プライバシーは一時的な概念です。  
–ロジャーニーダム  
「匿名化されたデータ」は、「健康的なアイスクリーム」や「選択的に破壊可能な暗号」などの神聖なグライルの1つです– CORY DOCTOROW  
  
Big Tobaccoが数十年かけて喫煙が肺癌を引き起こすことを否定し、Big Oilが数十年かけて気候変動を否定したのと同じように、Big Dataも数十年かけて機密情報を簡単に「匿名化」できるため、産業用原料として使用できると偽っていますデータ主体のプライバシー権を侵害することなく。  
その限界は4つの研究の波で探究されており、それぞれがその日の技術に対応しています。  
統計学者たちは、情報がどのように漏洩するかを研究し、推論制御の手段を開発し始めました。  
医療サービス管理者と医療研究者はどちらもこれを宝の山と見なし、患者の名前と住所を削除するだけでデータを非個人的なものにできることを期待していました。  
それ以来、匿名化が不十分なデータが漏洩したり、販売されたりしたときに、複数のスキャンダルが発生しています。  
2006年に理論の進歩があり、Cynthia Dworkらは、プライバシーの理論を開発しました。これは、クエリを制限してノイズを追加することで推論を防ぐことができる範囲を定量化し、必要な場所にノイズを追加できるようにします。  
2010年代後半、ソーシャルメディア、普及したゲノミクス、および電話アプリによって収集され、マーケティング担当者に広く販売されている個人の位置履歴の大規模なデータベースによって、4番目の波が訪れました。  
ますます多くのプレス記事は、そのような主張が通常どれほど偽物であるかを示しています。  
私たちは、匿名化と関連するプライバシー技術を使用して何ができるかと、医学研究者からマーケティング担当者から政治家に至るまでの利害関係者が何を信じたいのかというあくびのギャップに直面しています。  
再識別のリスクに関する私たちの知識がさらに詳細かつ確実になるにつれて、政府と産業界の両方の希望はますます非現実的になります。  
すべての政府が単に無知であったわけではないことを言わなければならない。  
、セクション26.6.1で後で説明します。  
  
推論制御は、個々の企業の貢献を隠す方法で経済データが編集された1920年代にさかのぼりますが、国勢調査データのコンテキストで体系的に最初に研究されました。  
この情報は、選挙区の決定、公共サービスへの政府資金のレベルの設定、およびその他のあらゆる種類の政策決定へのインプットとして使用されます。  
公開前にデータを一度完全にサニタイズするかどうか、またはプライバシーメカニズムが一度に1つのクエリを操作して、それが許容できるかどうかを判断するかどうかに応じて、2つの広範なアプローチがあります。  
特定のプライバシーの制約を受ける特定のタイプのデータの場合、特定の数のクエリのみが許可されます。問題は、これらを事前に決定するか、ユーザーの要求に応じて動的に決定するかです。  
千件に1件のレコードがテープで利用可能になりました-名前、正確な住所、その他の機密データは除外されました。  
 個人の追跡から。  
しかし、非常に高い収入などの極端な値を持つレコードは抑制されました。  
2番目のタイプの処理では、識別可能なデータがデータベースに格納され、プライバシー保護は、実行される可能性のあるクエリを制限することから生じます。  
人々は推論に基づいた新しい攻撃を思いつき続けるので、これに対する初期の試みはあまり成功しませんでした。  
たとえば、「50,000ドルから55,000ドルを稼いでいる2人家族の世帯の数を教えて」、「40歳から45歳までの男性が50,000ドルから55,000ドルを稼いでいる世帯の割合を教えて」、「割合を教えて」あなたがターゲットの個人に帰宅するまで、子供が成長して家を出た、50,000ドルから55,000ドルの間で稼いでいる男性が率いる世帯の割合です。  
関連する問題は多くの状況で発生します。  
低レベルのソースを組み合わせて高レベルの結論を導き出すことは、国家安全保障の文脈では集約攻撃として知られています。  
11.2.1推論制御の基本理論  
この研究の波は、AdamとWortmanによる1989年の調査論文に要約されています[17]。以下は基本的な考え方の概要です。  
 レコードのクエリセットを選択します。  
すべての属性（またはその否定）の論理ANDによって取得される最小のクエリセット  
クエリセットに対応する統計は、セットのサイズが小さすぎる場合、機密性の高い統計になる場合があります。  
Dを開示される統計のセット、Pを機密で保護する必要のある統計のセットとすると、プライバシーのためにD✓P 0が必要です。ここで、P 0はPの補数です。  
正確でない保護は、通常、データベースが応答できるクエリの範囲に関してある程度のコストがかかるため、その有用性が低下する可能性があります。  
11.2.1.1クエリセットサイズの制御  
しかし、これでは十分ではありません。  
また、攻撃者が1つを除くすべてのレコードにクエリを実行することを防ぐ必要があります。Nレコードがあり、クエリセットのサイズのしきい値がtの場合、tレコードとN ffiレコードの間が、許可されるためのクエリの対象でなければなりません。 。  
たとえば、私がこの本の最初の版を書いたとき、私たちの研究室の完全な教授の一人だけが女性でした。  
」と「平均給与男性教授？  
したがって、K⇢Lおよび| L |の場合、レコードセットKおよびLの連続クエリを回避する必要があります。 ffi | K | <t。  
11.2.1.2トラッカー  
一般的なトラッカー–機密性の高い統計を明らかにできるようにする数式のセットもあります。  
クエリセットの最小サイズnが統計情報の総数Nの4分の1未満であり、許可されるクエリの種類にそれ以上の制限がない場合、一般的なトラッカーを提供する式を見つけることができます[541]。  
このようなクエリ監査は、NP完全な問題であることが判明しました。  
11.2.1.3細胞抑制  
たとえば、2010年の国勢調査に関する英国の規則では、「自分自身を特定した統計単位が、その統計情報を推論によって使用して、国家統計の出力で他の統計単位を特定することはできない」[1416]。  
ここで、図11.1の表に、2つの理科科目を研究している学生の数が含まれているとします。  
したがって、「geology-with-chemistry」の平均を発表することはできません。  
したがって、化学行の他の少なくとも1つのマークを抑制する必要があります。同様の理由で、地質列のマークを抑制する必要があります。  
これで、表は図11.2のようになります。「D」は「開示目的で抑制された値」を意味します。  
データベーススキーマにさらに属性がある場合（たとえば、図も人種や性別で分類され、差別禁止法の遵守を示すため）、さらに多くの情報が失われる可能性があります。  
したがって、正確な保護を行っても、データベースはすぐに使用できなくなります。  
 全国的に集計され、地方自治体から除外されています。  
合計にノイズを追加できない限り、それでも十分ではない可能性があります。機密データの可能な値は、開示する情報によってさらに制限され、合計がマイナスにならないなどの副次的な情報もある可能性があるためです。 。  
11.2.1.4その他の統計的開示管理メカニズム  
その限界は、プライバシープロパティの数学的定義ではなく、プライバシーメカニズムの運用上の定義であることです。 k人すべてが同じ機密属性を所有している場合、あまり役に立ちません。  
別の方法は、問い合わせの種類を制限することです。  
ただし、効果的にするには、制限を厳しくする必要があります。  
より徹底的なアプローチ（可能な場合）  
前の章で、区画化されたセキュリティでラティスを使用して、コードワードの組み合わせを持つコンパートメント間で許可される情報フローの部分的な順序を定義する方法を説明しました。  
たとえば、A、B、Cの3つの属性（たとえば、居住地域、誕生年、病状）がある場合  
したがって、3つすべてについての問い合わせも許可されません。  
  
単純なクエリ制御にはいくつかの選択肢があります。  
その他の手法には、極端な値を持つデータを抑制することが含まれます。  
たとえば、その期間の英国の処方統計システムは、地域の統計によるエイズ薬の販売を抑制しました[1249]。 1990年代初頭のエイズ危機の間でさえ、そのような治療を受けている1人の患者だけの郡がありました。（C、A）  
 C A禁止許容B（B、C）  
クエリオーバーラップコントロールと呼ばれるこの方法では、ユーザーからのクエリを拒否し、ユーザーがすでに知っていることと組み合わせて、機密の統計情報を開示します。  
第1に、関連する処理の複雑さは時間の経過とともに、しばしば指数関数的に増加します。  
現在、ユーザーがすべて正直で明確な人物である場合でも、明日のユーザーが1人引き継がれる可能性は常にあります。  
11.2.1.6ランダム化  
そのため、クエリ制御は、さまざまな種類のランダム化と組み合わせて使用​​されることが多く、正当なユーザーのシグナルへの影響を最小限に抑えながら、攻撃者の視点からS / N比を低下させるように設計されています。  
彼らは摂動から始めるか、平均がゼロで既知の分散を持つノイズをデータに追加しました。しかし、これはサンプルセットのサイズが小さい場合に正当なユーザーの結果を正確に損なう傾向があり、サンプルセットが単純なクエリコントロールを使用するのに十分大きい場合はそのままにしておきます。  
次に、すべてのクエリセットを同じサイズにして、利用可能な関連統計からランダムに選択するランダムなサンプルクエリがあります。  
ランダムサンプルクエリは自然な保護メカニズムであり、調査対象の相関が十分に強いため、小さなサンプルで十分です。  
2006年以降、ランダム性を追加することでどの程度の保護を実現できるかについての確固たる理論があります。それは、差別的なプライバシーです。  
  
他の保護技術と同様に、統計的セキュリティは特定の環境で、特定の脅威モデルに対してのみ評価できます。  
1つの例は、1990年代半ばにSource Informaticsと呼ばれる会社が薬剤処方の傾向を分析するために開発したシステムであり、匿名データ1のプライバシーに関する英国の主要な訴訟で構成されました。  
プライバシーの目標は、特定可能な患者または個々の医師の処方習慣に関する情報を漏らさないことでした2。  
 薬局から、そしてさらに匿名化のさらなる段階で、医師のアイデンティティも取り除かれました。  
これを評価したところ、警戒薬担当者が処方パターンから医師を特定できることに気付きました。  
修正点は、処方の絶対数を、各特定の薬について行った各医師の処方のパーセンテージで置き換えること、一部の医師を無作為にドロップすること、および図を数週間前後にシフトすることによってタイミングをランダムに混乱させることでした[1249]。  
2Doctorsは常に麻薬営業担当者に追われており、多くの場合、手術から退出させるために何らかの製品を使用すると述べています。  
これは、従来の統計的セキュリティ技術が堅牢なソリューションを提供できるようなシステムの良い例です。  
それでも、英国保健省はデータベース運営者を訴え、データベースがプライバシーを侵害する可能性があると主張した。  
彼らは敗北し、これは（少なくとも英国では）前例を確立した  
ただし、一般的にはそれほど簡単ではありません。  
これは、実際にはソースインフォマティクスの場合に起こりました。 2015年までに、別のメカニズムを使用する別の競合システムが利用可能になり、人々は両方のシステムにアクセスできる製薬会社が時々一部の医師の処方行動を推測できることに気づきました。  
  
Source Informaticsシステムは毎週新しいトランシェを追加しましたが、ユーザーが単一の識別可能なレコードをデータベースに挿入できる場合があります。  
1990年代後半の顕著な事例は、アイスランドの医学研究データベースでした。  
政府は登録しましたが、アイスランドの医師は主にこの契約に反対し、それを患者のプライバシーと専門職の自治の両方に対する脅威だと見なしました。  
プライバシーコミッショナーが暗号化キーを制御しました。  
 首相の医療記録は、何らかの記録（アスピリンの処方箋など）を入力するだけで、1、2秒後に研究システムにポップアップ表示されます。  
多くの医師が患者にオプトアウトするようアドバイスし、人口の11％がオプトアウトしました。アイスランドは人口が非常に均一で、1,000年前の少数の開拓者の子孫であり、優れた家系図があるため、研究者にとって特に魅力的でした。  
医療記録を公開されている系譜にリンクすることにより、患者は、叔父、叔母、大叔母、大叔母などの数によって、事実上、家系図の形状によって識別できます。  
これは、豊富なコンテキストデータのより広い問題につながり、推論制御に関する作業の第2の波を引き起こしました。  
11.2.3豊富な医療データにおける推論制御  
世界中のウェブは新しく、ビジネス（および政府）としてテクノロジーに流入した大量のお金  
ヘルスケアITの人々は、安全性とプライバシーに関する多くの質問に苦しんでいました。記録はすでに紙からコンピュータに移っていたが、今ではすべてのコンピュータが互いに話し始めた[63]。  
 どのように暗号化し、誰が鍵を管理しますか？  
 研究者は以前に病院の図書館に座って紙の記録を読んで疫学を行っていましたが、これをあなたの机でできれば「明らかに」良いでしょう。  
それは匿名化するのがはるかに困難です。  
ニュージーランドは、暗号化された患者名に加えて、6件未満のレコードに関してクエリが回答できないというルールでデータベースを設定しましたが、それは十分ではなく、特別にクリアされた少数の医療統計家へのアクセスが制限されていることに気付きました[1422] 。  
レジストリは保護メカニズムを迅速にインストールする必要があり、これには匿名化と厳格な使用制御の両方が含まれます[266]。  
英国医師会は1995年から6月に集中研究データベースの提案に反対し、有望な精神科医であるフィームカルディコット卿の下に委員会が設置され、今後の方向性が示されました。  
彼女は、ヘルスケアファイナンス管理局の「公共使用」ファイルでさえ、それらを商用データベースと相互相関させることで識別できることが多いことを示しました[1849]。  
居住者は、生年月日と郵便番号で識別でき、自由形式のテキストを含むあらゆる種類のコンテキストデータを含む医療記録をこするのが非常に難しいことについて話し合いました[1849]。  
また、「バーモント州に住む70歳から74歳の白人女性」など、患者が一般的な用語でのみ識別されるレベルまで、識別子を取り除いたパブリックアクセスレコードもありました。  
スウィーニーは、マサチューセッツ州知事ウィリアムウェルドの記録を確認することで、これを世間の注目を集めました。  
セクション10.4で説明したように、クリントン政権は2000年にHIPAAに基づいてプライバシールールを発行し、データの公開共有に関する「セーフハーバー」標準を定義しました。その後、2002年にブッシュ政権はより緩和されたルールを採用しました。  
参加者の誕生年から10年の範囲を編集した後でも、住居のタイプなどのサイド情報により、名前で3％、住所で18％を正確に特定できました。  
Dame Fiona Caldicottの報告によると、60以上の違法情報が医療サービスに流れ込んでいます[367]。  
 その後、故意に再識別されたため、匿名性の約束のもとにデータが収集された人々とHIV慈善団体は騙されました。  
患者の同意はいくつかのケースで得られましたが、1998年から現在までのイングランドとウェールズでの10億を超える病院治療の記録が含まれている病院エピソード統計データベースに関する調査ではありませんでした。  
（暗号化キーは、データをライセンスする研究組織ごとに異なります。）  
トニーブレア前首相の記録を調べたいとしましょう。  
彼の暗号化されたIDを見つけて、彼が行った他のすべてを調べるにはそれで十分です。  
さらに、多くのシステムでは、平文の郵便番号と生年月日があります。この組み合わせでも、英国の居住者の約98％を特定するには十分です3。  
たとえば、「36歳の女性が14歳と16歳の娘を持つ母親の記録を見せて、母親とちょうど1人の娘が乾癬を持つようにする」などのクエリでは、数百万人から1人を見つけることができます。  
郵便番号と生年月日が一意ではない人々の1％程度は、ほとんどが一卵性双生児、または大学の寮や軍の兵舎に住んでいる若者です。そのようなクエリは、故意であれ、偶然であれ、個人を特定するような方法で作成できます。  
2007年、議会の健康選択委員会は電子患者記録を調査し、幅広い視点から証拠を聞き4、患者が研究でのデータの使用を禁止することを許可する必要があることを含め、多くの推奨事項を作成しました[925]。  
政府はこれを拒否しました。  
セクション10.4.4.3で詳しく説明したように、レッドテープを廃止し、英国を医学研究にとって世界で最高の場所にすることについての話し合いの中で、彼はテスト結果を追加する中央研究データベースである 'care.data'を立ち上げました。既存のHESデータベースへの処方箋とGPデータ。  
米国の大規模なコンサルタント会社の1つが、23 GBのデータをすべて「Googleのクラウドに」アップロードしました。これは、データを英国に残すことを要求する法律にもかかわらず、「Excelには大きすぎる」ため、クライアントが利用できるようにするためでした。  
健康と社会のデータを共有および分析できるのは「医療に利益がある」場合にのみ、他の目的には決して使用できないとの法律がすぐに可決されました。  
アカデミックケーススタディは、ストーリーを説明し、ヘルスケア法とデータ保護法の間の緊張を分析し、「この議論は患者データの匿名性を保護および維持する能力に集中し、簡単な答えはありません」と述べています[ 1548]。  
11.2.4第3の波：設定と検索  
その年の2つの事件により、この家は一般に公開されました。  
検索者の名前とIPアドレスは数字に置き換えられましたが、それは役に立ちませんでした。  
データは「調査目的」で公開されました。リークによりFTCに提出された苦情につながった後、会社のCTOが辞任し、会社は両方の従業員に関心を表明しました。 。  
検索履歴、または同等にあなたのクリックストリームは、あなたの考えや意図を反映しているため、非常に敏感です。  
当時、米国の顧客は600万人しかなく、物理DVDを出荷していたため、これは顧客の少数派でした。  
これは、「ロングテール」の影響が原因の1つです。誰もが視聴する100本程度の映画を無視すると、人々の表示設定はかなり独特になります。  
ヨーロッパとカナダのプライバシー規制当局の対応は、プライバシー強化技術（PET）を促進することでした。  
マイクロソフトの研究者たちは、彼らの言葉を彼らに受け入れ、差別的プライバシーの理論を発展させました。これについては、11.3で説明します。  
しかし、長年にわたり、政策立案者は、詳細なデータに対する研究者の要求とプライバシーに対するデータ主体の権利との間の緊張を解決できない理由をより詳細に説明していることを理解せずに、ソリューションとしてそれについて話しました。  
11.2.5第4の波：ロケーションとソーシャル  
この本の第2版の2008年の第23章では、FacebookがMyspaceから引き継いだばかりだったため、初期のソーシャルネットワークシーンについて説明しています。  
ソーシャルネットワークがヒットしたスイートスポットは、これをすべての人に展開することでした。  
ソーシャルコンテキストでは隠すのが難しいため、ソーシャルネットワークではあらゆる種類のプライバシー問題が直接発生すると予測しました。  
）  
私たちが見逃していた2つのことは、大量のデータがクラウドに移行したという事実と、人々に関するコンテキストデータから推定できる膨大な量の機密個人情報です。  
「その他のデータ」の例はロケーション履歴です。  
多くのスマートフォンアプリが現在地へのアクセスを要求するため、今日ではより多くの高解像度データが広く利用可能です。これにはGPSだけではなく（屋外での平均精度は8m程度）  
。  
データはマーケティング会社だけでなく、保釈金を飛んだ人を追跡するためにそれを使用する賞金稼ぎを含む私立探偵にも売られた[489]。  
あなたの毎日の軌跡は、あなたの家、あなたが去ったとき、あなたが仕事に旅行した方法、途中でコーヒーを飲みに立ち寄った場所、あなたのオフィスがある場所、あなたが昼食に行った場所などすべてを示しています。  
彼らは、Microsoftで働いていて、Amazonを訪れ、翌月Amazonで働き始めた男性を1人見つけました。  
このデータを公開市場で簡単に購入できることと、法執行機関が令状によってデータを取得するためにジャンプしなければならないこととの間には、明確な対照があります。低解像度の位置データを使用して、ラスベガスのブラックハットに行くと、オンラインギャンブル会社が広告を目の前に表示できます。  
それは彼らの家を追うこともできます。  
ここでの見出しの事件は、Michal Kosinskiと同僚が無料の心理テストを提供するFacebookアプリを作成し、それを数万人に使用するように説得したときに始まりました。  
彼らはまた、あなたが白人か黒人か、保守的かリベラルか、クリスチャンかイスラム教徒か、喫煙か、飲酒か、あなたが人間関係にあるかどうか、そして薬物を使用しているかどうかをさまざまな程度で正確に知ることができます。  
後の調査で、行動データがあると、パブリッシャーはコンテンツ広告に比べてわずか4％の広告収入しか得られないことがわかったため、この手法は単に禁止されている可能性があります[1239]。  
多くの場合、位置データとソーシャルデータの両方を取得して、大規模に取得できます。  
カードIDは匿名化されていましたが、通常、居住者がタッチオンとタッチオフの時間から自分のカードを識別するのに要する時間は1〜2回でした。研究者たちは、彼らが彼らの同行者を特定できることを発見した[502]。  
このデータセットにより、研究者は移動時間の感度を分析できました。  
現在、多くのソーシャルサイドチャネルと位置データがあります。  
ソーシャルネットワークは、私たちの連絡先、設定、および選択によりさらに豊かになり、これらの測定をより正確にすることができます。  
たとえば、2つのソーシャルグラフを正確にコピーしなくても、2つのソーシャルグラフを一致させるのはかなり簡単です。そのため、国の匿名の携帯電話の通話データレコードが与えられた場合、それらを（たとえば）と比較することで再識別できます。  
携帯電話のデータはすでに私たちの性格に関する多くの情報を漏らしています。外向的な人はより多くの電話をかけ、快適な人々はより多くの電話をかけ、電話間の時間の変動は良心を予測します[1334]。  
GoogleのAI子会社であるDeepMindは、2016年にロンドンの病院と共同で、腎障害を診断するアプリを開発することを発表しました。  
そのようなアクセスは製品の研究のためではなく、直接的な患者のケアに関与している企業にのみ与えられるべきであるので、プライバシー規制当局は病院を懲戒した。ただし、DeepMindにデータの削除を強制することはしませんでした。  
それは技術を制御すると主張する倫理委員会を設置し、病院のデータを親のGoogleに提供しないことを約束しましたが、倫理委員会の著名なメンバーは、それが窓隠しであると主張し、2018年に辞任しましたGoogleがDeepMindの健康管理を吸収していることが発表された[909]。  
それで、匿名化を適切に行うことは可能ですか？  
質問への回答に使用できる匿名のデータセットを作成することはできませんが、特定の一連の調査質問に回答しようとする場合、信頼できるプライバシーの指標を提供できることがあります。  
  
2006年、Cynthia Dwork、Frank McSherry、Kobbi Nissim、Adam Smithは、データベースに機密の統計情報が開示されないようにノイズを追加したプライバシーシステムを体系的に分析する方法を示す独創的な論文を発表しました[595]。  
理論的な暗号に関する論文として始まったものの、統計データベースのセキュリティと一般的な匿名化の両方のゴールドスタンダードと見なされるようになりました。  
したがって、クエリの数が限られている場合よりも多くなると、防御側はノイズを追加する必要があり、問題はどれくらいかです。  
プライバシーメカニズムは、すべてのデータベースXとX0が単一の行で異なる場合に、Xから回答が得られる確率が（1 +✏）の乗数因子内にある場合、✏区別不可能と呼ばれます  
したがって、ラプラス分布でノイズを使用すると、ノイズのある合計と区別がつかなくなり、すべてが数学的に扱いやすくなります。  
小さな値は強力なプライバシーを与えます。ただし、setting = 1000に設定すると、基本的に生データが公開されます。  
しかし、人口についての有用な情報を学びながら、個人については何も役に立たないという約束は、実用的なアプリケーションで実現できますか？  
11.3.1国勢調査に差別的なプライバシーを適用する国勢調査。  
最初に、国勢調査局は最新の分析ツールに照らして2010年の国勢調査の安全性を検討しました[752]。  
 米国の居住者から収集され、重複を取り除き、納税申告書などのデータから欠落したエントリを埋めるために編集されたデータの、各居住者に関する44ビットの機密データ（合計1.7Gb）  
問題は、microdataサマリーにこれよりもはるかに多くのデータが含まれていることです。すべてを書き出すと、数十億の連立方程式が得られ、理論的には機密データを解くことができます。  
 国勢調査のスタッフは、Kobbi NissimとIrit Dinurの研究に基づいてアイデアを実装し、人口の20％未満をカバーして、すべての変数が約38％の確率で正しいことを確認しました。  
ただし、統計データベースのセキュリティに対する従来のアプローチは実際には機能しないという教訓があります。  
すべての世帯を入れ替えたとしても問題はありませんでしたが、ユーザーはそれに耐えられなかったでしょう。彼らがブロックの正確な人口数を提供したという事実は、実際の脆弱性でした。  
大きな政策問題は、どこに✏を設定するかです。  
2018年、国勢調査スタッフは4つのテーブルを報告するエンドツーエンドのテストを行いました。  
 そこから表が導き出されます。  
差分プライバシーアプローチはすべてのユーザーを保護しますが、古いシステムはスワップされたユーザーのみを保護し、一度にすべて実行する必要があります。  
新しいトップダウンアルゴリズムは、地理的IDなしで国別ヒストグラムを生成し、次に、州の図が国別の図に合算されるように、地理的ヒストグラムのトップダウンを構築することに着手します（議会の再区画に必要です）。  
次に、州、郡、地区、ブロックグループ、ブロックを再帰的に構築し、マイクロデータを生成します。  
5人以上のレースに所属する100人を超える人はほとんどいない）  
トップダウンのアプローチは、ブロックごとにノイズを適用するよりもはるかに正確であることがわかります。これは、郡のデータはブロックよりもエラーが少なく、国のデータは基本的にエラーがないという点です。  
個人世帯の加入も困難です。あなたはブロック上の男性の数、または世帯の数を行うことができますが、一人の男性が率いる世帯の子供の数はより敏感です。  
アウトライン設計が完了したので、possibleの可能な値を探索するために使用できるシミュレータがあります。  
  
政治的側面では、医学研究であろうと市場調査であろうと、研究における軽く特定されたデータの使用は、長年にわたってプライバシー擁護者とデータユーザーの間の散発的なゲリラ戦を伴い、規制当局は通常、余波を除いてデータユーザーと側見をしました。スキャンダルの。  
これらの「ビッグデータ」への関心は、通常、規制当局をとらえることに長けています。  
ご想像のとおり、プライバシー法の専門家もセキュリティ学者も結果に満足していませんでした[96]。  
彼は、「科学者は、匿名化されたデータに隠された個人を驚くほど簡単に「再特定」または「匿名化」できることを実証した」と告白し、「私たちは間違いを犯し、根本的な誤解の下に苦労して、プライバシーをはるかに少なく保証しました。私たちが想定したよりも。  
ここにようやく、著名な弁護士が事実を詳しく説明し、AOLとNet ﬂ ixの物語を、法律ジャーナルで使用し、弁護士がアクセス可能な言語を使用しました。  
、データのバイナリの考え方を個人的かどうかにかかわらず非難し、プライバシーとデータ保護に関するより現実的な議論を求めた。  
 2012年、英国王立協会からの報告は、科学者が可能な限り公然とデータを公開することを要求しましたが、再識別のリスクの現実を認めました：「しかし、コンピューターサイエンスにおけるかなりの研究は、今や個人の安全がデータベースの記録は、アイデンティティが積極的に求められている匿名化手順を通じて保証することはできません '[1627]。  
最終的なコードでは、データユーザーがメカニズムを一般的な用語でのみ説明する必要があり、証明の負担を異議を唱えた人に委ねました[81]。  
このつま先は、戦術的な匿名性との関連性のある概念、つまりプライバシーセット、または私についてのいくつかの事実を知りたくない人々のセットをまとめています。有名人にとって、それは誰でもありえます。誰かが突然有名になったときに問題が発生する可能性があります。  
もう1つの便利ですがまったく異なる概念は匿名セットです。これは、混乱する可能性のある人々のセットです。  
差分プライバシーのような戦略的メカニズムは、匿名性セットを十分に大きく保つことに重点を置いていますが、多くの戦術的メカニズムは、一部のアプリケーションにアクセスできる人がプライバシーセットと重複するリスクを評価します。  
懸念が恥ずかしいときはプライバシーセットを心配するだけで十分かもしれませんが、詐欺師の場合は匿名セットを心配する必要があります。  
要するに、フィッシングに関しては、あなたのアイデンティティを何らかの関連するコンテキストに結びつけることができる誰もがあなたを攻撃することができるかもしれません。  
11.4.1戦術的な匿名性とその問題  
これは、学者と国家統計局によって調整されています。  
その作者は、守秘義務は義務ではなくリスクについてであると考えています。データ主体を特定する技術的な可能性だけでなく、これが試みられるかどうかを決定する制度的および社会的文脈に従って決定を行う必要があります。  
彼らは、サイドチャネルではなくガバナンスプロセスについて話します。彼らは、差別化されたプライバシーを「極限」として却下します。彼らは匿名化をプロセスと見なし、「匿名化」のような「成功条件」を使用しないようにアドバイスします。また、「データから直接再識別できない」ため、「非識別」と定義されています。  
したがって、そのようなメカニズムは、米国の国勢調査に注意深く組み込まれている戦略的匿名性とは対照的に、戦術的匿名性と見なす必要があります。  
その流れにもかかわらず、UKANフレームワークは、英国で戦術的または戦略的に匿名化に依存する場合、規制当局があなたに対して執行措置を取るかどうかを判断する基準となるため、注意が必要です。  
これは、EUを英国に拠点を置く企業にある程度の保護を提供するはずでしたが、英国がEUを離れたため、これはもはや成立しません。  
その結果、現在、データユーザーが準拠するのを支援することを目的とした製品やサービスを提供している企業がいくつかあります。  
会社は、顧客の携帯電話の位置を、出発地、目的地、交通手段を暗示する旅行に集約します。  
プライバシーメカニズムは、1つ目はすべてのサブスクライバーがオプトアウトを許可し、2つ目は暗号化された電話のIMSIがゆっくりと変化するキーを使用してデバイスごとに異なる仮名を与えることです。セルタワーは簡単に識別できます。  
。  
次に、プライバシーセットのサイズを確認する必要があります。  
評議会で誰かを採用する場合、孤立した家に住んでいる会社のスタッフを狙って、彼らやその家族を脅迫することができます5。  
この規模は、人を特定するために外部で照合される可能性のあるデータソースの数だけでなく、組織の内部データウェアハウスのサイズと複雑さの増大にも明らかです。  
データベーススキーマはなく、データが積み重なっているだけなので、リンケージのリスクについて何も考えられません。特に、会社にさまざまな子会社のあらゆる種類のものが含まれるマルチテナントクラスターがある場合はそうです。  
企業は責任を負うことを望んでいますが、開発およびテストチームにライブデータをどのように提供しますか？  
 データ製品をどのように販売できますか？  
摂取に関するデータをトークン化して明白な名前を取り除き、アクセスを制御して時系列と場所のストリームに特別なトリックを使用することができますが、ノイズの追加は軌跡では機能せず、場所を再識別する多くの創造的な方法がありますデータ（例  
。  
将来の問題はAIと機械学習に起因する可能性があります。これが現在のファッションであり、企業が大規模なデータウェアハウスを設置するように導いた2010年代半ばの「ビッグデータ」方式に倣っています。  
私たちはすでに失敗する可能性のある多くのことを知っています。  
機械学習システムは、既存の社会的偏見をトレーニングデータとともに取り入れます。機械翻訳システムが数ギガバイトのオンラインテキストを読み取ると、それらは大きくなります。2003年に私は大学の統治機関のメンバーに選出され、大学が医学研究に使用する動物用の新しい建物を提案した後、私たちは動物の権利活動家の標的になりました。だれでも、だれでも突然ターゲットになることができます。  
翻訳は得意ですが、人種差別、性差別、同性愛嫌悪にもなります（これについては、セクション25.3で詳しく説明します。  
繰り返しになりますが、MLシステムの機能を理解していないため、匿名性を主張する場合は懐疑的に扱う必要があります。  
情報コミッショナーのオフィスには、匿名サービスとアプリケーションを監視する機能や動機がないようであり、業界は自主規制しています。実際、企業は自分の宿題をマークします。  
すでに述べたように、きめ細かい位置情報やソーシャルデータなどを販売している企業の多くは、匿名であることがはっきりしなくても匿名であると主張しています。  
多くの場合、データユーザーは、問題が発生すると、実際の専門家と知りたがらないほど、修正にかかる費用が高くなることを認識したくないのです。  
匿名化がこのような邪悪な問題である理由の1つは、そのセキュリティの経済学が本当に恐ろしいことです。  
11.4.2インセンティブ  
しかし、それは、ラスカルがラスカルに物事を行う力があると感じさせるかもしれません（特にUKAN以降）  
したがって、統計的セキュリティでは、善を敵とするべきかどうかの問題は、他の場所よりも細かく判断する必要があります。  
ここでのインセンティブは何ですか？  
たとえば、医療のプライバシーは、人々が医療費を支払う方法によって条件付けられます。  
しかし、米国では、一般的に医療費は雇用主が負担します。そしてイギリスでは、政府がそのほとんどを支払います。  
このような競合は、匿名性に関する主張によってしばらくの間覆い隠される可能性がありますが、実現可能なプライバシー技術によって解決できる可能性はほとんどありません。  
  
1つのアプローチは、研究者に安全なサイトへの訪問を要求するかどうかにかかわらず、弱い匿名性とアクセス制御を組み合わせることです（ニュージーランドのように、そして英国の税データの研究のためにも）  
。  
きちんとしたセキュリティエンジニアリングにより、適切な能力2。  
法律の範囲内で、EUではデータ主体に尊重されるオプトアウトの権利を与えることを含みます。  
これにより、データへの完全なアクセスが可能になるだけでなく、やる気のある被験者と、通常の臨床活動の副産物として簡単に収集できるよりも高品質の臨床情報が提供されます。  
 患者とその家族の完全な同意を得て、12か国以上の医師と他の研究者の間で完全に識別可能な情報を共有します。  
同意モデルが広がっています。  
言うまでもなく、完全な同意を得たアクセスには、アクセスが研究者に制限されていることが条件となるため、堅牢なセキュリティエンジニアリングも必要です。  
セクション10.4.5.1で、米国におけるタスキーギー実験とドイツのナチス医師によって行われた実験における医療倫理の起源、および現在生じている安全策について説明しました：治験審査委員会（IRB）  
不安定な法的根拠に基づいて医療行為から収集された記録を使用する代わりに現実的な代替手段がなく、漏えいした匿名化メカニズムを使用して保護されている医学研究者は、IRBまたは倫理委員会に頼らざるを得ません。  
 機関の主要な原則は、そのような研究は研究者から独立した誰か、通常は調査の目的と提案された方法の両方を評価する1人以上の匿名の同僚によって承認される必要があるということです。  
  
倫理審査プロセスは、研究者に2つのレベルの責任の盾を提供します。  
他の人と同じプロセスに従い、各プロジェクトが「独立した」メンバー（実際には、実際のデータ主体の代表ではなく、他の大学の教授を意味する）を含む倫理委員会によって承認される場合  
第二に、最悪の事態が発生し、慣習法の国で「メンスレア」または不法な意図、ならびに「サクサスレウス」または禁止された行為の二重テストを含む刑事訴追の可能性に直面した場合。  
あなたが言ったとおりに行動し、独立した人々が承認した理由により、それがどのようにして不正な意図である可能性がありますか？  
これはビッグデータの注目を逃れていません。  
。  
AI倫理は学界でホットな話題であり、急速に成長している仕事の源です。倫理の洗浄とデータの乱用は今や密接に関係しています。  
2007年3月、歴史家のマーゴアンダーソンとウィリアムセルツァーは、1942年に国勢調査の機密性が停止され、ワシントンDCに住む日系アメリカ人に関するマイクロデータが1943年にシークレットサービスに提供されたことを発見しました[1699]。  
単一障害点は、国勢調査局長のJC Captであり、財務長官のヘンリーモルゲンサウからの要請を受けてデータをシークレットサービスにリリースしたようです。  
しかし、これは新しいことではありませんでした。  
他の場所では、ドイツ人は国勢調査データを使用して、ドイツだけでなくオランダや他の占領地域のユダヤ人を切り上げました。  
公に宣伝されているプラ​​イバシーメカニズムには、見かけよりも効果が低い例がたくさんあります。  
ただし、典型的なプライバシープランを見ると、流通ネットワークオペレーターが、その流通エリアであるミッドランド、サウスウェスト、ウェールズの30分ごとのメーターデータにアクセスしていることがわかります[2011]。  
配電業者は、このフィードを各フィーダーの30分ごとの合計に集約することを約束します。これらは、変圧器を出て多数の家に電力を供給するケーブルです。  
より強力なプライバシー規制当局は、彼ら自身の変圧器に彼ら自身のメーターを設置するように彼らに告げただろう。  
医学に関しては、米国  
大手テクノロジー企業は、誰がデータを処理していると言っているかによって、規制から逃れる場合があります。  
多くの国では、考えられるよりも多くの組織が完全に識別可能なデータにアクセスできます。  
11.5まとめ  
これは、国勢調査な​​ど、明確に定義されたいくつかの特別な場合にのみ機能します。この場合、差別的なプライバシーの形で確かな理論があります。  
しかし、政策立案者、マーケティング担当者、医学研究者などは、匿名性が個人データを使用する魔法の解決策を提供することを信じられないようにして、それを乱用することは困難です。  
プライバシー侵害の規模と範囲がますます明らかになるにつれて、深刻な問題が発生する可能性があります。  
研究の問題現在、匿名性とプライバシーに関するいくつかの活発な研究スレッドがあります。  
第2に、数学者は、相互に不信のある企業が保有するデータからの学習など、さまざまな状況で個人差のある機械学習を行う方法を検討しています。  
第4に、実用的な運動家（EPIC、Privacy International、Max Schremsなど）がいる  
理論、実践、奨学金、キャンペーンのこのエコシステムは、私たちの周りのものがより「スマート」になるにつれて、間違いなく進化し続けるでしょう。  
 限界では、差別化されたプライバシー以外に何もできないほど利用可能なコンテキスト情報がたくさんありますか？  
 最後に、最新の魔法の薬はプライバシーを保護する連合機械学習です。  
しかし、これは、過去40年間に匿名化に関して提供されてきたヘビ油の一種に過ぎないのではないかと思います。  
 それを暴くための最良の方法は何ですか？  
推論制御に関する古典的な参照は、ドロシーデニングの1982年の本[538]です。 Adam and Wortmanによる1989年の調査報告書は、当時の最新技術の優れた要約です[17]。  
政府の仕事は、統計手法に関する連邦委員会の「統計情報開示制限手法に関するレポート」であり、さまざまな米国の省庁で使用されているツールと手法を紹介しています[667]。これは2005年に遡るので、多少古くなっており、現在書き直されています。  
まったく異なるアプリケーションの例として、Mark AllmanとVern Paxsonは、[42]でネットワークシステム研究のためにIPパケットトレースを匿名化することの問題について説明しています。