データ流通時代を支える匿名加工技術

～機密性、完全性、可用性の3要件満たすデータセキュリティを目指す～

AI（人工知能）はこれまで以上に多種多様なデータを必要とし、内部データだけでなく外部のデータも活用の対象となる。そのためデータ流通環境の整備が欠かせない。データ流通の実現には、提供側の「個人情報のような機微データを安全に提供すること」と、利用側の「提供されたデータが有用な情報として利用できること」の両方が前提となる。この両立に有効なのが改正個人情報保護法で導入された匿名加工技術である。

大坪 正典
日鉄ソリューションズ株式会社
技術本部
システム研究開発センター
データ分析・基盤研究部
主務研究員

機械学習の精度を高めるため外部からの属性データも必要に

AIは、単純作業の機械化を目指す第1段階、複雑なプロの仕事を機械化する第2段階を経て、人間の思考を超えた知能を機械化する第3段階に移行するといわれている。第1段階では人間がルールを与え、第2段階では機械が人間のやり方を基にルールを見出す。これらの段階では、AIの学習に必要なデータを人間が判断し、1つひとつ用意してAIに供給すればよかった。しかしながら第3段階になると、学習に必要なデータは多岐にわたり、人間には必要性が判断できないデータもあるかもしれない。人智を超えた機械学習を実現するためには、これまで以上に良質で多種多様なデータ（いわゆるビッグデータ）が必要となる。

ビッグデータの定義を考える際に「6つのV[1)]」がよく使われる。優れたAIを実現するためのビッグデータには、大容量（Volume）、多様性（Variety）、高頻度（Velocity）、正確性（Veracity）が求められ、そこから価値（Value）を生み出すために投資（Venality）が必要となる。

私は以前に、ビッグデータによって機械学習の精度が高まるかどうかを検証した[2)]。具体的には、2値分類の機械学習について、学習するデータ量の増加に伴ってモデルの精度がどのように変化するかを調べた。ここでの「データ量の増加」には、①データ件数の増加と、②データ属性数の増加がある。

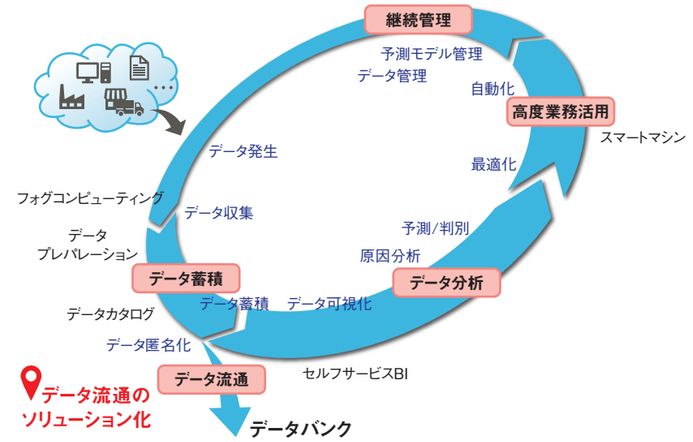
検証結果の概略は次の通りである。①データ件数の増加試験では、属性数が少ない（12個）データを学習した場合、データ件数が7500件を超えると精度の向上が頭打ちになった一方、属性数が多い（1万5000個）データを学習した場合、用意していた4万件まで精度が上がり続けた。②データ属性数の増加試験では、属性数が多いほど精度が上がることが分かった（用意していた属性数1万5000個に達するまで精度が上昇し続けた）。

これらの結果から、データ1件にひもづく属性データの数が多くなれば、機械学習の精度が高くなる可能性を秘めていると言える。このことはAI活用の場面において、内部データだけでなく外部からのデータも取り込んで属性数を増やすことが、AIの賢さを高める1つの手段として有効であることを示唆している。なお当時の検証では比較的シンプルな機械学習のアルゴリズムであるRandom Forestを用いたが、近年流行りのDeep Learningを用いれば、より一層ビッグデータの恩恵を受けられるかもしれない。

データ利用側にとって外部データに対するニーズが高まる一方、データ提供側の立場から見たデータに対する考え方は、単純な保有・保護の対象というだけでなく資産としての価値が認識され始めた段階にある。いかに「データを外に出さないか（漏洩させないか）」だけを考えていた時代から、活用に向けて「データを外に出す」動きが少しずつ活発化し始めている。

データを資産と見なすようになれば、「守るべき要素は守りながらも多くのデータを流通させ、使いたい人に使ってもらう方がよい」と考えるプレーヤーが増えてデータ流通時代がやって来る。その際には、企業のデータ活用ライフサイクルにもデータ流通が組み込まれることになる（図1）。

■図1 データ流通時代のデータ活用ライフサイクル――データを使いたい人に使ってもらうために



データを外部に出す前提で情報をコントロールする

従来、アプリケーション、データベース、ネットワークの各レイヤにわたって施されてきたアクセス制御や暗号化、VPNなどといった対策は、攻撃者の侵入や情報漏洩、不用意な情報閲覧などを防止する、すなわち「データを外に出さない（出るリスクを軽減する）」ための対策である。

しかしデータ流通時代になると、「データを外に出す前提で情報をコントロールする」ための対策が必要となる。つまり既存の対策に加え、匿名化や秘密計算といったデータの中身（情報）に着目した「データセキュリティ」が必要となる（図2）。

■図2 データセキュリティの重要性――データを外部に出す前提の対策が必要に

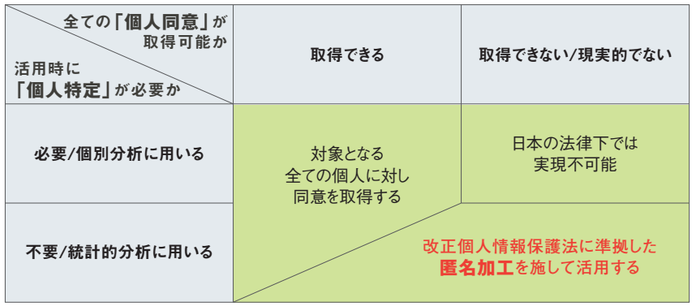
セキュリティの3大要素は、機密性（Confidentiality）、完全性（Integrity）、可用性（Availability）の3つであるとされている。データセキュリティについて考察してみると、機密性は「個人情報における匿名性が担保されること」、完全性は「匿名化処理後も分析に必要な情報量が保持されること」、可用性は「データが更新されても匿名性と情報量の両方が維持されること」と考えることができる。

データ流通を行う際には、機微情報の取り扱いが重要な課題となる。機微情報の中でも、利用ニーズが高く注目されているのが個人情報だ。個人情報はリスクと活用のバランスを制御しないと事故が発生しかねない。個人情報を流通活用させる手段としては、(1) 同意を取得したデータを提供する、(2) 匿名加工などの加工したデータを提供する、(3) データ分析結果の公開を制御する、という3つのアプローチがある。このうち (3) は今のところ研究段階であり、現時点では (1) か (2) が選択肢となる。

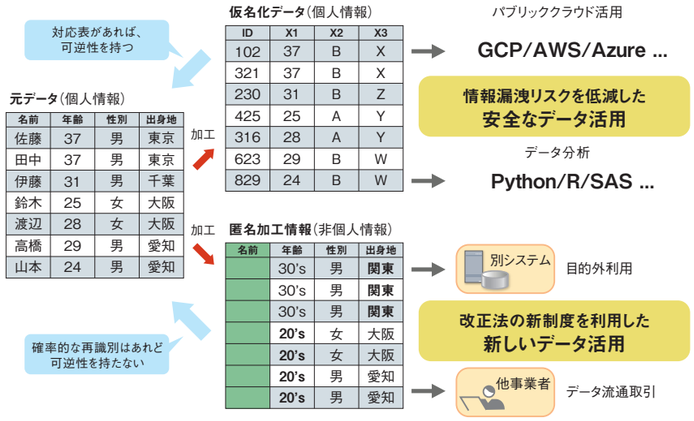
最近は「個人情報の利用権は当該個人のものである」という考え方が主流になってきているため、同意が取れる状況ならば「(1) 同意を取ること」をまず考えたい。しかしながら過去の膨大な蓄積データや、当該個人に連絡がつかない場合など、同意が取れないケースも多い。そこで次に考えるのは「利用時に個人の特定が必要かどうか」であ

1) ビッグデータのVには多様なパターンがあり、3V（Volume, Variety, Velocity）を基本として、4Vや5Vという場合もある。中には7V以上（6Vの他にVariability, Validity, Vulnerability, Volatility, Visualization）で語られる場合もある。本稿では、Thomas H. Davenport氏が唱える6Vを取り上げた
2)「すべてわかるビッグデータ大全」（日経BP：2014年）：2-5. ビッグデータ分析の精度を検証する

■図3 個人情報の受け渡し時における対応策――同意取得または匿名加工が必要



■図4 匿名化と仮名化は区別が必要――データ流通には匿名加工が有効



る。例えば、「商品Aを買った顧客がどういった年齢層かを知りたい場合」は個人特定不要だが、「商品Aを買った顧客にダイレクトメールを送りたい場合」には個人特定が必要となる。個別の同意取得が難しく、個人特定が不要である場合、改正個人情報保護法に基づいた匿名加工を施すことで、個人情報を流通活用することが可能となる(図3)。

「匿名化」は「暗号化」とは別物「仮名化」との違いに注意

匿名化と聞くと「暗号化と何が違うのか」と考える方が多いようだ。セキュリティ分野の技術という意味では近いのだが、加工の種別としては全く異なる。「暗号化」は復号を前提とした加工で可逆性を持ち、情報量は加工前後で変わらない。一方「匿名化」は復号ができない不可逆な加工で、情報量は加工前よりも減る。加工時の違いでいえば、「暗号化」や「圧縮」はデータしか見ない(意味のないByte配列でも加工できる)のに対し、「匿名化」や「統計化」はデータが持つ情報を見るため無意味なByte配列には加工を施すことができない。

また「匿名化」と「仮名化」の違い[3]にも注意が必要だ。よくある安全対策の加工として「氏名を削除しIDを置きかえる」というものがあるが、これはまさに「仮名化」である。

かつて、鉄道の乗車履歴情報が第三者提供され問題になった。この事案も「氏名やIDなどは加工されていたが、乗降駅名と乗降時刻(時分秒単位)が含まれて」おり、「仮名化以上、匿名化未満」であったと言える。なぜならば、例えば監視カメラ情報と突合すれば個人を特定できる可能性があるためだ。

近年では「仮名データは個人情報と見なすべき」というのが通説であるため、匿名化と仮名化を区別して使い分ける必要がある(図4)。

匿名化は、データの性質や提供先の用途に応じた加工を行う必要がある。このためNSSOLでは、(1) 匿名化スタートコンサルティング、(2) 概念実証(PoC)による匿名加工の適用検討、(3) 匿名加工基盤の導入構築、(4) 第三者視点による匿名化の安全性評価試験、の4つを実施している。

医療分野で始まったデータ流通の基盤づくり

近年、がん治療を中心としたゲノム情報に基づく創薬など、個別化医療(personalized medicine)の充実が急がれていることから、医療分野では他の分野に先駆け、官民を挙げてデータ流通の実現に向けた取り組みが始まっている。

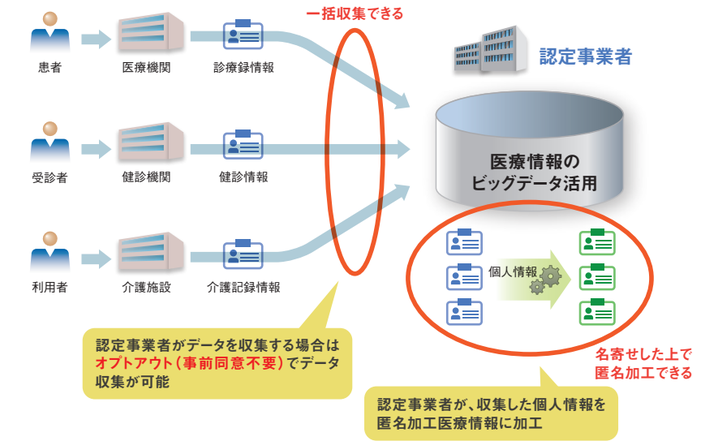
その背景にあるのが2018年5月に施行された次世代医療基盤法(医療分野の匿名加工に関する法律)だ。この法律では、国が「適切な匿名加工の能力を有する」として認定した事業者が、多数の医療機関から一括して個人情報を収集し、匿名加工を施した上でその情報(匿名加工医療情報)を研究機関や製薬などの民間企業に提供できるようになった(図5)。

当社は、日本医師会ORCA管理機構の下で医療情報や健診情報などを対象とした生涯保健情報統合基盤を構築した[4]。蓄積・加工基盤としてビッグデータ並列処理ミドルウェアであるHadoopを利用し、提供時の匿名加工エンジンとして当社システム研究開発センターで開発した「匿丸(とくまる)」を採用している。

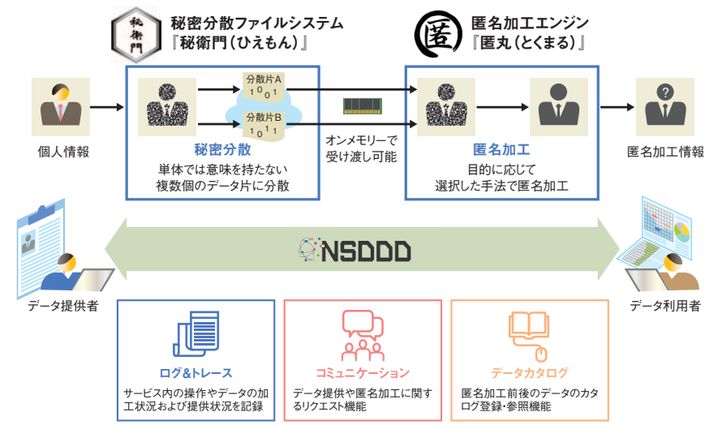
匿名加工データを流通させるには、匿名加工エンジンだけでは十分と言えない。データ提供者が個人情報を安心して渡すことができ、かつ、データ利用者が有用な匿名加工情報を受け取ることのできる環境が必要となる。

そこで当社では、社会公共ソリューション事業部とシステム研究開発センターが共同開発した「匿名加工データ流通ソリューション:NSDDD/エヌエスディースリー」を提供している。これは、匿名加工エンジンの匿丸に加え、データ提供者が安全に個人情報を保管できる秘密分散ファイルシステム「秘衛門(ひえもん)」、利用者が入手可能なデータを事前に知る手段である「データカタログ」、データの受領～加工～提供の全てを記録する「ログ&トレース」、提供者と利用者をつなぐ「コミュニケーションツール」など、データ流通に必要な要素をパッケージングしたものである(図6)。

■図5 次世代医療基盤法の下でのデータの流れ――認定事業者がデータを一括収集して匿名加工



■図6 匿名加工ソリューション「NSDDD」の構成――データ流通に必要な要素をパッケージ



システムインテグレーション(SI)からデータインテグレーション(DI)へ

今後、データ流通が活発化するに伴い、多様なデータ活用に関わるサービスが求められるはずだ。従来、システムインテグレーターが得意とするデータ基盤の構築や安定稼働をはじめ、近年注目されているデータ分析や、本稿で取り上げた匿名加工を用いたデータセキュリティの担保など、それぞれを担う様々な事業者が関わるデータビジネス経済圏の誕生が見込まれる。

SI業界は、おおよそ1980年からの20年間(HWやNWなどのITインフラ成長時代)で誕生から発展に至り、次の20年(2000-2020:SW/MWの発展やクラウド/デバイスの普及時代)で成熟し、存在を確かなものにした。多種多様なIT技術の溢れる現在において、幅広い知見や高いスキルを駆使して適切なIT技術を取捨選択し組み合わせ、顧客の複雑な業務のシステム化に応えている。

では、今後の20年間(2020-2040:AI時代・データ流通時代)には一体何が求められるのか。HW/NWの多くはクラウドに集約され、複雑なアルゴリズムはAIが担う可能性がある。高度なAIの実現には、高品質な大量かつ多様なデータが不可欠だ。すると、多種多様なデータの溢れる将来は、適切なデータを取捨選択し組み合わせる役割が求められることになるだろう。

SI業界が担う役割として、これまでの「顧客の複雑な業務をシステム化するシステムインテグレーション(SI)」だけでなく、「顧客の意思決定を支援・改善するデータインテグレーション(DI)」も加わることが予想される。

3)「仮名化」の概念を含む「個人情報の匿名性を多少なりとも上げる加工全般」を指して「(広義の)匿名化」と呼ぶこともあるが、本稿では「個人が特定できないレベルにまで匿名性を十分に上げた加工」を「(狭義の)匿名化」と呼ぶこととする

4) Key to Success 2019 Summer:Case226「医療情報の安心安全な研究利用を目指し、匿名化技術とHadoopで基盤を構築」