大統領への報告

ビッグデータとプライバシー：
技術的展望

大統領行政府

大統領科学技術諮問委員会

2014 年 5 月

大統領科学技術諮問委員会について

大統領科学技術諮問委員会（President's Council of Advisors on Science and Technology（PCAST））は、ホワイトハウス内、および政府省庁やその他の連邦機関から大統領に与えられる科学技術についてのアドバイスを増補することを目的に、大統領によって任命された米国の有力な科学者や技術者から構成される顧問団である。PCAST は、大統領が直面する政策選択に、科学、技術、イノベーションの領域からの理解が影響を与える可能性のある全問題について助言を求められ、しばしば政策提言を行う。

PCAST に関する詳細情報については、www.whitehouse.gov/ostp/pcast を参照のこと。

大統領科学技術諮問委員会

PCAST ビッグデータ・プライバシー作業部会

大統領行政府
大統領科学技術諮問委員会
ワシントン DC 20502

バラク・オバマ大統領
ホワイトハウス
ワシントン DC 20502

大統領殿

大統領科学技術諮問委員会（PCAST）が大統領閣下のために作成した本報告書『*ビッグデータとプライバシー：技術的展望*』をお送りできるのは、喜ばしい限りです。本報告書は、2014 年 1 月 17 日の閣下の要請に応じ、閣下の顧問であるジョン・ポデスタの主導の下で行われた、ビッグデータの政策への意味合いについての分析を補完し、情報を提供する目的で作成されました。PCAST はビッグデータを管理、分析するための現在の技術、およびプライバシーを保護するための現在の技術の性質について検討しました。また、これらの技術がどう進歩しているかを考察し、このような技術力や技術的動向が、ビッグデータの時代においてプライバシーを保護することを目的とした公共政策の作成や施行にどのような意味合いを持つのかを説明しました。

ビッグデータは、革新的な事業から新しい病気の治療法まで、大きな利益をもたらします。プライバシーに関する問題が発生するのは、あまりに大量のデータを（たとえば、電話から駐車場までのありとあらゆるものにあるセンサーから）収集し、それをあまりに効率的に（たとえば、データマイニングやその他のアナリティクスを通じて）分析できるため、ほとんどの人が期待していた、あるいは――進歩が続いていることを考えると――期待し得るよりもはるかに多くのことを知ることが可能になっているためです。このような問題は、プライバシーを保護するのに使用される従来の技術（非識別化など）の制限によって悪化します。PCAST は、技術だけではプライバシーを保護することができず、またプライバシーを保護するための政策は、何が技術的に実現可能か（また実現不可能か）を反映している必要があると結論づけます。

PCAST は、人々についての情報を収集、使用する方法が拡散し続けていることに鑑み、人々についての情報の具体的な*使用*がプライバシーに悪影響を与えるか否かに政策の的を絞ることを提言します。また、技術の進歩に伴って時代遅れにならないよう、結果――すなわち、「いかにして」ではなく「何」に政策の的を絞ることも提言します。政策の枠組みは、新しい技術的オプションの研究を含め、プライバシーへの悪影響を封じ込めるのに役立つ技術

の開発および商業化を加速するべきです。米国は技術をもっと効果的に使用することで、ビッグデータの利益を最大限に活用しつつ、ビッグデータがもたらすプライバシーに関する懸念を抑制するという面で、世界をリードすることができます。最後に、PCAST は、プライバシーに敏感な方法でビッグデータを開発、使用するのに必要な専門知識を持つ人材を十分に確保するための取り組みを求めます。

PCAST はこのような形で閣下および国家に奉仕できる機会に感謝申し上げ、この報告書を読んだ方々に私たちの分析が役立つことを願っております。

敬具

ジョン・P・ホルドレン
共同委員長、PCAST

エリック・S・ランダー
共同委員長、PCAST

目次

エグゼクティブ･サマリー

コンピューティングと電気通信技術の普及に伴い、デジタルとアナログの両方のソースからのデータが急増した。大量のデータの作成、分析、配布、保存を可能とする新しい能力が誕生したことで、プライバシーの性質や、個人のプライバシーを侵害または保護する手段に関する新しい懸念が提起されている。

本報告書では、まず本報告書の概略と起源を示し、次に第 1 章で、コンピューティング技術が進歩し、ビッグデータが大きくクローズアップされるようになるにつれ、プライバシーの性質がどう変化しているかについて説明する。プライバシーという用語は、有名な「ひとりにしておいてもらう権利」や、個人的な問題や関係を秘密にしておくということだけに限らず、情報を公にではないものの、選択的に共有できることも包含する。匿名性はプライバシーと重複するものの、両者は同一ではない。同様に、政府の干渉なしに個人的な意思決定を行うことは、特定の個人的特徴（たとえば個人の人種、ジェンダー、またはゲノムなど）に基づく差別からの保護と同じく、プライバシー権と見なされている。プライバシーとは単に秘密に関連するものではない。

プライバシーと新技術の対立は、米国の歴史の中で繰り返し発生している。19 世紀には新聞などのマスメディアの出現に対する懸念から、「隔離された状態への侵入」、私的な事実の公への暴露、商業目的のための氏名や肖像の盗用による損害や悪影響に対する保護が法律で定められるに至った。有線や無線通信が誕生すると、20 世紀には私信の盗聴や傍受を禁止する法律が定められたが、PCAST はこれらの法律が現在のデジタル通信技術がもたらす現実に必ずしも歩調を合わせたものではないと考える。

概してこれまでのプライバシーと新技術の対立は、現在では「スモールデータ」と呼ばれるもの――公民両セクターの組織によってデータセットが収集・使用され、データがそのままの形で配布されるか、従来の統計的手法で分析されること――に関連していた。現在のビッグデータに対する懸念は、収集されるデータ量の大幅な増加と、それに伴うデータの使用法に関する実際および潜在的な変化の双方を反映している。

ビッグデータは二つの異なる意味において「ビッグ」である。まずは、処理するのに使用可能なデータの量と多様性が大きい。次に、究極的には推測を行い、結論を導き出すことを目的に、これらのデータに適用することのできる分析（「アナリティクス（analytics）」と呼ばれる）の規模が大きい。データマイニングやその他の種類のアナリティクスによって、収集時にはプライバシーの問題をいっさい引き起こさない、または管理可能な問題しか引き起こさないと思われていたデータから、非自明で、場合によっては私的な情報が導

出され得る。そのような新しい情報は、適切に使用すると、個人や社会にしばしば利益をもたらすと考えられる（本報告書の第 2 章では、そのような事例の多くを紹介し、他の章でも追加の例に随時触れていく）。しかしながら、原則論からしても、特定のビッグデータの収集から、後にどのような情報が引き出されるかは決してわからない。というのも、一見したところ無関係なデータセットを単に組み合わせた結果、思いがけない情報が生み出されたり、データ収集時には発明さえされていなかったアルゴリズムによって、新情報が引き出されたりすることがあり得るからである。

適切に使用すれば個人や社会に利益をもたらすデータやアナリティクスが、同時に潜在的な害――プライバシーの規範に基づくと個人のプライバシーに対する脅威となるもので、広く共有される脅威と個人的な脅威の双方を含む――を生む可能性もある。たとえば、病気の研究の大規模分析は、電子カルテやゲノム情報と組み合わせると、個人にとってより質の高い適時な治療につながる可能性があるが、同時に保険や仕事に不適格であると不当に見なされる事態を招く恐れもある。個人の GPS 追跡は、共同体ベースの公共交通機関の改善につながる可能性があるが、同時に個人の位置情報の不適切使用を引き起こす恐れがある。個人の保護を必要とする悪影響や害にどのようなものがあり得るかについては、1.4 節において列挙する。ビッグデータ技術によってもたらされる利益は、新たに生じる害よりも大きい（あるいは大きくなり得る）と、PCAST は強く信じている。

本報告書の第 3 章では、個人データを一次ソースから、また事後の処理を通じて獲得する数多くの新しい方法について説明する。今日では、使用または誤用されることでプライバシーの問題を生じさせる可能性のある情報を、個人がそうとは気付かないままに、周囲に流出させている場合がある。物理的には、これらの情報の誕生場所には二つのタイプがある。「デジタル生まれ」と「アナログ生まれ」と呼べるものである。

「デジタル生まれ」の情報は、初めからコンピュータまたはデータ処理システムによる使用を意図して、私たち、またはコンピュータの代理が生成する情報である。データがデジタル生まれである場合、プライバシーの問題は過度な収集から生じる可能性がある。過度な収集は、プログラムの設計によって意図的に、またときにはひそかに、明示された目的と無関係な情報を収集することによって発生する。過度な収集は、原則として、収集時に把握することができる。

「アナログ生まれ」の情報は、物質世界の特徴から生まれる。そのような情報は、カメラやマイク、その他の工学機器などのセンサーに触れたときに、電子的にアクセス可能になる。アナログ生まれのデータは、その当面の目的に必要な最低限よりも多くの情報を含んでいる可能性が高いが、それには正当な理由がある。理由の一つは、様々な「ノイズ」が

ある中で、必要な「信号」を確実に検出しなければならないからである。別の理由としては、技術的収束、すなわち新製品（たとえば、ジェスチャーに反応することのできる住宅用警報装置など）に標準化コンポーネント（携帯電話のカメラなど）が使用されることの増加が挙げられる。

データフュージョンは、異なるソースのデータが接触し、新しい事実が浮上したときに発生する（3.2.2 節を参照)。個別には、それぞれのデータ・ソースには明確で限定された目的があるかもしれない。しかしながら、組み合わせると、新しい意味を明らかにする場合がある。とりわけデータフュージョンは、個人の識別、個人プロファイルの作成、個人の活動の追跡につながる可能性がある。もっと広く言えば、データ・アナリティクスは、次第に強力になる統計アナゴリズムを使用して、大きなデータ・コーパスの中のパターンや相関関係を発見する。これらのデータが個人データを含んでいる場合、データ・アナリティクスから導かれた推測を、今度は個人についての確実または不確実な推測に変換し直すことができるかもしれない。

データフュージョンにより、デジタル生まれのデータについては、収集時にプライバシー上の問題が認識できるとは限らない。アナログ生まれのデータについても、たとえ単一のソース（1 台のセキュリティ・カメラなど）からのデータであっても、信号処理のロバスト性と標準化により、同様のことが当てはまる。デジタル生まれのデータもアナログ生まれのデータも、データフュージョンによって結合可能で、データ・アナリティクスから新しい種類のデータが導出され得る。ほぼユビキタスなデータ収集には様々な有益な用途があり、徐々に重要度を増している一連の経済活動を活気づける原動力となっている。これらの点を合わせて考えると、データ収集の制限に焦点を絞った政策は、広く適用可能もしくは拡張可能な戦略ではなく、有益な結果と意図しないマイナスの結果（経済成長の阻害など)、との間の適切なバランスを実現する可能性も低いと言えそうである。

ほとんどの場合、収集を制限することが実際的でないとすると、どのような政策をとればよいのか。第 4 章において、これまでプライバシー保護のために用いられてきた数々の技術を紹介し、次に、程度の差こそあれ、将来の政策の基礎的技術要素として役立つ可能性のあるその他の技術について議論する。

一部の基礎的技術要素（たとえば、サイバーセキュリティ・スタンダード、暗号化に関連する技術、監査可能なアクセスコントロールの形式システムなど）については、すでに活用されており、市場において奨励される必要がある。一方で、かつては有望と思われていたプライバシー保護のための技術の中には、プライバシーのリスクを低減する補足的手段としては有用であるものの、現在では、ビッグデータに関して言えば、プライバシー保護

のための信頼できる基盤とするにはロバスト性が不十分と考えられるものもある。PCASTは様々な理由から、匿名化、データの削除、データとメタデータ（以下で定義）の区別がこのカテゴリーに当てはまると判断する。通知と同意の枠組みもまた、政策のための有益な基盤として機能しなくなりつつある。

匿名化は、まさにビッグデータを様々な形で合法的に使用するために開発されている技術によって、簡単に打ち破られるようになりつつある。概して、利用可能なデータのサイズと多様性が増すと、個人を再識別できる（すなわち、記録と氏名を再び結び付けられる）可能性が大幅に上がる。状況によっては、匿名化は付加的な保護手段として依然として有効かもしれないが、匿名化それだけを十分な保護手段と見なすアプローチは、変更が必要である。

価値がなくなったときに、あらゆる種類のデータを削除するというのは賢明なビジネス慣行であるものの、価値がないと見なされていた大量のデータにビッグデータの技法を適用することで、経済的または社会的な価値が生み出されることがしばしばある。同様に、保管データは、将来の歴史家にとって、あるいは学術研究者などによる後の縦断的分析において、重要になる可能性がある。上述のように、データ・ソースの多くは、個人についての潜在情報、すなわち所有者が分析リソースを費やして初めて判明する情報、あるいは新しいデータマイニング・アルゴリズムが開発されて初めて将来的に獲得可能となる情報を含んでいる。そのような場合、データ所有者は「個人についての全データ」を明らかにすることさえ実質上不可能で、ましてや指定されたスケジュールに沿って、あるいは個人の要請に応じて、それを削除することは全く不可能である。現在では、データストレージの冗長および分散的な性質を考えると、たとえスモールデータであっても、データを確実に破棄することが可能であるかさえ定かではない。

データセットが複雑になるにつれ、付属のメタデータも複雑になる。メタデータは、データが作成された日時や、それが作成されたデバイス、メッセージの宛先など、データの特性を記述する補助データである。データやメタデータには、多くの種類の識別情報が含まれている可能性がある。概して現在では、メタデータのほうがデータよりもプライバシーの問題を引き起こすことは少ないと断言することはできない。

通知と同意は、個々のアプリケーション、プログラム、またはウェブサービスの個人データの収集に対して、積極的な同意を与えることを個人に求める慣行である。しかし、どこかの空想の世界でない限り、ユーザーがこのような通知を実際に読んで、その意味を理解してから、同意をクリックすることはまずあり得ない。

通知と同意の概念的問題は、プライバシー保護の責任を基本的に個人に課している点である。通知と同意によって、プロバイダーとユーザーの間の暗黙のプライバシー交渉は不平等なものとなる。プロバイダーは複雑で交渉の余地のない一連の条件を提示し、ユーザーの側は実質上、その条件を評価するのに数秒の時間しか費やすことができない。これは一種の市場の失敗である。

PCAST は、ユーザーの選択に従って個人データを使用する責任は、ユーザー側ではなくプロバイダー側が負うべきであると考える。実際問題として、民間セクターでは、消費者によって選ばれた第三者（消費者保護団体、大きなアプリ・ストアなど）が仲介をすることが考えられる。消費者は仲介者が提示する複数の「プライバシー保護プロファイル」のうちの一つを選び、次に仲介者はこのプロファイルに照らしてアプリを吟味する。アプリを吟味することで、仲介者はプライバシーの共同体基準の交渉のための市場を創出することになる。連邦政府は、仲介者とアプリ開発者と販売業者の間の電子インターフェースのための基準の策定を奨励することができる。

データの収集後に、データ・アナリティクスが行われ、それによってプライバシーの問題が徐々に発生していく可能性がある。分析そのものは個人に影響を与えることはなく（分析そのものは収集ではなく、また追加の行動がなければ使用でもない）、外部から見られることもないかもしれない。対照的に、個人に悪影響を及ぼす可能性があるのは、企業、政府、報道機関、個人などによる分析の結果の*使用*である。

もっと広く言うと、影響を生むのは、データ（デジタル生まれ、またはアナログ生まれのデータ、およびデータフュージョンと分析の産物を含む）の使用の中心地であると PCAST は考える。この中心地こそが技術的にプライバシー保護の実現可能性が最も高い場所である。プライバシーポリシーを説明し、データの起源（出所）やデータへのアクセス、アナリティクスを含むプログラムによるデータの追加使用を記録し、そのような使用がプライバシーポリシーに準拠しているか否かを判断するための技術が研究され、商業界でも出現し始めている。その中にはすでに実用化されているものもある。

データ・アナリティクスの統計的性質を考えると、発見されたグループの特性がそのグループの特定の個人に当てはまるかは定かでない。個人について誤った結論を出すと、その個人に悪影響を与え、また特定のグループ（貧困層、高齢者、少数民族など）の構成員に偏った影響を及ぼす可能性がある。使用をベースとするアプローチに組む込むことができる技術的メカニズムには、データの正確性と保全性のための基準を課す方法や、個人が任意に付加情報を提供して記録を訂正することを可能とするインターフェースを組み込む方法などがある。

本調査を実施するにあたって PCAST に求められているのは、特定のプライバシー政策を提言することではなく、異なる幅広い政策アプローチの技術的実現可能性の相対評価を行うことである。したがって、第 5 章において、現在および将来の技術が政府のプライバシー保護政策にどのような意味合いを持つのかを議論する。プライバシー保護を実施するための技術的手段は、評判という圧力によって使用を促進することも可能であるが、そのような手段は、民事または刑事罰付きの規則や法律が存在するときに最も効果が高くなる。規則は有害な行動の抑止にもなれば、プライバシー保護技術を取り入れるインセンティブにもなる。プライバシー保護は技術的措置のみでは実現不可能である。

これらの議論から、以下の 5 つの提言が導き出された。

提言 1　政策の的はビッグデータの収集と分析ではなく、実際の使用に絞られるべきである。実際の使用というのは、個人または集団に悪影響または害を及ぼす可能性のある何かが発生する特定の事象を意味する。ビッグデータの文脈では、これらの事象（「使用」）はほぼ常に、未加工データまたは未加工データの分析の成果物のいずれかと相互作用するコンピュータ・プログラムまたはアプリケーションのアクションである。このような図式の中では、害を及ぼすのはデータそのものでもなければ、（データがない）プログラムそのものでもなく、両者の融合である。これらの（商業上、政府または個人による）「使用」事象は、規制の対象とするのに必要な具体性を持つ。対照的に、データの収集、ストレージ、保持、利用の先験的制限、および（データまたは分析の成果物の識別可能な実際の使用を伴わない）分析の規制に的を絞った政策は、プライバシーを改善するための効果的な戦略につながる可能性が低いと PCAST は判断している。そのような政策は時間の経過とともに拡張できる可能性も、厳格で経済的に有害な手段以外によって施行できる可能性も低いと考えられる。

提言 2　政府の全段階において、政策と規制は、特定の技術的ソリューションを埋め込むべきではなく、意図する結果という観点から明記されるべきである。

技術に後れを取るのを避けるため、プライバシー保護に関する政策は、メカニズム（「どのように」）を記述するのではなく、目的（「何」）を指定することが重要である。

提言 3　OSTP（科学技術政策室）との連携、後押しにより、NITRD（ネットワーキングおよび情報技術研究開発プログラム）機関は、プライバシー関係の技術、またこれらの技術の応用を助ける社会科学の関連分野に対する米国の研究を強化するべきである。

使用をコントロールするための技術は、すでにいくつか存在している。しかしながら、プライバシー保護を助ける技術、プライバシー保護の行動に影響を与える社会的なメカニズム、および技術の変化に対して頑健で、経済的な機会と国家的優先課題とプライバシー保護との間の適切なバランスを生み出す法的オプションについての研究（またそのような研究のための財政的支援）が必要とされている。

提言 4　OSTP はしかるべき教育機関や専門家団体と協力し、専門職の創出を含め、プライバシー保護に関連する教育やトレーニングの機会を拡大することを奨励すべきである。

（セキュリティの専門知識獲得のための教育プログラムに類似する）プライバシーの専門知識獲得につながる教育プログラムは重要で、奨励される必要がある。ソフトウェア開発の領域と技術管理の領域の両方のデジタル・プライバシーの専門職を創出することも可能かもしれない。

提言 5　米国は、現存する実用的なプライバシー保護技術の使用を奨励する政策を採用することで、国際的にも国内においても主導権を握るべきである。米国は、その集結力によっても（たとえば、基準作りや基準の採用を促進することによって）、またその調達慣行によっても（独自のプライバシー保護クラウドサービスの使用など）、リーダーシップを発揮することができる。

PCAST が承知している限りでは、米国外でより効果が高いイノベーションや戦略が生み出されているふしはない。PCAST が袋小路と考える道を進んでいるように見受けられる国もある。このような状況は、米国が国際的にプライバシー技術でリーダーシップを発揮する機会を提供しており、米国はこの機会を逃すべきではない。

1. 序論

バラク・オバマ大統領は広く注目された2014年1月17日の演説の中で、大統領顧問のジョン・ポデスタに対し、ビッグデータとプライバシーに関する包括的な調査を主導するよう指示した。この調査は、「プライバシーの専門家、科学技術者、ビジネスリーダーの見解を聞き、ビッグデータに固有の課題に公民両セクターがどのように取り組んでいるのか、ビッグデータを管理する方法について我々は国際的規範を確立することができるのか、またプライバシーとセキュリティの両方と両立する方法でどのように情報の自由流通を引き続き促進できるのかを明らかにする」ためのものであった。大統領とポデスタ顧問は、大統領科学技術諮問委員会（PCAST）に、調査の技術的側面を補佐するように依頼した。

この任務に関し、PCAST の作業指示書の一部には次のように書かれている。

> PCASTは、ビッグデータと個人のプライバシーが交差する部分について、関連する技術的能力とプライバシー問題の、現在の状態と将来考えられる状態の両方に関して技術的側面を調査する。
>
> 関連するビッグデータは、個人からまたは個人に関して、政府、民間企業、その他の個人を含めた主体者によって収集されるかその可能性のあるデータおよびメタデータである。そこには、所有権のあるデータと自由に利用できるデータの他にも、その他の活動（たとえば、環境モニタリングや「モノのインターネット」など）の間に偶然または付随的に収集された個人関連データも含まれる。

これは、とりわけ大統領が要請した野心的なタイムスケールに基づくと、難しい作業である。ビッグデータとプライバシーについての文献と公の議論は膨大で、産業界や学術界の科学技術者、プライバシー保護や消費者の擁護者、法学者、ジャーナリストなどの多様な有権者から、日々新しい考えや見識が提示されている。PCASTとは無関係に、ただし本報告書のための情報収集を目的とし、ポデスタ調査は全米の大学で三回の公開ワークショップを主催した。本報告書の責任範囲が問題の政策的側面ではなく、技術的側面に限定されていることで、PCASTの作業範囲は幾分狭まるものの、これは技術と政策の区別が難しい主題である。いずれにせよ、この主題の性質上、本報告書は科学技術の瞬間的なスナップショットに基づくものと見なされるべきであるが、主要な結論と提言は永続的な価値を持つと信じる。

1.1 本報告書の文脈と概要

コンピューティングと電気通信技術の普及に伴い、デジタルとアナログの両方のソースからのオンラインデータが急増した。大量のデータの作成、分析、配布を可能とする新しい技術力は、プライバシーの性質や、個人のプライバシーが侵害または保護される手段についての新たな懸念を生じさせている。

本報告書では、このいわゆる「ビッグデータ」にかかわる現在および将来の技術を、プライバシーに対する懸念との関連において論じていく。本報告書は、ビッグデータにかかわる技術の完全な要約でも、技術がプライバシーに与える影響についての完全な要約でもなく、ビッグデータとプライバシーが互いにどう影響し合うのかに焦点を当てたものである。たとえば、レスリーが秘密をクリスに打ち明け、クリスがその秘密を電子メールか携帯メールで広めた場合、それは情報技術のプライバシー侵害的使用になるかもしれないが、ビッグデータに関連する問題ではない。別の例を挙げると、海洋学データがリモートセンシングによって大量に収集されたならば、それはビッグデータであるが、プライバシー上の問題ではない。データの中には他と比べてプライバシーにセンシティブなものもあり、たとえば個人の医療データは、同じ個人が公に共有している個人データとは異なり、プライバシーにセンシティブである。異なる種類のデータには、異なる技術や政策が適用される。

本報告書で使用するビッグデータという概念と個人のプライバシーという概念は、意図的に幅広く、包括的なものにしている。ビジネスコンサルタントのガートナー社（Gartner, Inc.）は、ビッグデータを「見識と意思決定の向上のために費用効率の高い革新的な情報処理形態を必要とする高容量、高速、高多様性の情報資産」と定義しており、一方で様々な定義を検証しているコンピュータ科学者は、「NoSQL、マップリデュース（MapReduce）、機械学習などの一連の技法を用いた、膨大および／または複雑なデータセットの保存と分析を指す用語」という、より専門的な定義を打ち出している。（これらの専門用語の議論については、3.2.1節および3.3.1節を参照。）プライバシーの文脈においては、「ビッグデータ」という用語は、一個人または個人の集団に関するデータ、または分析を行うことで個人についての推論が可能なデータを意味することが一般的である。政府、民間企業、または個人によって収集されるデータまたはメタデータを含むこともある。そのデータおよびメタデータは、所有権があるものも自由に利用できるものもあり、意図的、偶然または付随的に収集されたものもある。テキスト、オーディオ、ビデオ、センサーベース、またはこれらの一部の組み合わせであることもある。あるソースから直接収集されたデータである場合も、また分析のなんらかの過程によって導出されたデータである場合もある。長期間にわたって保存されるものも、ストリーミングと同時に分析され

廃棄されるものもある。本報告書においてPCASTは、ほとんどの場合、「データ」と「情報」の区別をしない。

「プライバシー」という用語は、人目を避ける、あるいは個人的な問題や関係を秘密にするということに限らず、情報を選択的に、ただし公にではなく共有できるということも包含する。匿名性はプライバシーと重複するものの、両者は同一ではない。投票はプライベートなものと認識されているが、匿名と認識されているわけではなく、政治的なパンフレットの著述は匿名かもしれないが、プライベートなものではない。同様に、政府の干渉なしに個人的な意思決定を行うことは、特定の個人的特徴（たとえば個人の人種、性別、またはゲノムなど）に基づく差別からの保護と同様、プライバシー権と見なされている。よって、プライバシーは単に秘密に関連するものではない。

ビッグデータの収集と分析に対する期待は、導出されたデータを個人にも社会にも資する目的に利用できるところにある。プライバシーに対する脅威は、収集または導出された個人データの意図的または不注意による公開、データの悪用、および導出されたデータが不正確または誤っている可能性があることから発生する。これらの問題点のすべてに対処する技術が、本報告書のテーマである。

序論である本章の残りにおいては、プライバシーの法的概念が米国において歴史的にどのように発展してきたかの背景をさらに要約の形で示していく。興味深いことに、また本報告書に関連することであるが、プライバシー権と新技術の発展は、長年にわたって互いに切り離せない結びつきを持っていた。現代の問題も例外ではない。

本報告書の第2章では、シナリオと実例を挙げる。その中の一部は現在のものであるが、大多数は近い将来を予測したものである。ヨギ・ベラ（Yogi Berra）のしばしば引用される発言――「予測をするのは難しい。未来についてはなおさらだ」――は、当を得ている。しかしながら、本報告書のテーマに関して言えば、時代遅れの実例やシナリオに基づいた政策は失敗に終わる運命にあるということも確かである。ビッグデータ技術はあまりに急速に進歩しているため、いかに不完全であろうとも、将来についての予測を現在の政策展開の指針とせざるを得ないのである。

第3章では、ビッグデータの二本柱である収集と分析の技術的側面を検証する。ある意味において、ビッグデータはまさにこの二つの合流点――すなわち、「ビッグな」収集と「ビッグな」分析（しばしば「アナリティクス」と呼ばれる）が結合したものであると言える。「ビッグ」を可能とする大規模なネットワーキングとコンピューティングの技術インフラについても取り上げる。

第4章では、プライバシー保護のための技術と戦略について考察する。技術は問題の一部であるかもしれないが、解決策の一部でもなければならない。現在および予見可能な技術の多くはプライバシーの強化を実現できるもので、その他にも有望な研究領域が数多くある。

第5章には、それまでの章の議論を踏まえたPCASTの展望と結論を盛り込む。特定の政策を提言するのは本報告書の責任の範囲内ではないものの、ある種の政策は他と比べて技術的実行可能性が高く、新しい技術が出現しても不適切になったり実行不可能になったりする可能性が低いことが明らかである。そのような政策を強調した上で、その他の政策の技術的欠陥についてのコメントを添える。第5章にはまた、我々の責務に沿った、すなわち政策以外の領域についてのPCASTの提言も盛り込む。

1.2 長年にわたってプライバシーの意味を決定づけてきたのは技術

プライバシーと新技術の間の対立は新しいものではない。ただし、現在では以前よりも対立の範囲が広く、密接さや波及の度合いが大きいかもしれない。二世紀以上も前から、プライバシーに関連する価値観や期待は、新技術の影響に鑑みて、絶えず再解釈され、再提示され続けてきた。

ベンジャミン・フランクリンによって提唱され、1775年に設置された全国的な郵便制度は、各州間の通商を促進することを目的とした新技術であった。しかしながら、郵便は運搬中に定期的にも日和見的にも開封されたため、議会は1782年にこの行為を違法とした。合衆国憲法修正第4条は、自宅にいる人またはその人格に与えられるプライバシー保護の強化を成文化したものであるが（以前は英国のコモン・ローの原則）、プライバシー権の概念が電子を含むもっと抽象的なスペースへと拡大するのは、もう一世紀の技術的挑戦を経た後のことであった。電報、さらに後年の電話の発明によって、新しい緊張が生まれたものの、それが解消されるには長い歳月を要した。電報のプライバシーを保護するための法案は1880年に議会に提出されたものの、可決されることはなかった。

しかしながら、ウォーレン（Warren）とブランダイス（Brandeis）の1890年の論文『プライバシー権』（The Right to Privacy）を生むきっかけとなったのは、通信ではなく、消費者が操作できる持ち運び可能なカメラ（すぐにコダックとして知られるようになる）の発明であった。当時、この論文は議論を呼んだものの、今日では現代プライバシー法の基本文書と見なされている。この論文の中で、ウォーレンとブランダイスは、「瞬間写真と新聞事業がプライベートや家庭という神聖な領域に侵入し、様々な機械装置が『クローゼ

ットでのささやきが屋上から外に流される』という予言を現実のものとする脅威を生み出している」と述べ、さらに「私人の肖像の無断流通に対するなんらかの救済を法律によって与えるべきであるという意見が数年前からある……」と指摘している。

ウォーレンとブランダイスは、個人の間のプライバシー権を明確にすることを目指した（その基礎は民事不法行為法にある）。現在では、民事または刑事訴訟の訴因として、いくつものプライバシー関連の被害を認めている州が多い（1.4節でさらに詳しく論じる）。

ウォーレンとブランダイスの『プライバシー権』から75年経過した後に、最高裁判所は「グリズウォルド対コネティカット事件」（Griswold v. Connecticut）（1965年）において、（多数意見を著したウィリアム・O・ダグラス判事の言葉を借りるなら）その他の憲法上の保護の「半影」（penumbras）および「放射」（emanations）の中に、プライバシー権を見出した。現在の学者は、広い視野に立って、「プライバシー」の異なる法的意味を多数認めている。そのうち、以下の五つについては、このPCASTの報告書にとりわけ関連性があると考えられる。

(1) 秘密を守る、または隔離を求める個人の権利（1928 年の「オルムステッド対合衆国事件」（Olmstead v. United States）におけるブランダイスの反対意見の中の有名な「ひとりにしておいてもらう権利」）。

(2) とりわけ政治演説（ただし、それに限定されない）における匿名表現の権利（たとえば「マッキンタイア対オハイオ選挙委員会事件」（McIntyre v. Ohio Elections Commission）で示されたもの）。

(3) 個人情報が人の占有を離れた後に、他者によるその個人情報へのアクセスをコントロールする能力（たとえば、FTC（連邦取引委員会）の公正な情報取扱い原則（Fair Information Practice Principles）で明確にされているものなど）。

(4) 個人情報の使用によってある種類の悪影響が及ぶのを防止する（たとえば、2008 年に遺伝情報無差別法（Genetic Information Nondiscrimination Act）によって禁止された個人の DNA に基づく仕事上の差別など）。

(5) 健康、生殖、セクシュアリティーなどの領域において、政府の干渉なしに、私的な意思決定を行う個人の権利（グリズウォルド事件）。

これらは絶対的ではなく、主張された権利である。いずれも制定法と判例法の両方で支持されているが、同時に制限もされている。上のリストの(5)（「情報プライバシー」とは異なる「意思決定プライバシー」権）を除いて、どれも程度の差はあれ、市民と政府の相互

作用にも市民と市民の相互作用にも適用可能である。新技術とプライバシー権の衝突は、上の五つのすべてで発生している。これまでは州法と連邦法を寄せ集めて、多くの部門の問題に対処してきたものの、これらの問題を扱う包括的な法律はまだ存在しない。新技術とプライバシー権の衝突は、今後も発生すると考えられる。

1.3 現在における相違点

新しい技術力が急速に出現するのに伴って、技術とプライバシーの新たな衝突は顕著になってきている。上記で提起した五つのプライバシーに対する懸念、あるいはその現在の法解釈が、世論という裁判所において十分であるということはもはや明白ではない。

市民の懸念の多くは、個人データの単独使用または併用によって生じる害に向けられたものである。これまでは、本人の占有を離れた後の個人データに対するアクセスのコントロールが、潜在的な害をコントロールする手段と見なされていた。しかし現在では、個人データはその本人の所有下にはない、あるいは一度もなかったという場合があり得る。たとえば、公共のカメラやセンサーといった外部ソースから受動的に、あるいは本人が知らないままに、ソーシャルメディアを使用する他者による電子情報の公開から個人情報が獲得される場合がある。加えて、本人がその使用や結果について承知していない強力なデータ分析（3.2節を参照）から個人データが導出されることもある。このような分析は、その個人が開示を望まない正しい結論を導き出すこともある。さらに悪いことには、分析は誤検出や検出漏れ――すなわち分析の結果であるものの、真実でも正確でもない情報――を発生させる可能性がある。さらに、どのような目的で、どのような文脈において使用されるかによって、同じ個人情報の使用が有益な場合と有害な場合があり得る程度も以前よりはるかに増している。個人によって提供された情報が識別や相互関係などのその他の情報を導出するためだけに使用され、その後はその個人情報は不要となる場合もある。導出されたデータは、個人の支配下にあったことは一度もなく、導出後に良い目的にも悪い目的にも使用される可能性がある。

現在の議論の中で、プライバシー保護に関連する問題は、とりわけ公民権の領域においては、個人的なものも集団的なものもあると主張する者がいる。たとえば、ビデオから顔認証を用いて集会に出席している特定の個人を識別するという問題や、同じ集会に出席しており、同様にビデオから識別したその他の個人について、類似の意見を持つ、あるいは類似の行動をとると推測するといった問題がある。

現在の状況は、公共広場、あるいはパーティや教室などの準私的な集まりにプライバシー権をどのように拡大するかという問題も提起する。これらの場所を監視するのが単に人で

はなく、電子公開と分析に対する忠実度が高く、電子公開と分析に容易につながる記録装置――人目につくものと人目につかないものの両方――であったとしたら、原則は変わるのか。

同様に急速に変化しているのは、個人のプライバシーに対する潜在的な脅威としての政府と民間セクターの違いである。政府は単なる「巨大企業」ではない。権力を独占しており、市場での優位性をめぐって争い、それゆえに誤りを正すよう動機付けを与える直接のライバルもいない。政府には抑制と均衡があり、それによって人々の情報をどう扱うかに自主制限がかかることもある。企業は競争上の優位性やリスク、政府の規制、予測される訴訟の脅威と結果などの要素に鑑みて、そのような情報をどう使用するかを決定する。したがって、公共セクターと民間セクターには異なる制約があることが適切である。しかし、政府は――とりわけ法執行と国家安全保障の分野において――権限を有しており、それによって他に類を見ない強力な地位を占めているため、政府によるデータの収集と使用にどのような制約を課すかは特に注目に値する。実際、公開データと私的データの区別は次第にあいまいになっているため、政府への制約に対する注目はなおさら必要になっている。

このような違いは紛れもないものである一方、ビッグデータはある程度まで、政府と企業の間の違いを取り除く。政府も企業も同じデータ・ソースと同じ分析ツールに対する潜在的なアクセスを持つ。現在のルールでは、場合によっては政府自身が合法的に収集することができないデータを民間セクターから購入またはその他の方法で入手したり、政府自身では合法的に行うことができない分析を民間セクターに委託したりすることが可能かもしれない。適切なセーフガードなしに、政府が自身の独占力を行使し、同時に民間の情報市場に自由なアクセスを持つ可能性は不安を喚起する。

どのような行動を政府（連邦政府、州政府、地方自治体、また法執行機関を含む）と民間セクターの両方に対して禁じるべきか。どのような行動であれば、一方には禁じるべきで、一方には禁じるべきではないのか。現在の法的枠組みが現在の問題に対処する上で十分に頑健であるか否かは明らかでない。

1.4 価値観、侵害、権利

1.2節と1.3節で見てきたように、新しいプライバシー権は通常、学究的な抽象概念として出現するわけではない。むしろ、幅広く共有されている価値観が技術によって侵害されるときに生まれる。価値観についてコンセンサスがある場合、個人に対するどのような侵害がその価値観を傷つけるのかについてのコンセンサス形成も可能である。そのような侵害

のすべてが政府の行為によって予防または救済可能とは限らないが、逆に政府の行為が広く共有されている価値観にある程度まで基づいていない限り、それが歓迎される、または奏功する可能性は低い。

プライバシーの領域では、ウォーレンとブランダイスが1890年に（1.2節を参照）プライバシーについての対話を始めたことで、学術界および裁判所でのプライバシー権の発展につながり、後にこれをウィリアム・プロッサー（William Prosser）が法的保護の対象となる四つの異なった種類の侵害として結晶化させた。その直接的な結果として、現在では多くの州がプロッサーによって挙げられ、（州ごとに程度の差こそあるものの）プライバシー「権」となった四つの侵害を訴訟理由として認めている。その侵害とは以下のとおりである。

- 隔離に対する侵入。他者の孤独もしくは隔離、または私的な事情や問題に物理的または（現在では電子的を含む）別の形で意図的に侵入する者は、プライバシーの侵害に対する責任に問われ得る。ただし、それはその侵入が理性的な人にとって極めて不快に感じられるものである場合に限られる。
- 私的な事実の公への暴露。同様に、他者の私的な事実を公表した場合、たとえそれが事実であっても、告訴され得る。私的な事実とは、それまでは公にされておらず、正当な社会的関心事ではなく、理性的な人にとっては不快であると考えられる人の私生活についての事実である。
- 「虚偽の光（False light）」または公表。名誉棄損に密接に関連するこの侵害は、個人についての虚偽の情報が広く公表された結果として発生する。一部の州では、虚偽の光は単に虚偽の情報そのものだけでなく、虚偽の示唆も含む。
- 氏名または肖像の盗用。個人は商業的環境において自己の氏名または肖像を管理する「肖像権」を有する。

現在でも大多数の米国人は、たとえ法律用語（現在までに幾千もの判決文の中で磨かれてきた）が古風で奇妙に感じられたとしても、これらの侵害の中に潜在する価値観を共有し続けていると考えられる。しかしながら、新技術によるプライバシーへの実際または予想される侵害、加えて一世紀にわたる社会的価値の進化（たとえば、現在ではマイノリティーの権利やジェンダーに関連する権利の承認が拡大している）を考えると、1960年には十分であったプロッサーのリストをもっと長くせざるを得ない可能性がある。

本調査においてPCASTが焦点を合わせるべきは技術であって、法律ではないものの、プライバシーをテーマとする報告書は、PCASTの報告書を含め、すべてその時代の価値観に立

脚したものである必要がある。技術に詳しい米国の一部の者による断片的な見解に過ぎないものの、PCASTは議論の出発点として、既存の四類型リストに追加可能な項目を以下に提示する。これらの追加の害はそれぞれ、ビッグデータの時代における潜在的な権利を示唆している。

PCASTはまた、技術によってもたらされる利益は、新たに生じる害よりも大きい（あるいは大きくなり得る）と確信している。新しい害のほとんどは、同じ技術の有益な使用に関連、あるいは「近接」している。この点を強調するため、提示する新しい害のそれぞれについて、関連する有益な使用を併記する。

- **私信の侵害。**デジタル通信技術は、地理的境界を越えたソーシャル・ネットワーキングを可能にし、それまでは想像できなかった規模での社会的および政治的参加を実現する。しかしながら、私信に対する個人の権利は、郵便と有線電話についてはその配達インフラの隔離などによって保障されるが、デジタル時代においては再確認が必要かもしれない。デジタル時代においては、あらゆる種類の「ビット（bits)」が同じパイプラインを共有しており、傍受の障壁がはるかに低いことが多いからである。（これに関連して、4.2節において暗号化の使用と制限について取り上げる)。

- **人のバーチャル・ホームにおけるプライバシー侵害。**憲法修正第4条は、家の中にある私的な記録の保護など、政府による家への介入に対して特別な保護を与えており、不法行為法は非政府組織による類似の介入に対して保護を与えている。新しい「バーチャル・ホーム」には、インターネット、クラウドストレージ、その他のサービスが含まれる。クラウド上の個人データはアクセスし、整理することが可能である。クラウド上の写真や記録は、家族や友人と共有し、将来の世代に受け継がせることが可能である。基盤にある「家は自分の城」という社会的価値観は、論理的には「クラウド上の城」にまで広げられるべきであるが、この保護は新しいバーチャル・ホームでは維持されていない。（このトピックについては、2.3節においてさらに詳しく取り上げる)。

- **推測された私的な事実の公開。**強力なデータ・アナリティクスによって、一見したところ無害な入力データから私的な事実を推測できる場合がある。その推測は有益な場合もある。たとえば、ターゲットを絞った宣伝によって、消費者は自分が実際に欲求または必要としている製品にたどり着くことができる。また、人の健康に関する推測によって、より質の高い適時な治療や長寿につながる可能性がある。しかしながら、ビッグデータの出現前には、公開情報と個人情報との間には明確な違いがあると想定することができた。すなわち、事実は「そこにある」（そして、指し

示すことができる）か、そうでないかのいずれかであった。現在では、アナリティクスによって、昨日までは完全に私的な生活領域に属すると考えられていた私的な事実が突き止められる可能性がある。たとえば、購入パターンから性的嗜好を推測する、またはキーストリームおよびクリックストリームから初期のアルツハイマー病を推測するなどがある。後者の場合には、本人さえもその私的な事実を承知していないことがあり得る。（そのような推測を可能とするデータ・アナリティクスの背後にある技術については、3.2節で論じる）。そのような情報を公開する（そして、なんらかの非公共的な商業目的で使用する）のは、広く共有されている価値観に反することと考えられる。

- **追跡、ストーキングおよびロケーション・プライバシーの侵害。**今日の技術を使用すれば、個人の現在地または以前の位置を容易に特定することができる。便利な位置情報サービスには、ナビゲーション、より良い通勤ルートの提案、近くにいる友達を見つける、自然災害の回避、近くで商品またはサービスが入手可能であることを宣伝するなどが含まれる。公共の場で個人を見かけたという情報が私的な事実になることはまずあり得ない。しかしながら、ビッグデータによって、そのような目撃情報、あるいは他の種類の受動的または能動的データ収集を組み合わせることで、個人の私生活の絶え間ない位置追跡が可能であるとすれば、多くの米国人（その中には、最高裁判所のソトマイヨール（Sotomayor）判事が含まれる）は、広く受け入れられている「プライバシーの合理的な期待」からかけ離れている可能性があると考える。

- **ビッグデータ・アナリティクスから導出された個人プロファイルに基づく、個人についての誤った結論が生み出す侵害。**ビッグデータの力、またそれによってもたらされる利益は、相互に関係していることが多い。多くの場合、統計誤差による「害」は小さい。たとえば、映画の好みに関する誤った推測や、たとえ特定の事例については誤った警報であることが判明したとしても、平均的には有益であると考えられる分析に基づいた「健康問題について医者に相談するべき」という提案などがそうである。加えて、予測が統計的に妥当であった場合も、特定の個人については正しくない場合があり、誤った結論が害を及ぼす可能性がある。統計的に妥当なアルゴリズムにつきものの不確実性によって引き起こされる害に対して、社会は寛容でないかもしれない。そのような害が人種的少数派や高齢者などの特定の層の個人に不当な重荷を負わせる可能性がある。

- **個人の自律または自己決定の妨害。**大きな母集団についてのデータ分析により、その母集団内の個人に当てはまる特殊な事例が発見されることがある。たとえばビッグデータは、「学習法」の違いを特定することで、各個人の潜在能力を踏まえ、その個人の達成度を最大限に高められるように教育をパーソナライズすることを可能

にする。しかしながら、母集団のファクターを個人に投影することは、悪用を招く恐れがある。個人は自分で選択をし、必ずしも典型的ではない機会を追及できるということ、また誰もが統計的な期待よりも大きな功績を収める機会を拒まれないということは、広くコンセンサスを得ている。子供がどのテレビゲームを選択したかが後に教育的追跡（大学の入学など）に用いられたとしたら、それは私たちの価値観に反することである。スティーブン・スピルバーグの映画『マイノリティ・レポート』の原作であるフィリップ・K・ディックのSF短編小説で描かれた、「プレ・クライム（pre-crime）」を統計的に識別し、処罰の対象にするという未来も同様に私たちの価値観に反する。

- **匿名性と私的な交友関係の喪失。**匿名性は、詐欺やいじめ、サイバーストーキング、または児童との不適切な交流を行うことを目的とした場合には許容されない。しかし、不正行為を除けば、匿名を選択する個人の権利は、米国ではかねてより尊重されてきた（たとえば、フェデラリスト・ペーパーズを匿名で執筆するなど）。匿名を希望する個人を（再）識別するためにデータを使用することは、有害と見なされる（法の執行などの政府の合法的な機能の場合を除く）。同様に、個人はグループまたは他の個人と私的な交友関係を持つ権利を有しており、そのような交友関係の識別は有害になる可能性がある。

上のリストは完全なものであることを意図しないが、故意に掲載しなかった項目もある。たとえば、人を平等に扱うという意味において、ビッグデータが「公正に」使用されることを希望する人もいるかもしれないが、（すでに法律が定められて保護の対象となっている少数の集団を除き）これを十分に意味のある具体的な権利にすることは不可能と思われる。同様に、他者が自分について何を知っているかを知ることを望む人がいるかもしれないが、これはデジタル以前の時代から間違いなく権利ではなく、また統計分析が進んでいる現代では、「知っている」の定義もそれほど容易ではない。この重要な問題については3.1.2節で論じ、第5章でも再度取り上げるが、その際には、情報が「知られているか」否かといった論理的にあいまいな概念によって生じる害ではなく、情報の*使用*によって生じる実害に焦点を絞る。

2. 実例とシナリオ

本章では、いくつかの実例とシナリオを紹介することで、第1章の導入的議論をより具体的にすることを目指す。この中には現在使用されている技術もあれば、現在から10年後くらいまでの近未来に出現するとPCASTが予測する技術も含まれる。これらの実例とシナリオはすべて合わせて、ビッグデータがもたらし得る多大な利益と、これらの利益に伴って発生する可能性のあるプライバシーに関する課題の両方を明らかにすることを意図するものである。

以下の3つの節では、まず現在何が発生しているかを説明するためのごく短い実例を示してから、未来を舞台にした、より詳細なシナリオに移る。一部のシナリオについては他よりも詳細なものとなっている。

2.1 現在または非常に近い未来に発生すること

関連する実例をいくつか挙げる。

- 10年以上前に開発された電柱に取り付けるデバイスによって、電柱の脇を通過するドライバーが聞いているラジオ局を探知することができ、その結果を広告主に売却する。
- 2011年には、調査した警察署の4分の3で、ナンバープレート自動読み取り機が使用されていた。5年以内に、25%の警察署がこのデバイスをすべてのパトロールカーに搭載し、発行されている令状に関係した車両が視界に入ったら、警察に警告する仕組みにすると予想されている。同時に、ナンバープレート自動読み取り機は民間でも使用されるようになってきており、クラウドのプラットフォームを利用し、収集された情報を様々な用途に使用することが見込まれている。
- マサチューセッツ工科大学の専門家とケンブリッジ警察署は、機械学習アルゴリズムを用いて、どの窃盗が同じ犯罪者による犯行であった可能性が高いかを識別し、警察の捜査を補助した。
- 航空券や大学の費用などの領域では、差別的価格設定（事実上同じ商品に対して、顧客ごとに異なる価格を提示すること）が珍しくなくなってきている。ビッグデータは、この慣行が浸透し勢いを増すのを後押しし、さらにはその透明性のさらなる低下に拍車をかけるかもしれない。
- 英国の企業フィーチャースペース（FeatureSpace）は、オンラインプレーヤーのギ

ャンブル依存またはその他の異常行動の初期兆候を探知することのできる機械学習アルゴリズムをゲーム業界に提供している。

- CVS やオートゾーン（AutoZone）などの小売業者は、顧客の買い物のパターンを分析して、店のレイアウトを改善し、場所ごとに顧客のニーズに合った商品を置いている。オンライン業者がクッキーによって再来の顧客を認識するように、リテールネクスト（RetailNext）は携帯電話を追跡することで、ブリック・アンド・モルタル企業が再来の顧客を認識できるようにしている。類似の Wi-Fi 追跡技術を使えば、閉ざされた部屋に何人の人がいるのか（また場合によってはその人たちの身元）を突き止めることができるかもしれない。

- 小売業者のターゲット（Target）は、十代の顧客が妊娠していると推測し、善意でクーポンを郵送したところ、彼女の父親に意図せず妊娠の事実を知らせてしまった。

- 匿名の本、雑誌記事、ウェブ投稿などの筆者は、多くの無関係の個人の好奇心の結果、非公式のクラウドソーシングによって、しばしば身元を「暴露」されている。

- ソーシャルメディアや記録の公的なソースにより、ウェブを積極的に使用する人の大多数、およびそうでない人の多くについて、どのような交友関係があるのかを推測することが誰にとっても容易になった。

- ニューヨークのポキプシーにあるマリスト・カレッジは、予測モデリングを用いて、ドロップアウトのリスクのある大学生を特定し、助けを必要としている学生に集中的に追加の支援が行えるようにしている。

- 国防総省が資金を提供するデュルケーム（Durkheim）プロジェクトは、ソーシャルメディア上の行動を分析して、退役軍人の自殺念慮の初期兆候を察知している。

- カリフォルニアに拠点を置く新興企業レンドアップ（LendUp）は、ソーシャルメディアなどの非従来型のデータ・ソースを用いて、十分なサービスを受けていない個人に信用貸しを行うことを目指した。しかしながら、正確性と公平性を担保するのが困難なため、事業は進展していない。

- 大量の患者データと、未感染の患者および臨床スタッフに関する個人情報を組み合わせて使用することで、院内感染の広がりに関する見識を得た。

- 1 回の拍動に伴って顔色がわずかに変わることから、個人の心拍数を推測し、健康と情動状態についての推測を可能にする。

2.2 近未来のヘルスケアと教育の分野のシナリオ

以下に、ただちに組み立て可能な種類のシナリオの例を挙げる。

2.2.1 ヘルスケア：オーダーメイド医療

同じ病気の患者は全員が似ているわけではなく、治療の効果も患者によって異なる。研究者は近い将来に、何百万もの健康レコード（デジタルデータに加え、スキャンなどのアナログデータも含む）、大量のゲノム情報、臨床試験の成功および失敗に関する豊富なデータ、病院の記録などを利用できるようになる。病気の多様な兆候の中でも、特定の治療計画に反応する変異型を形成する一連の特性を一部の患者が持つことを識別できる場合もあると考えられる。

分析結果は特定の患者により良い結果をもたらす可能性があるため、そのコホートの個人を識別して連絡を取り、新しい方法で病気を治療し、その経験を研究の前進のために使用することが望まれる。しかしながら、そのデータは匿名で収集されたにすぎない場合や、非識別化されていた場合が考えられる。

解決策は、データベースのプライバシー保護のための専用の新技術によって提供されるかもしれない。これによって、保護されたクエリーメカニズムを作り、本人がコホートの中に自分が入っているかを確認できるようにする、あるいはコホートの特性に基づいた警告メカニズムを作り、医療専門家がそのコホートの患者を診察する際に通告が出されるようにすることができるかもしれない。

2.2.2 ヘルスケア：モバイル機器による症状の察知

団塊の世代の人の多くは、自分がアルツハイマー病であることを自分で察知できるかを気にしている。自分の行動を観察する手段としては、クラウド上のパーソナル・アシスタント（シリ（Siri）やオッケーグーグル（OK Google）など）に接続し、ナビゲートを助け、言葉の意味を教え、やるべきことを記憶し、会話を思い出させ、歩行を測定するなどして、配偶者でさえ気づけない従来の医学指標および新しい医学指標の漸減を察知することのできるモバイル機器があれば、それに勝るものはない。

同時に、そのような情報の漏えいは、有害な信頼の裏切りとなる。そのようなリスクに対する個人の保護はどのようなものか。個人の健康に関する推測情報は、追加の同意なしに第三者（たとえば製薬会社など）に売却することができるか。これがアプリの使用条件と

して明記されていた場合はどうか。当初の同意はあるが、事後の確認がない場合にも、個人のかかりつけの医師にその情報を伝達するべきか。

2.2.3 教育

大規模公開オンライン講座（MOOC）やそれよりも小規模のクラスを含めたオンライン講座の数百万ものログを利用することで、近い将来に数百万人の学習者の能力や学習法に関する時系列データを作成・維持することが可能になる。これには成績などの幅広い総合情報のみでなく、学習者の一人一人が複数の新しい教授法にどう反応したか、抽象度が異なる概念を習得するのにそれぞれどれくらいの助けを必要としたか、様々な状況の中で注意力がどれくらい持続したかなどの細粒度プロファイルも含まれる。MOOCのプラットフォームは、学習者が特定のビデオをどれくらいの時間見ていたか、あるセグメントを何度繰り返したり速度を速めたりスキップしたりしたか、小テストの結果はどうだったか、特定の問題を何度間違えたか、コンテンツを見る時間とテキストを読む時間のバランスはどうだったかを記録することができる。プラットフォーム上で異なる教材を異なる学習者に提供することが可能になるにつれ、目隠しの無作為化A/Bテストの可能性によって、実験科学の黄金基準がこの環境の中で大々的に実行できるようになる。

学習管理システム（キャンバス（Canvas)、ブラックボード（Blackboard)、Desire2Learnなど）の役割が拡大し、革新的な教授法が支援されるようになるにつれ、教室での授業についても類似のデータが入手できるようになってきている。現在では、多くの講座において、学習者の教材への取り組み方を逐一追跡し、その取り組み方を目標とされる学習の成果と相関させることが可能になっている。

このような情報があれば、教育を大きく改革できるだけでなく、どのようなスキルをどの個人に幼年期のどの段階で教えれば、成人してからの特定のタスクにおけるパフォーマンスの改善につながるか、あるいは成人してからの個人的および経済的成功につながるかの知見が得られる。このようなデータは教育研究に革命をもたらす可能性がある一方、プライバシーの問題は複雑になる。

この未来の教育の展望には、多くのプライバシー上の課題がある。初期の成績についての情報は、後の指導とカウンセリングに影響を与える暗黙のバイアスを生む。生徒を将来性の高い道か将来性の低い道へのいずれかに振り分けることのできる巨大な力は、表向きは社会のためであっても、悪用される大きなリスクをはらむ。保護者などは子供についてのセンシティブな情報へのアクセスを持つが、子供が成年に達したときに、これらの許可を変更するためのメカニズムはほとんど存在していない。

2.3 家の特別な地位に対する挑戦

家は個人のプライバシーの聖域として特別な重要性を持つ。憲法修正第4条の「身体、家屋、書類および所有物」というリストは、身体のみをレトリック的に目立つ位置に置いている。家はその他の三つの物理的入れ物であり、その境界の内側に強化されたプライバシー権が適用される。

しかしながら、現在では、憲法修正第4条のこれまでの解釈では不十分である。私たちも、修正第4条が意図する「書類および所有物」も、徐々にサイバースペースに移動するようになってきており、家という物理的な境界は意味がなくなりつつある。1980年には家族の財政記録は紙の書類で、おそらくは家の中の机の引き出しにしまわれていた。2000年には、家のコンピュータのハードドライブの中に移動したが、それでもまだ家の中にあることに変わりはなかった。2020年までには、そのような記録のほとんどはクラウド上に保存されるようになる可能性が高い。すなわち、単に家の外にあるだけでなく、複数の法域で複製される可能性が高いということである。というのも、クラウドストレージは、信頼性を実現するためにロケーション・ダイバーシティを使用するのが一般的であるためである。財政記録を「購入した政治的な書籍」「受け取ったラブレター」「鑑賞したアダルトビデオ」などに置き換えても、同様のことが言える。異なる政策的、法的、司法的アプローチがなければ、家の書類および所有物の物理的聖域は、たちまち空虚な法的容器になる。

家はまた、ブランダイスの「ひとりにしておいてもらう権利」の中心地でもある。しかし、この権利もまた、徐々にもろいものとなってきている。人は次第に自分の家にセンサーを導入するようになっている。センサーの直接的な目的は、便利さ、安全、セキュリティの確保である。煙や一酸化炭素の探知器は広く使用されており、安全規定で要求される場合も多い。米国の一部地域では、ラドン検出器が普及している。様々な種類の汚染物質やアレルゲンを探知することができる統合空気モニターは、近い将来に普及すると予測される。冷蔵庫は間もなく、腐食した食品が放つガスを「嗅ぎ取る」ことができるようになるかもしれない。あるいは、食品のパッケージの無線ICタグ（RFID）から賞味期限を「読み取る」ことができるようになる可能性も考えられる。現在の耳障りな警告音の代わりに、未来のセンサーは（現在のセンサーの一部もすでにそうなっているが）、モバイル機器やディスプレー画面の統合アプリを通じて、家族とインターフェースする。データは処理され、解釈されることになる。その処理はクラウド上で行われる可能性が非常に高い。よって、消費者が希望するサービスを実現するために、多くのデータを家の外に持ち出す必要が出てくる。

食品と空気の新たな安全を可能にする環境センサーは、タバコやマリファナの煙を探知し、識別することもできるようになるかもしれない。ヘルスケア提供者や健康保険業者は、非喫煙者であると自己申告している人が本当にそうであるかの保証を得たいと考えるかもしれない。保険料の引き下げの条件として、環境モニターのデータをチェックするための同意を自宅所有者に要求することは許されるのか。モニターがヘロインの煙を探知した場合、保険会社は警察に通報する義務を負うのか。保険会社は住宅所有者の損害保険を無効にできるのか。

一般的な家がどの部屋にもカメラやマイクを備えるようになるというのは、一部の人には突飛な考えに思えるかもしれないが、現にそのような方向に進んでいる可能性は高い。（すでに正面にも裏側にもカメラを搭載している）携帯電話をベッドの隣のナイトスタンドに置いておいた場合、それは何を聞き、何を見ることができるか。タブレットやラップトップ、また多くのデスクトップ・コンピュータには、カメラやマイクが搭載されている。自宅の侵入警報用の動作感知技術は、超音波や赤外線カメラから画像カメラに移行する可能性が高い。その結果として、誤認警報の回数が減り、ペットと人を区別できるようになるという利益が得られる。顔認証技術はさらなるセキュリティと利便性をもたらす。カメラとマイクが転倒や卒倒、あるいは助けを求める声を探知し、ネットワーク接続して救助要請をすることができるようになれば、高齢者の安全が高まる。

人は声やジェスチャーでコミュニケーションをとるのが自然である。人が電子アシスタントとコミュニケーションをとるときにも、必然的にこの二つの手段を用いることになる（結果として、カメラとマイクにアクセスすることが必要になる）。最近アップルに買収されたイスラエルのプライムセンス（PrimeSense）を始めとする企業は、ジェスチャーの解読のための高性能コンピュータ・ビジョン・ソフトウェアを開発中であるが、これは消費者向けコンピュータ・ゲーム・コンソールの市場においては、すでに重要な機能になっている（マイクロソフト・キネテクト（Microsoft Kinect）など）。消費者向けテレビは、ジェスチャーに反応する最初の「電化製品」の一つになっており、ネスト（Nest）煙探知機などのデバイスは、すでにジェスチャーに反応する。こめかみ部分をタップして、グーグル・グラスに音声指示のシグナルを送る消費者は、同じジェスチャーをテレビに、またさらに言えば家の中のすべての部屋のサーモスタットや照明のスイッチにも使いたいと考えるかもしれない。これは、家の中の至る所で音声と画像が収集される事態を示唆する。

これらの音声、画像およびセンサーのデータはすべて、聖域とされる家の中で生成されるのである。しかし、上述の「書類および所有物」と同様に、これらが家の中にとどまる可能性は低い。家の中にある電子デバイスは、複数の異なるインフラを通じて、目に見えな

い形で外界とすでに通信している。ケーブル産業の家庭への配線接続は、ブロードバンド・インターネットなど、複数の種類の双方向通信を提供している。有線電話は依然として、一部の家庭用侵入警報や衛星テレビ受信機によって、またDSLブロードバンド・サブスクライバの物理レイヤとして使用されている。家庭用デバイスの中には、携帯電話の無線インフラを使用するものもある。その他の多くのデバイスは、現代の生活の必需品となりつつある家庭用Wi-Fiネットワークに便乗している。現在のスマートな家庭用エンターテイメント・システムは、人がデジタルビデオレコーダー（DVR）に何を録画し、実際に何をいつ見ているかを把握している。2000年の個人の財政記録と同様に、今日ではこのような情報は、部分的に家の中で、DVRの内部のハードドライブにローカライズされている。しかしながら、現在の財務情報と同様に、これはクラウドへと移行しつつある。現在、ネットフリックスやアマゾンは、自分たちのプラットフォーム上での顧客の過去のキーストリームやクリックストリーム、閲覧履歴に基づいてエンターテイメントに関する提案を行うことができる。将来的には、テレビに搭載されたジェスチャー読み取りカメラに分刻みで読み取られた顔の表情を解釈することで、より良い提案が可能になるかもしれない。

これらのデータの収集は、消費者がそれとわかっていて要求している製品やサービスに必要という意味においては無害なものである。これがプライバシーの問題を引き起こすのは、アナログセンサーがその機能に必要な最低限よりも多くの情報を収集せざるを得ないためであり（3.1.2節を参照）、またそのデータが革新的な新製品からマーケティングの大成功、犯罪への悪用までの多岐にわたる二次使用を呼ぶためでもある。その他の多くの種類のビッグデータと同様に、データの所有権、データの権利、および許されるデータの使用にはあいまいさが付きまとう。コンピュータ・ビジョン・ソフトウェアは、その視野の中にある製品のブランド・ラベルをすでに読み取ることができる可能性が高い。これは顔認証よりもずっと簡単な技術である。人がフットボールを見ている最中にどのブランドのビールを飲んでいるのか、ビールの栓を抜いたのはそのビールの広告を見る前か後だったのかをテレビの中のカメラが知ることができたとすると、誰であれば（もし誰かいるとしたら）この情報をビール会社、またはその競合会社に売却することが許されるのか。テレビの電源がオフになっているはずのときに、カメラにブランド名を読み取らせてもよいものか。どのような雑誌や政治パンフレットがあるかを読み取らせてよいのか。冷蔵庫のRFIDタグ・センサーが賞味期限切れの食品を探知することができた場合、商品のブランドの選択を販売業者に報告してもよいのか。その結果として、あらゆるスーパーマーケットから関連クーポンが提供されたとしたら、それは不気味なことなのか、それとも消費者の金銭的利益になるのか。そのデータが、他者には割引価格を提供し、一方でその本人についてはブランドロイヤリティが強いことが判明しているため正規の値段を支払わせるのに使用されたとしたら、話は別か（差別的価格設定のジレンマ）。

米国人の約3分の1は、持ち家ではなく借家に住んでいる。この数値は、2007年の金融危機の長期的結果として、また米国の人口の高齢化に伴って、経時的に上昇する可能性がある。現在および予測可能な未来において、平均して賃借人は、自宅所有者よりも裕福ではない。法律は、家主の財産権と賃借人のプライバシー権との間に微妙な境界線を引いている。家主は様々な条件の下で、自分の不動産に立ち入る権利を有している。一般的にその条件には、賃借人が保健または安全規定に違反した場合や、修繕を行うため、などが含まれる。家の中でより多くのデータが収集されるようになると、賃借人と家主の権利は新たな調整が必要になる可能性がある。環境モニターが家主の不動産の付属品だったとすると、家主はそのデータに無条件の権利を有するのか。そのデータを売却することはできるのか。賃貸借契約でそう定められていたとすれば、モニターがタバコの煙を繰り返し探知したり、カメラのセンサーが禁止されているペットを識別したりした場合、賃借人を立ち退かせることはできるのか。

第三者が（おそらく暗号によるあらゆる種類の保護手段が付いた）顔認証サービスを家主に提供した場合、家主はこのデータを用いて、又貸しや追加の居住者を禁止する賃貸借契約の条項を施行することができるのか。家主はそのようなモニタリングを賃借の条件として要求することができるのか。家主のカメラは屋外にあるものの、その地所に出入りする人を残らず追跡するとしたらどうか。それは、地元警察が所有するセキュリティカメラが道路全体を監視する場合とどう異なるのか。

2.4 プライバシー、セキュリティ、便利さのトレードオフ

プライバシーの概念は世代ごとに変わる。現在では、「デジタル・ネイティブ」である若い世代とその親または祖父母の間には顕著な違いが見られる。現在のデジタル・ネイティブの子供たちは、自分たちの個人情報の流れに対して、さらに異なる態度をとるものと考えられる。自分たちのことについて何もかも知っているデジタル・アシスタントがおり、データの使用を規制する（願わくは）賢明な政策が施行されている世界で育つ未来の世代は、現世代がジョージ・オーウェル的とまでは言わないものの、脅威的だと感じるシナリオに、ほとんど脅威を覚えることはないかもしれない。予測できるぎりぎりの限界とも言えるPCASTの最後のシナリオは、この点を例証するために作成したものである。

テイラー・ロドリゲスは短期の出張に行く準備をしている。前夜に荷造りをし、ピックアップしてもらうために自宅の玄関の外に置いた。盗まれる心配はない。街灯のカメラが監視をしており、またいずれにしても、バッグの中のほぼすべてには小さなRFIDタグが付けられている。盗もうとしても、追跡され、数分以内に逮捕される。運送会社に具体的な

指示を出しておく必要もない。クラウドがテイラーの旅程と計画を知っているからである。バッグは夜のうちにピックアップされ、テイラーが目的のホテルの部屋に到着するまでにそこに届けられる。

テイラーは朝食を終え、玄関の外に出る。スケジュールを把握しているため、クラウドが自動運転車を手配しており、それが道端で待ち受けている。空港では、ゲートに直接進む――セキュリティを通る必要はない。また、ゲートでもいっさい手続きはない。乗客が飛行機に乗り込み、着席する（乗客はそれぞれ、自分のウェアラブル光学機器でハイライトされた席に座る）ために、20分の「扉を開ける」時間が設けられている。搭乗券もいらなければ、きちんと並ぶ必要もない。（空港内に入る他のすべての人と同様に）テイラーの身元は追跡されており、完全に判明しているのに、わざわざ手間をかける必要があるだろうか。判明している情報の誕生場所（電話、服のRFIDタグ、顔認証、歩行、情動状態）がクラウドに知られ、吟味され、事実上偽造が不可能であるとしたら、また万が一、テイラーが狂信的で危険になるというあり得ない事態が起こったとしても、探知可能な多くの兆候がすでに追跡、探知され、対処が行われているとしたら、わざわざ手間をかける必要などあるだろうか。

実際、テイラーのその日の持ち物はすでに、現在の大急ぎで行われる空港の検査よりもはるかに効果的に検査されている。また、テイラーの家にあるすべてのLED照明器具に搭載されたフレンドリーなカメラは、いつも通りその日も、彼女が服を着て、荷造りをしているところを観察している。通常、これらのデータは、注意を促したり、ファッションについての助言を与えたりするために、テイラーのパーソナル・デジタル・アシスタントのみが使用する。しかし、空港のトランジット・システムを利用する条件として、テイラーは、空港のセキュリティと治安の確保のため、このデータを使用することを許可したのである。

私たちにはテイラーの世界が不気味に思えるかもしれない。現在のほとんどの人が許容すると考える利便性、プライバシー、セキュリティの公共の利益の間のバランスとは異なるバランスをテイラーは受け入れたのである。テイラーは、クラウドとそのロボット・アシスタントがプライバシー面で信頼ができると無意識のうちに信じて（その認識は、施行されている政策の性質と有効性によって、正しい場合も正しくない場合もある）、行動している。そのような世界では、日々の生活の利便性とセキュリティの大幅な改善が可能になる。

3. 収集、アナリティクス、および支援インフラ

ビッグデータは二つの異なる意味において「ビッグ」である。まずは、処理するのに使用可能なデータの量と多様性が大きい。次に、究極的には推測を行うことを目的に、これらのデータに適用することのできる分析（「アナリティクス」と呼ばれる）の規模が大きい。いずれの種類の「ビッグ」も、大規模で幅広く利用可能なコンピュータ・インフラがあるか否かに依存するが、それは徐々にクラウドサービスによって提供されるようになっている。本章では、これらの基本概念について詳述する。

3.1 個人データの電子的ソース

コンピュータ時代が開始して以来、公共団体も民間団体も、人に関するデジタル情報を集めてきた。「バッチ処理」の時代には、個人情報のデータベースが作成された。実際、初期のデータベース技術についての説明では、給与支払いに用いられる人事記録に触れているものが多い。コンピューティングの能力が高まるにつれ、デジタル形式に移行するビジネス・アプリケーションが増えてきた。現在デジタル形式になっているものは、通話記録、クレジットカードの取引記録、銀行口座の記録、電子メールの保存など、枚挙にいとまがない。対話型コンピューティングが進歩するにつれ、個人はオンライン・サービスの自己識別のためや、財政管理システムなどの生産性ツールのために、ますます自分についてのデータを入力するようになってきた。

このようなデジタルデータには、「メタデータ」が付随するのが一般的である。メタデータとは、そのメタデータが記述するデータの形式および意味について説明する付属データである。データベースにはスキーマが、電子メールにはヘッダーがあるのと同様のことがネットワーク・パケットにも当てはまる。データセットが複雑さを増すにつれ、付随するメタデータも複雑さを増す。データまたはメタデータには、アカウント番号やログイン名、パスワードなどの識別情報が含まれていることもある。メタデータが記述するデータよりも、メタデータのほうがプライバシーに関する問題を引き起こすことが少ないと考える理由はない。

近年では、人に関する利用可能な電子データの種類が大幅に増加した。その原因には、ソーシャルメディアの出現や、モバイル機器、監視機器、様々なネットワーク・センサーの成長などがある。今日では、使用または誤用されることでプライバシーに関する問題を生じさせる可能性のある情報を、個人がそうとは気付かないままに、周囲に絶えず流出させている。物理的には、これらの情報の誕生場所には二つのタイプがある。「デジタル生まれ」と「アナログ生まれ」と呼べるものである。

3.1.1 「デジタル生まれ」のデータ

「デジタル生まれ」の情報は、初めからコンピュータまたはデータ処理システムによる使用を意図して、私たち、またはコンピュータの代理が生成する情報である。デジタル生まれのデータの例には以下がある。

- 電子メールおよび携帯メール
- 電話、タブレット、コンピュータ、またはビデオゲームのマウスクリック、タップ、スワイプ、またはキーストロークを通じたインプット、すなわち人が意図的にデバイスに入力するデータ
- GPS 位置データ
- 通話に付随するメタデータ、すなわち電話をかけた側とかけられた側の電話番号、通話の日時と継続時間
- ほとんどの商取引に付随するデータ、すなわちクレジットカードのスワイプ、バーコードの読み取り、（盗難防止や在庫管理に用いられる）RFID タグの読み取り
- 入り口へのアクセス（キーカードや ID バッジの読み取り）や有料道路へのアクセス（RFID タグの遠隔読み取り）に付随するデータ
- モバイル機器がネットワークとの接続を維持するのに用いる、デバイスの位置や状態を含むメタデータ
- 徐々に増えている車、テレビ、電化製品からのデータ（「モノのインターネット」）

消費者の追跡データは、経済的に重要になったデジタル生まれのデータの一例である。概して企業は、大量のデータを集め、そのデータをマーケティング、広告、またはその他の活動に使用することができる。従来のメカニズムはクッキー――ブラウザがユーザーのコンピュータに残すことのできる小さなデータファイル（20 年前にネットスケープが開発）を使用するものであった。この技術は、ユーザーが最初にサイトを閲覧するとクッキーを残し、その閲覧を後の事象と関連付けることを可能にするものである。このような情報は小売業者にとっては非常に貴重で、過去 10 年間の広告業の多くの基盤となってきた。そのような追跡を規制するために様々な提案が出されており、また、この追跡を行う前にオプトインの許可を要求する国も多い。クッキーに関連するのは、比較的シンプルな情報で、擁護者は悪用の可能性が低いと説明する。人は常にそのプロセスを承知しているとは限らないものの、無料または補助付きのサービスと引き換えに、このような追跡を受け入

れている。同時に、クッキーを使わない選択肢がある場合もある。クッキーがなくても、いわゆる「フィンガープリンティング」技術は、スクリーンの大きさやインストールされているフォントなどの公に露呈される情報によって、ユーザーのコンピュータやモバイル機器を一意に識別できる場合が多い。技術者の大多数は、アプリケーションがクッキーから移行すること、クッキーがあまりにもシンプルな考えであること、およびアナリティクスが向上し、より良いアプローチが発明されつつあることを信じている。しかしながら、消費者の追跡に対する経済的インセンティブは依然として残り、ビッグデータによってさらに適切な対処が可能になると考えられる。

追跡は不正使用を可能にする技術でもある。残念ながら、多くのソーシャルネットワーキングアプリは、人の連絡先リストを集め、全受取人にそのアプリの宣伝の入ったスパムを送信することでスタートする。この技術はしばしば悪用される。プライバシーを尊重するという評判が得られないことで失われる価値よりも、新しい顧客に接触することで得られる価値のほうが大きいと評価する小さな新興企業にとりわけその傾向が見られる。

デジタル生まれの情報はすべて、一定の特徴を共有している。いずれも特定の目的のため、識別可能な単位で作成される。この単位は、ほとんどの場合、なんらかの標準タイプの「データ・パケット」である。それは意図的に作成されるため、含まれる情報は、効率性および優れた工学設計により、収集の直接的な目的のみが果たせるように限定されるのが一般的である。

デジタル生まれのデータについては、プライバシーに関する懸念は、二つの異なる形で発生し得る。一つは自明なもの（「過度な収集」）で、もう一つはもっと新しく、微妙なもの（「データフュージョン」）である。過度な収集は、工学設計が意図的に、またときとしてひそやかに、明示された目標と無関係な情報を収集するときに発生する。スマートフォンは、テキストメッセージのキーストロークごとの人の顔の表情を写真撮影し、第三者に送信することが容易にできたり、すべてのキーストロークをキャプチャし、それによって削除したメッセージを記録したりできるものの、これはデフォルトのテキストメッセージアプリにとっては非効率的で不合理なソフトウェア設計の選択になる。そのような文脈においては、これは過度な収集の例と言える。

過度な収集の最近の例としては、超光懐中電灯無料（Brightest Flashlight Free）電話アプリが挙げられる。これは 5,000 万人以上のユーザーにダウンロードされたアプリで、懐中電灯が使われるたびに、その位置を販売業者に伝達するものであった。位置情報は懐中電灯の照明機能には不必要であるだけでなく、ユーザーが秘密にしておきたいと考える可能性のある個人情報を開示する。連邦取引委員会は、通知と同意の画面（4.3 節を参照）

の但し書きにおいて、位置情報が収集されることは明らかにされていたものの、それが広告業者などの第三者に販売されることは明示されていなかったとして、訴状を出した。この例から、通知と同意の枠組みには限界があることがわかる。超光懐中電灯無料アプリが当初にもっと詳細な但し書きによって情報公開をしていれば、たとえそのような但し書きを実際に読む人はほとんどいないとしても、FTC の措置を回避できていた可能性が高く、ダウンロード数に大きな影響を与えることもなかったと思われる。

過度な収集とは対照的に、データフュージョンは異なるソースからのデータが接触し、新しい、しばしば予期しない現象が明らかになったときに発生する（3.1 節を参照）。個別には、それぞれのデータ・ソースは、明確で限定された目的のために設計されたと考えられる。しかしながら、現在の統計データマイニング、パターン認識、共通の識別データに基づいた多様なソースからのレコードの結合などの技法を用いて複数のソースを処理すると、新しい意味が突き止められることがある。とりわけデータフュージョンは、個人の識別（すなわち、事象と個人の一意の ID との関連付け）、データが豊富な個人プロファイルの作成、数日または数カ月または数年に及ぶ個人の活動の追跡につながることが多い。

定義上は、データフュージョンによって発生するプライバシーの問題は、個々のデータストリームにあるわけではない。その収集、リアルタイムの処理、および保持は、その明確で直接的な目的のために完全に必要で、適切である可能性がある。むしろ、プライバシーの問題は、大規模で多様なデータセットを並置して分析し、新しい種類の数学アルゴリズムを用いて処理する能力が向上していることによる創発的な特性である。

3.1.2 センサーからのデータ

ここでは二つ目のタイプの情報の誕生場所に目を向ける。情報が物理世界の特性から発生する場合、それは「アナログ生まれ」の情報と言える。そのような情報は、「センサー」――物理的影響を観察し、それをデジタル形式に変換する工学機器――に触れて初めて電子的に利用可能になる。最も一般的なセンサーは、可視電磁放射を検知するビデオを含むカメラ、および音と振動を検知するマイクである。しかしながら、他にも多くの種類のセンサーがある。現在では、携帯電話はカメラ、マイク、無線のみならず、磁場（3D コンパス）センサーや運動（加速）センサーも搭載していることが一般的である。その他の種類のセンサーには、熱赤外線（IR）放射、化学汚染物質の識別を含む空気品質、大気圧（および高度）、低レベルのガンマ線放射、およびその他の多くの現象のためのセンサーなどがある。

個人情報を提供し、現在使用されているアナログ生まれのデータの例には、以下がある。

- 通話の音声および／または映像コンテンツ――アナログ生まれであるが、電話のマイクおよびカメラによってただちにデジタルに変換される
- 専用デバイス（これまで主たる提供業者はフィットビット（Fitbit））、または携帯電話のアプリで探知される、心拍、呼吸、歩行などの個人の健康情報
- ユーザーのジェスチャーを解釈するテレビおよびビデオゲームのカメラ／センサー
- セキュリティ監視カメラ、携帯電話、または頭上のドローンからのビデオ
- 人にとっては完全な暗闇と言える状況で物を見ることができる（また、過去の事象の消えゆく痕跡――いわゆる「熱の傷跡（heat scars）」――を見ることができる）赤外線イメージング・ビデオ
- 発砲を検知し、場所を特定するために、および公衆安全のために使用される都市のマイクロホン・ネットワーク
- 教室およびその他の集会室のカメラ／マイク
- 超音波運動探知器
- 医用画像、CT および MRI スキャン、超音波画像
- 機会に便乗して収集された化学または生体サンプル、とりわけトレース DNA（現在では時間のかかるオフライン分析が求められるが、近い将来にはもっと素早くなることが予測される）
- 雲があっても画像化でき、特定の条件下では、非金属の構造物の内部を見ることができる合成開口レーダー（SAR）
- 電気および電子機器からの意図しないマイクロ波放射

アナログ生まれのデータは、その当面の目的に必要な最低限よりも多くの情報を含んでいる可能性が高いが、それには正当な理由がいくつかある。理由の一つは、不要で関係のない情報（「ノイズ」）がある中で、必要な情報（「信号」）を検出しなければならないからである。センサーは通常、環境（「信号とノイズ」）を高精度で探知し、信号が最小で、ノイズが最大という考えられる最悪のケースでも、両者を区別する数学的手法を適用できるようにすることで機能する。

別の理由としては、技術的収束が挙げられる。たとえば、携帯電話のカメラが小型で安くなるにつれ、たとえ全画像が必要でない場合でも、他の製品に同一のコンポーネントを使用す

ることが好ましい設計上の選択肢になる。現在では、大画面テレビにIRリモートコントロール、室内輝度、動作感知（部屋に誰もいないときにテレビのスイッチを切る機能）のためのセンサーが別々にあり、加えてアドオンのゲーム・コンソールに本物のビデオカメラが付いている形であるが、将来的には、これらの機能をすべて、数ミリの大きさの一台の安価で高解像度のIR感受カメラに統合したモデルが出現しているかもしれない。

人についての情報を提供することを意図したデジタルおよびアナログ・ソースからの情報に加え、新しい「モノのインターネット」からの意図せぬ大量の情報開示が起きている。「モノのインターネット」とは、「スマートな」ネットワーク接続型計算能力によってその主たる目的を強化されたセンサーの融合である。例には、人がいることを検知し、それに従って空気温度を調整する「スマートな」サーモスタット、「スマートな」自動車イグニッション・システム、バイオメトリクスによって起動するロック・システムなどが含まれる。

アナログ生まれのデータが引き起こすプライバシーの問題は、デジタル生まれのデータの場合と少々異なる。（上で定義した）過度な収集は、節操のあるデジタル設計者にとっては不合理な設計の選択肢であり、プライバシーの問題に関して言えば、明らかなレッドカードであるが、アナログの領域における過度な収集は、ロバストで経済的な設計の選択肢である可能性がある。結果として、アナログ生まれのデータには、当初に予測されていなかった情報がしばしば含まれることになる。多くの場合、予測されていなかった情報は、予測されていなかった有益な製品やサービスにつながる可能性があるが、一方で予測されていなかった悪用の機会をもたらすことにもなる。

具体的な例として、ビデオ画像の三つの主要なパラメータ、すなわち解像度（画像のピクセル数はいくつか）、コントラスト比（画像が暗い領域をどれだけよく見通すことができるか）、測光精度（画像の輝度と色がどれだけ正確か）を考えてみるとよい。これまでにこの三つのパラメータはすべて、大幅に向上しており、今後も向上し続ける可能性が高い。現在では、特殊なカメラを用いれば、高い屋根の上から都市の景観を撮像する際に、数マイル以内のカメラの方向を向いている家およびアパートの窓から中の様子をはっきりと見ることができる。あるいは、前述のとおり、人の脈拍を遠隔測定し、健康状態や情動状態についての情報を提供することもできる。

このような能力があらゆる携帯電話やセキュリティ監視カメラ、またはあらゆるウェアラブル・コンピュータ機器に搭載されるようになると予想されており、それは不可避とさえ言えるかもしれない。（参加者のグーグル・グラス（あるいはセキュリティ・カメラやテレビカメラ）がリアルタイムでその他の全参加者の自律神経生理学的状態をモニターし、解釈することができたら、自動車の価格交渉や国際貿易協定の交渉はどうなるか想像するとよい）。

同じセンサーからの信号に他のどんな予期しない情報が隠れているかは予測のしようがない。

ひとたびデジタル世界に入ると、アナログ生まれのデータは、デジタル生まれのデータと共に融合し、マイニングすることが可能になる。たとえば、顔認証アルゴリズムは、単独では間違いを起こしやすいかもしれないものの、携帯電話（意図しない流出を含む）、店頭取引、RFID タグなどからのデジタル生まれのデータ、および（たとえば頭上のドローンからの）車両追跡やナンバープレート自動読み取りなどのその他のアナログ生まれのデータと組み合わせることができると、ほぼ完璧な ID 追跡が可能かもしれない。バイオメトリック・データは、さらに個人のプロファイルを強化する ID 情報を提供することができ、（ソーシャルメディアなどからの）行動データは、態度や感情を分析するのに利用されている（個人またはグループの「センチメント分析」）。要するに、キャプチャし、定量化フォーマットにした上で、表集計し、分析することのできる情報が増加しているのである。

3.2 ビッグデータ・アナリティクス

アナリティクスは、ビッグデータに生命を与えるものである。アナリティクスがなくても、ビッグデータセットはストレージが可能で、全面的または選択的に取り出しが可能である。しかし、出てくるものはまさに入ったものである。数々の異なる計算技術から成るアナリティクスこそが、ビッグデータ革命の原動力になる。アナリティクスはビッグデータセットに新しい価値を生むが、それは部分部分の価値の合計よりもはるかに大きい。

3.2.1 データマイニング

データマイニングは、大きなデータセットの中のパターンを突き止める計算過程を指し、アナリティクスとほぼ同一視されることもあるが、実際にはそのサブセットに過ぎない。データマイニングは、統計学、データベース、人工知能、機械学習を含む、応用数学とコンピュータサイエンスの両方に対する学術研究の多くの領域が収束したものである。他の技術と同様、データマイニングの進歩には、研究開発段階がある。それは新しいアルゴリズムとコンピュータ・プログラムが開発される段階で、その後に商業化と応用の段階が続く。

データマイニング・アルゴリズムは、教師あり学習――人によって処理された、認識すべきパターンの例がアルゴリズムにシード（seed）されているため、そう呼ばれる――によっても、教師なし学習――アルゴリズムが、事前のシードなしに、関連するデータを突き止めようとするため、そう呼ばれる――によっても、パターンを突き止めるように訓練できる。教師なし学習アルゴリズムの最近の成功例は、ウェブ上の数百万もの画像を検索し、「猫」が

投稿の多いカテゴリーであることを単独で突き止めたプログラムである。

データマイニングの望ましいアウトプットには、いくつかの形があり、そのそれぞれに専門のアルゴリズムがある。

- 分類アルゴリズムは、事物または事象を既知のカテゴリーに割り当てることを目指す。たとえば、病院が退院した患者について、再入院のリスクが高い、中程度、低いというカテゴリーに分類することが考えられる。
- クラスタ化アルゴリズムは、上の「猫」の例のように、事物または事象を類似性によってグループ化する。
- 回帰アルゴリズム（数値予報アルゴリズムとも呼ばれる）は、数量の予報を試みる。たとえば、銀行がローン申込書の詳細から、債務不履行の確率を予測することが考えられる。
- 連関法（association techniques）は、データセット内の項目と項目の間の関係を突き止めることを目指す。例としては、アマゾンのおすすめ商品やネットフリックスのおすすめ映画が挙げられる。
- 異常検出アルゴリズムは、データセット内の非典型的な例を探す。クレジットカード口座の不正取引を検出する、など。
- 要約法は、データ内の際立った特徴を見つけて、提示することを目指す。例には、単純な統計的概要（たとえば、学校および教師による生徒のテストスコアの平均）や、もっと高度な分析（ある個人について言及しているウェブ投稿のすべてから得た、その個人についての主要な事実のリストなど）が含まれる。

データマイニングは、機械学習と混同されることがあるが、後者は学術および産業研究におけるコンピュータサイエンスの幅広いサブフィールドである。データマイニングは機械学習やその他の専門領域を活用するが、機械学習はロボット工学など、データマイニング以外の領域にも応用される。

データマイニングで成し遂げられることには、現実面でも論理面でも限界があり、精度についても同様に限界がある。データマイニングはパターンや関係を明らかにするかもしれないが、通常これらのパターンの価値や意義を使用者に教えることはできない。たとえば、判明しているテロリストの特性に基づいた教師ありの学習は、類似した人物を見つけることはできるかもしれないが、その人物はテロリストである可能性もテロリストではない可能性もあり、加えて、そのプロファイルに合致しない異なるタイプのテロリストを見つけるこ

とはできない。

データマイニングは、行動および／または変数間の関係を識別するが、その関係は必ずしも因果関係を示すとは限らない。高圧の送電線の下に住んでいる人々の有病率が高い場合、それは送電線が住民の健康の害となっていることを意味している可能性も、送電線の下に住んでいる人々は貧しく、ヘルスケアへのアクセスが不十分な傾向があるということを意味している可能性もある。政策的含意は大きく異なる。いわゆる交絡変数（この例では収入）は、判明し理解されたときに修正可能であるものの、これらがすべて特定されたか否かを確かめる確実な方法はない。ビッグデータにおける真の因果関係の補定は、始まったばかりの研究分野である。

多くのデータ分析は、因果関係を反映している場合も反映していない場合もある相関関係を明らかにする。データ分析は、アルゴリズムの限界または偏ったサンプリングの使用により、不完全な情報を生み出すことがある。このような分析を無差別に使用すると、個人に対する差別、あるいは特定のグループへの不正確な関連付けにより公正さの喪失につながる可能性がある。データ分析を使用する際には、子供およびその他保護の対象となるグループのプライバシー保護に特別な注意を払わなければならない。

現実世界のデータは不完全でノイズが多い。このようなデータの品質問題は、データマイニング・アルゴリズムのパフォーマンスを下げ、アウトプットをあいまいにする。経済的に許すのであれば、インプットデータを慎重にスクリーニングし、準備することで、結果の質を高めることができるが、このデータの準備は労働集約的で、費用がかかる場合が多い。とりわけ商業セクターの使用者は、コストと正確性をトレードオフすることを余儀なくされ、そのデータで表される個人にマイナスの結果をもたらすことがある。加えて、現実世界のデータは極端な事象、または外れ値を含んでいる可能性がある。外れ値は、データにおいてたまたま過度に代表されている実際の事象である場合も、データの入力またはデータ転送エラーの結果である場合もある。いずれの場合にも、モデルをゆがめ、パフォーマンスを低下させる。外れ値の研究は、統計学の重要な研究分野である。

3.2.2 データフュージョンと情報統合

データフュージョンは、データマイニングやデータ管理のために処理がし易いよう、複数の異種のデータセットを融合して、一つの同質の表現にすることである。データフュージョンは、センサーネットワーク、ビデオ／イメージ処理、ロボット工学およびインテリジェント・システムなどの数々の専門領域で使用されている。

データ統合はデータフュージョンと区別される。統合はデータセットの結合のし方がもっと大まかで、維持される情報集合も大きい。データフュージョンには、たいてい削減または代替技術がある。データフュージョンは、データ相互運用性――二つのシステムが通信し、データを交換する能力――によって促進される。

データフュージョンとデータ統合は、ビジネス・インテリジェンスにとってキーとなる手法である。小売業者はオンライン、店舗およびカタログの売り上げデータベースを統合し、顧客の実態をより完全に把握しようとしている。たとえば、ウィリアムズソノマ（Williams-Sonoma）は、顧客データベースを 6,000 万の世帯に関する情報と統合した。追跡されているのは世帯収入、住宅の価値、子供の数などの変数である。この情報に基づいてターゲットを絞った電子メールは、ターゲットを絞っていないメールよりも回答率が 10 倍から 18 倍高いと言われている。これは、情報が多ければいかに推測が向上するかを表すわかり易い例である。プライバシー保護を助ける技術も出現している。

今日では、多重センサーのデータフュージョンに多大な関心が寄せられている。取り組みが進められている最大の技術的課題は、データ精度／分解能、外れ値およびスプリアスのデータ、矛盾するデータ、モダリティ（異種データと同質データの両方）および次元性、データ相関性、データ配置、データ内の関連付け、中央処理と分散処理、操作的タイミング、動的現象を扱う能力と静的現象を扱う能力に関連するもので、概して新しいより優れたアルゴリズムを開発することで対処されている。プライバシーに関する問題は、センサーの忠実度と精度、多重センターからのデータの相関によって生じる可能性がある。単一のセンサーのアウトプットはセンシティブではないかもしれないが、複数のセンサーからの組み合わせは、プライバシー上の問題を引き起こす可能性がある。

3.2.3 画像および音声認証

画像および音声認証技術は、静止画像、ビデオ、および録音または放送された音声の巨大なコーパスから情報を抽出することができ、限られた事例では人間の理解に近づいている。

都市のシーン抽出は、写真やビデオから地上の LiDAR（レーザーを用いたリモートセンシング技術）までの様々なデータ・ソースを用いることで実行できる。政府部門では、都市モデルが都市計画と視覚化に必要不可欠になりつつある。同様に、歴史、考古学、地理学、コンピュータグラフィックス研究などの学問分野にも都市モデルは重要である。デジタル都市モデルは、グーグル・アースやビングマップ（Bing Maps）などの人気の高い消費者向けマッピングおよび視覚化アプリケーションや、GPS ベースのナビゲーション・システムの中核を成している。シーン抽出は、個人情報を意図せずに捉える例であり、個人情報を明ら

かにするデータフュージョンに用いることができる。

顔認証技術は、商業的応用も法執行機関での応用も現実的になりつつある技術である。動的なシーンの中で動いている顔を捉え、正常化し、認識することができる。単一カメラ・システムを用いたリアルタイムのビデオ監視（および物体の認識と活動の分析の両方が可能な多重カメラ・システムを用いたリアルタイムのビデオ監視の一部）は、本土防衛、犯罪防止、交通規制、事故防止および検出、家庭での患者、高齢者および子供のモニターなど、公共および私的な環境のいずれにおいても幅広い用途がある。用途によって、ビデオ監視の導入の程度は異なっている。

画像認識がほかに可能なことには、以下が含まれる。

- ビデオ要約およびシーン変化検出（すなわち、ある期間を要約する少数の画像を選び出す）
- 衛星またはドローンからの画像の正確なジオロケーション
- 画像ベースのバイオメトリクス
- ヒューマン・イン・ザ・ループ監視システム
- 人および車両の再識別、すなわちセンサーからセンサーへと移動する同じ人または車両の追跡
- 様々な種類の人間の活動の認識
- セマンティック・サマリー（すなわち、画像をテキスト要約に変換）

カメラの視界内において対象を追跡し、複数のソースからの情報を組み合わせることで広いエリアでの異常な活動を検出することが可能になると予想されるものの、対象の再識別は依然として難しく（カメラ間の追跡の課題）、混雑した環境でのビデオ監視も同様に難しい。

使用されるデータは公共エリアで収集されることが多いものの、グーグル・ストリートビューなどのシーン抽出技術は、プライバシーに関する懸念を引き起こしている。ストリートビューで使用するために撮影された写真は、観察され写真を撮られていることに気づいていない人についてのセンシティブな情報を含んでいる可能性があるからである。

ソーシャルメディアのデータは、シーン抽出技術のインプット・ソースとして利用可能であ

る。しかし、このようなデータを投稿したとき、ユーザーは自分のデータがこのような集約的な方法で用いられ、（公開したものとはいえ）自分のソーシャルメディアの情報が合成されて新たな形に生まれ変わったように見えるかもしれないことに気づいていない可能性が高い。

自動音声認識は少なくとも 1950 年代から存在していたが、新たな可能性を持つようになったのは、ここ 10 年間の進歩によるものである。現在では、最先端の技術を用いれば、読み上げられた文章（ニュースのアナウンサーが原稿の一部を読み上げる、など）は、95 パーセント以上の精度で認識が可能である。自発語については、正確に認識するのがはるかに難しい。ただし最近では、研究者が利用できる自発語データのコーパスが劇的に増加したため、精度は上昇している。

今後数年で、音声認識インターフェースが用いられる場所は大幅に増加すると考えられる。たとえば、複数の企業が、テレビや車のコントロール、テレビ番組の検索、DVR の予約録画に音声認識をどう使用するかを研究している。ニュアンス（Nuance）の研究者は、音声技術をどう設計すればウェアラブル・コンピュータで使用可能になるかを積極的に検討しているという。グーグルは、グーグル・グラスにこの基本的な機能の一部をすでに導入しており、マイクロソフトのエックスボックスワン・システムは、システム機能のコントロールのためにマシンビジョンとマルチマイク音声入力を統合している。

3.2.4 ソーシャルネットワーク分析

ソーシャルネットワーク分析とは、相互接続する多様なユニットの関係が重要であり、これらのユニットが自律的には機能しないという想定のもとで、ユニットから情報を抽出することを指す。ソーシャルネットワークは多くの場合、オンラインのコンテキストで現れる。最も顕著な例は、フェイスブック、リンクトイン（LinkedIn）、ツイッターなどの専用ソーシャルメディア・プラットフォームである。これらはユーザー同士をインターネット上で直接接続し、コミュニケーションや情報交換ができるようにすることで、社会的相互作用への新たなアクセスを提供する。オフラインでの人のソーシャルネットワークも、たとえばどの電話が通話やテキストを交換したか、継続時間はどれくらいであったかを記録する通話メタデータ・レコードの中などに、分析可能なデジタルトレースを残すことがある。ソーシャルネットワーク分析は、人々を関連付けるデジタルデータの収集が増加することで徐々に可能になっているが、そのようなデジタルデータがその個人についての他のデータやメタデータと相関された場合にはとりわけそうである。そのような分析のためのツールは、オンラインのソーシャルメディア・プラットフォームに接続するオープン・アプリケーション・プログラミング・インターフェースを通じてアクセス可能なソーシャルネットワークのコ

ンテンツ量が増加していることなどを受け、開発および利用が進んでいる。この種の分析は、活発な研究領域である。

ソーシャルネットワーク分析は、従来のデータベースの分析を補完する。（たとえば、交流ネットワークのクラスタリングなど）使用される技術の中には、いずれの分析にも使用可能なものもある。ソーシャルネットワーク分析は、多様な種類の情報の関連付けが容易（すなわち、かなりのデータフュージョンが可能）なため、より強力である。この分析は結果の視覚化に向いており、それによって分析結果の解釈を助ける。人は自分自身となんらかの共通点がある人と交流する傾向があることに鑑み、どのような人と交流しているかを通じて人について知るためにこの分析を使用することができる。

ソーシャルネットワーク分析は、人を驚かせる可能性のある結果を生み出している。とりわけ個人の一意的識別は、データベース分析のみの場合よりも容易である。加えてそれは、多くの人が理解しているよりもさらに多様な種類のデータによって行われており、匿名性の低下につながる。個人のネットワーク構造は独特で、それそのものが識別子となる。時間と空間の共起は重要な識別の手段であり、本報告書の中で繰り返し述べてきたように、様々な種類のデータを組み合わせることで、識別が助けられる。

ソーシャルネットワーク分析は、罪を犯したかもしれない人がどのようなつながり、手段、動機を持っているかを理解するために、犯罪法医学調査で使用されている。とりわけソーシャルネットワーク分析は、テロリストの隠れたネットワーク――そのようなネットワークの力学は、公のネットワークのものと異なっている可能性がある――に対する理解を深めるために用いられてきた。

商業の分野では、人は自分の友人が何を好み、何を買うかによって、自分自身も何を買うかが左右される可能性があることが理解されている。たとえば、アイフォンを持っている友人が一人いると、そうでない場合に比べて、アイフォンを所持する確率が 3 倍になることが 2010 年に報告された。アイフォンを持っている友人が二人いる場合には、アイフォンを持つ確率は 5 倍になった。そのような相関がソーシャルネットワーク分析で明らかになると、製品動向を予測し、個人が欲しがる可能性のより高い製品の販促のためのキャンペーンを用意し、ソーシャルネットワークにおいて中心的な役割を果たす（よって影響力が大きい、「ネットワーク・バリュー」が高いと言われる）顧客にターゲットを絞るためにその情報を用いることができる。

一般に病気は個人（人間または動物）間の直接的な接触によって広まるため、利用できるあらゆるプロキシを通じてソーシャルネットワークを理解すれば、直接的な接触の可能性を

突き止め、それによって病気の発生のモニタリングと阻止を助けることができる。

フェイスブックの研究者たちは最近の調査において、個別のユーザーの地理的位置とその友人たちの地理的位置との間の関係を分析した。研究者たちはこの分析から、個別のユーザーの位置をそのネットワークの中の少数の友人の位置に基づいて、単にユーザーの IP アドレスを見ただけの場合よりも高い精度で予測するアルゴリズムを作成することができた。

ビジネス・インテリジェンス用にソーシャルネットワーキング・フィードからデータをマイニングする商業用「ソーシャル・リスニング」サービス――ラディアン 6／セールスフォース（Radian6/Salesforce）クラウド、コレクティブ・インテレクト（Collective Intellect）、リチウム（Lithium）など――が数多くある。ソーシャルネットワーク分析に加え、このような情報を利用することで、個人やコミュニティの間の影響の変化やトレンドの広まりを評価し、マーケティング戦略の策定につなげることができる。

3.3 ビッグデータの裏側のインフラ

ビッグデータ・アナリティクスは、アルゴリズムとデータだけでなく、データを保存し、分析する物理的プラットフォームも必要とする。個人データに用いる関連セキュリティ・サービス（4.1 節および 4.2 節を参照）も、インフラの必要不可欠な要素である。かつては巨大な組織のみしか利用できなかったこの種のインフラは、現在では「クラウド」を通じて、小企業や個人にも利用可能になっている。ソフトウェア・インフラが幅広く共有されれば、プライバシーを保護するインフラ・サービスももっと利用し易くなる。

3.3.1 データセンター

ビッグデータのプラットフォームについて考える方法の一つが、「データセンター」の物理的ユニットで考えるというものである。最近では、データセンターはほぼ基準商品になった。典型的なデータセンターは、フットボールのフィールド数面と同じ大きさのコンクリート平板上に建つ巨大な倉庫のような建物である。安い電力や光ファイバーのインターネット基幹接続へのアクセスが良い場所に位置し、農村地域か隔離された地域にあるのが一般的である。典型的なデータセンターは、20～40 メガワットの電力（人口が 2 万人から 4 万人の都市に相当）を消費し、合計で数十ペタバイトになる数万のサーバーやハードディスク・ドライブを収納している。世界にはこの規模のデータセンターが約 6,000 あり、そのうちの約半数が米国にある。

データセンターはあらゆる形のビッグデータの物理的中心である。大規模なデータ収集は、

パフォーマンスとロバスト性の両方を向上させるため、複数のデータセンターで複製されることが多い。データセンター・サービスを販売する市場は成長市場である。

専用ソフトウェア技術によって、複数のデータセンターにある（また、数万のプロセッサやハードディスク・ドライブに広がっている）データが協力してデータ・アナリティクスのタスクを遂行することが可能になり、それによって規模拡張とパフォーマンス向上の両方がもたらされる。たとえば、マップリデュース（もともとはグーグルの専有技術であったが、現在は一般名詞化した用語）は、事実上無制限の数のプロセッサでの並行処理のためのプログラミング・モデルである。ハドゥープ（Hadoop）は、同様の考えに基づいた、人気の高いオープンソースのプログラミング・プラットフォームおよびプログラム・ライブラリーであり、NoSQL（not Structured Query Language（非構造化照会言語）に由来する名称）は、従来の「リレーショナル」データベースの制限の多くを緩和し、1 または複数のデータセンターにある数多くのプロセッサにまたがって、拡張可能性を高める一連のデータベース技術である。。現在の研究は、Hadoop の次世代をターゲットにしている。研究の一つの方向性は、国家安全保障局が着手し、オープンソースのアパッチ・コミュニティに移行されているアキュミュロ（Accumulo）によって代表される。別の方向性は、バークリー・データ・アナリティクス・スタック（Berkeley Data Analytics Stack）に代表される。これはメモリー集約的データ・アナリティクスにおいて、Hadoop よりも 100 倍パフォーマンスが優れたオープンソース・プラットフォームで、フォースクエア（Foursquare）、コンビバ（Conviva）、クラウト（Klout）、クォンティファインド（Quantifind）、ヤフー、アマゾン・ウェブサービスなどの企業に使用されている。（SQL から NoSQL への動きに合わせて）NoHadoop と呼ばれることもある、この動向に合致した技術には、グーグルのドレメル（Dremel）、（概してスーパーコンピューティングに使用される）MPI、（グラフ向けの）プレーゲル（Pregel）、（リアルタイムのアナリティクス向けの）クラウドスケール（Cloudscale）などがある。

3.3.2 クラウド

「クラウド」は（一般の人の多くにはそう思われているかもしれないが）、単に世界中の全データセンターのことを指すわけではない。クラウドを理解する一つの方法は、データセンターの物理的共有化によって*可能になる*一連のプラットフォームおよびサービスとして捉えるというものである。データが「クラウド上に」あると言うとき、それは単にそのデータと共に（どこかに！）存在している物理的なハードドライブ・ディスクのことを指すのではなく、アプリケーション・プログラム、ミドルウェア、ネットワーク・プロトコルの複雑なインフラ、および（この点も重要であるが）すべて競合的に割り当てられた費用で、そのデータを取り込み、アクセスし、活用することを可能にするビジネスモデルのことも指すので

ある。全体としてクラウドをセットアップする商業事業体は、付加価値の多くの階層レベルと多くの異なる共存モデルを持つエコシステムに存在している。エンドユーザーと物理的データセンターの間で幾度か責任のハンドオフがあるかもしれない。

現在のクラウド・プロバイダーは、従来の法人データセンターやスモールビジネス・コンピュータに比べて、セキュリティ面での利益（およびそれによって、プライバシー面での利益）を提供する。これらのサービスには、より優れた物理的保護およびモニタリング、集中サポート・スタッフ、トレーニング、監督が含まれる場合がある。クラウドサービスは同時に、本調査のテーマであるセキュリティの新しい課題も突き付ける。利益もリスクも資源の集中化がもたらす結果――すなわち（複数のサーバーやサイトに分散されてはいるものの）事業体が所有するデータの量が増え、人およびシステムの登用や管理に高い基準を適用することで、クラウド・プロバイダーが別々に所有されているデータセンターよりもパフォーマンスを高める結果である。

クラウドの使用および個人のクラウドとのインタラクション（意図的なものか、そうでないものかにかかわらず）は、今後数年で急増することが予想される。モバイル・アプリが増加することでプラットフォームとしての携帯電話やタブレットの使用が強化されること、またセンサーが広く普及することの双方に伴って、分散するデバイスがもたらす情報を保存、処理、またはその他の形で扱うことを目的にクラウドシステムを使用する機会が増加していく。モバイル環境の進歩はモバイル・クラウド・アプリケーションの有用性を高めるものの、情報交換をより効果的にユーザーの目から隠すという意味において、プライバシーに害を及ぼす可能性がある。もっとコアなモバイルの機能性がクラウドに移行するにつれ、交換される情報量も多くなり、ユーザーは自分の携帯電話にローカライズされたままではなくなる情報の性質に驚くことになるかもしれない。たとえば、携帯電話のクラウドベースのスクリーンレンダリング（または「仮想化スクリーン」）により、携帯電話の画面には、クラウド上で計算されてモバイル機器に送信された画像が表示されることになる。このことは、モバイル機器の画面に表示される画像はすべて、クラウドからアクセスし、操作することが可能であるということを意味する。

クラウドのアーキテクチャは、ビッグデータのアナリティクスをサポートするためにも使用されるようになっている。使用者には大企業（グーグル、アマゾン、イーベイなど）のみでなく、自分自身のインフラを獲得する代わりに、公共のクラウド・プラットフォーム（アマゾン・ウェブサービス、グーグル・クラウド・プラットフォーム、マイクロソフト・アジュール（Microsoft Azure）など）を臨時または定期的に使用する小企業や個人も含まれる。フェイスブックやツイッターなどのソーシャルメディア・サービスは、クラウドシステムを使用しているプロバイダーによって展開、分析されている。このような使用法は、アナリテ

ィクスの一種の民主化を表しており、新事業などを促進する潜在力がある。今後期待されるのは、クラウド・アプリケーションを連合または相互接続するためのオプションや、クラウド・アプリケーション用のアプリケーション・プログラミング・インターフェースの不均一性を一部低減するためのオプションなどの研究である。

4. プライバシー保護のための技術と戦略

データは誕生し、収集され、ただちに処理される（「メタデータ」の追加を含む）場合も、やりとりされる場合も、（ローカル、リモート、またはその両方で）保存される場合も、コピーされる場合も、分析される場合も、ユーザーに伝えられる場合も、アーカイブに保管される場合も、破棄される場合もある。これらの段階のうちのいずれかにおける技術が、プライバシーにプラスの影響やマイナスの影響を与える可能性がある。

本章では、プラスの面に焦点を当て、プライバシーの保護に役立つ主な技術の一部を検証する。また、プライバシーと（サイバー）セキュリティの間の重要な違い、ならびに暗号化技術が果たすことのできる重要な、ただし限られた役割について明らかにすることを目指す。匿名化など、かつては有益であった従来の技術の中には、将来性が限られていると見られているものもある。新しい技術――すでに市場に参入しているものもあれば、さらなる研究が必要なものもある――について簡潔に説明する。

4.1 サイバーセキュリティとプライバシーの関係

サイバーセキュリティは、コンピュータの使用と電子通信のいくつかの異なる側面に関連するポリシーの実行を目指す専門領域、または一連の技術のことである。そのような側面の典型的なリストは、以下のとおりである。

- ID および認証：あなたはあなたが名乗るとおりの人物か。
- 権限付与：あなたは何をすることを許されているか。
- アベイラビリティ：権限を付与されている機能を攻撃者が妨害することができるか。
- 機密保持：データまたは通信は、権限を付与されていない者によって（受動的に）コピーされ得るか。
- 完全性：データまたは通信は、権限を付与されていない者によって（能動的に）変更または操作され得るか。
- 否認防止、監査能力：後に、行為（支払いが最良の例かもしれない）が行われたことを証明することは可能か。

優れたサイバーセキュリティは、正確で明白なポリシーを実行する。実際、数学的に表現可能なポリシーの明確さは、サイバーセキュリティの至高の目標――「安全であることが

証明可能な」システム――の実現のために必要な前提条件である。現在のところ、証明可能なセキュリティは、ある種のコンピュータ・チップの特定の機能など、非常に限られた領域にしか存在しない。安全であることが証明可能なシステムの範囲をさらに広範な領域に広げていくことがサイバーセキュリティの研究の目的である。実用サイバーセキュリティは、そのような研究の新しい原理を利用するが、それよりもさらに重要な手引きとなるのが、明らかになっているサイバーセキュリティの失敗から得る現実的な教訓である。サイバーセキュリティの慣行を絶えず改善し、ほとんどの場所でほぼ常に、次々と出現する脅威に先駆けることができるようにすることが現実的な目標となる。

不十分なサイバーセキュリティは、明らかにプライバシーへの脅威である。プライバシーは、データの機密保持の失敗、ID および認証プロセスの失敗、またはアベイラビリティを損なうなどの、より複雑なシナリオによって侵害される可能性がある。

セキュリティもプライバシーも共通して、悪意に焦点を合わせる。データのセキュリティは、不注意や偶発事故によって損なわれる可能性があるが、同時になんらかの当事者が損なおうとする意図を持って行動する――セキュリティの言葉を使えば、攻撃を行うことでも損なわれる。「損なう」や「攻撃」という言葉を「侵害」や「侵入」に置き換えれば、同じことがプライバシーにも当てはまる。

しかしながら、たとえ完璧なサイバーセキュリティがあったとしても、プライバシーは依然として危険にさらされる。コンピュータ・セキュリティに失敗がなかったとしても、プライバシーの侵害は起こり得る。権限を付与された個人がデータを悪用（たとえば開示）することを選択した場合、その個人が違反したのはプライバシーポリシーであって、セキュリティポリシーではない。あるいは、すでに論じたように（3.1.1 節を参照）、プライバシーはデータフュージョン――たとえ権限を付与された個人が安全なコンピュータ・システムで行った場合でも――によっても、侵害される恐れがある。

プライバシーは、その他の側面でもセキュリティと異なる。第一に、プライバシーポリシーは正確に成文化するのが難しくなっている。これはほぼ間違いなく、コンピュータ・セキュリティについての主張の有効範囲よりも、人間の前提や選好のほうが多様であることに起因する。実際、人間のプライバシーに関する選好をどう成文化するかは、誕生したばかりの重要な研究領域である。

コンピュータ・システムが安全であるという（なんらかのレベルの）保証を与えたとしたら、それはまだ発明されていないアプリケーションについて主張すること――すなわち、現在の機械にすでにある技術的設計特性により、そのようなアプリケーション・プログラ

ムがその機械の関連セキュリティポリシーに違反することを明日でさえも予防できると主張することである。プライバシーについての保証は、それよりもはるかにおぼつかないものである。まだ発明されていないアプリケーションは、まだ想像もされていない新しいデータ・ソースと、まだ発見されていない強力なアルゴリズムにアクセスするため、明日のプライバシー侵害の新しい手段に対する技術的な防護対策を今日提供することは、はるかに難しい。セキュリティは、今日のプラットフォームに対する明日の脅威に対処する。それだけでも十分に難しいが、プライバシーは明日のプラットフォームに対する明日の脅威に対処する。というのも、そのようなプラットフォームは単にハードウェアとソフトウェアからだけでなく、新しい種類のデータと新しいアルゴリズムからも構成されるからである。

多くの場合、コンピュータ・サイエンティストは、セキュリティの形式的ポリシーを基盤にして作業をする。エンジニアが何かを明確に記述することで、純粋に技術的な手段でそれに対処する特別な方法を設計できるようにするのと同様である。プライバシーについて考え始めるコンピュータ・サイエンティストが増えるにつれ、プライバシーポリシーの形式化に関心が集まっている。誇張的に言えば、自分がしていることが正しいことをしているかどうかを知るためには、自分が何をしているかを知らなければならないということである。規制やポリシーをソフトウェア仕様と合わせるという課題に対処する研究には、ポリシーやシステム要件を表現する形式言語、ポリシーやソフトウェア仕様の中およびそれらの間の対立、矛盾、あいまいさについて推論するためのツール、要求確認エンジニアやビジネスアナリストやソフトウェア開発者がポリシーを分析して改善し、経時的にモニターができる測定可能なシステム仕様にするための方法、監査およびアカウンタビリティ制度を通じたプライバシーの形式化および施行、ビッグデータ・システムにおけるプライバシー・コンプライアンス、目的の制限の形式化および施行が含まれる。

4.2 暗号学と暗号化

暗号学は、データを保護するための一連のアルゴリズムとシステム設計原理――その中には十分に発達したものも、原始的なものもある――から成る。暗号学は、暗号化技術を生み出す知識領域である。優れた設計のプロトコルがあれば、暗号化技術はプライバシー侵害を抑止するが、「銀の弾丸（確実な方法）」ではない。

4.2.1 定評のある暗号化技術

暗号学を用いれば、いかなる種類の読み取り可能なデータ――プレーンテキストと呼ばれる――も、ランダムであることが証明可能なビットの事実上理解不能なストリング、すな

わちいわゆる暗号原文（cryptotext）に変換することができる。暗号原文は、いかなる種類の機密保持も必要としない。クラウドに保存することも、都合の良いどの場所に送ることもできる。NSA やロシア連邦保安庁（FSB）に送られたとしても、暗号原文しかない場合――かつそれが厳密な数学的意味において適切に生成されたものであった場合には、NSA や FSB にもどうすることもできない。データを読むことも、演算することも不可能である。復号する、すなわち暗号原文を最初のプレーンテキストに戻すのに必要なのは、「鍵」である。「鍵」というのは実質上、権限を付与されたユーザーのみが知っている（あるいは計算可能である）はずのビット・ストリングである。鍵があって初めて、暗号化されたデータを使用する、すなわちその価値を読み取ることができる。

プライバシー保護という文脈では、主として関心事は暗号学ではない。むしろデータの侵害は、以下の主たる二つの方法のうちのいずれかで発生する。

- データはそれが暗号化される前、あるいは復号された後に、盗まれたり、誤って共有されたりする可能性がある。暗号化されているはずのデータが攻撃されたという場合、実際には攻撃されたのは、暗号化されていないプレーンテキストを――たとえ束の間であれ――含んでいた機械である。たとえば、2013 年にターゲットが 1 億件のデビットカードの番号と暗証番号（PIN）を盗まれたが、その際 PIN は、ほんのわずかな間のみ、暗号化されていない状態にあった。それでも、盗まれたのである。

- 鍵は認証、生成、配布され、使用されなければならない。鍵は、これらのすべての段階において、その鍵が保護することを目的としているデータを最終的に侵害することにつながる侵害または悪用を受ける可能性がある。秘密鍵へのアクセスを持つ人がそれを共有するように強要されたとしたら、当然、暗号化に基づくシステムは安全ではない。

1970 年代までは、鍵は紙やコンピュータ・メディアを用いて物理的に配布され、書留郵便や、極端な場合には武器を持った警備員によって保護されていた。すべてを変えたのは「公開鍵暗号」であった。公開鍵暗号は、その名前が示唆するように、個人が自分の鍵を公開することを可能にする。しかし、この公開鍵は暗号鍵に過ぎず、プレーンテキストを他者には無意味な暗号原文に変換するのに使われる。それに対応する「秘密鍵」は、暗号原文をプレーンテキストに変換するのに用いられ、それについては受取人によって秘密にされたままである。よって、公開鍵暗号は鍵配布の問題を ID 判定の問題に変える。ボブに対するアリスのメッセージ（暗号化されたデータ送信）は、ボブの公開鍵によって完全に保護されるが、それは自分が使用しているのは本当にボブの公開鍵であって、ボブになりすましている他の誰かの公開鍵ではないことをアリスが確かめられる場合に限られる。

幸いなことに、公開鍵暗号は ID 確認を助ける手段――すなわち真正性を証明するためのメッセージの電子「署名」――も提供する。電子署名は、「X として知られる権限ある私は、以下が本当に、従属する Y という者の公開鍵であることを認証する。（署名）X」という形のメッセージを可能にする。このようなメッセージは認証と呼ばれる。認証は、A が B の ID を認証し、B が C の ID を認証するというように、次から次へとつなげることができる。認証によって、ID 確認の問題は事実上、Y の可能性のある数百万の ID を認証するというものから、はるかに少数のトップレベルの認証局（CA）の ID を認証するという問題に変化する。しかしながら、100 以上ものトップレベルの CA が広く認められている（たとえば、ほとんどのウェブブラウザーから受け入れられる）というのは憂慮すべき事態である。というのも、CA からユーザーに至るまでの認証のヒエラルキーには、いくつかの中間段階がある可能性があり、その各段階で、秘密鍵はいずれかのコンピュータ上でいずれかの署名者によって保護されなければならないからである。この秘密鍵が侵害されると、ID の偽造認証を作ることができるため、それよりも下の段階の全ユーザーのプライバシーが侵害される可能性がある。そのようなセキュリティ上の弱点が付かれたこともある。たとえば、2011 年にオランダの CA の秘密鍵が盗まれたと考えられ、オランダ政府の潜在的にすべての記録のプライバシーが侵害された。

最近では主要企業の多くが、データを送信するために暗号を使用するか、暗号の使用を強化している。現在では「前方秘匿性（(perfect) forward secrecy）」を使用している企業もある。これは、個人の秘密鍵が侵害された場合に、その人がその後に受け取るメッセージのみが侵害され、一方で、たとえ盗まれた秘密鍵を現在保持している同じハッカーが以前に暗号原文を保存していたとしても、そのような過去のやり取りについては機密が保持されるようにする公開鍵暗号の変種である。

4.2.2 暗号化の最先端

これまで言及してきた技術は、保存中のデータと送信中のデータの両方の保護を可能とし、(i) すでに正しい鍵を持っているユーザー（自分自身が後に使用するためにデータを保存する場合などが考えられる）、あるいは (ii) データ所有者に権限を付与されており、そのデータ所有者から信頼されている CA に ID を認証されているユーザーのいずれかがこれらのデータを完全に復号することを許すものである。暗号研究の最先端では異なる種類の鍵――様々な種類の制限付きアクセスのみを与える鍵、または事前に誰であるかは正確にはわからないまま、あるグループに属する個人にメッセージを送信することを可能にする鍵――を作成する方法が研究されている。中には実用化されつつある発明もある。

たとえば、「ID ベースの暗号化」や「属性ベースの暗号化」は、「1980 年 5 月 23 日生まれのラモーナ・Q・ドウという名前の人」、あるいは「職種がオンブズマン、行政監察官、または消費者の擁護者である人すべて」のみによる使用を意図して、メッセージを送信、またはデータファイルを保護する方法である。このような技術は、信頼された第三者（事実上は認証局）を必要とするが、メッセージそのものはその第三者の手を通ることを必要としない。このようなツールは、採用の初期段階にある。

「ゼロ知識（Zero-knowledge）」システムは、低いレベルのデータを公開することなく、抽象度が一定レベルよりも高い情報について、暗号化データにクエリーを行うことを可能にする。たとえば、ウェブサイトの運営者は、ユーザーが 21 歳以上か否かを、ユーザーの実際の生年月日を知ることなく確認することができる。これが画期的なのは、ユーザーが年齢を偽っていないことを数学的に証明する方法で行われる点である。運営者は、実際に認証を見ることさえもせず、（言うまでもなく、いずれかの CA によって署名された）認証がユーザーの生年月日を証明していることを数学的確信度で知ることができる。ゼロ知識システムは、シンプルなケースにおいて商業化され始めたばかりである。近い将来に、たとえば、同意をしていない患者からの健康レコードデータのリサーチマイニングに必要などの、複雑で構造化されていない状況に拡張できる可能性は低い。

たとえばロケーション・プライバシーなど、もっとシンプルな領域では、実用的な暗号保護のほうが現実に近い。典型的なケースとしては、友人同士のグループが互いに近くにいるときにそのことを知りたいが、第三者とは自分たちの実際の位置を共有しないという場合が考えられる。当然ながら、このようなアプリケーションは、信頼されている第三者がいる場合のほうがずっとシンプルで、実際、現在のそのような商業用アプリケーションのほとんどはそうである。

準同型暗号（Homomorphic encryption）は、暗号化されたデータベースの単なるクエリーを超えて、暗号化データを復号しないままに実際の計算（統計収集など）に使用することを目指す研究領域である。このような技術は実用から遠く、本報告書の時間枠においては政策オプションになる可能性は低い。

準同型暗号に関連しており、金融部門においてとりわけ関心が高いセキュア・マルチパーティ計算法（secure multi-party computation）では、計算は暗号化された分散データ蓄積に対して行われるかもしれない。個々のデータについては、「結託に対してロバストな」暗号化アルゴリズムを用いて秘密にされるものの、データを用いて一般統計を計算することができる。なんらかの個人データを知っている当事者は、その当事者が知らない情報については開示しないまま、その人が知っている情報とそうでない情報の両方に基づい

て有益な結果を生み出すプロトコルを使用する。

暗号化に関連しているもの、暗号化とは異なる比較的新しい発展である差分プライバシー（differential privacy）は、データベースクエリーまたは計算の正確性を最大化し、一方で一般的にはクエリーの結果の難読化を通じて（たとえばスプリアスの情報、または「ノイズ」を付加することで)、データベースの中にレコードのある個人の識別可能性を最小化することを目指すものである。その他の難読化の方法と同様に、データの匿名化とクエリーのアウトプットの正確性および有用性の間にはトレードオフがある。このようなアイディアは、クエリーをそもそも許すというリスクの評価を向上させるという点を除けば、実用的応用から程遠い。

4.3 通知と同意

現在では通知と同意は、消費者のプライバシー保護のために最も幅広く使用されている戦略である。ユーザーが自分のモバイル機器に新しいアプリをダウンロードする際、またはウェブサービスを受けるためにアカウントを作成する際に、通知が表示され、ユーザーはそのアプリまたはサービスを使用するために積極的な同意を示さなければならない。どこかの空想の世界であれば、ユーザーがこのような通知を実際に読んで、(必要であれば弁護士に相談の上）その法的な意味を理解し、より良いプライバシーの扱いを受けるために類似サービスの提供者と交渉を試み、その上で最終的に同意をクリックするかもしれないが、現実は異なっている。

通知と同意は、プライバシー保護の責任を基本的に個人に課している。これは、通常「権利」が意味することとは真逆である。さらに悪いことには、プロバイダーが個人データを共有する権利を有することがそのような通知の中に隠れていた場合、たとえデータの使用法が異なっていたとしても、ユーザーは共有先である別の企業から通知を受けることはなく、ましてや同意を与える機会が与えられることはない。さらに、プロバイダーがプライバシーの通知を改悪した場合、ユーザーは有益な形で通知されることがないのが一般的である。

通知と同意は、まさにビッグデータによってもたらされる利益、すなわち新しく非自明的で予期せぬほどに強力なデータの使用法のせいで、有効な政策ツールとして機能しなくなっている。個人が新しい状況またはアプリのそれぞれについて細かな選択をするのは、あまりに複雑すぎる。とはいえ、通知と同意は現在の慣行に深く根付いているため、どうすればその有用性を高められるかについては一考の価値がある。

通知と同意の問題の一つとして挙げられるのは、プロバイダーとユーザーの間の暗黙のプライバシー交渉が不平等なものになるという点である。プロバイダーは巨大な法的な力を後ろ盾に、複雑で交渉の余地のない一連の条件を提示し、一方のユーザーは、実質上わずか数秒でその条件を評価せざるを得ない。というのも、ユーザーが目指す取引を完了するためには受諾が必要で、かつ提示される条件は、短時間で理解するのには難しい場合がほとんどであるからである。これは一種の市場の失敗である。別の文脈では、このような市場の失敗は、かなりの数のユーザーを代表して交渉を行うことができる第三者の仲介によって軽減することができる。以下の 4.5.1 節で、どうすればそのような仲介を実現できるかについて提案する。

4.4 その他の戦略と技法

4.4.1 匿名化または非識別化

ヘルスケアの研究や被験者が関与するその他の研究エリアで長年にわたって用いられてきた匿名化（非識別化とも呼ばれる）は、データが単独で使われ、特定の人と結び付けられることがなく、プライバシー規範に違反しない場合に適用される。たとえば、単に患者 X として識別され、実際の名前や患者の識別子が記録から除かれていれば、医療記録を研究で用いることは問題ないと言えるかもしれない。

データレコードの匿名化は、実行が容易であるように思えるかもしれない。残念なことに匿名化は、ビッグデータを様々な合法的な形で使用するために開発されている技術によって、簡単に打ち破られるようになってきている。概して、使用できるデータのサイズと多様性が増せば増すほど、個人を再識別できる（すなわち、レコードと個人の名前を再び結び付ける）可能性が大幅に高まる。

スウィニー、アブ、ウィン（Sweeney, Abu, and Winn）から説得力のある例が示されている。この三人は最近の論文の中で、郵便番号、生年月日、性別を含む公の個人ゲノム計画（Personal Genome Project）のプロファイルを公の有権者名簿と融合し、付属文書に隠れている名前をマイニングすることによって、名前が提供されたプロファイルの 84～97%が正しく識別されたことを示した。

匿名化は付加的な防護手段としては、依然として幾分有効であるものの、近い将来の再識別手段に対してはロバスト性がない。PCAST の見解では、匿名化は政策の有益な基盤ではない。残念なことに、匿名化はすでに法律に根付いており、特定の識別子がないデータは個人の識別が可能な情報ではないと見なされ、そのために家族の教育権およびプライバ

シー法（Family Educational Rights and Privacy Act (FERPA)）などの法律の対象外になるなど、プライバシーについて誤った印象を与えることがある。

4.4.2 削除と非保持

いかなる種類のデータも、価値がなくなった時点で削除するというのは明らかに望ましいビジネス慣行である。実際、経営のしっかりした企業では、特定の時間が経過した後に、ある種の（紙および電子の両方の）記録の破棄を義務付けている。それは多くの場合、そのような記録を保存する利益がほとんどないのに加え、取り出すコストが発生する可能性があるためである。たとえば、従業員の電子メールは、（たとえば）離婚弁護士による法的手続きの対象となる恐れがあり、マイナスの保持価値があると見なされることが多い。

しかし、ビッグデータは価値がないと見なされていたデータ・マスの経済的または社会的価値をしばしば突き止めることができるという新しい見解によって、この慣行に変化が生じてきている。保持の物理的費用が時間の経過とともに（とりわけクラウド上では）大幅に減少するのに伴い、政府も民間セクターもより多くのデータをより長く保持する傾向が高まると考えられ、それはプライバシーに影響を与えることが明らかである。保管データはまた、未来の歴史家、または学術研究者による後の縦断的分析において、重要になる可能性がある。

この流れに対抗できるのは政策介入のみである。政府は保持ポリシーを義務付けることができる。民間セクターに影響を与えるためには、規制の権限を持つ領域（たとえば消費者保護など）にポリシーを義務付けてもよいかもしれない。しかし同時に、アーカイブのデータも含め、個人に害を及ぼすデータを持つ企業に対して、より厳格な法的責任基準の策定を促すこともできる。そのような基準が策定された場合、民間セクターがとる合理的な対応は、保持するデータを減らすか、使用を保護するかであると考えられる。

上記の点は、公然と所持されているプライバシーに対してセンシティブな個人データ、すなわち自分がそのデータを所持していること、またそれが誰に関連するものかを所持者が知っているデータについて当てはまる。しかしながら、3.1.2 節で述べたように、データ・ソースはますます個人についての潜在情報、すなわち所持者が分析資源を（場合によっては経済的に実行可能な以上に）費やすことで初めて判明する情報、または今後新しいデータマイニング・アルゴリズムが開発されることによって初めて知り得る情報を含むようになっている。そのような場合、データ所持者が「個人についての全データ」を明らかにすることさえ実質上不可能で、ましてや指定されたスケジュールに沿ってそれを削除することは全く不可能である。

短命（データをオンザフライ、または短期間のみ保存すること）と透明性（自分についてのどのようなデータが保持されているかを個人が知ることができるようにすること）の概念は密接に関連しており、同じように実質的な制限がある。データがストリーミングされるだけで、アーカイブに保管されず、将来的な使用のリスクが低いと考えられても、侵入者が想定された法則に沿って行動するという保証はない。ターゲットの 1 億件のデビットカードの PIN が盗まれたときにも、PIN はそれぞれ束の間保存されていたに過ぎなかった（4.2.1 節を参照）。

現在では、データストレージの冗長および分散的な性質を考えると、データを確実に破壊することが可能であるかさえ定かではない。データ破壊についての研究は現在進行中であるが、データはユーザーの目または耳に（「アナログ」で）表示された瞬間に、技術的な保護なくコピー（「再デジタル化」）され得るというのが基本的な事実である。データが暗号化されていない形で不正コンピュータ・プログラム――技術的防護手段を出し抜くように設計されたもの――に一度でも利用された場合にも、同様のことが当てはまる。誤解に基づいた公開議論もあるが、自動的削除データなどというものは、規則を遵守する完全にコントロールされた環境以外においては、存在しない。

最新の例としては、スナップチャット（SnapChat）が、受信者の指定されたモバイル機器に、数秒間だけ表示される短命のスナップショット（画像）を配信するサービスを提供している。スナップチャットは、過去のスナップショットをサーバーから削除すると約束しているが、それは単に約束に過ぎない。加えて、対象とする受信者がコントロールされない期限切れのないコピーを作成しようと試みることはないという約束は慎重にも避けている。実際、スナップチャットの成功は、まさにそのような複写アプリを開発するインセンティブになっている。

政策策定の観点から言うと、現在および予測可能な将来における持続可能な唯一の想定は、データはひとたび作成されると永続的であるということである。*使用*は規制できるかもしれないが、データが存在し続けることは、控えめに言っても不変の事実であると見なすのがベストである。

4.5 将来のロバストな技術

4.5.1 通知と同意の後継

通知と同意の目的は、ユーザーが自分にとって受け入れ可能な明示された目的のために、

個人データの収集と使用に同意することである。個人データを収集して使用するプログラムおよびインターネットが利用できるデバイス――その中には見えるものと見えないものの両方がある――の数が多いことにより、この枠組みは次第に効力を失い、機能しなくなってきている。PCAST としては、個人データをユーザーの選好に従って使用する責任はユーザーではなく、できれば双方に受け入れ可能な仲介者による支援の下で、プロバイダーが負うべきであると考える。

それはどのようにすれば可能か。個人に対して、第三者が任意で提供するプライバシーの選好プロファイル（すなわち設定または選択）の標準セットの一つに加入するように奨励するのもよいであろう。たとえば、ジェーンは個人の権利に特別な重きを置く米国自由人権協会（American Civil Liberties Union）が提供しているプロファイルに加入し、ジョンは消費者にとっての経済的価値に重きを置くコンシューマ・リポーツ（Consumer Reports）が提供しているプロファイルに加入する選択をするかもしれない。評判の価値を重視する（アップル・アプリ・ストア、グーグル・プレイ、マイクロソフト・ストアなどの）大きなアプリ・ストア、または金融などの大きな商業セクターが、競合するプライバシーの選好プロファイルを提供するようになる可能性もある。

最初の事例では、プロファイルを提供する組織は、新しいアプリが各々のプロファイルの範囲内で容認可能か容認可能でないかを吟味する。基本的にこれらの組織は、ユーザーがつぶさに読むべきであるが、実際にはそうはしないプロバイダーの通知をつぶさに読む。これは面倒なことのように思えるかもしれないが、実際にはそれほど面倒ではない。アプリは数百万もあるが、最も人気のあるダウンロードは比較的少数で、限られた数のポータルに集中しているからである。当初、それぞれ顧客が少ないアプリの「ロングテール」は、「未評価」のまま残されることになるかもしれない。

単純にアプリを吟味することで、第三者組織はプライバシーの共同体基準の交渉のための市場を自動的に創出することになる。マーケット・シェアを伸ばすため、プロバイダー（とりわけ小さなプロバイダー）は、自己の通知が可能な限り様々な第三者が提供する、できる限り多くのプライバシー選好プロファイルにおいて適格とされるように努めると考えられる。連邦政府は（米国標準技術局を通じるなどして）、プロバイダーと評価者がプライバシーの意味合いと設定について意思疎通するための標準的な機械可読インターフェースの開発を奨励することができる。

現時点では人間の専門家が自然言語で表されたポリシーを使って吟味してもよいが、将来的には、そのプロセスを自動化するのが望ましい。そのためには、プライバシーポリシーを特定するための形式と、ソフトウェアを分析してこれらのポリシーに適合しているかを

判定するためのツールがある必要がある。しかし、それは取り組むべき課題の一部でしかない。もっと大きな課題は、ポリシーの言葉が表現として十分であるか、ポリシーの内容が充実しているか、適合性試験は威力が十分であるかを確認することである。このような要件から、コンテキストと使用についての考慮が必要になる。

4.5.2 コンテキストと使用

これまでの議論、とりわけ 3.1 節および 3.2 節の議論は、電子個人データの収集、ストレージ、保持に焦点を絞っても、将来の政策の技術的に頑健な基盤にはならないという PCAST の見解を示したものである。これらの問題に触れた多くの著者の中でも、ケーガン（Kagan）とアベルソン（Abelson）は、アクセスコントロールがなぜプライバシーを保護するのに十分でないかを説明している。マンディ（Mundie）は、この問題について説得力のある、より完全な説明をした上で、収集をコントロールするよりも、メタデータやアナリティクスから導出されるデータを含む、広義での個人データの使用をコントロールしたほうが、プライバシー保護に役立つと主張している。これを補足する形で、ニッセンバウム（Nissenbaum）は、いかなる使用が許容可能かは、使用のコンテキストと浸透した社会規範の両方に左右されると説明している。

特定の目的（すなわちコンテキスト）のための個人データの使用にプライバシーポリシーを意味のある形で適用できるようにするためには、プライバシーポリシーをデータとそのデータのオペレーションコードの両方に関連付ける必要がある。たとえば、必ず一定の特性を持つアプリのみが特定のデータに適用できるようにしなければならない。プライバシーポリシーをコンピュータ科学者が言う自然言語（わかりやすい英語またはそれに相当するもの）で表し、ユーザーが関連付けを行ってもよいし、プライバシーポリシーを形式的に表し、関連付けと施行を自動的に行ってもよい。いずれの場合にも、計算のアウトプットと関連付けられたポリシーがなければならない。これもまたデータであるからである。アウトプット・データのプライバシーポリシーは、インプットと関連付けられたポリシー、コードと関連付けられたポリシー、およびアウトプットの意図する使用（すなわちコンテキスト）から計算されなければならない。このようなプライバシーの特性は、一種のメタデータである。合理的な水準の信頼性を獲得するためには、その実行は、データがコピーされる際に、耐タンパ性があり、「粘り強く（sticky）」なければならない。

そのような能力の開発に資する領域の研究が多くなされており、中には商業化が開始しつつあるものもある。これまでも、使用をコントロールするために、データベース・システムのメタデータ（「タグ」または「属性」）が使用されてきた歴史がある。プライバシーポリシーとその統合の形式化は研究課題であるものの、最新の製品の中には、そのようなポ

リシーを人が解釈し、使用タグを人が決定するものもある。（ユーザーおよびその役割、すなわちコンテキストを認証する）ID 管理システムについても、研究と実用化の両方が進んでいる。

使用コントロールを実行するための商業用プライバシー・システムは現在、信頼データ・フォーマット（Trusted Data Format(TDF)）という名前で存在している。これは主として米国の諜報コミュニティのために開発されたものである。TDF はファイルのレベルで機能する。主としてカスタム・ベースで大手のコンサルティング会社によって実行されており、多くの場合、オープンソースのソフトウェア・コンポーネントから構築される。主として現在の顧客は、連邦情報機関や地方自治体の犯罪情報班などの政府機関、または金融サービスなどの垂直に統合された業界の大手営利企業や、アカウンタビリティおよび監査能力の向上を目指している製薬会社である。このようなシステムを構築する専門技能を持つコンサルタント会社の中には、ブーズ・アレン、アーンスト・アンド・ヤング、IBM、ノースロップ・グラマン、ロッキードなどが含まれ、パランティア（Palantir）などの製品主体の企業や、内部使用監査、政策アナリティクス、政策推論エンジンを開発する新興企業もまた、そのような技能を持つ。十分な市場の需要があれば、今後 5 年で市場浸透が広まると考えられる。アマゾンやグーグル、マイクロソフトなどの主要なクラウド・プラットフォーム・プロバイダーがその提供物の中で使用がコントロールされたシステム技術を実装すれば、市場浸透はさらに加速すると考えられる。政府を通じて使用を拡大すれば、既製の標準ソフトウェアを作成する動機付けになる。

4.5.3 施行と抑止

プライバシーポリシーとコンテキストにおける使用のコントロールは、実現され施行される限りにおいてのみ有効である。違反者が捕まる確率を上げる技術的手段は、違反者を抑止する民事または刑事罰付きの規則や法律があって初めて奏功する。そうなって初めて、有害な行為の抑止とプライバシー保護技術を展開するインセンティブの両方が生まれる。

現在ではメタデータをデータと関連付けるのは技術的に容易で、粒度は個々のデータ、レコード、全収集で様々である。このようなメタデータは、出所、詳細なアクセスおよび使用ポリシー、認証、実際のアクセスと使用のログ、破棄の年月日などの多様な監査可能情報を記録することができる。プライバシー・アウェア・ロギングに加え、そのようなメタデータを派生または共有データ（二次的使用）に拡大すれば、監査を容易にすることができる。最先端技術は依然としていささかその場しのぎのもので、監査は自動化されていない場合が多いものの、いわゆる監査可能なシステムは徐々に配備され始めている（4.5.2 節）。とりわけ監査が自動化され、持続的なものになれば、プライバシーポリシー違反を

突き止める能力は、プライバシー侵害を抑止するのにも、違反者が確実に処罰されるようにするためにも使用できる。

次の 5 年で、規制または市場主導の奨励によって、大規模なクラウドベースのインフラ・システム（グーグル、アマゾン、マイクロソフト、ラックスペース（Rackspace）など）がアカウンタブル・システムのデータの出所および使用遵守の側面をクラウドのアプリケーション・プログラミング・インターフェース（API）に統合し、ポリシー周知のために API を付加的に提供するようになる可能性がある。そうなると、このような能力は、（ラックスペースと関連する）オープンスタック（Open Stack）などのオープンソース・ベースのシステムやその他のプロバイダー・プラットフォームに容易に取り込める。そのようなクラウドベースのシステムで実行することを意図したアプリケーションは、たとえ小企業や個人の開発者に開発されたものでも、プライバシーの概念を「焼き付けた」上で構築できることになる。

4.5.4 消費者プライバシー権利章典の運用

2012 年 2 月に、オバマ政権は消費者プライバシー権利章典（CPBR）について説明した報告書を発表した。CPBR は個人データの商業的（公共セクターではなく）使用を取り上げたもので、米国のプライバシーに関する価値観の強力な声明である。

ここでの議論の目的上、CPBR で表された原則を二つのカテゴリーに分類することができる。一つはデータ保持者、分析者、商業的使用者の義務である。これらは消費者の見地から言えば受動的――消費者が知っている、気にかけている、または行動するかに関係なく、果たされなければならない義務である。もう一つは消費者のエンパワーメント――消費者に権利を与え、積極的に開始させるべきこと――である。ここでは、CPBR の原則をカテゴリーごとに整理し直すことが有益である。

義務のカテゴリーには以下の要素がある。

- コンテキストの尊重：消費者は、自分がデータを提供したコンテキストに整合するやり方で企業が個人データを収集、使用、開示することを期待する権利を持つ。
- 的を絞った収集：消費者は企業が収集し保管するデータに合理的な制限を課する権利を持つ。
- セキュリティ：消費者は個人データが安全かつ責任あるやり方で取り扱われることに対する権利を持つ。

- アカウンタビリティ：消費者は消費者プライバシー権利章典を遵守するための適切な措置を導入している企業に個人データを取り扱わせる権利を持つ。

消費者のエンパワーメントのカテゴリーには以下の要素がある。

- 個人によるコントロール：消費者は、企業が自分に関するどのようなデータを収集し、どのように使用するかをコントロールする権利を持つ。
- 透明性：消費者はプライバシーとそれを保護する対策についての理解しやすく、アクセスしやすい情報に対する権利を持つ。
- アクセスと正確性：消費者は使用可能な個人データにアクセスし、データの機密性およびデータが不正確な場合に消費者に不都合な結果が生じるリスクに応じたやり方でデータを訂正する権利を持つ。

PCAST は CPBR の基盤となる原則を健全なものとして支持する。しかしながら、ビッグデータに付随する技術が急速に変化しているため、CPBR の効果的な運用は危険にさらされている。現在まで CPBR の運用方法に関する議論は、データの収集、ストレージ、保持に焦点を絞り、CPBR の策定の動機付けとなった「スモールデータ」のコンテキストに力点を置いていた。本報告書の複数箇所（3.1.2 節、4.4 節、4.5.2 節など）で述べたように、そのようなアプローチは、ビッグデータにも適用される将来の政策の基盤として技術的に頑健でないというのが PCAST の見解である。加えて、アプリケーションやデータの使用がますます複雑さを増していることから、「通知と同意」のような単純な概念さえも揺らいでいる。

PCAST は、データの有害な使用の認識とコントロールに基づいたより頑健な体制に CPBR の原則を適応させることは容易であると考える。以下に具体的な案を提示する。

まずは上でデータ保持者の義務に分類した権利について見ていく。

コンテキストの尊重の原則は拡大の必要がある。本報告書で繰り返し述べてきたように、個人データが顧客から提供されていない場合もある。そのようなデータは、データが収集されてから相当な時間が経過した後に、また場合によっては複数の手を経た後に行われた分析の産物である可能性がある。この権利の意図、すなわち個人に不都合な結果または害を及ぼすことのない正当な目的のためにデータを使用させるという意図は適切であるが、「消費者がデータを提供」という CPBR の表現はあまりにも限定的である。この権利は、個人についてのデータは――どのような方法で獲得されたかにかかわらず――その個人に不都合な結果または害が生じるような形で使用されるべきではないことをなんらかの方法で明記し

たものである必要がある。（なんらかの規制の対象となり得る不都合な結果や害のリスト案については、1.4 節を参照のこと。）

当初の考えでは、的を絞った収集の権利は、非識別化やデータの削除などの技術で実現される予定であった。しかしながら、4.4.1 節で論じたように、データフュージョンによって、非識別化（匿名化）はビッグデータにとってロバストな技術ではなくなった。有益な目的のためにデータを保持するやむを得ない理由がある場合もある。この権利は収集ではなく、使用に関するものであるべきである。非識別化に頼るのではなく、データの全ライフサイクルを通じてデータの不適切使用を防止するベストプラクティスを活用することを強調するべきである。消費者に関するどのようなデータを保持しているかを企業が「すべて」認識できていると考えるべきではない。そのようなことは、技術的に不可能になりつつある。

CPBR のセキュリティとアカウンタビリティの基盤となる原則は、使用をベースとする体制でも依然として有効である。データ収集、分析、使用を含むバリューチェーンの全体に適用される必要がある。

次に消費者のエンパワーメントに分類した権利について見ていく。

消費者が有意に行使することが事実上不可能になった消費者のエンパワーメントについては、データまたはデータ分析の成果物を実際に使用する商業団体の義務として作り直す必要がある。これは CPBR の個人によるコントロールおよび透明性の原則に当てはまる。

ビッグデータの分析の成果物の非自明的な性質のため、個人が新しい状況またはアプリのそれぞれについて、詳細なプライバシーの選択を行うことがほぼ不可能になったことを 4.3 節において説明した。個人によるコントロールの原則が意味を持つためには、接触する企業のすべてについて、「通知と同意」などの枠組みによってプライバシーを管理する責任を消費者に負わせるべきでないと PCAST は考える。むしろ、企業の側がそれぞれ、消費者が指定し、その企業に提供した（消費者が指定した第三者から提供される場合もある）個人のプライバシー・プロファイルに個人データの使用法を合致させるべきである。4.5.1 節において、このような責任の所在の変更のためのメカニズム案を示した。

（プライバシー慣行の開示という意味においての）透明性にも、同様の問題の多くが付きまとう。今日、消費者は対処しきれないほど大量のプライバシーポリシーの通知を受け取るが、その多くは事実上、「私たちプロバイダーは好きなことを何でもできる」という内容である。個人によるコントロールと同様、消費者が指定した個人のプライバシー・プロファイルに適合する責任は企業が負い、企業がそのプロファイルを受け入れられない場合には消費者に

通知をするべきである。企業は顧客を失いたくないため、これによってプライバシー慣行を競い合うポジティブな市場力学が生まれることになる。

アクセスと正確性の権利が意味を持つようにするためには、個人データは単に収集ではなく、データ・アナリティクスの成果物も含まれなければならない。しかし、本報告書ですでに説明したように（4.4.2 節）、企業が消費者について「何を知っているかを知る」ことは常に可能であるとは限らない。データの中でその情報は認識されていない場合や、将来的にデータセットを新しいアルゴリズムを使って結合して初めて識別可能になる場合があるからである。しかしながら、データを使用することで、そのデータの個人的特性が企業に明らかになったときには、誤りを修正する手段を提供する義務が発動されるべきである。企業が分析から生じるデータを認証、修正し、また誤りはすべて修正されることはないため、不正確なデータの使用によって消費者に不都合な結果が生じるリスクを最小限に抑えるための措置を講じるという期待を消費者が持てるようにするべきである。ここでも、主たる責任は消費者ではなく、ビッグデータの商業的使用者に課せられなければならない。

5. PCAST の展望と結論

プライバシーの侵害は、個人およびグループに害を及ぼす可能性がある。可能な場合にはそのような被害を予防し、被害が発生したときには救済策を推し進めるのが政府の役割である。プライバシーの技術的な強化は、規制や法律が伴っているときのみ有効である。というのも、なんらかの罰則が適用されない限り、違反者と保護者の間で際限なく、対抗措置をぶつけ合う「ゲーム」がエスカレートしていくからである。規則は有害な行為の抑止にも、プライバシーを保護するソフトウェア技術を展開するインセンティブにもなる。

これまでの議論から、ビッグデータの新しいソースが豊富であること、それが増加し続けること、また多大な経済的および社会的利益をもたらす可能性があることが明らかなはずである。同様に重要なのは、新しいアルゴリズム、ソフトウェアおよびハードウェア技術によって、データ・アナリティクスの威力が予期せぬ形で増し続けるという点である。このようなデータ集約と処理の新しい能力を考えると、個人についてのバルクデータと細粒度データの両方の意図せぬ漏えいの可能性、また意図する者によるプライバシーへの新しい体系的な攻撃の可能性が新たに生じることは避けられない。

カメラ、センサー、およびその他の観測またはモバイル技術は、プライバシーに対する新たな懸念を提起する。個人はそうとは気づかぬうちに、データの提供に同意していることがしばしばある。このようなデバイスは、その主たる目的とは無関係なデータを自然に引き寄せる。データ収集は目に見えないことが多い。分析技術（顔、シーン、音声、言語認識技術など）は急速に進歩している。モバイル機器は、その他の形では自発的には提供されないと考えられる位置情報を提供する。これらのソースからのデータを結合することで、影響を受ける個人は気づかぬまま、プライバシーの脅威となる情報が生成される可能性がある。

しかしながら、プライバシーに対してセンシティブなデータは、最初の収集時に確実に認識できるとは限らないことも確かである。というのも、プライバシーにセンシティブな要素は、データの中に隠れていて、アナリティクス（これから発明されるものも含む）またはその他のデータソース（まだ知られていないものも含む）との融合によって初めて、明らかになる可能性があるためである。よって、プライバシーに対してセンシティブなデータの収集を抑制することは徐々に難しく、また非生産的になると考えられ、ビッグデータの社会的に重要な利益や経済的利益の発展を妨げることになる。

加えて、複数のソースおよび複数の種類のデータの結合を抑制することも望ましくない。ビッグデータの威力の多くは、この種のデータフュージョンによってもたらされる。とはいえ、かなりの量の個人データがデータフュージョンから導出される可能性があるというのは、

憂慮すべき問題であることに変わりない。換言すると、そのようなデータは、意図的な個人情報の開示なしに、獲得または推測できるということである。

ビッグデータの特定の収集および特定の種類の分析には、有益な用途もプライバシーにとって不適切な用途もあることは紛れもない事実である。データおよび分析の双方の用途が適切かどうかは、文脈に大きく左右される。

具体的な害または悪影響は、データまたはその分析の成果物が、バリューチェーンの中の三つの異なる種類の主体の支配を経る結果、発生する。

第一の主体は*データ収集者*である。データ収集者は、個人または環境へのインターフェースを支配する。データは明らかに私的な領域（健康に関するアンケートやウェアラブル・センサーなど）から収集されることも、どっちつかずの状況（パーティで撮影された携帯電話の写真やグーグル・グラスのビデオ、または外部に中継放送するために教室に置かれたカメラとマイクなど）から収集されることもある。あるいは、プライバシーに対してセンシティブなデータが隠れていて、当初には認識できない可能性がある「公共広場」から収集されることもあり、このようなデータは量が増加し、質も向上している。

第二の主体は*データ分析者*である。ここでビッグデータの「ビッグ」が重要になる。分析者は多くのソースからのデータを集約することもあれば、他の分析者とデータを共有することもある。収集者と区別される分析者は、大規模なコンピュータ環境にアルゴリズムとデータセットを寄せ集めることで、用途を生み出す（「分析の成果物」）。重要なのは、個人がデータフュージョンまたは統計的推測によってプロファイルされる可能性のある中心地が分析者であるという点である。

第三の主体は*分析されたデータの使用者*――企業、政府、または個人である。使用者は分析者と商業的関係を持つのが一般的である。すなわち、分析者の分析の成果物の購入者やライセンシー（など）である。望ましい経済的および社会的成果を生み出すのは使用者である。しかしながら、実際の悪影響や害が発生する場合、それを生み出す中心地となるのも使用者である。

5.1 政策介入の技術的可能性

新しい法律によって、あるいは既存の規制の範囲内で生み出される政策は、原則として、上述のバリューチェーンの様々な段階に介入することができる。ビッグデータの社会的および経済的利益を実現しようとする場合、そのような介入のすべてが技術的な観点から等し

く実行可能なわけでも、等しく望ましいわけでもない。

第 4 章で述べたように、収集のコントロールに的を絞った政策は、明確に私的なコンテキスト（健康データの測定または開示）や、現在はまだ存在しないものの、「意味のある」明示的または黙示的な通知と同意（たとえば、プライバシーの選好プロファイルなどによる通知と同意など、4.3 節および 4.5.1 節を参照）の可能性がある非常に限定された状況を除けば、成功する可能性は低い。

技術的には、「忘れられる権利」や類似する保持の制限が有意義に定義または実行できる可能性は低い（4.4.2 節を参照）。今後は、個人についてのデータの「すべて」を表に出すことは、技術的に可能ではなくなってくる。匿名化による保護に基づいた政策は、追加データ量の増加に伴って再識別の実行可能性が急速に高まるため、無意味である（4.4.1 節を参照）。データとメタデータの有意義な区別は徐々になくなってきている。データフュージョン、データマイニング、および再識別の能力が発達することで、メタデータも問題を生む可能性という点でいえばデータとそれほど変わらなくなっている（3.1 節を参照）。

しかしながら、収集を直接的にコントロールすることがほとんどの場合不可能であるとしても、収集の慣行に注意を向けることで、リスクの低減に役立つ状況もあるかもしれない。出所の追跡、アクセスおよび使用の監査、継続的な監視およびコントロール（4.5.2 節および 4.5.3 節を参照）などのベストプラクティスは、政府と業界の提携（「飴」）に加え、不法行為法を明確にし、何が過失に相当するかを定義すること（「ムチ」）によって、推進することができる。

次にデータ分析者に目を向ける。一方においては、分析者を規制することは難しいかもしれない。というのも、その行動は直接的に個人に触れるものではなく（収集でも使用でもない）、外からは見えない可能性もあるからである。個人について単に推測するだけで、それを公表したり使用したりしなかった場合には、規制の対象とはなり得ないかもしれない。他方で、データ・アナリティクスを適用することで、より多くのプライバシーに関する問題が浮上する。多くのプライバシーに関する問題は、意図せずに収集した――すなわち収集時には特定の個人も、さらには特定のグループさえもターゲットにしていなかった――データを分析することで発生する。これは、多くのソースからのデータの結合が徐々に威力を増していくためである。

個人についてのデータの「特定の瞬間（moment of particularization）」に、またはこれが最小限の数の個人に同時に行われたときに、規制を導入することは可能かもしれない。そのような規制を有効なものとするためには、同時にデータの進化の全段階で、出所の追跡、ア

クセスおよび使用の監査、セキュリティ措置（ロバストな暗号化インフラなど）の使用を義務付け、また特定の瞬間に透明性および／または通知を義務付ける必要がある。

ビッグデータの「分析の成果物」は、価値のあるものを生み出すためにアルゴリズムとデータをまとめるコンピュータ・プログラムによって作られる。法的な意味において、そのようなプログラムまたはその成果物を特定し、その商取引を規制することは可能かもしれない。たとえば、個人のプライバシーの選択またはその他の共同体的価値の表明に一致していない限り、商業使用（販売、リース、ライセンス供与など）を許可しないということが考えられる（4.3 節および 4.5.1 節）。使用および作成するデータについて、出所、監査能力、正確性などの適切な基準への遵守を義務付けたり、誤りを正す責任を負い、成果物によって引き起こされる様々な種類の害または悪影響に対する法的責任を負うのは誰か（ライセンサーかライセンシーか）を有意義に特定することを義務付けたりすることがあり得る。

しかしながら、悪影響を引き起こす可能性があるのは、単なる分析の成果物の作成ではない。悪影響は、それが企業、政府、報道機関、または個人などに実際に使用されて初めて発生する。将来的には、規制を適用することが技術的に最も可能なのはこの部分であると考えられる。すなわち、害が（全くとは言わないまでも）ほとんど特定できないかもしれない遥か上流ではなく、害が生まれる可能性のある中心地に焦点を絞るのである。

分析の成果物が、個人を誤って分類し、悪影響を発生させる可能性がある不完全な情報を生み出す場合、データの正確性および保全性に関する基準を満たしていること、個人が任意に付加情報を提供して記録を訂正することを可能とするインターフェースがあること、また悪影響が一定の段階に達したときに、金銭的救済を含む救済のための効率的なオプションを準備することを義務付けてもよいかもしれない。

一部の害は、特定可能な個人よりも集団（貧困層、少数民族など）に影響を与える可能性がある。そのような場合の救済のメカニズムを編み出す必要がある。

プライバシー侵害による悪影響があった場合の法的責任の基準を明確にしておく必要がある。現在では、時代遅れの州法や判例の寄せ集めがあるのみである。技術的見識を豊富に盛り込んだサイバー不法行為についてのモデル法案を起草し、各州の検討のたたき台にすることを奨励してもよい。

最後に、たとえ民間セクターでは使用可能だとしても、政府については、特定の種類の使用を禁止してもよいかもしれない。

5.2 提言

本調査を実施するにあたって PCAST に求められているのは、特定のプライバシー政策を提言することではなく、異なる幅広い政策アプローチの技術的実現可能性の相対評価を行うことである。この問題に対する PCAST の総体的結論は、我々の提言のうちの最初の二つに反映されている。

提言 1　政策の的はビッグデータの収集と分析ではなく、実際の使用に絞られるべきである。

実際の使用というのは、個人または集団に悪影響または害を及ぼす可能性のある何かが発生する特定の事象を意味する。ビッグデータの文脈では、これらの事象（「使用」）はほぼ常に、未加工データまたは未加工データの分析の成果物のいずれかと相互作用するコンピュータ・プログラムまたはアプリケーションのアクションである。このような図式の中では、害を及ぼすのはデータそのものでもなければ、（データがない）プログラムそのものでもなく、両者の融合である。これらの（商業上、政府または個人による）「使用」事象は、規制の対象とするのに必要な具体性を持つ。プログラムとデータの融合の目的は、求められる識別可能なタスクを遂行することであるため、使用事象は、データ収集そのもの、またはプログラム開発そのものでは不可能と考えられる形で、なんらかの意図を浮き彫りにする。規制を必要とする段階にまで高じる悪影響または害とはどのような種類のものかという政策問題は、PCAST の調査対象外であるものの、米国共通の価値観に根ざしていると考えられる実例を 1.4 節で示した。

データの収集、ストレージ、保持、利用の先験的制限、および（ビッグデータまたはその分析の成果物の識別可能な実際の使用を伴わない）分析の規制に的を絞ったもう一つのビッグデータ政策は、プライバシーを改善するための効果的な戦略につながる可能性が低いと PCAST は判断している。そのような政策は、特定のデータセットに関して、どのような個人情報が中に隠れているか――または、現在もしくは将来の考えられる他のすべてのデータセットと融合させることでどのような個人情報が浮上する可能性があるか――を確認するのが徐々に困難になってきていることを考えると、時間の経過とともに拡張できる可能性も低い。これに関連して、収集と保持を制限する政策は、厳格で経済的に有害な手段以外によって施行できる可能性が徐々に低くなってきているという問題がある。定義可能なある種のデータについては、社会にあまりにも嫌悪を抱かせるため、単なる所持が刑事罰の対象になっているが、通常の商業もしくは政府の機能の範囲内にある大量のデータ、あるいは公共広場での情報収集についても、プライバシーに対する懸念を引き起こす可能性のあるビッグデータの情報と区別しにくくなっている。この情報の二重使用という

性質もまた、収集よりも使用の規制を主張する理由である。

提言 2　政府の全段階において、政策と規制は、特定の技術的ソリューションを埋め込むべきではなく、意図する結果という観点から明記されるべきである。

技術に後れを取るのを避けるため、プライバシー保護に関する政策は、メカニズム（「どのように」）を記述するのではなく、目的（「何」）を指定することが重要である。たとえば、匿名化の使用を定めることで健康情報の公開を規制しても、データフュージョンの威力への対処にはならない。また、学校が保有する学生の成績の記録の閲覧を規制することで未成年者についての情報の保護を規制しても、オンライン学習技術によって学生の情報が獲得されることへの対処にならない。データをどのように入手するかに関係なく、健康情報または学生の成績の不適切な開示を規制したほうが、確実である。

PCAST はさらに以下の提言を行うことで、課せられた責任を果たす。以下の提言は、強力なプライバシーの価値観と、それを支えるのに必要な技術的ツールという課題を推進することを意図したものである。

提言 3　OSTP との連携、後押しにより、NITRD 機関は、プライバシー関係の技術、またこれらの技術の応用を助ける社会科学の関連分野に対する米国の研究を強化するべきである。

使用をコントロールするための技術は、すでにいくつか存在している。しかしながら、プライバシー保護を助ける技術、プライバシー保護の行動に影響を与える社会的なメカニズム、および技術の変化に対して頑健で、経済的な機会と国家的優先課題とプライバシー保護との間の適切なバランスを生み出す法的オプションについての研究（またそのような研究のための財政的支援）が必要とされている。

プライバシー関連の研究を増やすべきであるという PCAST からの提言を受けて、情報技術に関する研究を支援している全連邦機関において、プライバシーに焦点を絞った研究を検証する政府内部調査が 2013～2014 年に行われた。その結果として、プライバシーの強化に明確に焦点を絞った研究か、その他のなんらかの目標（典型はサイバーセキュリティ）に付随してプライバシー保護に取り組む研究のいずれかに 8,000 万ドルの資金援助をすることが提案された。資金援助を受けた研究で取り組むのは、本人による情報のコントロール、透明性、アクセスと正確性、アカウンタビリティなどである。これらは、保健分野または（比較的新しい）消費者のエネルギー使用に焦点を絞った研究を除き、通常は一般的な性質を持つ。個人やセンターへの助成という形で、プライバシー研究に最も幅広く

多様な支援を行っているのは国立科学財団（NSF）で、対象は社会科学、コンピュータサイエンス、工学の分野にまたがる。

セキュリティの延長または補完としてのプライバシー研究は、国防総省の様々な機関（空軍研究所、陸軍の遠隔治療および先端技術研究センター、国防総省国防高等研究事業局、国家安全保障局、海軍研究事務所）と諜報コミュニティ内の情報先端研究プロジェクト活動（IARPA）によって支援されている。たとえばIARPAは、様々な暗号化技術を調査するセキュリティおよびプライバシー保証研究（Security and Privacy Assurance Research）プログラムを主催してきた。米国標準技術局（NIST）での研究は、プライバシーを強化するための暗号とバイオメトリック技術の開発、およびID管理のための連邦基準とプログラムへの支援に焦点を絞っている。

将来を見据えると、セキュリティに付随するプライバシーのトピックのみでなく、あらゆるソースからのデータの使用の最も幅広い側面に対するプライバシー保護の自動化にも、継続的な投資が必要である。関連トピックには、暗号学、プライバシーを保護するデータマイニング（格納データのみならずストリーミングの分析も含む)、プライバシーポリシーの形式化、ソフトウェアのプライバシーポリシーおよび法政策への適合を自動化するためのツール、コンテキストの中で使用を監査し、ポリシー違反を識別する手段、様々なビッグデータ分析の結果に対する人々の理解力を高める研究などがある。人々が複数の場所に保存しているデータを利用するようになると予想されていることに鑑みると、セキュア・マルチパーティ計算法など、分散データに対する質の高いアナリティクスとプライバシー保護の双方を支援する技術の開発がさらに重要になる。国家、州、地域、および国際規定の間の矛盾や相違を分析するツールを作れば、新しい規定の策定に役立つだけでなく、市場ごとにサービスをカスタマイズする必要のあるソフトウェア開発者にとっても有益である。

提言4　OSTPはしかるべき教育機関や専門家団体と協力し、専門職の創出を含め、プライバシー保護に関連する教育やトレーニングの機会を拡大することを奨励すべきである。

（セキュリティの専門知識獲得のための教育プログラムに類似する）プライバシーの専門知識習得につながる教育プログラムは重要で、奨励される必要がある。ソフトウェア開発の領域と技術管理の領域の両方のデジタル・プライバシーの専門職を創出することも可能かもしれない。プライバシーに焦点を絞った仕事（個人情報保護管理責任者など）が増加している産業界（およびあらゆるレベルの政府）の中だけでなく、個人に対して「毎年のプライバシー検査」を行う仕事など、消費者や市民への擁護や支援の分野にも、雇用の機会を創出するべきである。過去20年間で、技術者の世界では、サイバーセキュリティに

ついての教育とトレーニングが充実するようになった。それと同様に、現在では、コンピュータ・サイエンス・プログラムに焦点を絞った従来の小さなニッチエリアを超えて、プライバシーの意味合いとプライバシー強化についての教育とトレーニングを提供する機会が生まれている。プライバシーはまた、技術専門家の倫理教育の重要な要素でもある。

提言 5　米国は、現存する実用的なプライバシー保護技術の使用を奨励する政策を採用することで、国際的にも国内においても主導権を握るべきである。米国は、その集結力によっても（たとえば、基準作りや基準の採用を促進することによって）、またその調達慣行によっても（独自のプライバシー保護クラウドサービスの使用など）、リーダーシップを発揮することができる。

4.5.2 節において、現在の米国市場にすでにある一連のプライバシー強化のためのベストプラクティスを紹介した。PCAST が承知している限りでは、米国外でより効果が高いイノベーションや戦略が生み出されているふしはない。PCAST が袋小路と考える道を進んでいるように見受けられる国もある。このような状況は、米国が国際的にプライバシー技術でリーダーシップを発揮する機会を提供しており、米国はこの機会を逃すべきではない。公共政策は、政府調達を通じても、またそれよりももっと大きな、民間セクターの技術研究を動機付ける政策的枠組みを通じても、プライバシー保護技術の商業的潜在力を培うのを助けることができる。

セキュリティの分野と同様に、クラウド・コンピューティングはプライバシーに新たな好機をもたらす。米国政府は、契約を結ぶクラウド・サービス・プロバイダーにプライバシー強化サービスを要求することで、これらのプロバイダーに対し、中小企業が独力では入手できないと考えられる高度なプライバシー強化技術を中小企業とその顧客に提供するよう促すことができる。

5.4 結びの言葉

プライバシーは人間にとって重要な価値である。技術の進歩は、個人のプライバシーを脅かすと同時に、プライバシー保護を強化する機会ももたらす。米国政府と、それよりも大きな米国と世界の共同体にとっての課題は、現代世界においてプライバシーがどのような性質を持つかを理解し、プライバシーを維持、保護するための技術的、教育的、政策的手段を見つけることである。