資料3

匿名加工情報の利活用に向けて

2015年12月18日

ニフティ株式会社

概要

・個人情報を匿名化して、その情報を利活用する研究は、個人情報保護法改正前から各企業、研究機関等で行われており、ニフティは、情報処理学会での活動を通じて、安全性評価を経た匿名化済み情報の利活用を可能とする流通プラットフォーム事業を模索してきた。

・個人データには、識別子等（ID,QID）だけでなく、行動履歴などのセンシティブ属性（sensitive attribute;**SA**）が記録される。匿名加工をしてもSAが一定程度利用可能であれば有用性は高いが、いざ再識別されれば、SAの加工の程度に応じプライバシー侵害リスクが高まる。したがって、SAの復元リスク低減加工に関する定性的な考え方、評価方法の確立が必要。

・また、取り扱うデータ種類や業種により、再識別リスクに関する定量的な安全性指標（確率等で表現された閾値）を決めることも必要。

※説明にあたり、情報処理学会主催の匿名加工・再識別コンテスト“PWSCUP 2015”(https://pwscup.personal-data.biz)及び論文（菊池，山口，濱田，山岡，小栗，佐久間，「匿名加工・再識別コンテストIce & Fireの設計」）からの引用または再構成をしたものを含んでいます。

※文中における引用する条文番号はすべて改正個人情報保護法（「個人情報の保護に関する法律及び行政手続における特定の個人を識別するための番号の利用等に関する法律の一部を改正する法律」）です。

改正法で新設された「匿名加工情報」

特定の個人を識別することができないように、個人情報を加工（個人情報に含まれる記述等の一部又は個人識別符号の全部を削除）したものとし、当該個人情報を復元することができないようにしたものをいう。（2条9項）。

個人情報取扱事業者は、匿名加工情報（略）を作成するときは、特定の個人を識別すること及びその作成に用いる個人情報を復元することができないようにするために必要なものとして個人情報保護委員会規則で定める基準に従い、当該個人情報を加工しなければならない。（36条1項）

匿名加工情報取扱事業者は、あらかじめ匿名加工した項目の公表、第三者提供先に匿名加工情報である旨を明示（37条）、識別行為の禁止（38条）、安全管理措置等（39条）を行う義務が課されている。

認定個人情報保護団体は、個人情報保護指針に、匿名加工情報に係る作成の方法、その情報の安全管理のための措置その他の事項を定める努力義務が課されている（53条1項）。

いかなる個人情報に対しても、識別非特定情報や非識別非特定情報となるように加工できる汎用的な方法は存在しない。

汎用的な匿名化方法は存在しないものの、ケースバイケース、つまり個人情報の種類・特性や利用の目的等に応じて技術・対象を適切に選ぶことにより、識別非特定情報や非識別非特定情報に加工することは不可能ではない。

パーソナルデータに関する検討会 技術検討ワーキンググループ（技術検討WG） 報告書 (2013/12/10) より引用https://www.kantei.go.jp/jp/singi/it2/pd/dai5/siryou2-1.pdf

参考：匿名加工情報の取扱い

改正後

出典：内閣府「第8回投資促進等ワーキンググループ」内閣官房情報通信技術（IT）総合戦略室　提出資料

匿名加工方法に関する疑問

@nifty

- 匿名加工情報は、
  - 個人情報に含まれる記述等の一部又は個人識別符号の全部を削除（＝仮名化？）
  - して得られる「個人に関する」情報（＝識別非特定情報？）

  AND
- 当該個人情報を復元することができないようにしたもの(2条9項)

※　個人情報保護委員会規則で定める匿名加工方法の基準は
　　１．特定の個人を識別すること AND ２．個人情報を復元することができないようにするために必要なこと（36条1項）

⇒　匿名加工結果は、あくまで「個人に関する」情報なので、同じ属性の個人が2人以上必ず含まれるK-匿名化まで加工するのは必要無し、と解釈するのが素直なのではないか？

⇒　個々の個人情報の加工方法は「削除」（又は不可逆な情報の置き換え）のみ明示され（＝非特定化）、当該個人情報を「復元することができない」ことで規律。

⇒　個人情報を「復元することができないようにする」とは、削除した情報を復元することに限定されているようにも読める。単純に仮名化した個人情報のID変換テーブルを破棄するだけで「復元することができないようにする」ことにならないか。

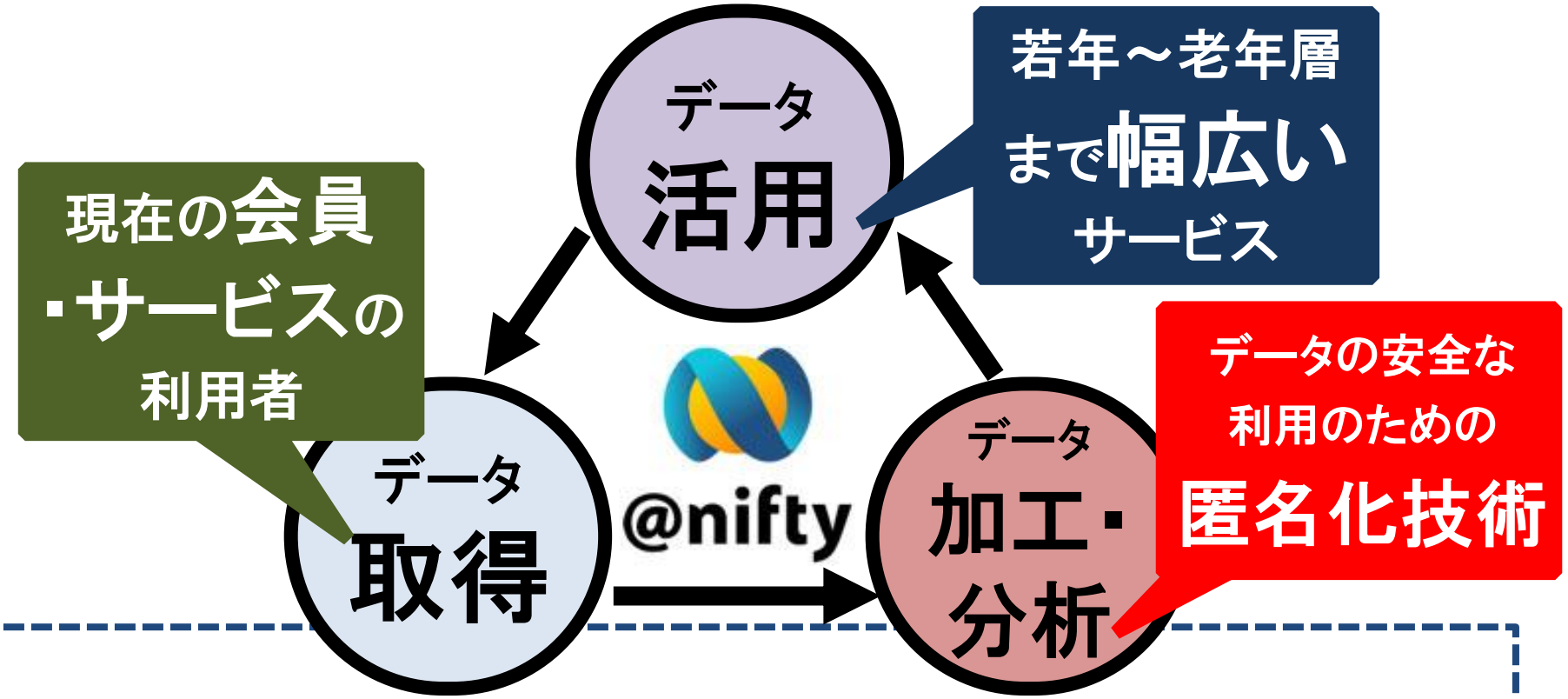
　たとえば、国会答弁等で例示されている、トップコーディング・ノイズ負荷・データの入れ替え等の処理が行われていれば、当該加工方法を知らずに元データから正確に復元するのは困難。

・以下、改正法が想定した匿名加工方法を越える（個人情報保護委員会が基準を制定できる範囲外）と思われるk-匿名化技術を中心とした取り組みと課題を紹介する。（匿名加工情報・匿名加工方法・基準等の用語は必ずしも法令上の用語ではない）

匿名加工情報に対するニフティのアプローチ

・ニフティは回線事業者/サービス事業者/クラウドサーバ運営者として、利用者のプライバシー情報を安全に取り扱う義務を負っている。

・顧客データを分析、活用し、適切にフィードバックするためには、個人の再識別可能性を低下させる匿名化処理技術が不可欠であると考える。



データ利活用とプライバシー保護を両立させる事業開発

＜参考＞ニフティが保持するデータとデータ量

・ニフティが展開するサービスは、国内では一定のシェアを確保していたとしても、自社データのみの分析では、顧客傾向の特徴検出やノイズ判定の検証ができず、限界がある。

・法定された匿名加工情報の加工基準が明確になり、ニフティ社内のみならず、多くの蓄積されているデータの流通と利活用が促進されることに期待する。

事業分野	主なデータ種類	データ説明
インターネットプロバイダ事業	インターネットＤＮＳログ	１日あたり約50G
	MVNO利用者ログ	NifMo利用ユーザログ
	顧客登録情報	約1000万ユーザ
	コールセンター対応履歴	9万件/月
クラウド事業	クラウドサーバ監視ログ	約4000案件/企業の稼働状況
	ＴＣＯ管理ログ	VM稼働状況と回線負荷
ウェブサービス事業（ポータル・検索・サービス・メール等）	WEBサービスアクセスログ	1日あたり約７Ｇ
	シュフモ会員情報	190万会員のユーザ情報
	チラシ・買い物情報分析	1万以上のスーパーのチラシ情報
	サービス決済情報	1月あたり1000万回以上
	検索エンジンログ	1日あたり約200万回
		1日あたり約70万語
情報分析サービス・ソーシャル事業	個人用データ蓄積事業	ストレージ利用状況
	ＥＣ検索事業	ＥＣ検索ログ
	リサーチ・マーケティング事業	顧客アンケート・広告等
スマートネットワーク	家電利用ログなど	IoTセンサーデータログ
コーポレート	人事・経理ログ	社内分析に適宜利用

蓄積されているが、分析・活用されていない

データ分析事業に展開

これらのデータを匿名加工することが求められるのか **「どの基準値まで」**

匿名加工情報の流通イメージ

・ニフティのように、自社だけでは利用できない情報を持つ企業は多く存在する。そこで、共通の事業目的を持つ企業群を対象にした、オープンな匿名加工情報の流通プラットフォームが必要であると考える。

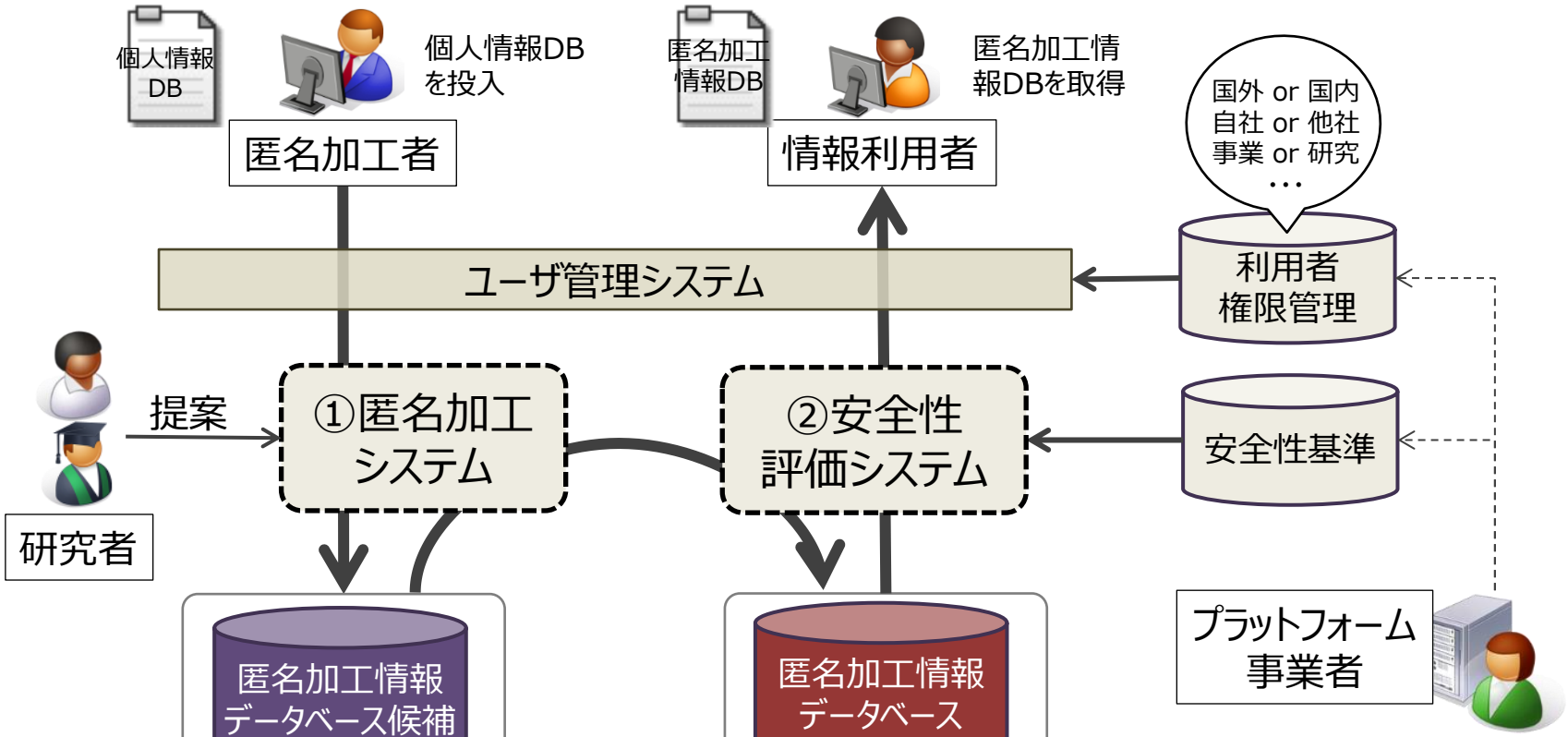
・限定的なパーソナルデータを持つ、地域行政、地場企業、ナショナル企業などが、匿名加工情報を共有し、相互送客ソリューション等に展開することを期待する。

匿名加工情報　流通プラットフォーム概念図

@nifty

・以下の図は、匿名加工・再識別コンテストでも用いられたニフティで開発中の匿名加工情報の流通プラットフォームの概念図である。

・適切な匿名加工アルゴリズムによる処理（①）と、データ種類に合わせた再識別リスクの安全性評価システム(②)を、安全性の高いプラットフォーム上で両立できるシステムを目指す。



個人情報委員会の匿名加工基準、認定個人情報保護団体の匿名加工方法に従い、安全性基準を調整する。

仮名化処理とk-匿名化処理の違い

@nifty

・仮名化処理では、元データと1対1の関係であり、下図の例では2,4,5番は1人しかいないことから、この3ユーザは識別されている。(1,3番はその1/2のリスク)

・再識別の危険性を示す指標の基本である ｋ-匿名性は、データに含まれる全ユーザが **１／ｋ 以下の確率（1人にはならない）で識別**されないことを示し、本質的なリスクが仮名化処理とは異なっている。

個人情報DB

ＩＤ	氏名	性別	年齢	購入
A001	山田	男	36	ジュース
A002	里中	男	34	水
A003	山岡	男	36	ジュース
A004	夏川	女	21	雑誌
A005	山田	女	28	雑誌
‥	‥	‥	‥	‥

氏名とIDを削除

仮名化DB

性別	年齢	購入
男	36	ジュース
男	34	水
男	36	ジュース
女	21	雑誌
女	28	雑誌

情報を抽象化

匿名加工情報DB？

性別	年齢	購入
男	30代	飲料
男	30代	飲料
男	30代	飲料
女	20代	本
女	20代	本

仮名化処理

1,3番は識別できない
（１/２のリスク）が、
2,4,5番は1人しかいない.

k-匿名化処理

全ユーザは、最大でも
１/２の確率でしか
識別されない

匿名加工によるSA属性の利活用について

@nifty

・準識別子(QID)をk-匿名化処理して安全性を高めた場合でも、年収、位置情報といったセンシティブ属性(SA)が付与されている場合、その値から識別性が高まり（識別子化）識別された場合のプライバシー侵害リスクが大きくなることが知られている。

しかし、SAを加工しすぎるとデータの有用性が失われる。

・QIDの安全化処理に加え、SAの属性に応じた識別リスク低減の考え方・基準と、それら全てを含めた安全性の評価システムが必要とされているが、現状では定番と言えるシステムは存在しない。

識別子 (ID)		準識別子：QID (説明変数，特徴量)			センシティブ属性：SA (目的変数)			
ID	名前	性別	年齢	職業	年収	位置情報1	位置情報2	・・
A001	山田	男	39	会社員	1200万円	歌舞伎町1丁目	新宿3丁目	・・
A002	里中	男	34	自由業	540万円	西新宿5丁目	北新宿2丁目	・・
A003	山岡	男	32	アルバイト	180万円	浦安市舞浜1	丸の内1丁目	・・
A004	夏川	女	29	会社員	710万円	六本木6丁目	虎ノ門3丁目	・・
A005	山田	女	25	主婦	70万円	浦安市舞浜1	丸の内1丁目	・・
・・	・・	・・	・・	・・		・・		・・

識別子は匿名加工ルールに従って処理する

QIDは、k-匿名化処理、Pk-匿名化処理 等の安全化処理を行うことで再識別リスクが低減される

SAの脅威例①
データの分散傾向からの逆推定

SAの脅威例②
属性値の追加/蓄積によるデータの詳細化

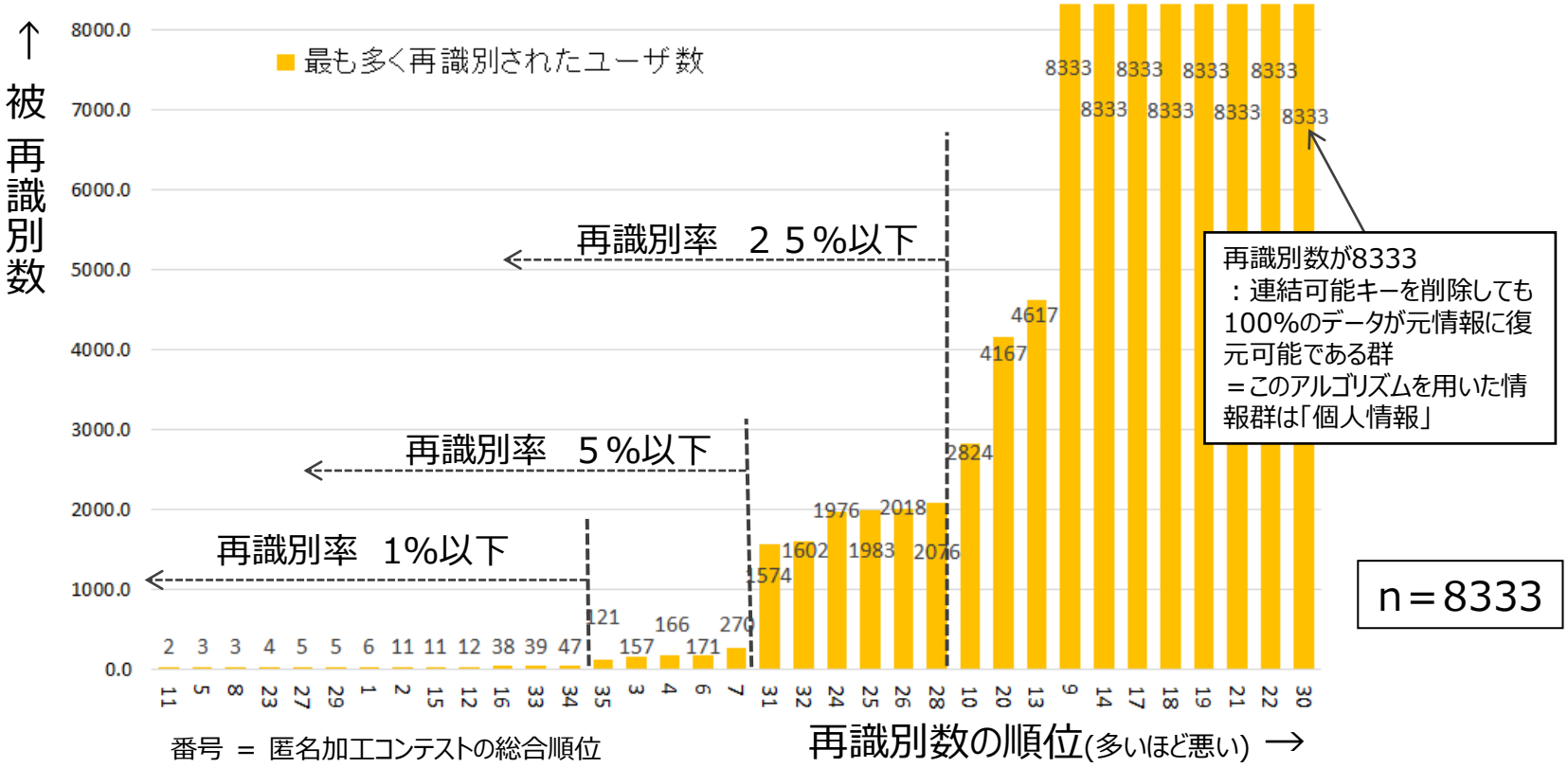
SAの流通による多様なリスクを全て解消することは困難

「再識別」の許容範囲とは

これまで、匿名加工技術の評価は、理論的な安全性の研究にとどまっていたが、情報処理学会によって、匿名加工技術と再識別技術を競うコンテスト（PWSCUP）が開催され、具体的な再識別リスク数値が計測できるようになってきた。

以下は，PWSCUPの結果として、再識別された情報人数の分布である。

被 再識別数＝再識別者が、匿名加工情報とその元となった情報を対照して、匿名加工情報のｎ行目が、元の個人情報のm行目であったことを当てた数.

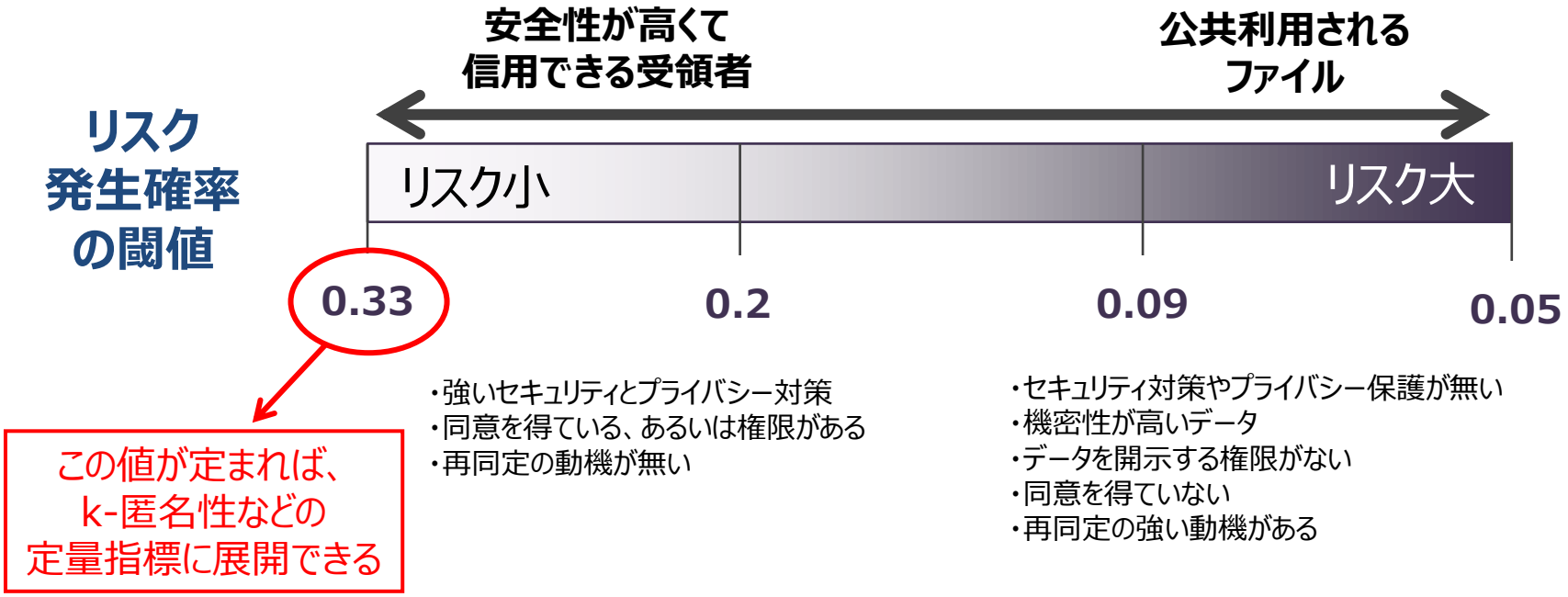


参考:カナダの小児医療における再識別リスクの定量性基準

・カナダの小児医療研究所CHEOにおける再識別リスクは以下の図のように区分されている[1] 。

・実際には、データへの攻撃者や関係者、関係機関などから漏洩する確率を全て合計し、最も高い確率を示している [2]。

・リスク発生確率の閾値が明確化される事で、k-匿名化手法を選択することが容易になる。



[1]Khaled El Emam,Luk Arbuckle,「データ匿名化手法 –ヘルスデータ事例に学ぶ個人情報保護」株式会社オライリー・ジャパン　2015.5.2

[2]例えば、一般的な医療機関にて情報漏洩が発生する確率は、0.27と決まっており,それらのリスクを含めた総合評価でリスク発生確率を算出する

匿名加工方法について解決すべき課題

事業者が匿名加工方法を検討する際に、以下の点が未解決である。

① QID/SAの有用性を維持可能な（加工）方法とは（定性的基準）

→特に、位置情報、IPアドレス、購買履歴等の本人特定リスクが高い属性情報を利用する場合、再識別リスクが高くなる。有用性を維持しつつ再識別された際のプライバシー侵害リスクを低める加工方法のあり方。

② 匿名加工情報の再識別率（定量的基準）

→匿名加工技術は多く存在するが、再識別性が存在しないこと(=再識別率0%の達成)を保証する匿名加工方法は無い。匿名加工後の情報から再識別されるリスクの許容範囲を定めることの是非とその閾値となる基準。

→個人の再識別リスク（発生確率）を「**%以下」と表現できるか？その場合に基準（閾値）はどうあるべきか。

③ 匿名加工情報の復元リスクの基準（定性・定量的基準）

→匿名加工情報 ⇒ 個人情報の「復元」とは？（全項目の復元？）

まとめ

汎用的な匿名化方法は存在しないが、QID/SAの有用性を維持可能な加工基準（定性的基準）と匿名加工情報の再識別率（定量的基準）の組み合わせ、復元リスクの明確化をすることで、個人情報の種類・特性や利用の目的等に応じて技術・対象を、認定個人情報保護団体が適切に選ぶことが可能となる。

QID/SAの有用性を維持

↓

再識別リスクと復元リスクの定性・定量化

↓

認定個人情報保護団体により利用可能な
匿名加工方法（技術/システム）の安全性評価

ニフティとなら、きっとかなう。

With Us, **You Can.**

参考： PWSCUP 2015とニフティ

・ニフティは、個人情報を安全に活用するための匿名化技術に早くから着目し、匿名化処理を簡便化するためのプラットフォームの研究を進めている。

・匿名化処理プラットフォームは、情報処理学会 コンピュータセキュリティシンポジウム2015(CSS2015)内で実施された**「PWSCUP 2015」**にて、国内のプライバシー研究者による匿名加工/再識別コンテストに利用された。

個人データ活用ビジネス、勝負の始まり

第1回　NTT、ニフティなどが狙う匿名加工データビジネス

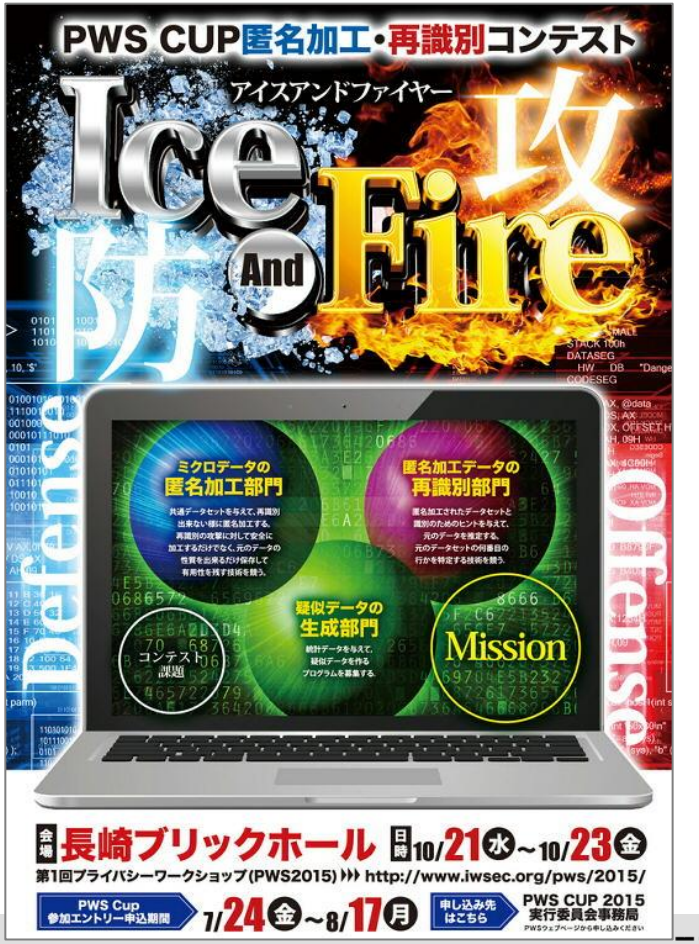
2015/11/09
大豆生田 崇志＝日経コンピュータ

2015年10月、匿名加工情報の技術を競う「匿名加工・再識別コンテスト」（PWS CUP）が初めて開催された。企業や大学の参加17チームのうち賞を総なめにしたのが、NTTセキュアプラットフォーム研究所の若手チームだ（**写真1**）。

写真1●NTTセキュアプラットフォーム研究所研究員の濱田浩気氏（左）、長谷川聡氏（中央）、正木彰伍氏。チーム名は「ψ沈黙のジャスティスψ」
[画像のクリックで拡大表示]

コンテストの狙いは、安全な匿名加工の技術の開発と評価方法を確立すること。コンテストでクラウド上の「匿名化処理プラットフォーム」を提供したニフティもノウハウの普及を狙う。

出典：ITpro 2015年11月9日

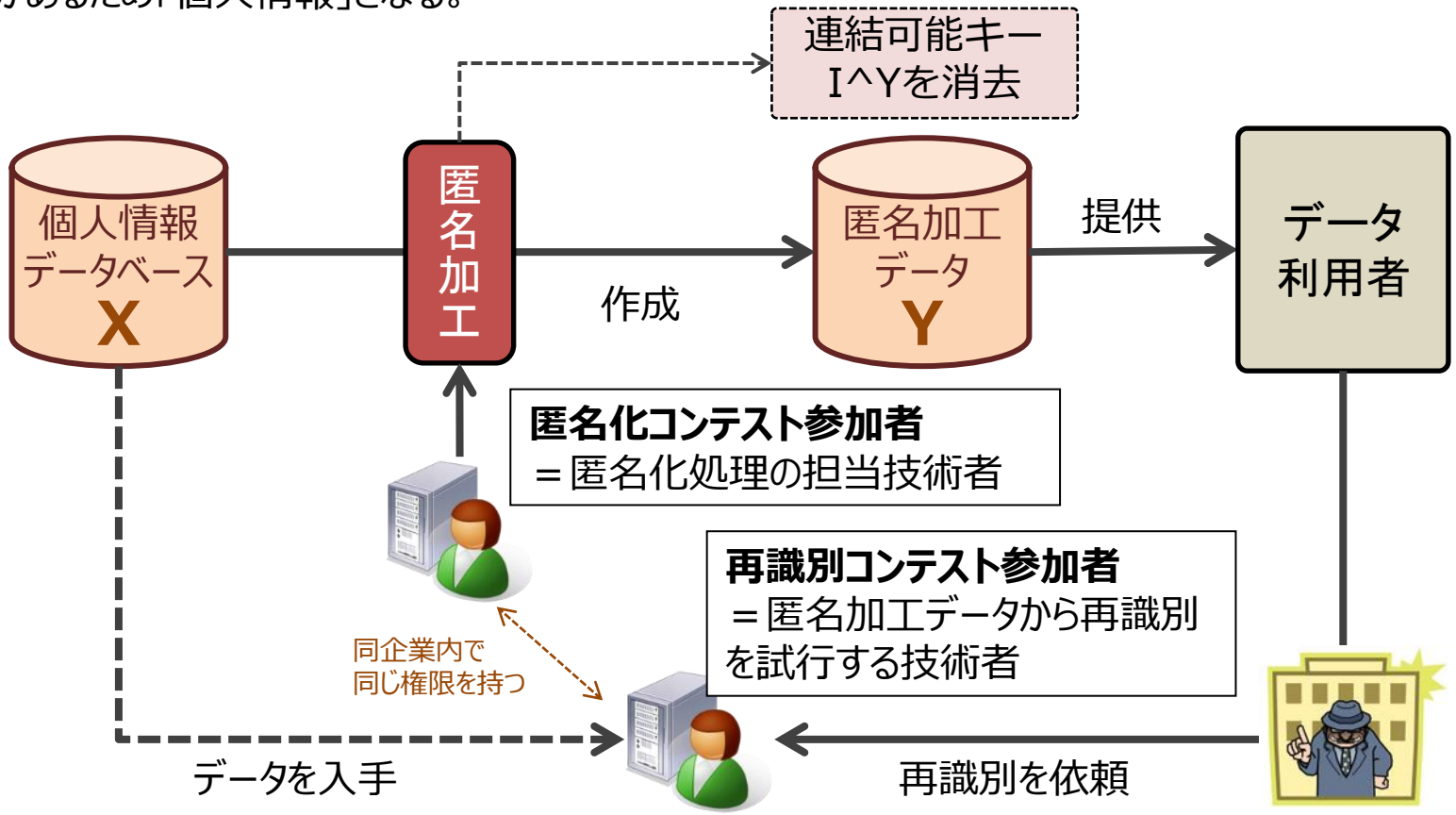


PWSCUPの技術的な位置づけ

<前提>

匿名加工者、再識別者の両者が同じ個人情報を保持していると考える。これによって、作成した匿名加工情報が、完全な背景知識を持つ人間によって、何%が復元可能であるかを擬似的に再現する。

・例えば、再識別者によって、元の個人情報に100%復元可能である場合、この情報は容易照合性があるため「個人情報」となる。



PWSCUPにおける評価指標について

@nifty

・コンテストでは、国内のプライバシー/セキュリティ専門家が、競技用の擬似データに対して適切な評価指標を13項目選択し，その総合数値で順位を決定した。

・限定的な条件下において、匿名化技術とその安全性評価を同時に実現した。

評価指標

区分	指標名	指標説明	実行環境	評価アプリ	作成者
有用性指標	有用性指標1　meanMAE	SA={14,…,25}についての平均絶誤差	R	meanmae.R	菊池 浩明
	有用性指標2,3　cross	A={7, 8, 9} (性別,年齢,就業), B=15(消費支出)のクロス集計値の平均絶対誤差	Ruby	cross.rb[1]	濱田 浩気
		上記クロス集計数の平均絶対誤差		cross.rb[2]	
	有用性指標4 corMAE	SA={14,…,25}についての全相係数の平均絶対誤差	R	cormae.R	菊池 浩明
	有用性指標5　IL	匿名加工データの各値の平均絶対誤差	Ruby	il.rb	濱田 浩気
	有用性指標6　nrow	個人データと匿名加工データのレコード数の差	システム	-	–
安全性指標	安全性指標1,2　k-anony	カテゴリ属性のk-匿名レベル	R	kanony.R[1]	菊池 浩明
		カテゴリ属性k-匿名平均値		kanony.R[2]	
	安全性指標3 E_1　re-id^{IdRand}	QI={1,…,13}が同じレコードの中から, ランダムにレコードを識別する方式	Ruby	id.rb[idrand]	濱田 浩気
	安全性指標4 E_2　re-id^{IdSA}	QI={1,…,13}が同じレコードの中から, SA={15} (消費支出)について識別		id.rb[idsa]	
	安全性指標5 E_3　re-id^{Sort}	SA={14,…, 25}の総和でソートする方式	Python	fire.py	山口 高康
	安全性指標6 E_4　re-id^{SA21}	SA={21}でソートして対応するレコードを識別	Ruby	20.rb	濱田 浩気
	安全性指標7 E_5　re-id^{AYA}	山岡攻撃を実施した行を再識別されたと判定	Ruby	20.rb	濱田 浩気

大会ルール等は、論文として提出されており、また、参加者向けのWEBサイトにも詳細は記載されている。(https://pwscup.personal-data.biz)
以下論文は　菊池，山口，濱田，山岡，小栗，佐久間，「匿名加工・再識別コンテスト Ice & Fire. の設計」，プライバシーワークショップ 2015

PWSCUP 匿名加工・再識別コンテスト

@nifty

・情報処理学会 コンピュータセキュリティ研究会主催のシンポジウム CSS 2015において、プライバシーワークショップ(PWS)が開催された。

・PWS内イベントとして、セキュリティ/プライバシー研究者による、個人データの匿名加工による情報量の維持と、攻撃者による再識別化を防ぐ、“PWSCUP”が開催された.

大会ルール等は，論文として提出されており，また，参加者向けのWEBサイトにも詳細は記載されている．(https://pwscup.personal-data.biz)
以下論文は　菊池，山口，濱田，山岡，小栗，佐久間，「匿名加工・再識別コンテスト Ice & Fire. の設計」，プライバシーワークショップ 2015

匿名加工・再識別コンテスト Ice & Fire の設計

菊池 浩明 †　山口 高康 ‡　濱田 浩気*　山岡　裕司**　小栗 秀暢¶
佐久間 淳 §

† 明治大学総合数理学部　‡(株)NTT ドコモ先進技術研究所
〒 164-8525　東京都中野区中野 4-21-1
* NTT セキュアプラットフォーム研究所　** 株式会社富士通研究所　¶ニフティ(株)
§ 筑波大学

あらまし　個人情報の保護とビッグデータの活用を両立させるために，個人情報の匿名加工の法整備が進み，様々な方式が導入されようとしている．その一方，匿名加工の為の方式には様々な提案が行われており，適用する分野やデータの種別に適した方式をどのように選定するか定まっていない．加えて，有用性と安全性を正しく評価する方法が確立していない．そこで，共通の疑似データを用いて，匿名加工とその再識別の技術を競うコンテストを企画する．本稿では，このコンテストの目的，提供する疑似データの選定，サンプルの匿名加工と再識別アルゴリズム，有用性評価方法，安全性評価方法，および，コンテストを支援する評価プラットフォームの設計について述べる．

PWSCUPでの匿名加工処理について

@nifty

匿名化プラットフォームの役割

使用したデータセット

独立行政法人 統計センター NSTAC

予備戦 CSV
擬似ミクロデータ（小：8333行）

本戦 CSV
擬似ミクロデータ（大：32027行から8333行）

データセット X

行番号	A	B	‥
1	1	2	‥
2	1	4	‥
3	2	5	‥
‥	‥	‥	‥

匿名加工処理

匿名加工者

行番号データ I^Y

行番号
3
2
1
‥

対照・判定

推定/再識別 行番号データ I^E

行番号	A	B	‥
1	1	2	‥
2	1	4	‥
3	2	5	‥
‥	‥	‥	‥

情報損失量の算出

匿名加工データ Y

A	B	‥
2	5	‥
1	5	‥
1	2	‥
‥	‥	‥

匿名情報の公開

再識別アルゴリズムEを用いて復元

再識別者

最大知識攻撃者モデル(maximum-knowledge attacker) を想定し、元データセットを共有

安全性指標の評価（「十分な匿名化」とは）

・以下のグラフは、PWSCUPの予選において提出された匿名加工情報に対して、実行委員が考案した再識別アルゴリズムによる再識別攻撃と、他の参加者による再識別攻撃によって、元の情報に再識別されたデータの再識別された割合を示すものである。

・このような状況を踏まえて、匿名加工情報と認定される定量的な基準の明確化が求められる.

再識別成功率分布