それは人々が神がすべてを見ることができると信じるのをやめたときに始まり、政府が埋めるべき空席があることに気づいたときに止まった。  
1はじめにビッグタバコが何十年もかけて喫煙が肺癌を引き起こさないことを否定し、ビッグオイルが何十年もかけて気候変動を否定してきたため、ビッグデータも数十年かけて機密情報を簡単に「匿名化」できるので、侵害することなく産業原料として使用できますデータ主体のプライバシー権について。  
その限界は4つの研究の波で探究され、それぞれがその日のテクノロジーに対応しています。  
 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の両方がこれを宝物庫と見なし、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。  
2。  
 匿名化が不十分なデータが漏えいしたり、販売されたりしたときに、スキャンダルが何度も発生しています。  
 理論の進歩は2006年に行われ、Cynthia Dworkらは、プライバシーを制限する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。  
第4の波は、2010年代後半に、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で私たちに起こりました。  
 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、あくびをする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ますます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がますます非現実的になるようになっています。  
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。  
6。  
 発効後、希望に満ちた思考のゆがみの余地は少なくなりました。ただし、ヨーロッパの機関でさえ、匿名化によって何がもたらされるかについて、時々バラ色の見方をしてきました。  
2推論制御の初期の歴史推論制御は、1920年代に遡り、個々の企業の貢献を覆い隠す経済データが蓄積されたが、国勢調査データの文脈で体系的に最初に研究された。  
2。  
 この情報は、選挙区を決定するため、公共サービスのための政府の資金調達のレベルを設定するため、および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。  
データが完全に無害で公開前かどうか、またはプライバシーメカニズムが一度に1つのクエリを実行して、それが許容でき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。  
 与えられた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。  
1000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。  
 サンプルレコードに加えて、さまざまな属性のローカル平均も与えられました。  
 そのような抑制がなければ、小さな村に住んでいる裕福な家族は、自分の家族の収入が推定されるのに十分なだけ、平均の村の収入を増やすかもしれません。  
 たとえば、単純なルールは、結果が3つ以上のデータ主体のデータを使用して計算されない限り、質問に答えない、いわゆる3つのルールです。  
 典型的な攻撃は、標的となる個人を含むサンプルに関する多数のクエリを作成し、いくつかの秘密の事実を推測するために働きます。  
 クエリコントロールを無効にするためにコンテキストを連続して追加するクエリは、トラッカーと呼ばれます。  
たとえば、ニュージーランドのジャーナリストは、軍隊および外交官のリストを時系列での掲載パターンについて精査することにより、その国の信号インテリジェンスサービスであるGCSBの多くの関係者の身元を推定しました[849]。  
11。  
1推論制御の基本理論推論制御の基本理論は、主に米国国勢調査の問題に対応して、1970年代後半から1980年代初頭にドロシーデニングらによって開発されました[538]。2。  
 多くの最新のプライバシーシステムの開発者は、この作業に気づいていないことが多く、1960年代の過ちの多くを繰り返しています。  
特性式は、レコードのクエリセットを選択する（一部のデータベースクエリ言語の）式です。  
 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は、基本セットまたはセルとして知られています。  
 推論制御の目的は、機密情報の開示を防ぐことです。  
 D = P 0の場合、保護は正確であると言えます。  
11。  
1。  
 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をクエリできないようにする必要もあります。Nレコードがあり、クエリのsetsizeしきい値がtの場合、tとN�tの間のレコード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、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。  
 したがって、K⇢Lかつ| L |の場合、レコードセットKおよびLのクエリの連続を回避する必要があります。 �| K | <t。  
2。  
2TrackersThatは、個別のトラッカーの例です。これは、禁止されたクエリへの回答を間接的に計算できるカスタム式です。  
ドロシーデニング、ピーターデニング、メイヤーシュワルツが原因で1970年代後半に行われたやや憂鬱な発見は、一般的なトラッカーは通常見つけやすいということでした。  
 したがって、クエリセットのサイズを制限したり、他の方法で許可されたクエリを制御したりしない限り、トラッカー攻撃は簡単です。  
セキュリティエンジニアリング354ロスアンダーソン11。  
 推論統制の初期の歴史11。  
1。  
 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、全国統計出力のその他の統計単位を特定することはほとんどあり得ない」と求められていました[1416]。  
 ここで、図11のテーブルがあるとします。  
Major：BiologyPhysicsChemistryGeologyMinor：Biology-161711Physics7-3218Chemistry3341-2Geology9136-図11。  
 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、化学行の少なくとも1つの他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。  
 テーブルは図11のようになります。  
Major：BiologyPhysicsChemistryGeologyMinor：Biology-D17DPhysics7-3218Chemistry33D-DGeology9136-図11。  
データベーススキーマに他の属性がある場合（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。  
2。  
 ただし、上記の試験マークの表のように、マイクロ統計を公開する場合は、しばしば必要になります。  
11。  
1。  
 その制限は、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、効果を上げるには、制限を厳しくする必要があります。  
 （可能な場合）より徹底的なアプローチは、サンプル母集団を非常に多くのセットに分割するクエリを拒否することです。  
 また、一部のデータベースでクエリコントロールを体系化するために、少し異なる方法で使用することもできます。  
 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。  
3。  
2。  
5より洗練されたクエリコントロール単純なクエリコントロールに代わる方法はいくつかあります。  
その他の手法には、極端な値を持つデータの抑制が含まれていました。  
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たった1人の患者を持つ郡がありました。  
2。  
3：–誰が何にアクセスしたかを追跡することにより、3つの属性を持つデータベースのテーブルラティス。  
 これは良い考えのように聞こえるかもしれませんが、実際には、2つの通常は克服できない欠点があります。  
次に、ユーザーが共謀しないこと、または1人のユーザーが2つの異なる名前で登録したことを確認するのは非常に困難です。  
11。  
1。  
 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、攻撃者の視点からS / N比を低下させながら、正当なユーザーの信号対比をできるだけ低下させない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が、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に正確に影響を与え、サンプルセットが十分に大きい場合はそのままにしておく傾向がありますとにかく単純なクエリコントロール。  
 次に、ランダムなサンプルクエリがあり、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します。  
2。  
 ランダムサンプルクエリは、調査対象の相関が十分に強く、小さなサンプルで十分であるという自然な保護メカニズムです。  
2006年以来、ランダム性の追加からどれだけの保護を得ることができるかについての確固たる理論があります。それは、差分プライバシーです。  
11。  
2従来の統計的セキュリティの制限あらゆる保護技術と同様に、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ています。  
 プライバシーの目標は、特定可能な患者または個人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。  
週：1234Doctor A17261922Doctor B2531929Doctor C32303927Doctor D16191813図11。  
4。  
 そして医者Cはおそらく彼女をカバーしていたマービン・スミスだろう」  
1完全な開示：私は評価者であり、英国医師会に代わって行動しました。  
 このような重要なプライバシーのケースが、プライバシーの目的として、医師が白い嘘をつき続ける能力を持っていることは奇妙です。  
2。  
 アプリケーションは適切に定義されており、データベースはそれほど豊富ではありません。許容されるクエリはかなり単純であり、長期にわたって安定しています。  
 同省の動機は、このようなデータの産業界への供給を独占することでした。  
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、異なる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。  
11。  
2。  
 その場合、アクティブな攻撃は特にパワフルになります。  
 スイスの製薬会社がレイキャビック政府との契約を結ぶために地元の新興企業に資金を提供しました。研究のために私たちに調べさせてくれれば、現代の健康カードシステムを構築します。  
彼らの提案された設計では、医療記録が生成されるたびに、それはアイスランドのプライバシー委員会に送信され、そのシステムは患者の名前と住所を取り除き、それを彼らの社会保障番号の暗号化バージョンに置き換え、それを研究データベースに渡します。  
 ただし、首相の医療記録を（たとえば）見つけたいシステムの誰もが、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不均一で、少数の入植者から数千年前に生まれ、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統図にリンクすることにより、形状セキュリティエンジニアリング359ロスアンダーソン11の影響で、叔父、叔母、大叔母、大叔母などの数の要素によって患者を特定できます。  
 彼らの家系図の推論制御の初期の歴史。  
 これにより、リッチコンテキストデータのより広範な問題が発生し、推論制御に関する作業の2番目の波が押し寄せました。  
2。  
 世界中のWebが新しくなり、企業（および政府）が事業をオンラインに移行する方法を考え出そうとするにつれて、急流のお金がテクノロジーに流れ込みました。  
 メールを使用して病院から医師の手術に検査結果を送信できますか、それともWebフォームですか？どのように暗号化し、誰が鍵を管理しますか？また、個別の処方箋などのエピソードデータだけではなく、名前と住所を削除して、研究で使用するのに十分な完全な医療記録を作成できますか？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応答しないというルールでデータベースを設定しましたが、それでは不十分で、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ました。これには、匿名化と厳密な使用管理の両方が含まれます[266]。  
 BritishMedical Associationは1995–6年に集中型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DameFiona Caldicottの下に委員会が設置され、前進する方法が提案されました。  
 彼女は、HealthCare Finance Administration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によって再識別できることが多いことを示しました[1849]。  
S  
当時、Medicaresystemは、患者の名前と社会保障番号が暗号化された受益者暗号化レコードを個人データであると見なしていたため、Security Engineering360Ross Anderson11のみが使用できました。  
 推論統制の初期の歴史信頼できる研究者。  
 それにもかかわらず、研究者は、パブリックアクセスレコードを商用データベースと相互相関させることにより、多くの患者を特定できることを発見しました。  
 これにより、医学研究データの匿名性が米国の政治課題に加わりました。  
4、クリントン政権は2000年にHIPAAに基づいてプライバシー規則を発行し、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やかな規則を採用しました。  
 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のため、名前で3％、住所で18％を特定できました。  
 Dame Fiona Caldicottの報告は、60以上の違法情報が医療サービスに流れていることを特定しています[367]。  
その後、議会は大臣に医療データの二次利用を規制する権限を与える法律を可決しましたが、幅広い方向性は信頼できる研究者でした。委員会はデータアクセスのためのアプリケーションを吟味しました。  
 HESデータは、患者の名前と住所が削除され、暗号化された識別子に置き換えられて、研究者が利用できるようになります。  
）しかし、患者名を暗号化するだけでは十分ではありません。  
 2003年10月19日と2004年10月1日、彼はロンドンのハマースミス病院で不規則な心臓の鼓動のために治療を受けていたことがわかりました。  
 このようなリークは誰にとっても邪魔になる可能性があります。有名人にとっては、それは報道価値があるかもしれません。  
 生年月日が生年に置き換えられたとしても、記録が詳細である場合、または異なる個人の記録がリンクできる場合、私は依然として患者のプライバシーを危うくする可能性があります。  
クエリセットのサイズ制御はこの種のトラッカーを停止する可能性がありますが、研究者は、多くの条件で複雑なクエリを実行して、a3UK郵便番号を含む疾患クラスターを見つけたいと考えています。米国郵便番号よりも解像度が高く、通常、郵便番号ごとに30の建物があります。  
セキュリティエンジニアリング361ロスアンダーソン11。  
 推論制御の初期の歴史数百人、あるいは数十人の患者さえも。  
2006年、英国のプライバシーグループは、人々にリスクを警告し、二次的なデータの使用をオプトアウトする権利を行使するよう招待するキャンペーンを組織しました。  
 プライバシーの問題は、患者が自分のデータを使用しないように合理的に要求する唯一の理由ではありません。たとえば、敬虔なカトリックの女性は、自分のデータを中絶や避妊のための薬の開発に使用しないよう要求するかもしれません。  
2010年に選出されたデビッドキャメロン政権は、ジョージブッシュが10年前に行ったように、プライバシー保護を弱体化させました。  
4。  
3、彼は「ケア。  
 2013年11月、BTを介してHESデータをオンラインで販売できるようになりました[948]。2014年2月に、HESデータベースのコピーが、学術研究者だけでなく商業企業を含む世界中の1,200の組織に販売されたことが明らかになりました。製薬会社の倒産[774]。  
 データは健康以外の目的で、具体的には保険料を調整するためにアクチュアリーによって使用されました。  
 別のコンサルタントが別のレポートを作成するために雇われ、オプトアウトした人々はもう一度オプトアウトするように言われました。11。  
4 3番目の波：設定と検索次の波は2006年に破綻し、その頃にはかなりの数のトランザクションがオンラインに移行し、eBayとAmazonのおかげでレコメンダーシステムが出現し、検索エンジンによって干し草の山から針を見つけやすくなりました。  
まず、AOLは、657,000人が3か月間に作成した2000万件の検索クエリの匿名の記録をリリースしました。  
 調査ジャーナリストは検索を調べ、プライバシー侵害でショックを受けた一部の検索者を迅速に特定しました[167]。  
セキュリティエンジニアリング362ロスアンダーソン11。  
 推論制御の初期の歴史は、データとその監督者を発表しました。  
第2に、Net ﬂ ixはより優れたリコメンダーアルゴリズムに100万ドルの賞金を授与し、50万人の購読者の視聴者評価を公開し、名前を削除しました。  
 Arvind Narayananand Vitaly Shmatikovは、インターネットムービーデータベース[1384]で公に表明された嗜好と匿名のレコードを比較することにより、多くの加入者を識別できることを示しました。  
 米国の法律は映画レンタルのプライバシーを保護しているため、攻撃はNet ﬂ ixにとって深刻な恥ずかしいものでした。  
 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、プライバシー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で説明します。  
 これは、匿名化の制限を明確にするため、プライバシー規制当局がフックすることはありません。  
11。  
5第4の波：ロケーションとソーシャル2010年代、世界はスマートフォンとソーシャルネットワークによって変化しました。  
 ロバート・パットマンの著書「ボウリング・アローン」は、1960年代にテレビが登場した教会、クラブ、社会などの自主的な協会を通じて社会的関与が衰退したことを記録している[1563]、そしてインターネットの初期のUsenetニュースグループとメーリングリストがなんとかその裏をかいた。  
どのように興味を再確認しても、世界中のどこにいても、それらを共有する人々とつながることができます。  
 （暗号化技術者、デジタル著作権活動家、200年前のダンスミュージックに興味のある人々と一緒にたむろする私以外の誰か？）十代の若者たちのセックスや薬物についての誇りが後に彼らを悩ませるために戻ってくるとき、永続性はさらに危険を加えます就職の面接で。  
 2011年までに、Googleはそのコアコンピタンスを「クラウドソーシングデータの統計データマイニング」として説明していました。データセットが大きくなり、基本的な統計手法が機械学習で強化されたため、学習できる量が増えました。  
 Yves-Alexandre deMontjoyeと彼の同僚は、2012年までに、携帯電話のセキュリティエンジニアリング363Ross Anderson11しか入手できない場合でも、一般に4つの携帯電話の場所で誰かを特定できることを示しました。  
 推論制御の初期の歴史場所[1333]。  
 ほとんどの人は、何も考えずに同意するためにクリックします。現在、ロケーション追跡データを売買する企業のエコシステム全体があり、数百ではなく数メートルに正確になりました。  
2019年12月に、ニューヨークタイムズは数か月間で1200万人のアメリカ人の位置追跡を把握し、人々を追跡できる程度をグラフィカルに示しました。  
 ジャーナリストたちはデータベースで大統領トランプのための教会の礼拝で歌った有名人を見つけた。国防総省とCIAで働いている何百人もの人々、ならびに大統領のシークレットサービスボディーガード。彼ら全員が家について行くことができます。  
 彼らは暴動を見て、暴動と警察の家の両方を追跡できることを発見しました[1885]。  
 ロケーションデータの会社はすべて、データが匿名であると主張しています。しかし、実際には電話帳や有権者のロールを使用して住所からあなたの名前を検索していない場合でも、ブラウザの1つ以上のCookieに基づいて、広告IDに関連付けられた位置データをいくつか販売します。  
 外国の諜報機関は高解像度のデータを使用して、国防総省で働いている人々を探し出し、ゲイのクラブや売春宿を訪問することもできます。  
「より良い推論」の例は、ソーシャルネットワークデータの行動分析にあります。  
 彼らは、誰かがまっすぐであるかゲイであるかをFacebookの4つの好きなものから知ることができると考えました。 60が好きなら、彼らはユーザーの「ビッグファイブ」の性格特性を評価することができます。あなたが経験にオープンであるか、慎重であるか、良心的であるか、寛大であるか、外向的か内向的か、同意できるか分離されているか、神経症的または信頼できるかどうかです[1086]。これにより、一部の同僚がマーケティングと政治運動のためにFacebookのデータを産業規模で収集し、Cambridge Analyticaスキャンダルにつながりました。これについては、パート3で説明します。  
ただし、業界オブザーバーは、プラットフォームが広告収入の大部分を獲得しているため、これよりも多くの収入が得られることに気づきます。そのため、セキュリティエンジニアリング364ロスアンダーソン11に抵抗することが期待できます。  
 推論制御の初期の歴史このようなプライバシー法[1181]。  
 たとえば、オーストラリアのビクトリア州政府は、2015年から8月にかけて、1500万チケットで10億回の旅をカバーする交通チケットの使用に関するデータベースを公開しました。  
 次に、オーストラリアの連邦議会のパスを使用している人々を特定しました。国会議員のツイートから仮説を確認することができます。  
 彼らは、移動時間が時間に切り捨てられたとしても、時間と分が捨てられたとしても、4つの場所が3分の1以上の旅行者を特定することを発見しました。  
 ロケーションヒストリーは、私たちが住んでいて、一緒に働いて、パーティをしている人を明らかにするほど多くのデータをリークします。  
 そして、社会分析はスタックの最下層にまで到達することができます。  
 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、快適な人はより多くの電話をかけます、そして電話間の時間の変動は良心性を予測します[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臓損傷を診断するためのアプリを開発することを発表しました。  
同意を得ることなく、すべての患者の6m完全に識別可能な記録[1542]。  
 同社は診断アプリの開発に米国のVAデータを使用しました。  
 この遅い列車事故に続いて、Googleはすでに5,000万人の米国の患者の記録を取得するために燃えているというニュースが続いていました[121]。  
 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、プライバシーの信頼できる手段を提供できることがあります。  
セキュリティエンジニアリング365ロスアンダーソン11。  
 差動プライバシー11。  
 彼らの理論である差動プライバシーは、無制限の計算能力と豊富なサイド情報を持つ敵が存在する場合でも、セキュリティエンジニアが開示の確率を制限できるようにし、したがって、ワンタイムパッドと無条件に安全な認証と同等と見なすことができます。 -暗号化における表記コード。  
 開始点は、2003年にデータベースのifqueriesがそれぞれ情報のプライベートビットの線形関数の近似を返すことを示したKobbi NissimとIrit Dinurによる以前の論文であり、エラーが十分に小さい限りデータベースを再構築するために必要なクエリはあまり速く成長しません;そのような再構築攻撃は結局のところ、線形代数に基づいているため、攻撃者は注意深く対象を絞った追跡攻撃を行うのではなく、ランダムなクエリの全ロットを作成し、代数を実行して取得することができますすべてアウト[562]。  
プライバシーの違いに関する重要な洞察は、不注意による開示を回避するために、クエリの結果に対する個人の寄与があまりに違いを生じさせないようにするため、データの感度に応じてノイズの標準偏差を調整することです。  
 したがって、ラプラス分布でノイズを使用すると、ノイズの多い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？  
1国勢調査に差分プライバシーを適用する2020 Uでは、差分プライバシーが本格的なテストを受けています。  
 国勢調査。  
 まず、CensusBureauは2010年の国勢調査のセキュリティを最新の分析ツールに照らして検討しました[752]。  
3。  
7 Gb）。  
実際にはどうですか？国勢調査は、KobbiNissimとIrit Dinurの研究に基づいてアイデアを実装し、すべての変数が約38％の確率で正しく得られ、人口の20％未満をカバーしていることを発見しました。ただし、教訓は、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ことです。  
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、ユーザーはそれに我慢しませんでした。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、国勢調査局は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。  
 予測可能な問題には、数値が加算されないことが含まれます。したがって、別々のネイティブアメリカン部族のメンバーの数は、ネイティブアメリカンの合計に加算されません。そのことは、一般に説明する必要があります。  
 すべてのレコードは、全体的なプライバシー予算に従って変更される可能性があるため、CEFとMDFの間の正確なマッピングはありません。  
 次に、州、郡、地区、ブロックグループ、およびブロックを再帰的に構築して、マイクロデータを生成します。  
g。  
 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ことです。  
 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。  
アウトラインデザインが完了したので、,の可能な値を調べるために使用できるシミュレータがあります。  
セキュリティエンジニアリング367ロスアンダーソン11。  
 ギャップを気にする11。  
 セクション26で説明するように、規制当局はどちらも圧倒され、対立しています。  
1、そしてほとんどの場合、大規模なインターネットサービス会社や政府部門に対抗する政治的サポートはありません。  
 たとえば、2008年にゴードンブラウン首相は、英国の情報コミッショナーと英国最大の医療研究慈善団体の責任者に、研究でのデータの使用に関するガイドラインを作成するよう依頼しました。彼らはプライバシー権を無視し、費用と便益の計器的見解をとり、データの二次利用を「データ共有」としてスピンした。  
2009年、非常に影響力の大きい論文「プライバシーの破られた約束」は、著名な米国の法律教授であるポールオームによって書かれました[1465]。  
 この間違いは、ほぼすべての情報プライバシー法、規制、および議論に浸透していますが、規制当局や法学者はこれにほとんど注意を払っていません。  
 ここにようやく、著名な弁護士が事実を詳細に説明し、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。  
 これにより状況が変わる可能性はありますか？2012年、英国王立協会からの報告により、科学者はデータを可能な限り公に公開するよう求められましたが、再特定のリスクの現実は認められました：「しかし、コンピューターサイエンスのかなりの部分で、データベース内の個人記録のセキュリティは、アイデンティティが積極的に求められる匿名化手順では保証できません '[1627]。  
 最終的なコードでは、データユーザーはメカニズムを一般的な用語でのみ説明する必要があり、証明の負担を反対した人に移しました[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。  
 有名人にとっては、それは誰でもかまいません。と問題canSecurity Engineering368Ross Anderson11。  
 MIND THE GAP？誰かが突然有名になったときに発生します。  
もう1つの有用である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。  
 差別的なプライバシーのような戦略的メカニズムは、匿名性のセットを十分に大きく保つことに焦点を当てていますが、多くの戦術的なメカニズムは、一部のアプリケーションへのアクセス権を持つ人々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。  
4。  
 2016年、UKANはICOによって正式に署名された匿名化に関する企業の決定方法に関するガイダンスの本を作成しました[626]。脅威モデルは、もっともらしい侵入者シナリオに基づくべきです。  
再識別リスクを管理するための対策は、リスクとその影響の可能性に比例する必要があります。そして、匿名化の測定は、他のデータセットからの最終的な三角測量のために限られた寿命を持っているかもしれません。  
 UKANの作者は、プライバシーの違いを真剣に検討していないようです。  
 匿名化の効果が着実に低下するにつれて、データユーザーと規制当局の両方にショックアブソーバーと責任の盾を提供する可能性があります。  
 ただし、データと環境の両方が十分に理解されているアプリケーションでの戦術的な匿名化のリスクを評価するための合理的な量の実用的なアドバイスは含まれています。  
セキュリティエンジニアリング369ロスアンダーソン11。  
 ギャップを気にしていますか？このフレームワークのもとで公然と事業を展開している企業の例は、「位置情報」製品を販売するモバイルネットワークオペレーターVodafoneです。  
 起点と終点のマトリックスは地方自治体と輸送会社に販売され、幹線道路と鉄道に沿った流れが流れます。  
 実際、ここでのリスクは低いと主張することができます。地方自治体やバス会社のアナリストがあなたを特定できるかもしれません。特に小さな集落に住んでいる場合は（私のように、最寄りの村から200m離れた4つの家）。  
 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。  
 彼らが評議会で誰かを募集する場合、彼らは彼らまたは彼らの家族を脅かすために孤立した家に住んでいる会社の従業員を標的にすることができます5。  
 この規模は、人々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、リンケージリスクについての考えはありません。特に、会社にさまざまな種類の異なる子会社のマルチテナントクラスターがある場合はそうです。  
 会社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？学者や新興企業とどのように連携できるのか？どのようにデータ製品を販売できるのか？匿名化テクノロジーはすべて、この規模ではかなり初歩的なものであり、何が起こっているのかわからないので、差別的なプライバシーやその他の分析できる範囲を超えているきれいに。  
g。  
 人々が部分的に許可され、部分的にアクセスできる場合、事態はさらに困難になります。  
 あなたは今、あなたが本当に理解していないデータについて、彼らが何をしているか説明できないシステムを訓練しています。  
 保険制度は、少数派地域の保険料を引き上げ、差別禁止法に違反しています。  
 一部の同僚は、彼らの家で怒鳴り声を上げるために活動家に立ち向かわせ、2人の活動家は後にオックスフォードでの同様のキャンペーンの後にテロ行為の罪で有罪判決を受けた。  
6サーバーのクラスターにペタバイト規模でデータを保存し、NoSQLを使用してデータにアクセスするためにYahooによって最初に開発されたオープンソースソフトウェア。  
4。  
3。  
 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だけなので、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Oceには、匿名サービスとアプリケーションを監視する能力や動機がないように思われるため、業界は規制しています。実際には、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てい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、実際の専門家とは知りたくないほど、修正するのに費用がかかることに気づき、本当の専門家と話をしたくありません。  
 匿名化がそのような邪悪な問題である理由の1つは、そのセキュリティ経済学が本当に恐ろしいことです。  
4。  
 しかし、それは、（特にUKAN以来）悪質なことをする能力があると、悪意者に感じさせるかもしれません。  
 経済学の章で述べたように、多くの利害関係者がいる大規模システムでのセキュリティ障害の最も一般的な原因は、インセンティブが間違っているときです。Aliceがシステムを保護し、Bobが失敗のコストを支払うときです。  
 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ています。  
 しかし米国では、医療費は一般的に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。  
セキュリティエンジニアリング371ロスアンダーソン11。  
 ギャップを気にする11。  
3AlternativesOneアプローチは、弱い匿名性とアクセス制御を組み合わせることです。これは、研究者に安全なサイトを訪問するように要求し（ニュージーランドのように、英国の税データについても調査するため）、非開示契約に加えてアクセスと（ドイツのように）被験者を特定する試みを禁止するコントロールを使用します。  
 きちんとしたセキュリティエンジニアリングを使用して、有能である; 2。  
 EUでは、尊重されるオプトアウトをデータ主体にオプトアウトする権利を与えることになる法律の範囲内。  
 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や、通常の臨床活動の副産物として単純に収集できるよりもはるかに高品質の臨床情報が提供されます。  
 このネットワークにより、非常に強力なプライバシー法が定められたドイツと、ほとんど法規制のない日本との間のデータ共有が可能になります。そして、USAFがセルビアを爆撃したときでさえ、データは米国とセルビアの研究者の間で共有され続けました。  
 2番目の例は、数十万人のボランティアが残りの人生の記録への完全なアクセスだけでなく、広範なアンケートに回答し、血液サンプルを提供して後の人生で興味深い病気を発症する人が遺伝的およびプロテオミクスメイクアップ分析。  
信頼できる研究者のルートであれ、完全同意のルートであれ、研究へのアクセスも倫理的な承認に依存します。  
4。  
1私たちは、米国のタスキーギー実験とドイツのナチス医師が行った実験における医療倫理の起源と、現在生じているセーフガード：アメリカの制度審査委員会（IRB）とヨーロッパの倫理委員会について議論しました。  
 正確なプロセスは機関間（および機関内）で異なりますが、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、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。  
セキュリティエンジニアリング372ロスアンダーソン11。  
 ギャップを気にする11。  
4ダークサイド倫理レビュープロセスは、研究者に2つのレベルで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た場合、あなたが従った強い事例を作ることができますそれらの基準。  
 倫理的な承認プロセスは、男性の領域がなかったという証拠を提供するように設計されています。  
これはビッグデータの注目を逃れていません。  
2。  
 当然のことながら、特に企業が大規模なデータウェアハウスで人工知能と機械学習技術を投入し始め、結果がどうなるかについて明確な考えがほとんどないため、倫理ボードは急増しています。  
 シニカルオペレーターは、UKANの推奨事項のいくつかに準拠するという動機をたどり、失業者の哲学者を雇って道徳哲学と知性の性質について話しながら、最も親密な個人情報をスパマーに販売するビジネスに取り掛かります。  
さらに、公に宣伝されたプライバシーメカニズムの存在は、基礎となる個人データの悪用から注意をそらす可能性があります。  
 ブロックレベルのデータはカリフォルニアの社会人に与えられ、そこで日系アメリカ人を収容のために切り上げました。  
 局はその後公に謝罪した[1319]。  
 1914年に第一次世界大戦が勃発したとき、イギリス政府は1911年の国勢調査を使用して外国人を追放の対象としました。1941年の国勢調査は1939年に持ち込まれ、徴兵、配給、および収容の基礎として機能しました。セキュリティサービスは、1980年代まで国勢調査への裏口を持ち続けました。  
 より最近では、Cambridge Analyticaとその親会社であるSCLは、現政府が再選に勝利するのを支援した多くの国々から、国勢調査の全データへの秘密のアクセスを許可されました[2052]。5。  
 英国は、中央のクリアリングハウスを介してすべてのガスと電力の消費量を報告する「スマートメーター」のシステムを構築しています。他の企業がデータにアクセスするには、承認されたプライバシー計画が必要です。  
 目的は、変電所変圧器をいつ交換する必要があるかを予測することです。  
 しかし、データを見ると、0であることがわかります。  
67％が3つ以下のサービスを提供しています。  
 実際、より賢明な公共政策は、スマートメータープロジェクトをまったく行わないことでした。これについては第14章で説明します。  
S  
 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は、規制から逃れる場合があります。  
 多くの国では、想像以上に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。  
5まとめ多くの人が、機密性の高い個人データを、名前などの明白な識別子を取り除くことによって産業用の原材料に変えることができると信じたいと考えています。  
 ほとんどの場合、データは多すぎてデータ主体の再識別は簡単です。  
 ビッグデータと機械学習を巡る絶え間ない誇大宣伝は、これらのテクノロジーが匿名性をさらに一層難しくしているように、教育タスクをより困難にします。  
 本当の変化をもたらすにはおそらくスキャンダルが必要であり、これが最終的に発生した場合、その混乱はささいなことではないでしょう。  
まず、既存の公開データから機密データを導き出す新しい方法を探す、またはマーケティング担当者やサイバー犯罪者が実行するエクスプロイトを理解しようとする実践的な研究者がいます。  
5。  
第3に、法律と慣行のギャップを埋める方法を模索しようとしているプラ​​イバシー法学者がいます。  
 この理論、実践、奨学金、キャンペーンのエコシステムは、私たちの周りのより多くのものが「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。  
 差別的なプライバシーの場合のように、そのようなことが機能するエッジケースを見つけることができるのは間違いありません。  
 （ねえ、ヘビ油を水酸化ナトリウムで沸騰させると、ヘビ石鹸が出るはずです。  
 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。 1989年のAdamとWortmanによる調査論文は、当時の最新技術の優れた要約です[17]。  
S  
 UKANbookは、英国の管轄内で業務を行うクライアントの匿名化を行う場合は必読です[626]。  
最後に、米国、特に第二次世界大戦中の国勢調査データの乱用に関するマーゴアンダーソンとウィリアムセルツァーの論文は、[52]にあります。