Topics

プライバシー保護技術に関する動向と医療ヘルスケアデータの利活用における示唆

医薬産業政策研究所　主任研究員　佐々木隆之

はじめに

IoTの普及やAI（人工知能）の進化、通信技術の進展等により、日常生活で取得・活用されるデータの量は爆発的に増え、個人のデータをもとにその人の嗜好や状態にあった価値提供が為される時代を迎えている。しかしながら、社会で流通する多種多様なデータの中には、個人に関わるデータも多く含まれており、プライバシーを確保することは大前提にある[1)]。特に医療や健康に関する情報は機微性が高く、情報流出への懸念と保護への期待は、程度の差はあれ、誰しもが抱いているところであろう。

2020年に発生したCOVID-19パンデミックにおいても、各国で緊急事態宣言が発せられるなか、感染者数や病床数といったデータだけでなく、携帯電話の位置情報、通信アプリ、交通系ICカード等から集められるデータの活用や、感染者との接触情報を収集・提供する「コンタクトトレーシングアプリ」の普及など、データに基づく様々な公衆衛生上の施策が展開された[2)]。最近では、米国におけるCOVID-19の症状に関する「検索傾向」のデータも、Google Cloud Platformから公開されている[3)]。しかしながら、これらの情報に個人情報が含まれていないかという不安や、あるいは公共の利益のためであれば個人の情報コントロール権の制限が許容されるのかといった議論もまた再燃した。

一般に、データは統合分析することによりその価値をさらに増す。例えば米国の医療は、病院、介護者、製薬企業等がデータを共有することで、毎年3,000億ドル以上の価値を創出することが可能である、とされている[4、5)]。また、ビッグデータから質の高い知見を得るためには、そのデータ数（n）ではなく属性数（p）が重要であることも指摘されている[6)]。こうした背景もあり、医療健康分野のビッグデータの利活用においては、医療情報、PHR、IoTデータ、検診・健診データ等の統合解析が目指されているが[7)]、個々のデータのプライバシーに配慮しても、それらのデータを重ね合わせることで新たにプライバシーの開示が発生してしまう「モザイク効果（mosaic effect）問題」も、オープンデータの充実に向けた課題として指摘されている[8)]。

デバイス、アプリやインフラの提供者は当然、こうしたプライバシーの保護や同意の取得について慎重に検討し、それに適した技術を適用してい

1）日本経済団体連合会「Society 5.0 －ともに創造する未来－」（2018.11.13）
2）「新型コロナウイルス感染流行下におけるデータ利活用　～接触確認アプリの事例を中心に～」政策研ニュース No.60
3）“COVID-19 Search Trends symptoms dataset”（Google Cloud Platform）
https://console.cloud.google.com/marketplace/product/bigquery-public-datasets/covid19-search-trends（2020.9.12閲覧）
4）McKinsey Global Institute, Big data: The next frontier for innovation, competition, and productivity, 2011年５月
5）McKinsey Global Institute, The ‘big data’ revolution in healthcare ― Accelerating value and innovation, 2013年１月.
6）日本オミックス学会「ビッグデータ医療時代における人工知能への期待」（2019.11.29）
7）「医療健康分野のビッグデータ活用研究会報告書 vol.5」医薬産業政策研究所
8）寺田雅之「差分プライバシーとは何か」システム／制御／情報、Vol.63、No.2、pp58-63（2019）

るが、一方でそれらの技術の特徴が注目されることはあまりない。プライバシー保護に関する技術に関し理解を深めることもまた、個人データの活用に対する社会の懸念を払しょくしていくために必要なステップのひとつである、とも言えるのではないか。そこで本稿では、プライバシー保護を目的に開発、実装されている技術について概要を紹介する[9)]。

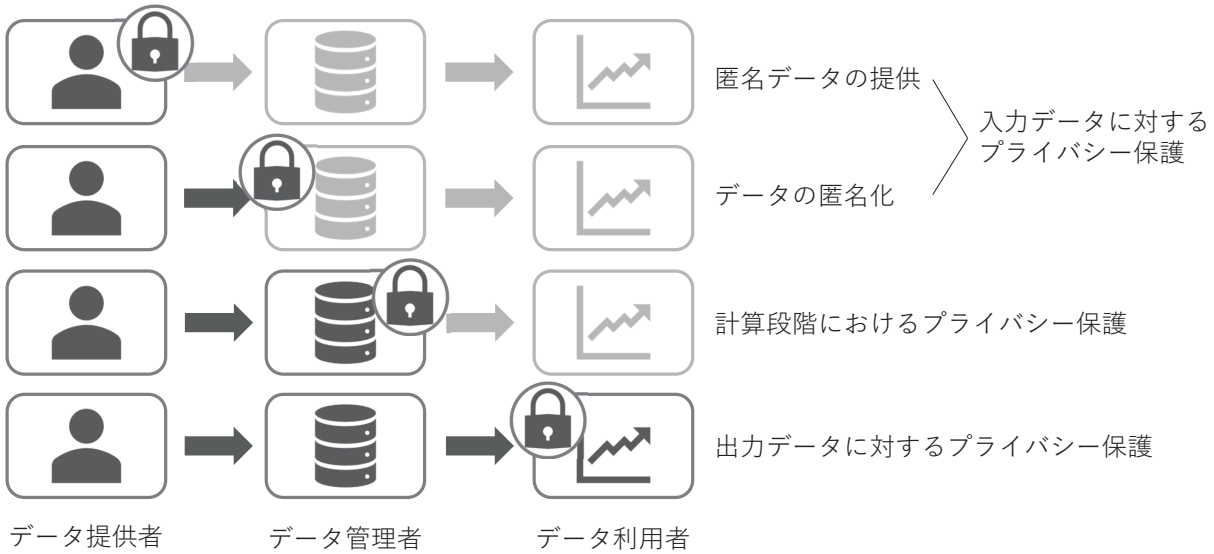
プライバシー保護技術の例

プライバシーを保護しながらそのデータから有用な知見を得る手法として、プライバシー保護データマイニング（Privacy Preserving Data Mining；PPDM）という語を聞いたことがある方もおられるかもしれない。PPDMは、データに関する個人のプライバシーを守りながらデータ分析を行うための技術の総称である、とされている[10)]。プライバシーを保護しながらデータを解析する技術にはさまざまなものがあるが、一般に、データの観点からは３つに大別可能であろう（図１）。すなわち、入力データに対するプライバシー保護（例：匿名でのデータ提供、あるいは提供された生データの匿名化）、計算段階におけるプライバシー保護（例：秘密計算）、出力データに対するプライバシー保護（例：差分プライバシー）である。

現時点で実装されているプライバシー保護技術は、ほとんどが匿名化であり、プライバシーの保護性は高いが、一方で統合解析の観点からはデータの質は落ちてしまう。そこで本稿では、データ利活用とプライバシー保護を両立しうる「次世代のプライバシー保護技術」として期待が高まっている、計算段階や出力データにおけるプライバシー保護について、暗号化や機械学習の観点から紹介したい。

なお前提として、データに含まれる情報の秘匿性（安全性）と、データの性質や特徴から知見獲得（有用性）はトレードオフの関係にある、という点がある[8)]。すなわち、プライバシー保護技術は、このトレードオフの関係において、「適切な安全性（完全な安全性ではない）」を保持したうえで「より高い有用性」を備えたデータを出力できることが求められる、という点を理解しておく必要がある。

図１　プライバシー保護技術の段階



出所：「プライバシー保護データマイニング」高橋勝巳、システム／制御／情報、Vol.63、No.2、pp43-50、(2019) をもとに筆者作成

9）個人の医療健康情報等に関するプライバシー保護に限らず、例えば新薬創出に関する企業内データの共有など、秘匿しておきたい情報を組織をまたいで活用する際にも、こうした技術は役立つ可能性がある。
10）高橋勝巳「プライバシー保護データマイニング」システム／制御／情報、Vol.63、No.2、pp43-50（2019）

秘密計算（秘匿計算）

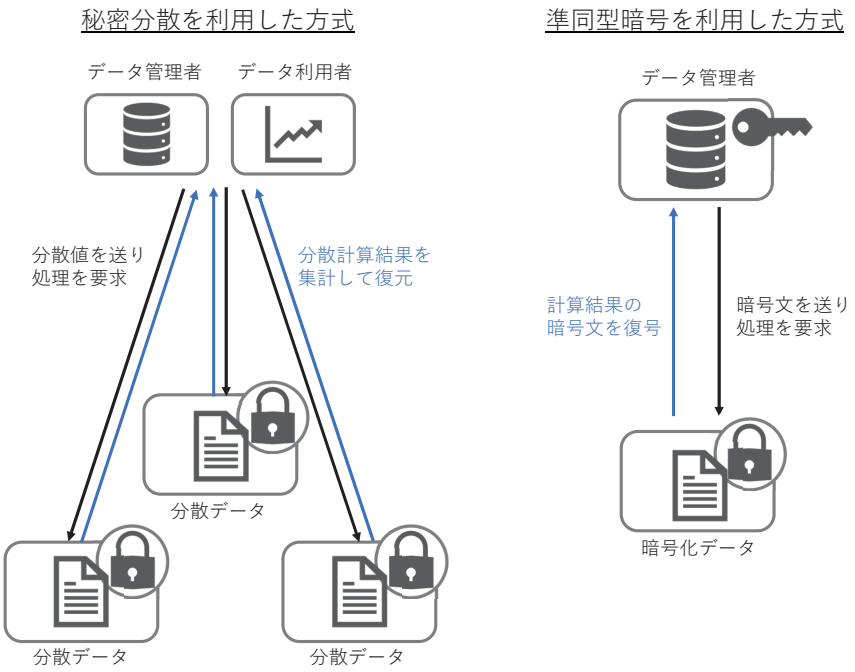
秘密計算とは、データを暗号化したまま、一切復号せずに様々な処理をする暗号技術である[11]。この技術を活用することで、元データを一切開示することなく、データの結合･分析が可能となる、とされている。バイオ分野では、ゲノムの相同性評価、アラインメント等の配列解析や、ロジスティック回帰分析による疾患リスクの予測等への応用が期待されており、DNA 編集距離計算では現実的な性能を達成している[12]。

秘密計算の代表的な方式には「秘密分散方式」と「準同型暗号方式」がある（図２）。秘密分散方式は、各組織の機密データを単体では意味を持たない複数の情報に分散し（秘密分散）、それに対応した複数のデータ管理者は機密データが開示されない状態で分散化されたデータを結合・分析し、分析者は各管理者の分析結果を集めて復元することで最終的な分析結果を得る、という手法である。これに対し準同型暗号方式は、機密データを暗号化し、そのまま結合計算処理、暗号化された処理結果を得た後、分析者は鍵を用いてこの結果を復号する、という処理を経る。

差分プライバシー

差分プライバシー（differential privacy）は、データセットに意図的に統計的ノイズを付加することによりプライバシーを保護する手法であり、暗号技術のひとつである。この技術は既に広く普及しているものであり、例えば Apple は iOS10（2016年リリース）から、quick typeや絵文字の提案、ヒントの参照、Health Type Usage のインテリジェント化を実現するためにこの手法を活用している[13]。COVID-19パンデミック対応においても差分プライバシーは活用されており、例えば上述

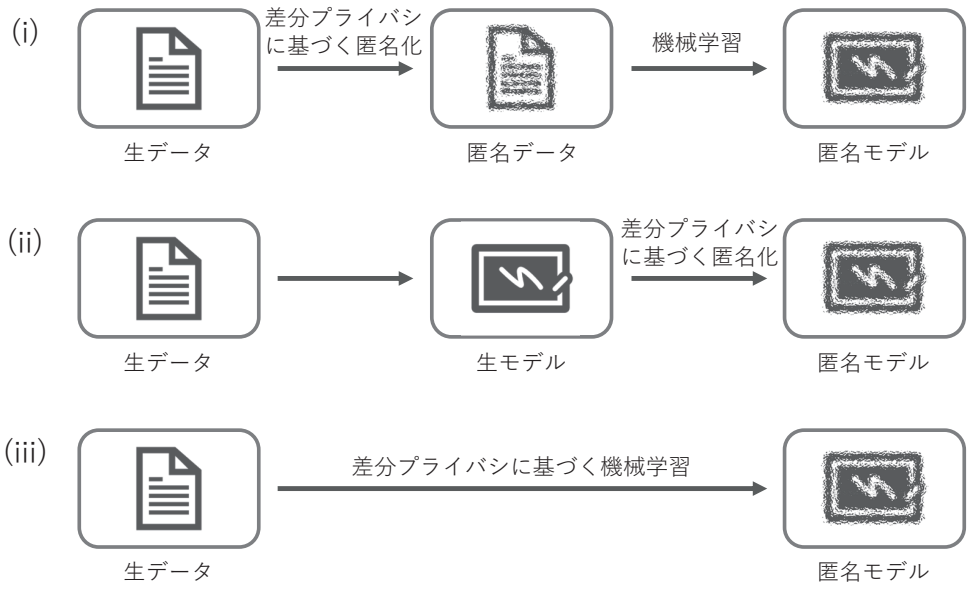
図２　秘密分散と準同型暗号



出所：NEC セキュリティ研究所資料をもとに筆者作成

11）竹ノ内隆夫「秘密計算の PWS2018での議論と最近の動向」（2019年３月７日 PWS Meetup NEC セキュリティ研究所発表資料）
12）S.Laur,e et al.“From Oblivious AES to Efficient and Secure Database Join in the Multiparty Setting”*Applied Cryptography and Network Security*, 84-101（2013）
13）apple社“Differential Privacy”https://www.apple.com/jp/privacy/docs/Differential_Privacy_Overview.pdf（2020.9.12閲覧）

図３　差分プライバシの適用段階

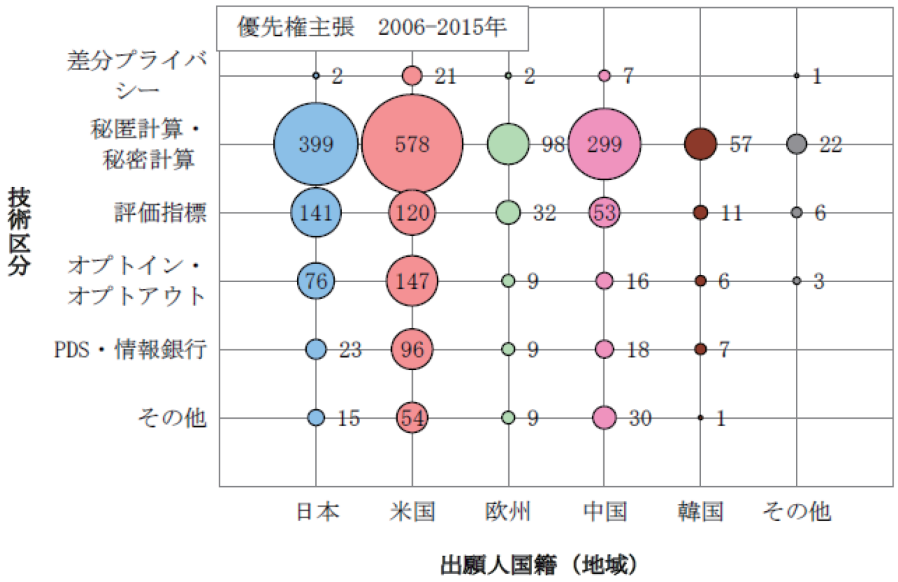


出所：清雄一ら、日本ソフトウェア科学会第32回大会（2015年度）講演論文集をもとに筆者作成

のCOVID-19の症状に関する「検索傾向」データや、同じくGoogle社が提供しているCommunity Mobility Reportsなどが該当する[14]。

機械学習において差分プライバシーを適用する対象は、図３のとおり、主に３種類が考えられる[15]。すなわち、(i)生データを匿名化する場合、(ii)生データから学習することで得られたモデルを匿名化する場合（匿名モデルの生成）、(iii)生データから学習する際に差分プライバシーに基づく機械学習を行う場合（例えばニューラルネットの活性

図４　匿名化技術　出願人国籍（地域）別ファミリー件数



出所：特許庁　平成29年度　特許出願技術動向調査報告書（概要）匿名化技術

14）“google mobility report” https://www.google.com/covid19/mobility/（2020.9.12閲覧）
15）「差分プライバシを満たすニューラルネットワークモデル構築手法の提案」清雄一ら、日本ソフトウェア科学会第32回大会（2015年度）

化関数の出力におけるノイズ付加）であり、上記の例は生データの匿名化に相当する。

差分プライバシーを活用したソリューションは、上述のような米国のデータプラットフォーマーのものが目立っており、日本企業が提供している例はあまりないようである。少々古いデータだが、2006～2015年の間に差分プライバシーに関して出願された特許はほぼ米国からのものであり、日本は秘密計算・秘匿計算に注力していた点も影響があるかもしれない（図４）[16]。

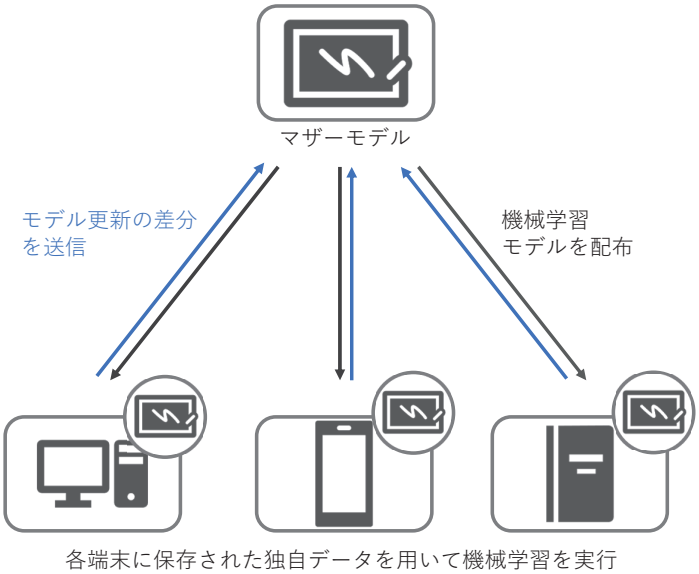
連合学習

一方、プライバシーを保護しながらデータを解析することそのものではなく、複数組織にまたがるデータをもとに機械学習の精度を向上させることが目的の場合に適用される技術として、連合学習（federated learning）が注目されている。この技術はGoogle社が2017年にその枠組みを発表した、深層学習の一手法であり[17]、各サイトに分散保管された学習用データに対して、まず機械学習の共通モデル（マザーモデル）を配布し、各サイトで独自に保存されている学習用データを用いてマザーモデルを更新、更新前と更新後のモデルの差分をマザーモデルに返すことで、マザーモデルをアップデートする（図５）。

この手法は、データをデバイス間で共有したり転送したりする必要がなく、常にローカル環境に保存される点に特徴がある。すでに、代表的な機械学習フレームワークであるTensorFlowやPyTorchでは、連合学習を実行できるようになっている[18、19]。なお、類似の機械学習手法に「分散学習」があるが、分散学習はデータセットが同一であることが前提なのに対し、連合学習では異なるさまざまなデータを使用できる点に特徴がある。

連合学習は、金融や医療・ヘルスケアのような機微性の高い情報を扱ってAIをトレーニングする分野に特に適しているとされる。例えば、米国のスタートアップで、医学分野のデジタルリサーチプラットフォームを提供するOwkin社は、データを移動させずにアルゴリズムを強化する手法として、連合学習を採用している[20]。また、今年に入って、連合学習によりプライバシーを保護しな

図5　連合学習のコンセプト



16）平成29年度　特許出願技術動向調査報告書（概要）匿名化技術
17）https://ai.googleblog.com/2017/04/federated-learning-collaborative.html（2020.9.12閲覧）
18）https://developers-jp.googleblog.com/2019/03/tensorflow-federated.html（2020.9.12閲覧）
19）https://pytorch.org/（2020.9.12閲覧）
20）Owkin社HP https://owkin.com/federated-learning/（2020.9.21閲覧）

がら電子医療記録（EMR）を解析する手法が検討され、MRI画像から健康な脳組織とがん性を区別するタスクにおいて、連合学習は従来のデータ集約モデル（Centralized Dada Sharing）と同等のパフォーマンスを発揮することが報告されている[21]。

さらに連合学習は、中央の共通モデルのアップデートを主眼に置いたこれまでの仕組みから、完全な分散型「Decentralized Federated Learning」も検討され始めている[22]。この手法はいわゆる「中央集権」を脱却するものであり、ブロックチェーン技術との相性が良く、対改ざん性や透明性を高める連合学習の一手法として注目が高まるだろう。

日本では、類似の技術として「プライバシー保護機械学習」が存在しており、金融業界での取り組みが進んでいる。例えば、深層学習技術を用いた不正取引の自動検知システムにおいては、単独の金融機関では十分量の教師データを準備することが難しいため、より多くの銀行のデータをもとに学習した結果を統合することで、検知精度の向上を目指す取り組みがある。一例として、情報通信研究機構（NICT）は、三菱 UFJ 銀行、三井住友信託銀行など金融機関５行と連携し、プライバシー保護深層学習技術（DeepProtect）による不正送金検知の実証試験を2020年より開始している（図６）[23]。この実証試験では、準同型暗号技術により暗号化された学習モデルのパラメータ（重み）を用いたプログラムの更新が可能となっている。

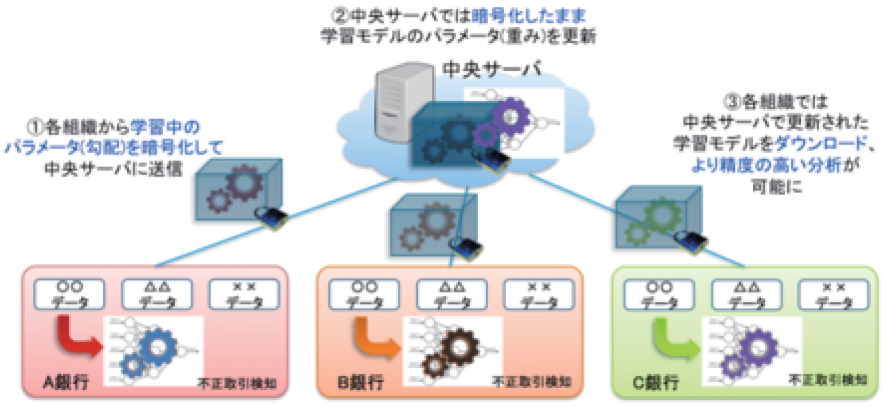
プライバシー保護技術に関する課題

ここまでみてきたプライバシー保護技術は、プライバシー保護とデータ利活用を両立させうるものとして有望とされているが、課題もある。

例えば暗号化技術は、暗号文を攻撃者に解読されると秘密情報が漏洩してしまうリスクを常に抱える、という点がある。秘密分散では、データ管理者に結託があった場合、元の機密データを復元できてしまう可能性があるため、結託がないこと、すなわち秘密分散したデータが完全に分離された状態が保たれることが前提になる。また、準同型暗号では暗号化鍵を用いるため、鍵の安全な管理（例えば暗号文と鍵の分離）が必要となる。

計算能力や処理速度に関する課題も存在する。例えば、秘密計算に共通する課題として、多くの通信を行うため、通常の処理よりも数十～数千倍程度処理が遅くなる、という点が指摘されている[24]。また、連合学習でも、各ローカルノードに大規模な機械学習演算を実行できる計算能力が必要となる場合がある。エッジコンピューティング

図６　プライバシー保護深層学習技術「DeepProtect」の概念



出所：情報通信研究機構プレスリリース（2020.5.19）

21）Sheller et al. *Scientific Reports* 10, 12598（2020）
22）Yuzheng Li, et al. “A Blockchain-based Decentralized Federated Learning Framework with Committee Consensus” https://arxiv.org/pdf/2004.00773.pdf（2020.10.11閲覧）
23）情報通信研究機構（NICT）プレスリリース　https://www.nict.go.jp/press/2020/05/19-1.html（2020.9.12閲覧）
24）「NECの秘密計算技術のご紹介」https://jpn.nec.com/rd/technologies/201805/pdf/mpc_introduction.pdf（2020.9.12閲覧）

が進展するなかで、ローカルノードの計算能力が向上したとしても、機械学習、特に計算能力を必要とする深層学習を常にエッジ側に強いるのは、必ずしも効率的ではない局面も想定される。

また、秘密計算や連合学習では、アルゴリズムを強化する側は教師データの質を確認することができず、データセットの質はエッジノードでのクオリティ管理に依存する。AIやデータは“FAT（Fairness, Accountability and Transparency. 公平性、説明責任、透明性）”が重視されるようになっており、どういったデータセットがどこまでAIの質の向上に貢献したかを見える化することも必要であろう[25]。

さらに、法制度面の課題として、秘密計算技術を利用するに際しても、暗号化した個人データを第三者に提供することになるので、個人情報保護法23条の規制が障害となって技術を利用できないのではないかとする点も懸念されている。現行の個人情報保護法では、「個人情報を暗号化しても個人情報に該当する」とされており、そもそもこれまで、どのような意味で「暗号化しても個人情報である」との説が唱えられてきたのかを整理した上で、秘密計算技術に基づくデータ交換の個人データ該当性について検討し提言にとりまとめる必要性が訴えられている[26]。こうした先進的なプライバシー保護技術の活用に向けては、関連法制度の改正も視野に幅広い視点から議論することが必要であろう。

おわりに

このように、プライバシー保護技術は、個人情報保護に関する機運の高まりもあり、ここ数年で急速に進歩し、研究段階から実用段階へシフトした技術も多くみられるようになった。一方で、一般社会で広くこうした技術の概念や、メリットやデメリットが浸透しているか、また実際にプライバシー保護が高まった実感が市民にあるかというと、必ずしもそうでないのが現状ではないか。

また、いくつかのプライバシー保護技術は、計算リソースや通信速度などに課題を有するケースも散見されるほか、汎用性が高くないことにより開発コストが膨大になる、といった点もある。こうしたコスト面の要素は、まわりまわってユーザー（個人、企業、政府等）に費用負担の増加として跳ね返ってくるものであり、プライバシー保護にも費用対効果の側面があることを忘れてはならないだろう。

更には、これらのプライバシー保護技術を用いても、例えばデバイスから送信される更新データに個人データが含まれていないかを確認するすべが個人にない、といった透明性に関する課題や、送信される計算結果や差分、あるいはノイズを付加したり暗号化されたりした個人情報をどこまで個人情報保護法の対象とするかなど、法制度面の課題があることは前述のとおりである。またそもそも、前述のとおり情報の秘匿性と知見獲得はトレードオフの関係にあり、秘匿性に応じた利用用途の設定、あるいは反対に利用目的に応じた秘匿性の確保を都度考える必要にあるだろう。

こうした技術の実装に向けた環境整備の必要性については、経済産業省商務情報政策局が発表した「IoT進展に立ちはだかる中期的課題への新たなアプローチ」でも言及されており、政策の方向性として「秘密分散・計算技術の活用によるデータ協調環境整備の検討」も掲げられている。しかしながら、例えば厚生労働省「保健医療分野のAI開発加速コンソーシアム」など、実際にデータを使う側も交えた諸会議では、プライバシー保護が大切であるという発言は見受けられるが、具体的にどう技術面からこの課題を解決するか、国民の理解を深めるか、といった展開には必ずしもなっていないのが現状である[27]。医療健康データ利活

25）本稿を執筆している2020年9月現在、ECML（the European Conference on Machine Learning and Principles）の“BIAS 2020”にて、連合学習の公平性がトピックとして挙げられている（https://sites.google.com/view/bias-2020/）。こうした議論が国内でも進展することを期待したい。
26）情報法制研究所（JILIS）秘密計算技術応用研究タスクフォース　https://jilis.org/taskforce/（2020.9.12閲覧）
27）厚生労働省　保健医療分野AI開発加速コンソーシアム 第1回～第11回資料および「議論の整理と今後の方向性を踏まえた工程表について」

用の関係者（研究者だけでなく、政策立案者、医療関係者、データ利用企業・団体等の者も含む）が、こうした技術を知り、興味を持ち、互いに知恵を提供しあうことが、さまざまなレベルの検討体で必要と思われる。

2021年に設置が予定されているデジタル庁でも、いわゆる電子化に代表されるような「守りのデジタル化」だけでなく、データの利活用を見据えた「攻めのデジタル化」を議論していくことが、産業振興には必要であろう。特に、新たな革新的医薬品の創出に医療健康データ等を活用していくためにも、製薬産業はこうしたプライバシー保護技術の有用性を理解したうえで、実証実験の企画・参加や手法の評価などを進め、活用を提案していくような動きがあっても良いのではなかろうか。

プライバシー保護技術の進化はまさに「秒進分歩」であり、データ利活用に向けた計画をたてた当初の想定から大きく技術が進むことも起こりうる。こうした「技術のスピード感」を政策に反映できるよう、技術の専門家を予め多く巻き込んでおくことが重要であろうし、時には金融や交通など他産業で進んでいる技術や社会変化を迅速に取り入れ、ロードマップや研究計画を柔軟に変更していくことも大切ではないかと考える。