SEMINAR REPORT

パーソナルデータ安全活用のためのプライバシー保護技術の開発と制度設計への貢献

日本電信電話（株）
セキュアプラットフォーム研究所
主席研究員／チーフ・セキュリティ・サイエンティスト
高橋 克巳 氏

皆様、こんにちは。NTT の高橋克巳です。今 SCAT の新宿にある会議室にお伺いしているのですが、広い、ちょうど正方形っぽい部屋の中に、本日講演をされる田中さんと SCAT の方のちょうど４人で広い部屋を四角く囲んで巨大な雀卓のようなところにいます。完全に３密も避けられていて、そもそもここにたどり着くまですごく御苦労があったと思いますが、開催にこぎつけられて、誠にありがとうございました。

改めまして、１月に SCAT の会長大賞を受賞しました。ありがとうございました。SCAT の皆様、それからこの仕事は数多くの方々の御支援、御鞭撻、また一緒に仕事をしてくださったことによるものですので、こちらからで失礼いたしますが、心より御礼申し上げます。

私は NTT の研究所でずっと研究開発に従事しています。今回の業績はここ 15 年近くの仕事に対するものですので、自己紹介的なものも話の中でお伝えしたいと思っております。

パーソナルデータを活用するとは

それでは本題に入る前に、パーソナルデータを活用するということについて考えてみたいと思います。皆さんがお考えのパーソナルデータの活用は何でしょうか。例えばコンビニで買い物をするとポイントがたまります。そういった自分に対していろいろ便益があるというのがひとつの基本です。さらに進んでくると、例えば自分が帰宅していろいろしゃべると自分のことを分かってくれていて、電気がついたりテレビがついたりという話があろうかと思います。これも自分のために役立つ例です。

次に、今は新型コロナによるパンデミックですから、世の中のことが気になります。どう行動していいのか分からないというのが皆さんたくさんあるのではないでしょうか。コロナ関連のパーソナルデータ活用で有名なものに、人出の空間統計、新宿の人出は昨年比 80%でしたとか、こういったことがニュース等で報道されると思いますが、これはまさにパーソナルデータの活用、携帯電話の位置情報を使ってそれが可視化されています。このことで世の中の動きが分かります。

先ほど自分の情報で自分の便益というような話をしましたが、コロナの場合は他人のことが気になるのかもしれません。３密という話がありますが、おとなしく散歩あるいはお出かけをしたいとき、人に迷惑をかけたくない、あるいは人混みのところに行きたくないようなときは、他人がどこにいるか、混んでいるところが分かれば、それが避けられるわけです。

そうすると普段、我々はあまり他人の行動に興味がないとしても、コロナの場合は、例えば密を避けるというようなことで、他人の行動を知り自分の行動を変え、戦略的に振る舞いを決めることができる可能性があります。行動再開で経済を回すみたいなこともありますが、何が本当に良いのかなかなか分からない話です。気がついたら馴染みの飲食店が潰れてしまったという御経験も皆さんもあるかもしれません。しかし応援しようにもどういう振る舞いがいいのかということがなかなか分からないと思います。

パーソナルデータを活用するとは、自分が便利になるだけではなく、コロナの環境下においては世の中全体で他人の振る舞いも知りながら密を避けたり、あるいは助け合ったりというような活用が期待されていると考えます。

本日、お話しすることは２つです。テーマ A は、技術についての話です。テーマ B は、技術が法制度の一部となりました。匿名化技術ですが、その経緯についてお話をしていきたいと思います。それではいきなりまとめです。このまとめ２枚で本日の話はご理解いただくことができます。

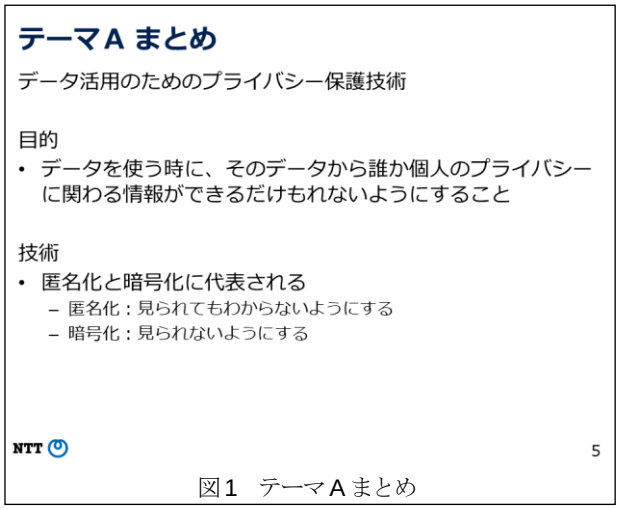

図1　テーマAまとめ

まずテーマA、「データ活用のためのプライバシー保護技術」をまとめるとこのようになります(図1)。データを使う時にそのデータから誰か個人のプライバシーに関わる情報ができるだけもれないようにすること、これがデータ活用のためのプライバシー保護技術だと考えています。そしてその技術は匿名化と暗号化に代表されると考えています。匿名化というのは、そのデータを見られても分からないようにする。一方、暗号化は、御存じのとおり見られないようにするということになります。

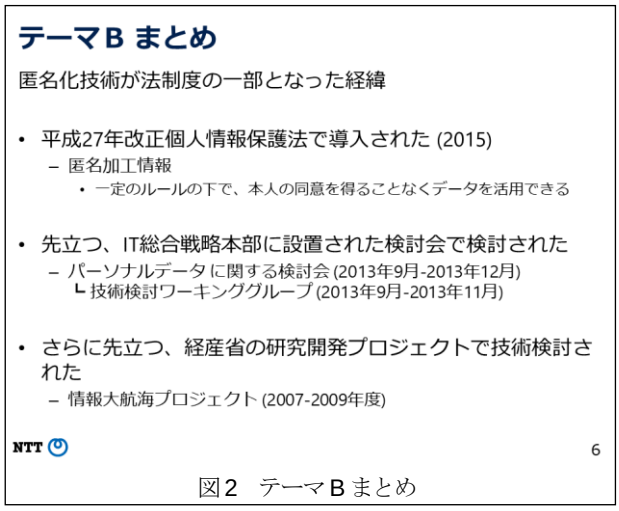

図2　テーマBまとめ

続いてテーマBです(図2)。匿名化技術が2015年に改正された個人情報保護法で法制度の一部となりました。これはどういう経緯で作られたかというと、先立つ内閣官房のIT総合戦略本部に設置された検討会で、私も参加させていただいて検討されました。ここに書いてある技術検討ワーキンググループはもう7年も前になってしまったのかという感じですが、非常に面白い議論をたくさんさせていただきました。さらに先立つ経産省の情報大航海プロジェクトで技術検討されたという流れであると考えています。

データ活用のためのプライバシー保護技術とは何か

データ活用のためのプライバシー保護技術とは、「データを使う時に、そのデータから誰かのプライバシーに関わる情報ができるだけもれないようにする技術」と述べました。ここをもう少し細かく見てみます。

「プライバシーに関わる情報」、これは何か。なかなか定義が難しいものですが、「その人が他人に知られたくない情報」と考えることが多いです。

ただし、それがどこからプライバシーかということに決まりがあるわけではなく、個人差もありますし、感じ方は非常にいろいろあります。我々事業者目線で言うと、多様なプライバシー感に対応する必要があり、ここまで踏み込んでしまってプライバシーに迷惑をかけているのではないかと疑心暗鬼になって悩んでいるところもあります。

それから「情報ができるだけもれない」ことが大事です。できるだけというところがポイントですが、実はパーソナルデータは何かの役務に附帯して渡されるものですから、例えば誰かに住所をもらさなければ荷物の1つも届かないことになります。何ももらさないわけにはいかないというので、「利用目的に応じた必要最小限」というような言い方がされます。できるだけ伝えない、できるだけもらさないということが目標になります。

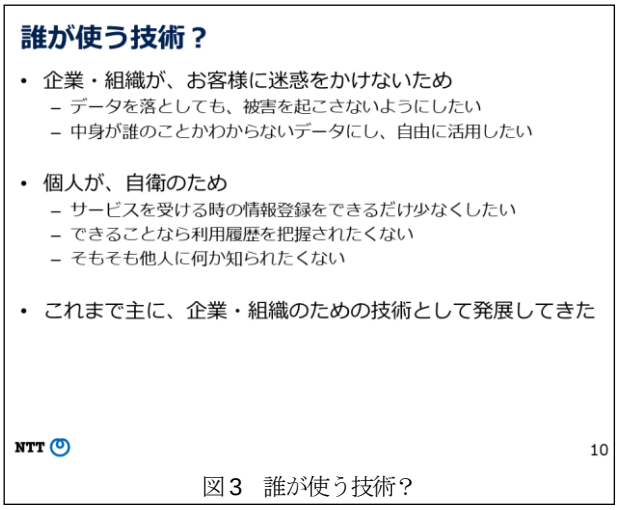

図3　誰が使う技術？

続いて、誰が使う技術かを考えます(図3)。まず企業が使うことが中心です。「お客様に迷惑をかけないため」と書いてありますが、お客様に迷惑をかけてしまう最たるものは情報漏洩です。企業は当然防ぎたい。企業のもう一つのモチベーションはデータ活用です。個人単位のデータ活用が必要でない場合は、誰のことか分からないデータにして自由に活用したいというニーズがあります。

次に、プライバシー保護技術は個人が自衛のために使うことができると考えられます。例えばサービスを受けるときに情報登録をできるだけ少なくしたいとか、そもそも他人に何か知られたくない等いろいろなモチベーションがあって、自衛に使うことができます。これまでプライバシー保護技術は主に企業向けに発達してきた経緯がありますがこちらも重要ですね。

プライバシー保護技術は匿名化と暗号化の2つに代表されます。匿名化とは、データを見られても分からないようにするということです。個人が分かるデータも、「20代・女性」であれば誰だか分かりません。他方、暗号化とは鍵を渡した人にしか見られないので、ほかの人には何も分からないという効果があります。

匿名化と暗号化

それではこの匿名化と暗号化を１つずつ、少し細かいところを御紹介していきたいと思います。

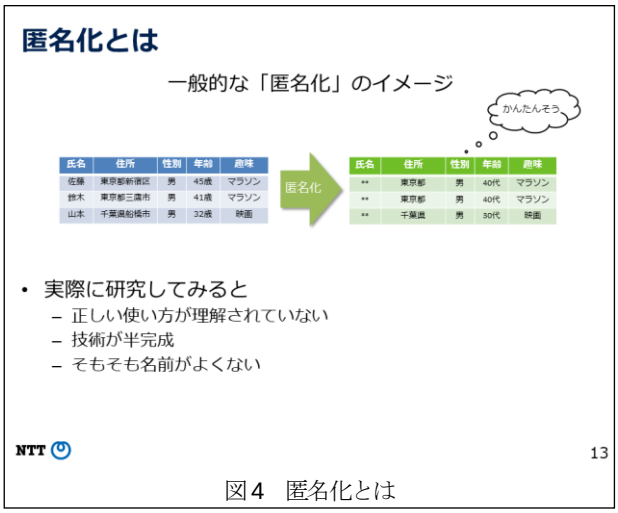

図４　匿名化とは

上に描いてある図が一般的な匿名化のイメージです(図 4)。佐藤さん、鈴木さん、山本さんといった名前を消して、住所等々も加工していくと匿名化ができます。私も簡単そうだと思いました。しかし実際に研究してみると、技術は簡単なものも確かにありますが、実は正しい使い方が理解されていない。

そして技術が半完成、そもそも名前がよくないのではないかなど、いろいろなことに気づきました。

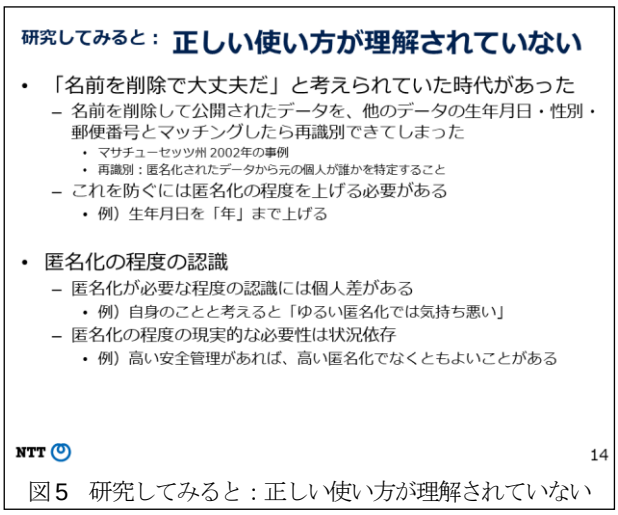

図５　研究してみると：正しい使い方が理解されていない

順番に見ていきます(図 5)。まず、「正しい使い方問題」ですが、つい最近まで、「名前を削除しておけば大丈夫だ」という考え方が実際にありました。マサチューセッツ州で 20 年近く前にあった事例として、名前を削除して医療データを公開したケースがありました。別の名前が削除されていないデータを持ってきて、そのデータの生年月日、性別、郵便番号とマッチングしたら誰か分かってしまったという事故になりました。これを防ぐためには匿名化の程度を上げる必要があります。程度というのは、例えば生年月日を「生年」にするというようなことです。

匿名化の程度の認識は非常に差があります。例えばデータを使われる個人の気持ちで考えると、「ゆるい匿名化では気持ちが悪い」と考えますし、一方でデータを使う側があまり考えずに「名前を削除しているだけで大丈夫だ」などという間違った認識が起きる可能性があるわけです。さらに、匿名化をどの程度行うかということもあります。匿名化の程度の現実的な必要性は結構、状況依存です。例えばデータ自体に高い安全管理が施されていれば、匿名化を一生懸命やらなくてもよい場合があると技術的には考えられています。

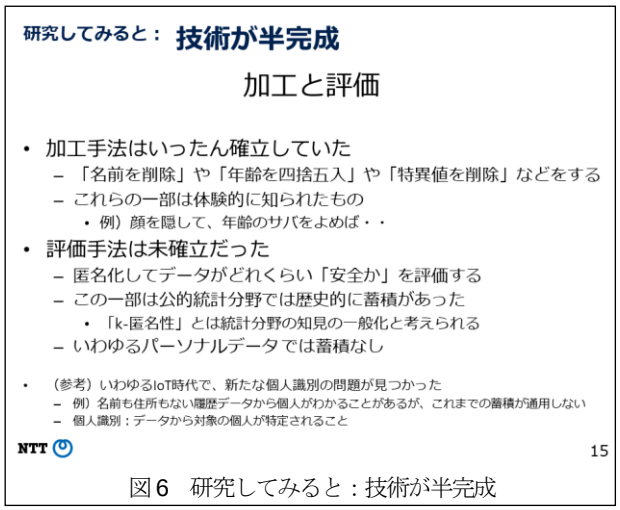

図６　研究してみると：技術が半完成

続いて、技術が半完成という話をしたいと思います(図 6)。半完成とはどういうことかというと、実は匿名化は加工することと、それがうまくできているか評価することの両方がないとだめなのです。ざっくり言うとその加工の技術はほぼ確立していますが、評価のほうがあまり確立していないという状況でした。

加工手法は普通に確立していました。例えば名前を削除して年齢を四捨五入して云々ということが体験的に知られていました。

一方、評価手法はまだ確立していませんでした。評価とは匿名化したデータがどの程度安全かということを確認することです。これはなかなか難しい問題です。この考え方は公的統計、いわゆる統計局で歴史的に蓄積がありました。これくらいの塊にして統計を公開すれば問題が起きないだろう、といったものです。

後ほど説明しますが、k－匿名化性というものがあります。この考え方は実は統計分野で蓄積されてきた知見の一般化と考えられます。ただし、統計的な形式のデータに対しては知見がありましたが、いわゆるパーソナルデータ、履歴データといった複雑で細かいものに関して言えば、蓄積がありませんでした。さらに現在は IoT で新たな個人の識別が起きる問題も認識されています。

３つ目、名前がよくないという話です(図 7)。「匿名化しました、いやできてないだろう」という落語のような話があります。匿名化は行為なのか結果なのかということを冷静に見てみる必要があります。世の中で「匿名化」と呼んでいるものの多くは単に加工行為だけのことなので、加工行為だけで匿名化と言っても匿名化されていないよ、ということが起きてしまいます。

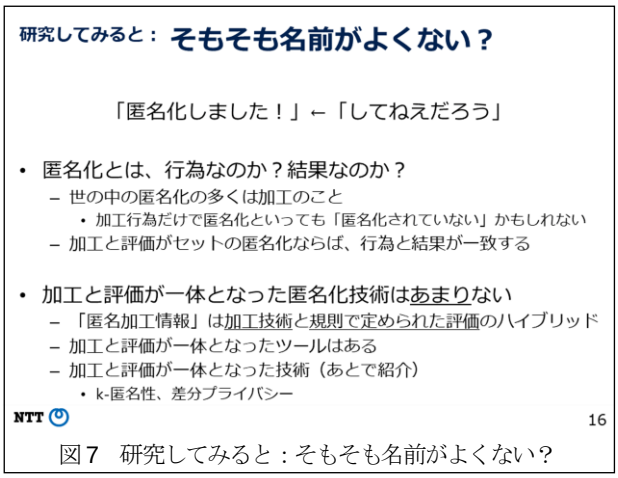

図7　研究してみると：そもそも名前がよくない？

他方、加工と評価がセットで匿名化を行うのであれば、行為と結果が一致するということが言えます。しかし、加工と評価が一体となった匿名化技術はあまりないです。「匿名加工情報」という制度については後半説明しますが、これはこれまでに知られていた加工技術に評価を規則ではめ込むというハイブリッド型になっていると考えられます。なお、加工と評価が一体となったツールはあります。NTT でも開発して販売しています。また、加工と評価が一体となった技術があります。これは面白いもので、後ほど紹介してまいります。

匿名化：加工

ではその加工のところをもう少し説明していきたいと思います。加工とはデータを元のものと変えて誰か分からなくする、正確に言うとプライバシー情報を得る手がかりをなくすというようなことです。

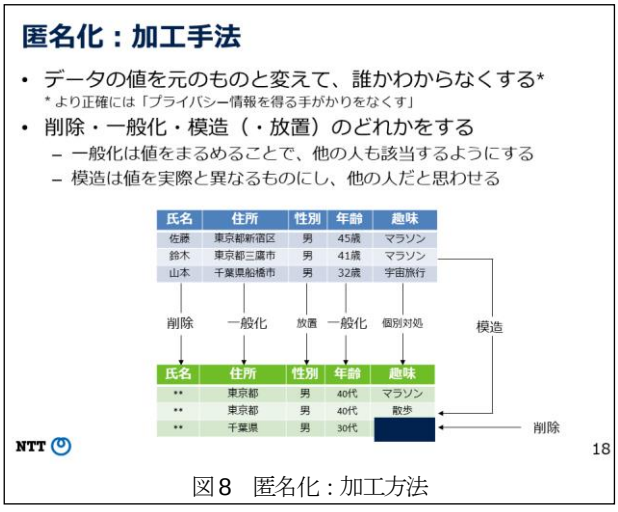

図8　匿名化：加工方法

このようなデータの項目(図8)があったときに、削除、一般化、そして模造と書いてありますが、模造というのは例えば「マラソン」を勝手に「散歩」と変えることです。マラソンと散歩で個人の識別性が変わるかどうか微妙だとは思いますが、嘘をつくことでその人が誰か足がつきにくくするテクニックがあります。あとは匿名化といっても全部の値をいじるわけではなく、そのまま放置しておくといったこともあります。加工はこれだけです。

匿名化：評価

評価は先ほど言ったように、どの程度安全であるか、というようなことです。安全性を確かめるためにはどういったリスク、どういう状況で困ったことが起きてしまうのかということを知る必要があります。

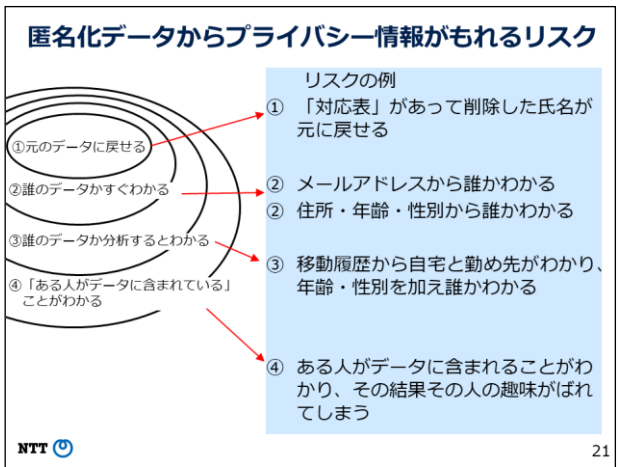

図9　匿名化データからプライバシー情報がもれるリスク

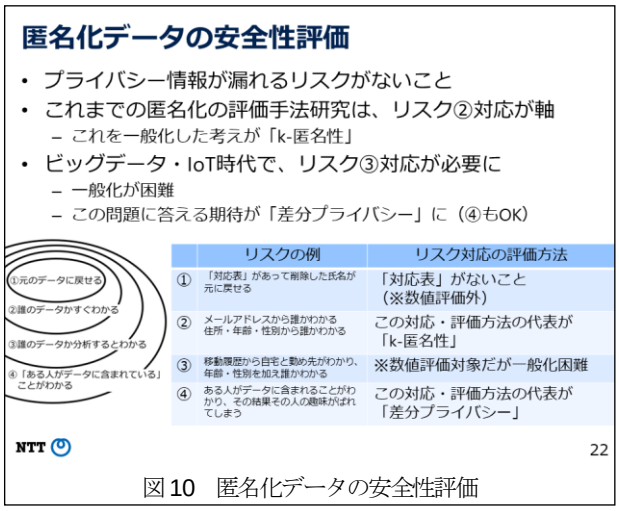

図10　匿名化データの安全性評価

ここに「匿名化データからプライバシー情報がもれるリスク」を書いてみました(図9)。リスクを便宜上４つに分けてみます。

まずは①、真ん中に書いてありますが、元の個人データにすぐ戻せるものです。これはあっという間にプライバシー情報がもれてしまいます。

②は元のデータにすぐには戻せないものの、メールアドレスなどがあれば誰のデータかすぐ分かってしまうものがあります。

③はすぐには分からないが一生懸命分析をすると分かるというものです。例えば移動履歴を調べていくと自宅と勤め先が分かってしまうので、あとは年齢・性別が分かると誰か分かってしまうかもしれないというシナリオです。

④は誰か分からなくてもある人がそのデータに含まれていることが分かってしまうことによるリスクがあります。これはどういう都合が悪いかというと、趣味がばれてしまうと例に書きましたが、例えばある特定の狭い地域でコロナが 500 人発生したということが分かると、そこに住んでいる人は高い確率でコロナにかかっているのではないかということが、誰のデータか

分からなくても知られる可能性があることが危惧されます。

では、その安全性評価はどうしているかというと、基本的にはこういったリスクがないようになっていることが安全性評価ということになります(図 10)。これまでの匿名化の手法の研究はリスク②の対応が中心でした。すなわち、誰のデータかすぐ分からないように加工するというものです。ビッグデータや IoT 時代になってきて、リスク③の対応が必要になってきたと考えられます。

例えばたくさんの履歴データがあれば、そこから何かいろいろなことが分かってしまうので、それらを対応する必要があります。しかし、その対応方法を一般化するのはかなり難しい問題だと考えられています。この問題に答える技術の候補として「差分プライバシー」が期待されています。

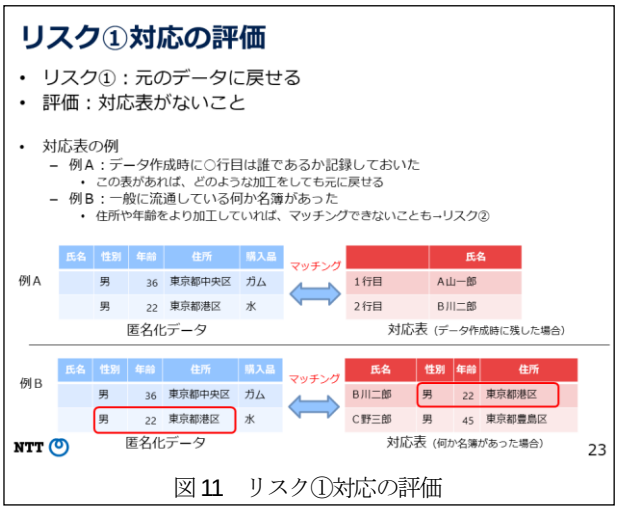

図 11　リスク①対応の評価

それでは少し細かいところを説明してまいります(図 11)。まず、元のデータに戻せることへの対応。これは例のように、匿名化したデータと対応表をマッチングさせると元にもどりますので対応表がないようにします。

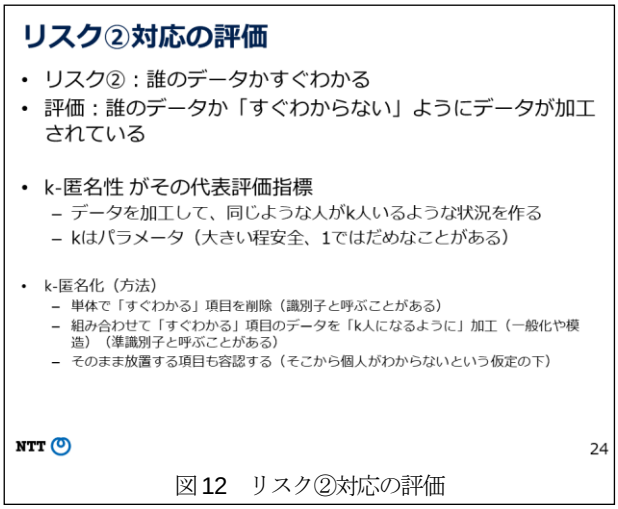

図 12　リスク②対応の評価

続いて２番目ですが、突き合せでぱっと分からないにしても、誰のデータか住所と年齢ですぐに分かってしまうケースです(図 12)。これに関しては、どのように評価していくかというと、誰のデータかすぐに分からないようにデータが加工されているかを見ます。それを実現するものに k−匿名性というものがあります。これが代表的な評価指標になります。

これは何かというと、データを加工して同じような人が k 人いる状態を作ります。k というのはパラメータで、５や 10 や 100 といったことを想定します。大きいほど安全です。「同じ人が 100 人いれば」ですので、おっさんが 100 人いればおっさんはその中に紛れて誰か分からないであろうというロジックになります。

k−匿名化の方法はここに書いてあるとおりで、例えば住所と年齢のようなものを組み合わせて、１人にならないように、何人かになるようにデータを一般化していくことが肝になります。

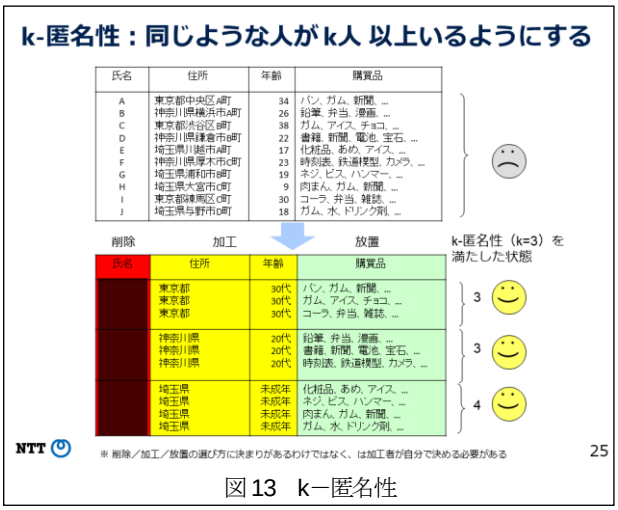

図 13　k−匿名性

これが k−匿名化を説明した図になります(図 13)。住所と年齢を加工してそれぞれ３人、３人、４人というようなグループができました。したがって最低でも３人未満には絞り込むことができないので、このデータは３匿名性が達成されていることになります。

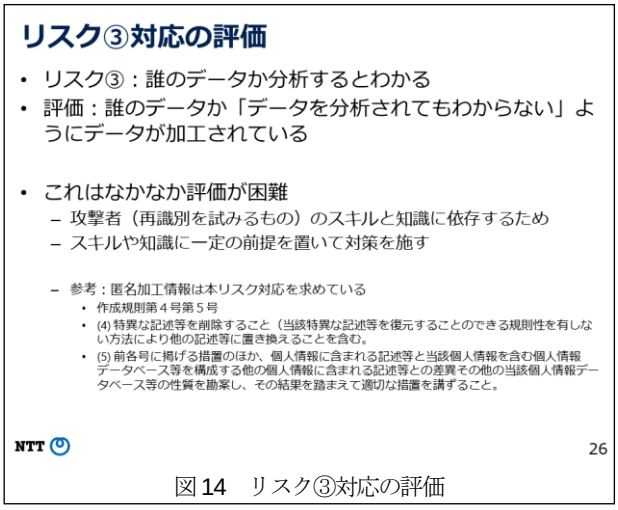

図 14　リスク③対応の評価

続いて３番目のビッグデータ的なものに対する「データを分析されても分からないようにする」という課題への対応を考えます(図 14)。これはなかなか評価が困難です。困難な理由は自明で、攻撃者―そのデータは誰のものか暴こうとする人―のスキルと知識に依存するために一般化が困難です。したがって対応する場合はスキルや知識に「こういう攻撃者を想定しましょう」という一定の前提を置いて、実際に加工の程度を決めてい

きます。なかなか厄介なものですが、匿名加工情報はこのリスクへの対応も求めています。

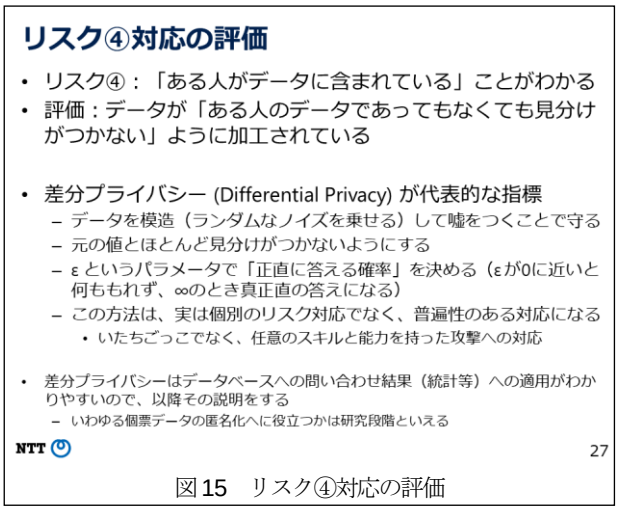

図15　リスク④対応の評価

続いて最後の「ある人がデータに含まれている」ことが分かる状態が避けられているかということに対応していく技術について、少し説明したいと思います(図15)。これはどういう評価をするかというとなかなか難しいのですが、文字通りある人がそのデータに含まれているかどうかということを評価するものです。この手法は十数年前に考え出されました。対象のデータがある人に対して、その人が含まれていても含まれていなくても見分けがつかないように加工されている場合、そのデータはその人が含まれていることに耐性があるだろうという考え方です。この考え方を定式化したものが差分プライバシーという指標です。

これはある意味、嘘のデータを作って、そのデータのプライバシーを守ろうという仕組みです。元の値とほとんど見分けがつかないようにすることがテクニックになります。この中で先ほどの k−匿名化は k というパラメータがありましたが、差分プライバシーには ε (イプシロン)というパラメータがあって、ε の大小で正直に答える確率を決めていきます。学術的には定義がスパッと言い切れているので研究のしがいがあり、非常に研究が進んでいます。

この技術の非常に良いところは、実は個別のリスク対応ではなくて、そのデータとある人の関係がないということが言えてしまうので、攻撃者の知識やスキルに依存した対応といういたちごっこから解放されて普遍性のある対応になると考えられるところです。

差分プライバシーは統計結果への適用がとても分かりやすいので、その例を中心にして次のページから説明をしたいと思います。

ここでは差分プライバシーを理解するために、統計結果からプライバシーが守られるかという問題を考えてみます(図16)。

「テストの落第人数を先生が発表しました」という問題を考えてみたいと思います。以下の人数が落第であった―男子５人、女子５人―ということで、ここから誰か分かってはいけないという問題です。

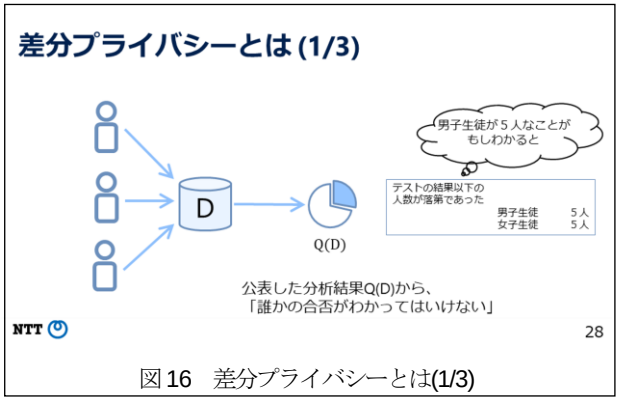

図16　差分プライバシーとは(1/3)

この問題は一見大丈夫そうな気がしますが、実はこのクラスは女子は35人いますが、男子は５人しかいないという状態で、そのことがもし分かると、落第した男子が分かってしまうという不都合が起きます。

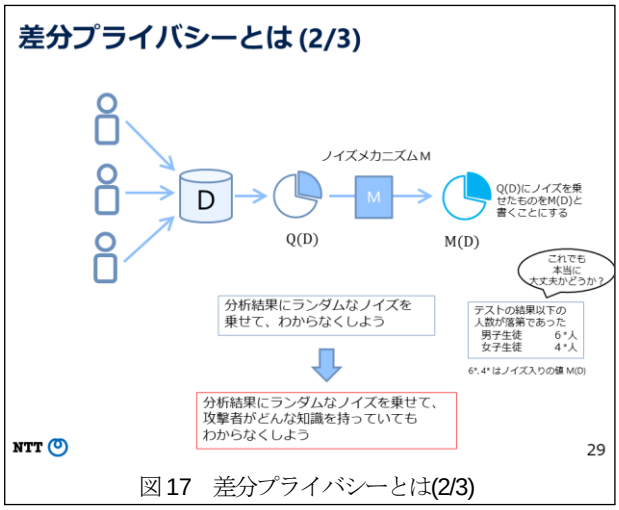

図17　差分プライバシーとは(2/3)

差分プライバシーの基本的な考えは、分析結果にランダムなノイズを乗せて嘘を作って分からなくしようというものです。

右に例が書いてありますが(図17)、男子が６人ぐらい、女子は４人ぐらい落第したという嘘を作るわけです。これで本当に大丈夫かどうかというと、やはり先ほどの「男子が５人」ということを知っていると、これでも大丈夫かなというふうに思うわけです。この技術の目標は、ノイズを乗せて攻撃者がどういう知識を持っていても分からなくしようというものです。

この上の図の説明を少ししておきますと、データベース D を作ったときに、それに対する答えが Q(D)としましょう。そこにノイズメカニズム M を提供してノイズを乗せて６人とか４人を作るわけですが、そうしたときにその M(D)を評価しているわけです。

このページ(図18)が差分プライバシーの定義となります。差分プライバシーは誰がやっても結局、数式を使わなければ説明ができないようなところがあって、これはいい技術ですが、数式に落ちてしまうから普及はできるかな、大丈夫かなとずっと考えています。このページだけは数式で説明をします。

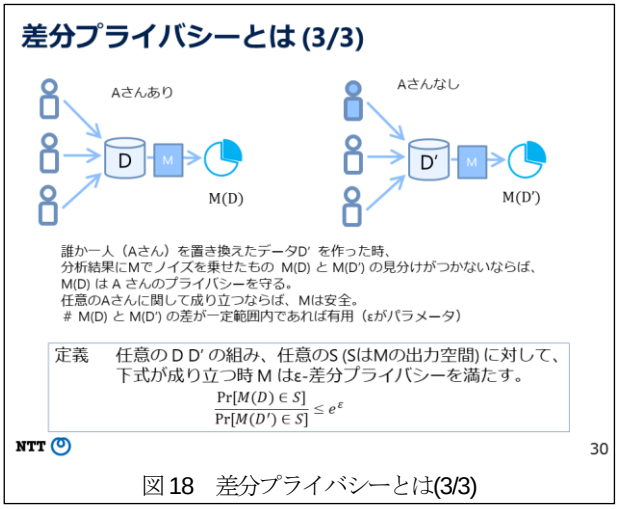

図 18　差分プライバシーとは(3/3)

A さんを含んだデータ D と A さんを除いたデータ D´ というデータを２つ作ってみました。それに関して両方にノイズを乗せたデータを作ったとき、「ノイズを乗せた A さんありのデータと A さんなしのデータがこのような不等式で押さえられるときにε差分プライバシーを満たす」が定義になります。こういう技術が今、流行ろうとしています。

こちらの性質は先ほど言ったように普遍的な対応ができてプライバシー保護のいたちごっこから解放されるという利点があります。問題点は少しとらえにくい所ですね。k－匿名性は分かりやすいですね。なんだかんだ言っても同じような人、おっさんを９人寄せれば何とかそこに紛らわせるだろうという話ですが、これはデータにノイズを乗せて、そのノイズの乗せ方を評価してあげて、かくかくしかじかでどう安全である、と評価するようなものですので、ちょっと直感的にはいかんとも説明しがたいところはあります。いずれにせよこういった技術が出てきていて、私たちはこの研究もしています。

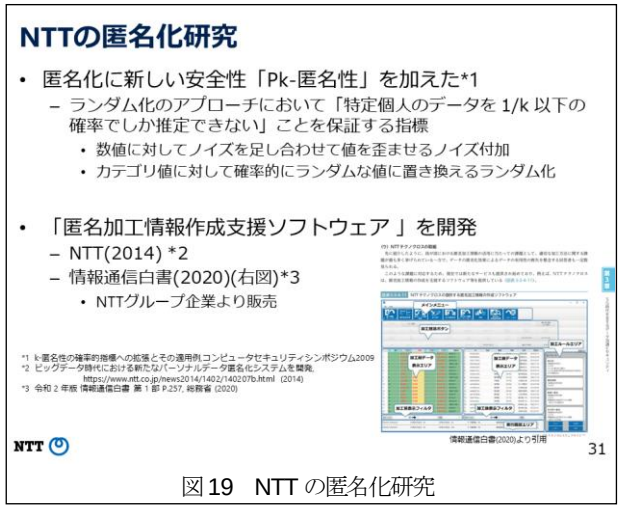

図 19　NTT の匿名化研究

匿名化技術の最後のパートですが、NTT でも匿名化の研究をしています(図 19)。

暗号化

続いて暗号の話です。暗号は通信の秘匿などにいろいろ使われるわけですが、私はプライバシー保護のために暗号を使うという、ちょっと変わったシチュエーションでの話をします。

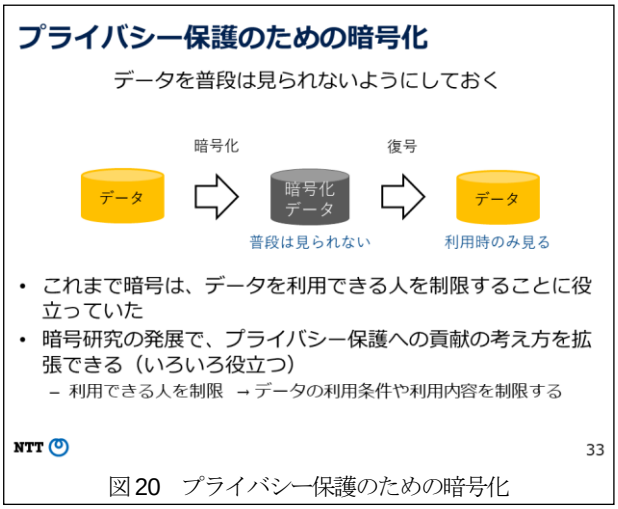

図 20　プライバシー保護のための暗号化

暗号というのは、データを普段見られないようにしておく、という考えです。これまでデータは鍵の管理も含めて考えると、データを利用できる人を制限することに役立っていたと考えられます。データ、個人情報を暗号化しておきましょう。鍵をあげると開けて中を見ることができる、鍵がないと見られないという単純な話です。

最近の暗号研究の発展で、いろいろな拡張ができるようになりました。利用できる人を制限するというところから、データの利用条件や利用内容を制限するというようなことができるようになってきました。(図 20)

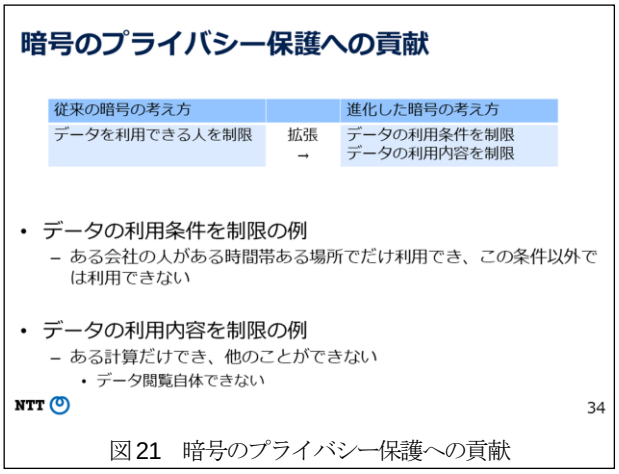

図 21　暗号のプライバシー保護への貢献

ここに先ほど言ったことを表現しました(図 21)。条件や内容というふうに、もっと細かいいろいろなことが暗号によってできるようになりました。

データの利用条件を制限するとプライバシー保護的にどのようにうれしいかというと、例えばこのデータはある会社に所属している社員が、ある時間帯、会社の建物の中だけで利用できて、それ以外では利用できないというようなことができます。

次に利用内容の制限の例では、「このデータを使ってある計算だけできてほかのことはできない」、「データの閲覧自体できない」ということができます。これが非常に面白い性質ですから、プライバシー保護に生かすいろいろな実例が作られていますし、もっともっと使われてほしいと考えています。

図 22　「データ最小化」と暗号の関係

ここで Data Minimization という考え方を少し説明しておきたいと思います。プライバシーフレームワークという標準にもなっているデータのプライバシー保護の原則があります。Data Minimization とはその原則の一つで、処理されるデータを必要最小限にして、かつそれを触る人も必要最小限にするという考え方です。これが非常にプライバシー保護に効くと言われています。(図 22)

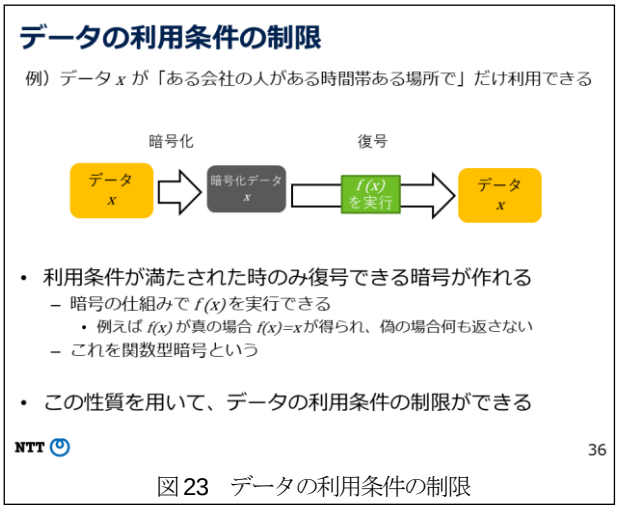

図 23　データの利用条件の制限

では、利用条件の制限の話をしたいと思います(図 23)。例えばデータが「ある会社のある時間帯だけ」利用できるというようなものは、関数型暗号というものを使って実現することができます。

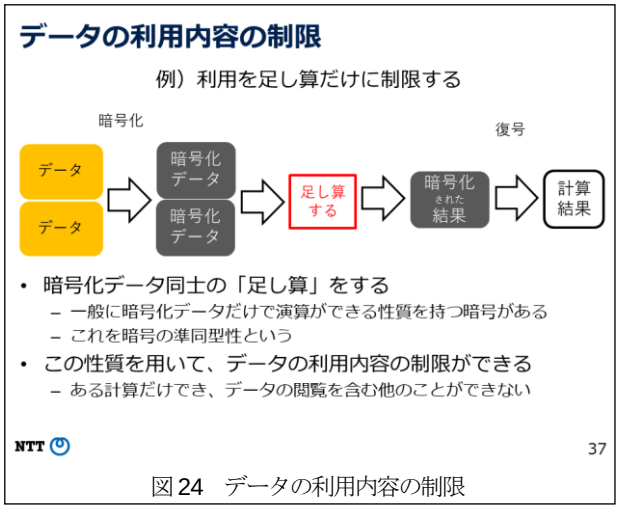

図 24　データの利用内容の制限

また、利用内容の制限ということで(図 24)、実は暗号の一部は暗号化データ同士の「足し算」をするようなことができます。これを暗号の準同型性といいますが、この準同型性を使うと暗号化したままで計算ができますし、それをうまく使うことで利用内容の制限ができるようになります。

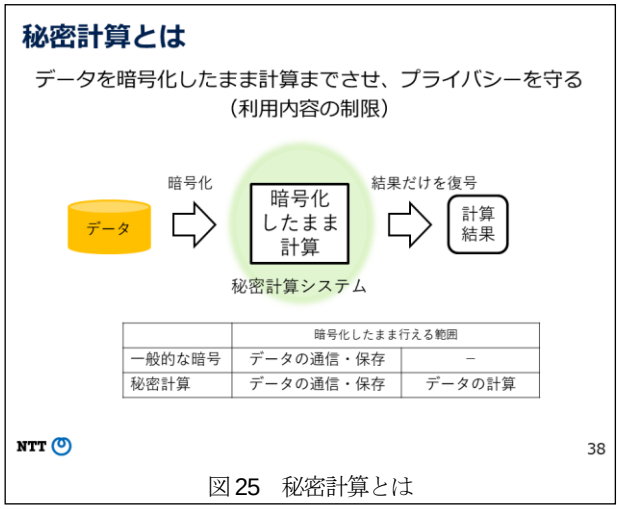

図 25　秘密計算とは

秘密計算という技術を紹介します。秘密計算というのは暗号化のアプローチの１つで、利用内容の制限をするというようなことに役に立つことができます。秘密計算は何ができるかというと、ここに書いているように暗号化したまま計算して、その結果だけを元に戻すというようなことができます。(図 25)

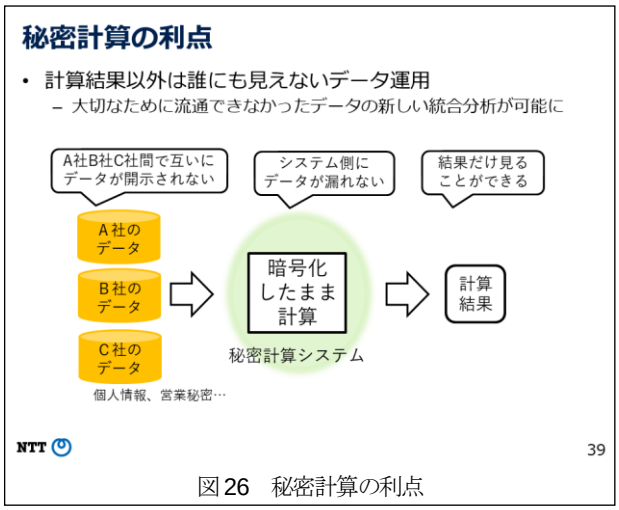

図 26　秘密計算の利点

秘密計算の利点にはいろいろなことがあります。とにかく暗号化したままデータを集めることができるので、例えば複数の違う会社のデータを集めることもできます。A 社・B 社・C 社で暗号化してデータを持ち込むので、互いにデータがばれてしまうことはありません。また、計算するシステムで、例えばクラウドの中で計算するとしても、クラウド側にデータがもれることが想定されないという利点もあります。

そして結果だけ見ることができる性質を持っています。例えば何か統計を取りたいという目標がはっきりと決まっているのであれば、途中の個人情報は誰も見る必要がないはずですから、結果だけ見ることができるというのは望ましい性質ではないかと考えられます。(図 26)

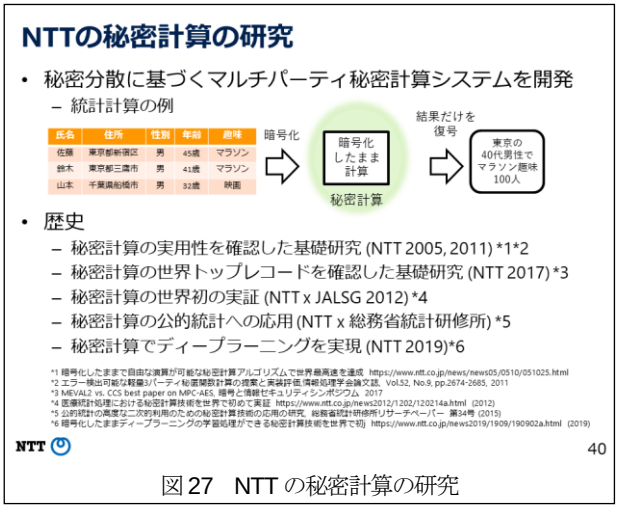

図 27　NTT の秘密計算の研究

少し細かい話になりますが、私たちは秘密分散に基づくマルチパーティー秘密計算システムという秘密計算を作っています(図 27)。このように個人データを入れると、暗号化したまま計算して結果だけ、「東京でマラソンが趣味の人は 100 人いる」というようなことがぽこんと出てきます。下半分は NTT の業績になりますが、NTT 研究所では 15 年以上前、この技術があまり知られる前から継続して秘密計算の研究を行なっています。

匿名化技術が法制度の一部となった経緯

それではテーマ B で匿名加工情報に関する話をしていきたいと思います。

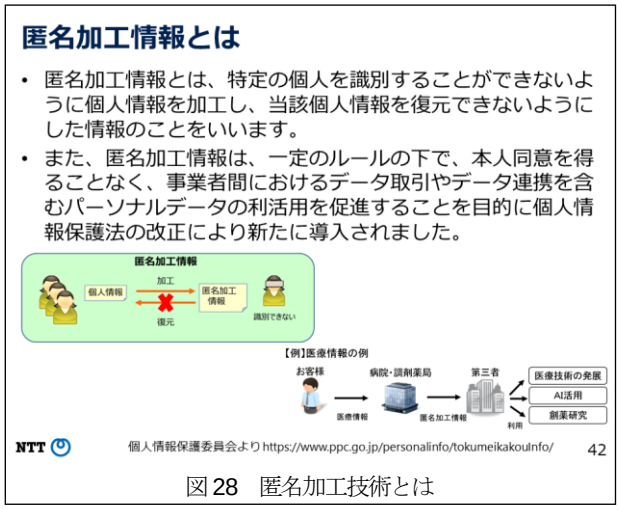

図 28　匿名加工技術とは

匿名加工情報とは何かというと、図(図 28)の一番下を御覧ください。例えば医療情報が集まったとき、第三者と書いてありますが、いろいろな医療研究をしたり薬を作ったりというときにデータを集めたいので、それを匿名化して自由に使う仕組みはないかというニーズに応えて作られた制度になります。匿名加工情報の内容の説明はしませんが、個人情報を加工して誰か分からないようにするというのが基本的な考え方です。

もう少し法律の話をします(図 29)。匿名加工情報制度には、匿名加工情報を作る事業者に対して実際に何々をしなさいという 4 つの義務があります。

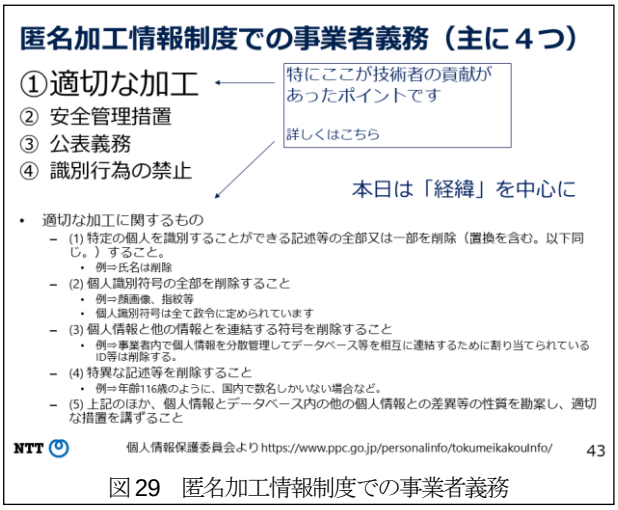

図 29　匿名加工情報制度での事業者義務

①は適切な加工をしてくださいということですが、この「適切な加工」が私たち技術者の貢献があったポイントです。技術者が法律の改正に貢献できるというのはなかなかない機会でしたので、非常に得がたい経験をさせていただきました。①の内容は下の(1)～(5)になります。これは法律からのコピペに近いものですが、この中身ではなく経緯の話を今日はしていきたいと思います。

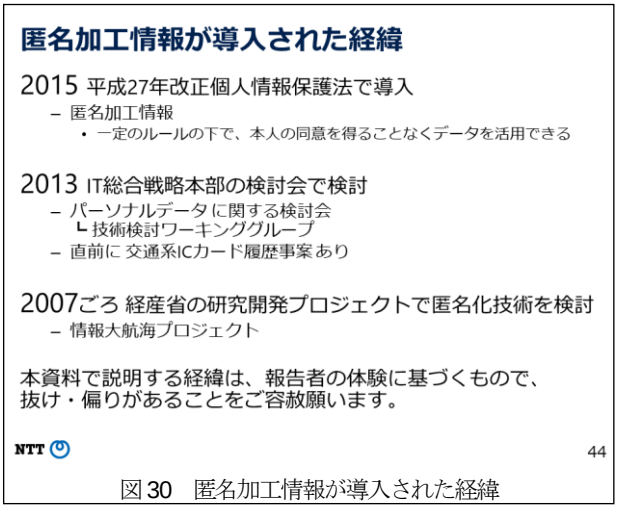

図 30　匿名加工情報が導入された経緯

ざっくり言って 2013 年の「パーソナルデータに関する検討会」が中心にあり、その前にも色々な検討があったというような話をしてまいります。この経緯についてはいろいろと調べてそれなりに客観性がある情報の提供を心がけましたが、所詮私の体験に基づくもので、賞をいただいた受賞講演だから自分が中心でもいいよねみたいな甘えがありますので、抜け、解釈の違いがあることを御容赦いただければと思います。

このページ(図 30)で大事なことは、2013 年に政府が本格的に検討を始めたということ、さらにその前にはビッグデータを使おうという世の中の動きがあったということです。後者は役所的には 2007 年の経産省の研究開発プロジェクトがあったのですが、2010 年代前半にビッグデータ活用の取り組みが色々行われました。

その中で、交通系 IC カードで個人情報保護の是非が問われた問題が起きたということもありました。語弊があるかもしれませんが、こういったことも制度を作る議論を深め、広い意味で

貢献しました。この交通系 IC カードの件は、その後の事業社の皆さんの真摯な努力もあって、歴史的に悪い意味ではなくポジティブに振り返る必要がある事案になったと思います。

こういった流れで導入された匿名加工情報の制度をこれから残りの時間を使って順番に説明してまいります。

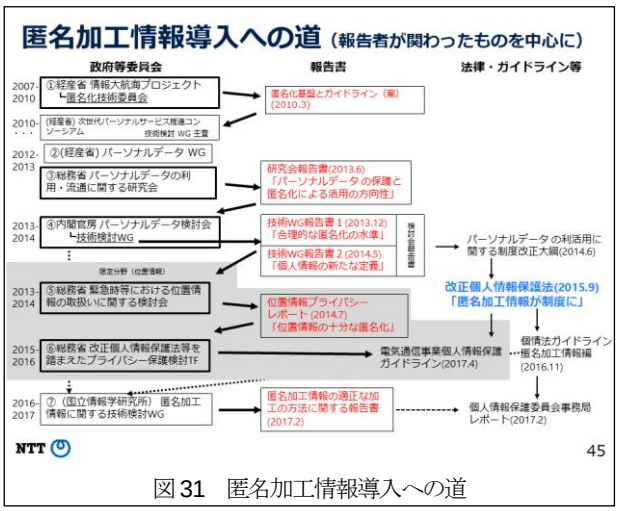

図 31　匿名加工情報導入への道

この図「匿名加工情報導入への道」が本日のメインコンテンツです(図 31)。これは恐ろしく時間がかかりました（笑い）。なんの意味があるのかというと、SCAT 大賞をいただいた業績を振り返る場ですので、自分では非常に楽しかったですが、謎の地図になってしまいました。

見方ですが、右側の青字「2015 年の個人情報保護法改正」がマイルストーンで、ここで「匿名加工情報」という技術込みの制度ができた流れを表現しています。そのことを遡ってどうやって技術が制度になったかを正確に評価するのはなかなか難しいところで、それこそ「諸説あります」になります。しかし私が関わらせていただいたものを中心にというエクスキューズを入れて、こういった流れで捉えていくととても面白いと思います。左側の①～⑦までそれぞれ何が起きたかということを残りの時間で紹介していきます。

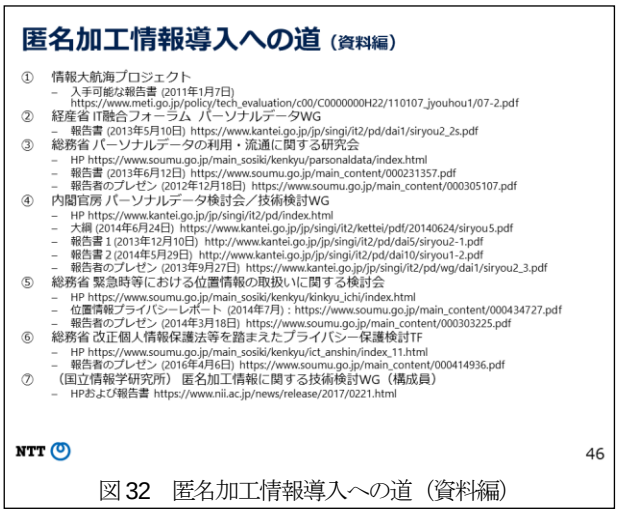

図 32　匿名加工情報導入への道（資料編）

これは自分の備忘録のようなもので、それぞれの報告書をチェックし直して現在ウェブで取れるものをあげています。(図 32)

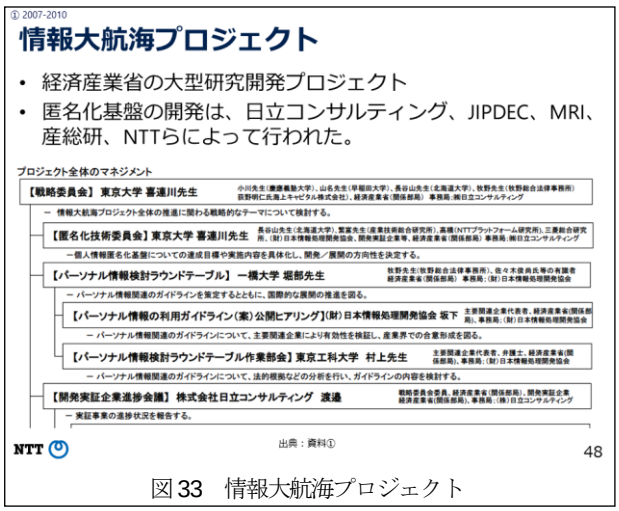

図 33　情報大航海プロジェクト

では時系列に、時代の古いところから入っていきたいと思います。

まず経産省の情報大航海プロジェクトです(図 33)。先の一覧（図 31）の大半はいわゆる国の政策を決めるための委員会ですが、一番上の情報大航海プロジェクトは俗に国プロという、研究開発をするものでした。ただし、研究開発の中で制度のことも一緒に議論していました。匿名化が注目されて期待の対象になった流れはここからスタートしているのではないかと考えています。

匿名化に関して言えば、ここに書いてある皆様と一緒に仕事をさせていただきました。プロジェクト全体は今回 SCAT 会長大賞を受賞された東大の喜連川先生がヘッドとなって行われたものであります。

図 34　匿名化基盤の開発

いろいろ書いてありますが、とにかくプロジェクト全体の目的はビッグデータを扱えるようにするということでしたので、その基盤となる匿名化技術を開発しました。(図 34)

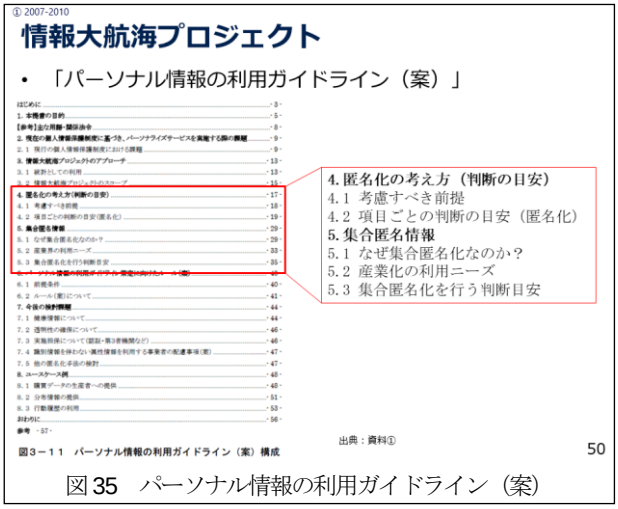

図 35　パーソナル情報の利用ガイドライン（案）

同時にパーソナル情報の利用ガイドラインというものが作られていました(図 35)。その中には匿名化の考え方や集合匿名化がいいということが打ち出されています。

集合匿名化というのは別名 k－匿名化というものですが、集合匿名化をすることでデータの安全な活用ができるという方針が打ち出されました。したがって、ここが後々非常に影響を与えたと思います。集合匿名化という考え方が、匿名加工情報の考え方に影響を与えています。

ただし、このガイドライン（案）の（案）のところが重要で、このガイドラインは私の記憶の限り、結局日の目を浴びることはなく未完成のまま終わってしまいました。ですので、後続のいろいろな検討会に部分的に拾われていく、というような形になりました。

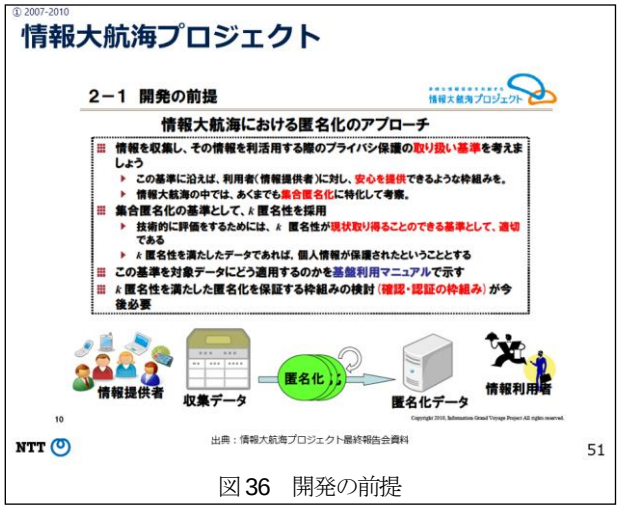

図 36　開発の前提

これはその当時の古い資料を発掘してまいりましたが、このような形で匿名化でデータを活用しようという話です。(図 36)

このk－匿名性の説明の図(図37)は10年以上前に描いた図ですが、これはめちゃくちゃコピペされまくって人様に使っていただいているのを私も何度も見たことがありますが、これは私が描いたものです。この図は非常に有名になりまして、「k－匿名性推し」みたいなものがここで確立しました。

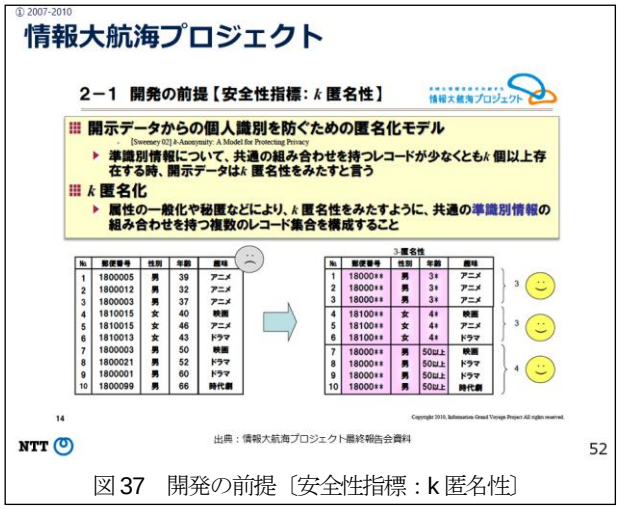

図 37　開発の前提〔安全性指標：k 匿名性〕

ここで理論的な基盤として k－匿名性に着目したのは正しかったと思っていますが、後々これに苦しめられることにもなります。

続いて(図 31)②、経産省のパーソナルデータのワーキンググループがありました。そこに至るまで２年ぐらい時間がたっています。その間、実は情報大航海はコンソーシアムを立ち上げて、私もそこの検討ワーキングの主査などをさせていただきましたが、結構苦戦していました。

そんな中、2012 年に総務省と経産省がほぼ同時に検討会を始めました。これが個人情報保護法改正のための内閣官房の検討につながっていきました。

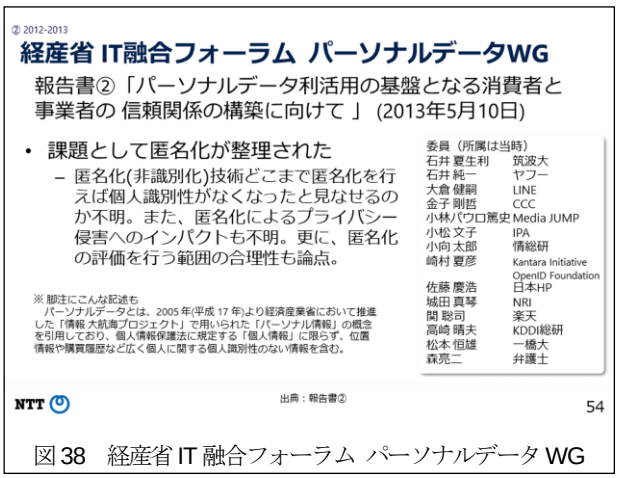

図 38　経産省 IT 融合フォーラム パーソナルデータ WG

まず(図 31)②の経産省のほうですが、こういう報告書が出ていました(図 38)。こちらの先生方が委員をやっていて、これには私は参加していません。課題として匿名化が整理されていて、匿名化について議論する必要があると、はっきりと書かれています。

次に、同時に開催された総務省の研究会(図 31 ③)の話をします。

こちら(図 39)のメンバー表に私が入っています。これは役所の研究会のデビュー戦で 40 代でしたが、ほとんど大先生ばかりで、あまり技術者もいないというアウェー感満々というところでやってきました。

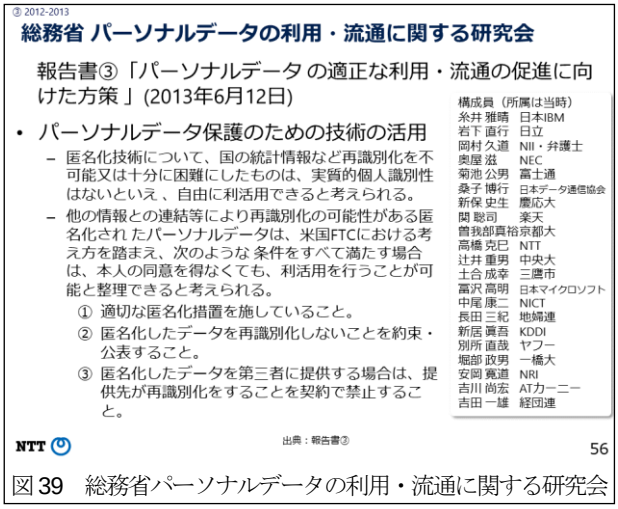

図 39　総務省パーソナルデータの利用・流通に関する研究会

ここで何が議論されたかというと、匿名加工情報のスキームのアイデアの卵がここで書かれています。どういうことかというと、「再識別の可能性がある匿名化されたパーソナルデータ」を「再識別化しないことを約束すること」を条件付きで活用させたらいいのではないかというようなことが、この報告書で出てまいります。ここが後々の匿名加工情報への議論につながっていきます。

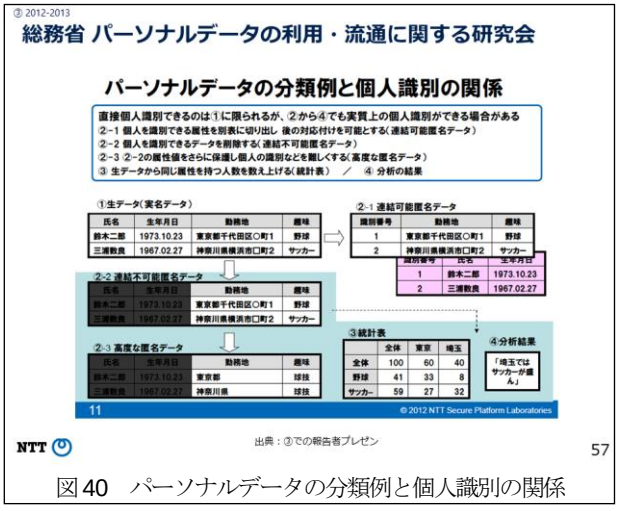

図 40　パーソナルデータの分類例と個人識別の関係

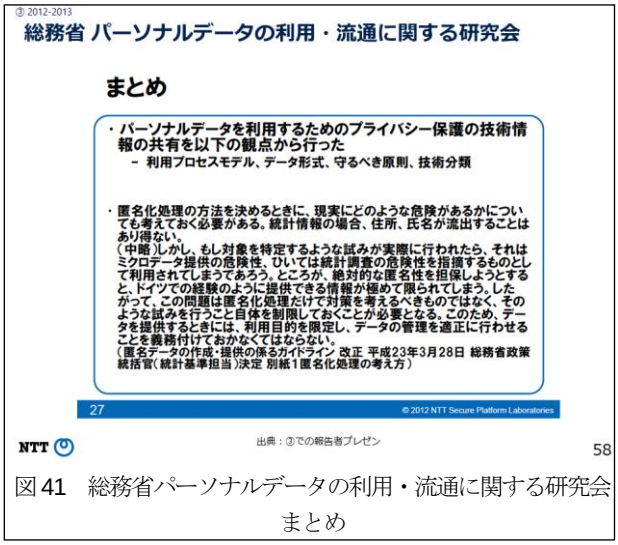

図 41　総務省パーソナルデータの利用・流通に関する研究会まとめ

これはその当時、私がその委員会でプレゼンさせていただいたものです(図 40)。今はだいぶ慣れてきましたが、テンパってプレゼンしたのを覚えています。書いている内容はどうということはなく、いろいろな技術的整理をたくさん行いました。

これも(図 41)同プレゼンからですがドイツの統計局の言葉を引用して、実際に匿名化という水準を決めることは結構困難だよということを述べています。

そんなこんなで報告書が出ました。この報告書に、2013 年 6 月に先ほどのことが書かれていて、それを受ける形で内閣官房のほうで、総務省だ、経産省ではなくてオールジャパンでこの検討会(図 31 ④)が立ち上がりました。この話をしてまいります。ここが今日の中心になります。

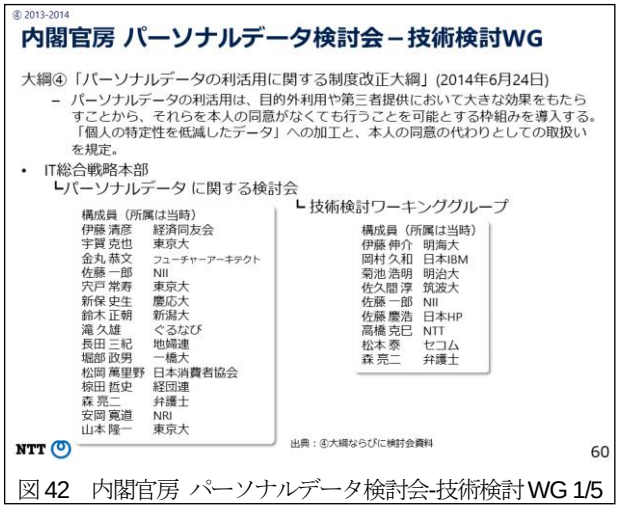

図 42　内閣官房 パーソナルデータ検討会-技術検討 WG 1/5

私は技術検討ワーキンググループに入れていただきました。親会がこのような形であり、佐藤一郎先生、弁護士の森先生が両方に参加して議論していくという形で、この検討は本当に楽しい検討でした。結局、「個人の特定性を低減したデータ」を加工することで何か新しい取り扱いをしておこうということが決まっていきます。(図 42)

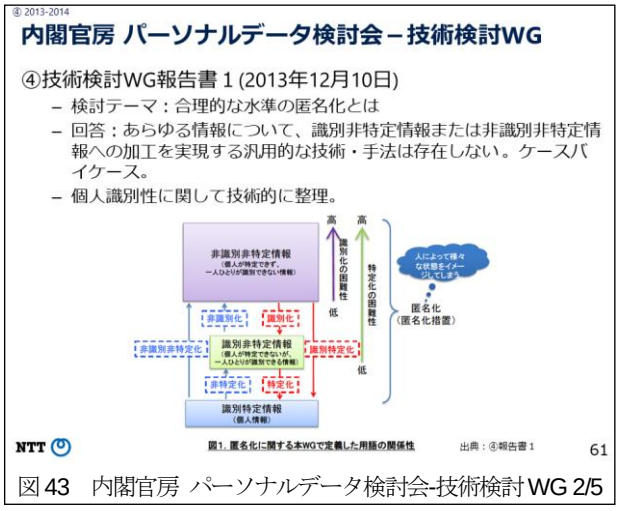

図 43　内閣官房 パーソナルデータ検討会-技術検討 WG 2/5

どういうことをやったかというと、技術検討ワーキングに問いかけられた課題は、「合理的な匿名化の水準を決めてください」ということが来たわけです。数年前に k－匿名性を押した

ことで、ここで苦しめられることになります。「そういういいものだったら合理的な水準を決められるだろう」と言われたわけです。

それに対して私たち WG が返した答えは、「あらゆる情報について汎用的な技術・手法は存在しない」ので、「何か基準は決められない」ゼロ回答をしました。つまり、いい技術があると言いながらも、この分野に関してはそう簡単にはいかないと押し返す、非常につらい受け答えでした。内容としては下の図のような細かい検討もして、個人識別性に関するいろいろな認識を高めることができました。(図 43)

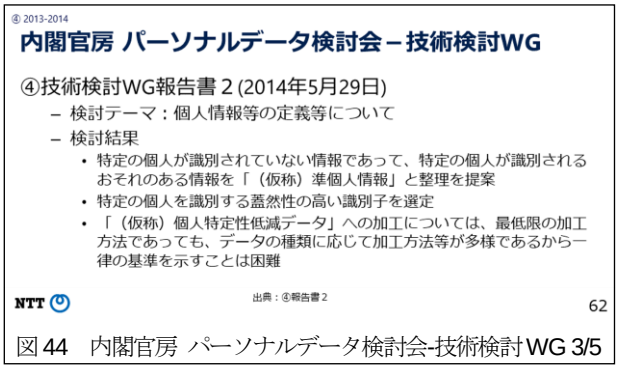

図 44　内閣官房 パーソナルデータ検討会-技術検討 WG 3/5

この技術検討ワーキングはもう１つ報告書を書いていて、これは第２弾です(図 44)。ここでは「(仮称) 準個人情報」という考えを整理しました。

これは最終的に固まった匿名加工情報よりも個人情報に近いもので、これに対する必要な技術や規律の検討が親会より指示されました。最終的に（仮称）準個人情報は採用に至りませんでしたので、「準個（じゅんこ）はどこへ行った」というネタがありましたが、この検討の蓄積が匿名加工情報の規則類に活かされることになりました。

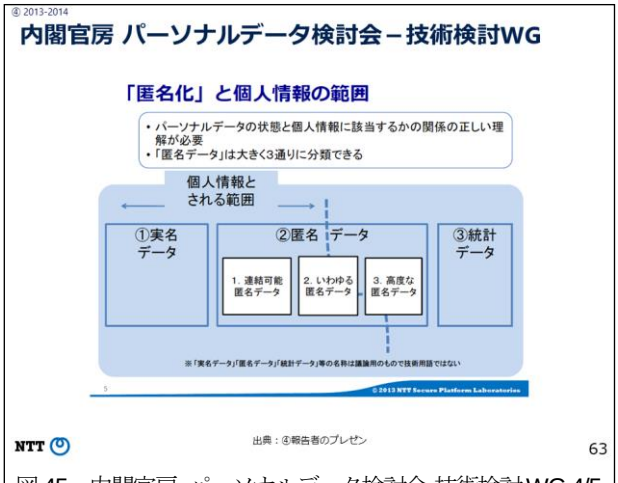

図 45　内閣官房 パーソナルデータ検討会-技術検討 WG 4/5

当時、委員会内でもプレゼンをさせていただきましたがどういう意味があったかというと、啓蒙なんですね。一筋縄ではいかないというようなことを細かい技術を説明しながらも、制度化は乱暴に行わないでくださいというような議論をしていました。(図 45)

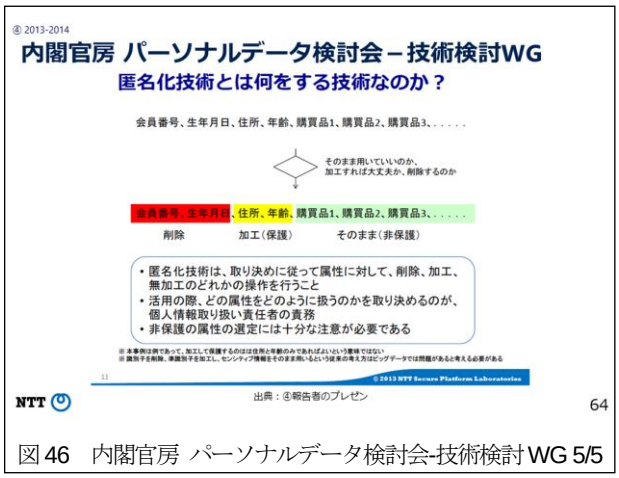

図 46　内閣官房 パーソナルデータ検討会-技術検討 WG 5/5

これも匿名化技術の説明でよく引用された図(図 46)だと思いますが、匿名化の基本的な考え方の整理をしています。

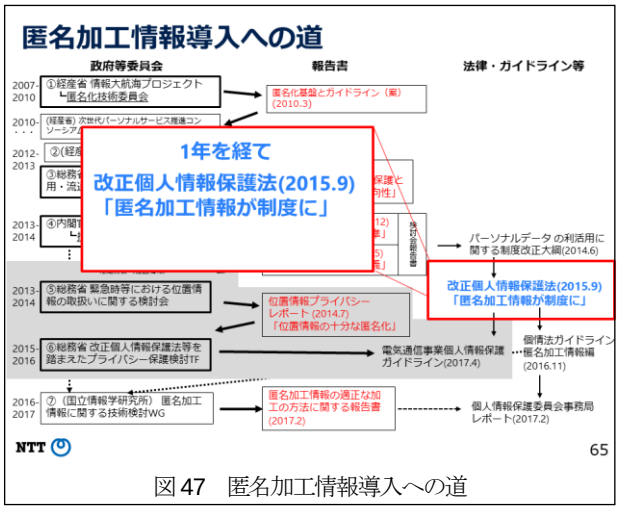

図 47　匿名加工情報導入への道

このように技術検討ワーキンググループの貢献もあって、このパーソナルデータの大綱が作られて、そこから１年を経て 2015 年に匿名加工情報が出来ましたということになります(図 47)。この灰色のところ(図 47 ⑤ ⑥)は全体の流れを受けて、早速並行してやっていこうということで、総務省的な電気通信事業のところでは、特に位置情報に関するパーソナルデータの分析をさんざんやりました。これは最終的には電気通信事業のガイドラインの改正までつながっています。

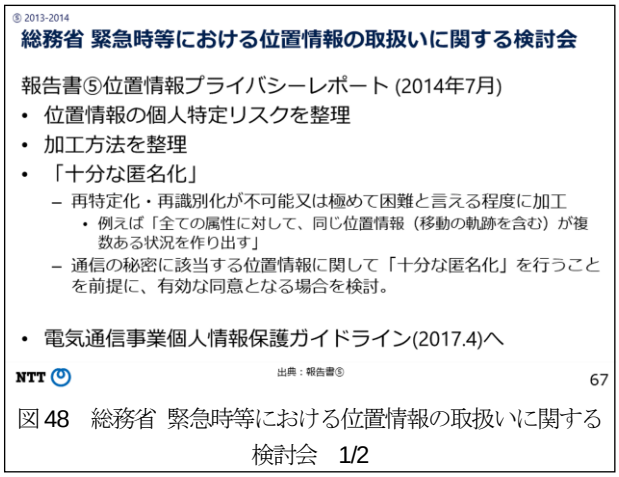

図 48　総務省 緊急時等における位置情報の取扱いに関する検討会　1/2

これは位置情報の新しい使い方が匿名化に関してできたという画期的なもので、この整理に基づいて現在のコロナに関するの人出の「新宿が何人でした」という空間統計が提供されていると言ってよいと思います。(図48)

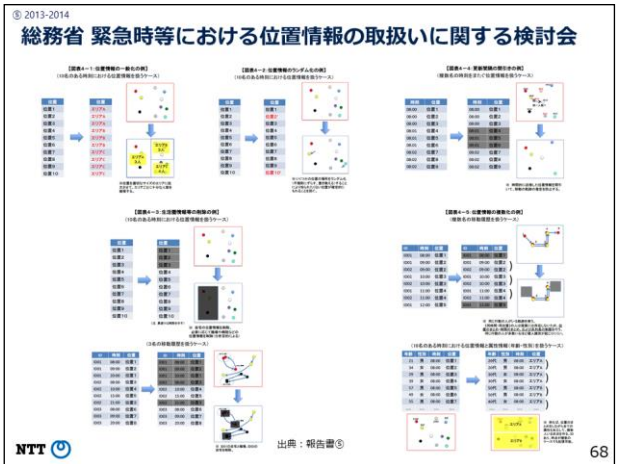

図49　総務省 緊急時等における位置情報の取扱いに関する検討会　2/2

これはいろいろな図(図49)が報告書にあって、どういうときにはどういう匿名化が考えられるというバリエーションが山ほどあって、報告書、位置情報プライバシーレポートに描かれています。これは全部私が描いた図です。総務省のエディターの方が本当によく理解している人で、色々なリクエストをくださって苦労したことを覚えています。

ということで、個別分野に関することもめでたく議論でき、こういった物ができてまいりました。

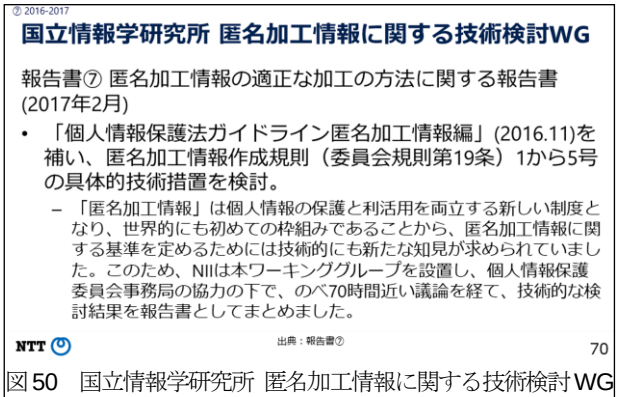

図50　国立情報学研究所 匿名加工情報に関する技術検討WG

最後に、国立情報学研究所(NII)で検討が行われました(図50)。これは何をしたかというと、個人情報保護法改正に関する匿名加工のルール、ガイドラインができてきましたが、そこの深掘り、細かいところを埋める作業を個人情報保護委員会の協力のもと、自主的にやりましたという記録です。

まとめと展望

以上で私の話はほぼ終わりです。まとめてまいります。

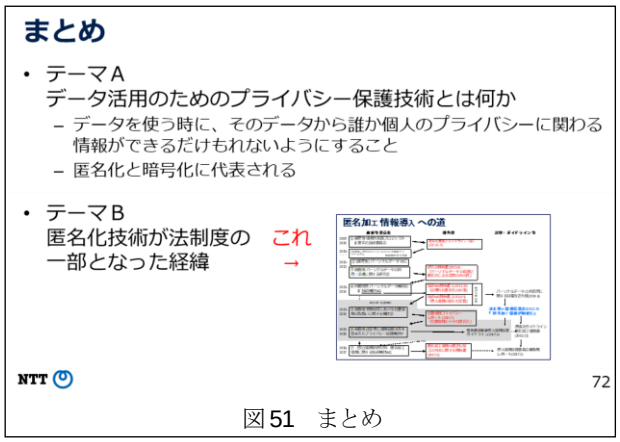

図51　まとめ

２つの話をしました。テーマAは、データ活用のプライバシー保護技術とは何かという話をしました。これは匿名化と暗号化に代表されます。テーマBは経緯ということで、これは一言で言うとこの地図ということになります。(図51)

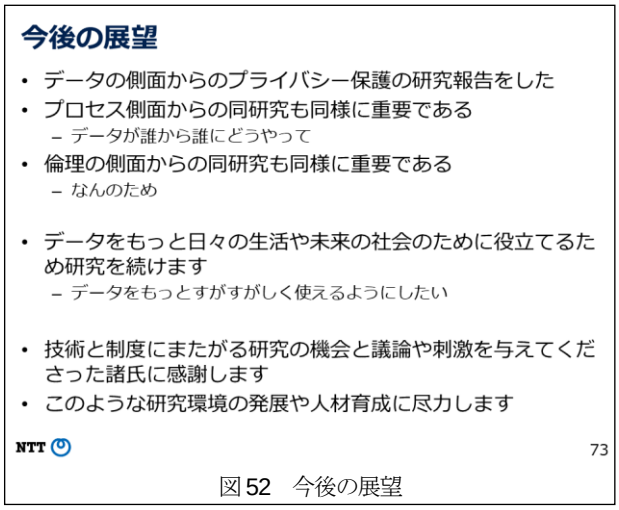

図52　今後の展望

今後の展望です(図52)。今日はデータの側面からプライバシー保護の研究報告をしました。個人データをどのように加工すればよいかというものです。しかし、プロセスの側面からの研究も重要だと考えています。データを一発で安全に加工して流通できればよいですが、それだけでできることは限られます。より高度なデータ利用やプライバシー保護をするためには、データが誰から誰にどうやっていくかということをちゃんと調べていくことが大事です。それから倫理的な側面も大事だと考えています。データをもっと日々の生活や未来の社会のために役立てるための研究を続けたいと思います。

改めて、技術と制度にまたがる研究の機会と議論や刺激を与えてくださった皆様に感謝して、さらに研究を続けるとともに、これからの研究環境の発展や人材育成に尽力してまいりたいと思います。今日は機会をいただきましてありがとうございました。御清聴ありがとうございました。

本講演録は、令和2年10月27日に開催されたSCAT主催「第108回テレコム技術情報セミナー」のテーマ、「Society5.0時代における国民の安心安全を支える研究業績」の講演内容です。