計算	HW 内で復号化される瞬間がある
匿名化技術	匿名加工	法的に認められた匿名化による秘匿性強化	情報損失と有用性のバランス
	匿名合成データ生成	元データの特徴や統計的性質は保持したデータ生成	有用性の評価は研究段階
Privacy-Preserving Machine Learning (プライバシー保護 ML 技術)	Federated Learning	分散環境で生データを共有せずに機械学習実行 非構造化データでの利用	アグリゲーターの設置 ノード間の通信環境、通信コスト
	データコラボレーション	統計的特徴を保ってデータ統合、機械学習実行 秘密計算に比べ計算コストが小さい	中間データ変換による情報損失検証 アグリゲーターの設置
	差分プライバシー	ノイズ付与によりプライバシー保護した状態で他の機械学習手法と組み合わせる	ノイズ付与による情報損失検証 個人情報保護法への対応検討