NTT

資料3

パーソナルデータ利用・流通のための プライバシーを制御する技術

2012年12月18日

日本電信電話株式会社
NTTセキュアプラットフォーム研究所
高橋克巳

パーソナルデータ利用・流通における プライバシーに関する技術の背景にある問題

・情報セキュリティの場合
  - 情報の機密性、完全性、可用性を維持すること
  - 上記が損なわれるリスクに対して対策技術が定義される
    - 情報を「もらさない」技術

・プライバシーの場合
  - 守るべき原則の理解共有はこれから
  - 守るべき原則に対応する技術の明確化はさらにその先
    - プライバシー技術への期待は？

・個人情報を「もらさない」ことではない（単純な保護との違い）
  - 購買データを利用する →《誰かが何かを買ったこと》はもれる
  - 「どんなプライバシーの制御ができるか？」

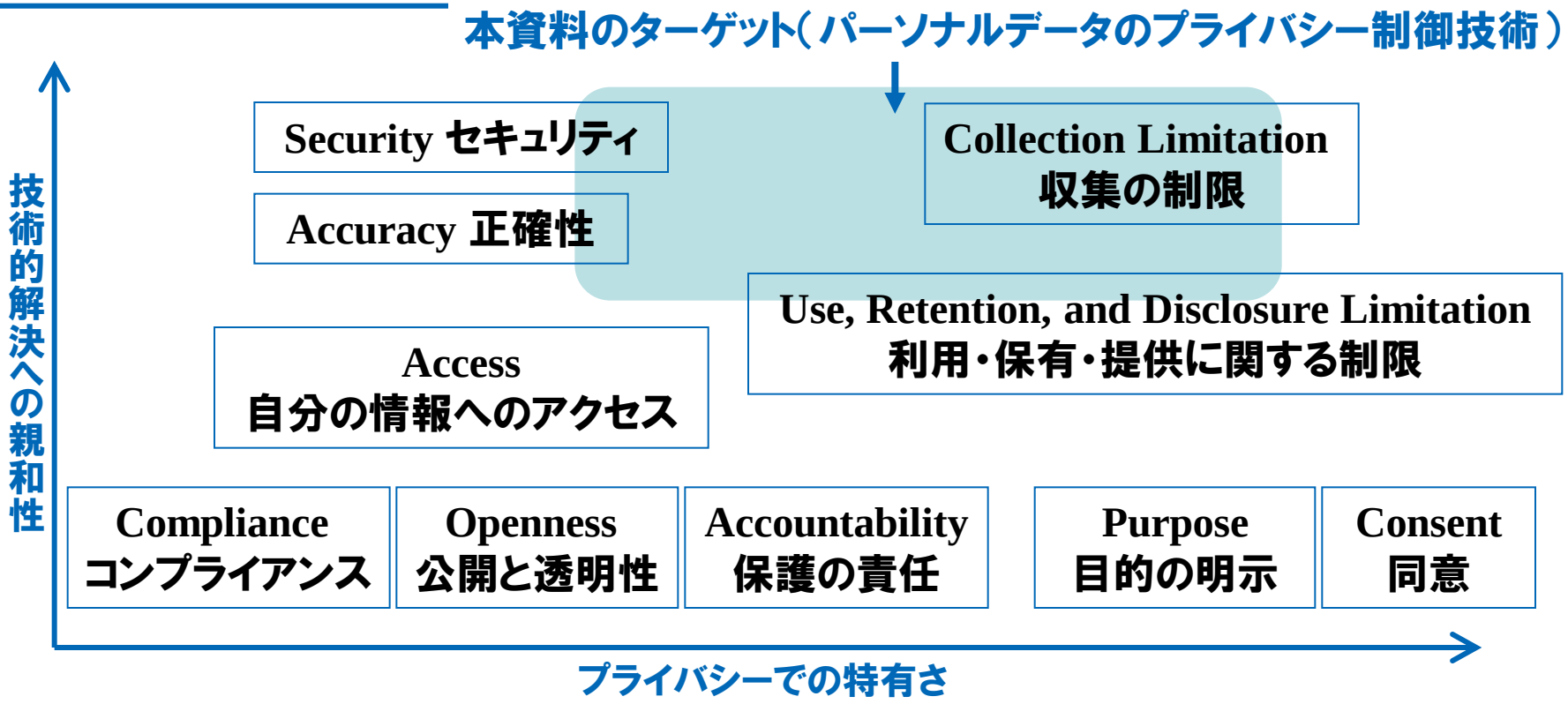
・本資料の目的
  - プライバシーを制御する技術の種類と性質の情報共有
  - なにができて、なにができないか

本資料の目的と構成

- 現在知られているプライバシーを制御する技術の種類と性質の情報共有
- 構成
  - 基本概念の整理
    1. 利用プロセスモデルの定義
    2. パーソナルデータの形式の分類
    3. 守るべき原則の分類
  - 要素技術の紹介

パーソナルデータ利用流通の原則における本資料のスコープ

- パーソナルデータを正しく扱うためには様々な原則に配慮する必要がある
- 本資料のスコープをカブキアン博士のプライバシー原則*にあてはめると以下になる



＊ Privacy Principles from “Creation of a Global Privacy Standard” Ann Cavoukian (2006)
10原則の分類・配置は筆者による（ISO/IEC 29100 Privacy frameworkも参考にした）

基本概念の整理

1．利用プロセスモデルの定義

利用プロセスモデルの定義とプライバシー

- パーソナルデータを誰が扱うか、誰から誰へ提供するかを明確に理解することが重要

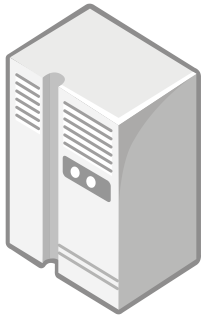
パーソナルデータの
提供者

パーソナルデータの
処理者

パーソナルデータの
活用者

Aさん

Bさん



← 出力プライバシー →

← 処理プライバシー →

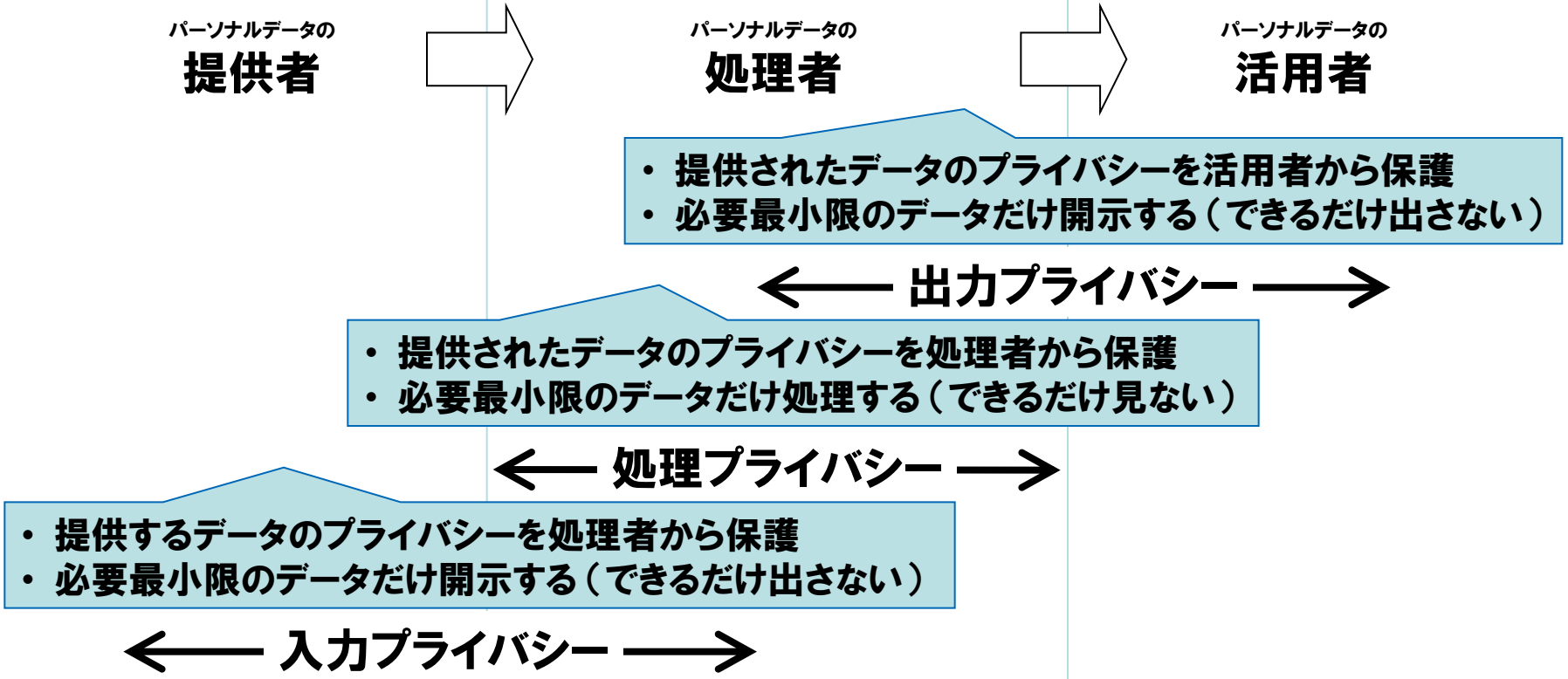
← 入力プライバシー →

利用プロセスの3つの主体

- パーソナルデータ提供者
  - 購買やウェブ閲覧やアンケート回答を行うことにより、その履歴を提供する主体
- パーソナルデータ処理者
  - データを受け取り、処理加工し、活用者に開示する主体
    - 「匿名データ」へ加工、「統計表」へ加工、分析結果の導出、・・・
- パーソナルデータ活用者
  - 加工結果を受け取り、業務に利用する主体
    - 顧客分析の実施、商品推薦サービスの提供、・・・
    - 活用者から更に二次的な活用者への開示が起こる場合は、一次活用者と二次活用者間に処理者と活用者の関係が継承される

各プライバシーに求められる基本性質

- 扱う／提供するパーソナルデータを必要最小限にすることが基本
- 提供の出元・出先間で 意識のギャップ を作らないことが重要
  - 意識のギャップ ： こう扱われているはず ⇔ こう扱っていいはず
  - ギャップ断は同意が前提 → 同意結果の正確な技術実装／技術前提の同意



基本概念の整理

2. パーソナルデータの形式の分類

パーソナルデータの形式の分類

分類	性質	代表的な利用のされ方と解釈
①生データ （実名データ）	個人を識別することができる氏名、生年月日などに代表される記述を含むデータ	・利用には多くの制約 ・法の「個人情報」に該当
②匿名データ	・個人を識別することができる記述を取り除いたデータ（形式的定義） ・個人の識別ができないようにしたデータ	・利用・流通で用いられる ・形式的定義から個人識別性は判断できない
③統計表	匿名データから同じ性質を持った個人を数え上げて表にしたもの（形式的定義）	・利用・流通、公開データに用いられることも ・個人識別性の根本は②に準拠
④分析結果	パーソナルデータを分析した結果の値やそこから得られる知識	・業務アプリケーションと共に用いられる②③の最終利用形態 ・個人識別性の根本は②に準拠
（参考） 暗号化データ	権限（鍵）を持つもの以外には閲覧・処理が不可能なデータ	・安全管理措置のひとつとして望ましい ・暗号化の有無と個人情報の判断は独立

パーソナルデータの分類例と個人識別の関係

直接個人識別できるのは①に限られるが、②から④でも実質上の個人識別ができる場合がある

②-1 個人を識別できる属性を別表に切り出し 後の対応付けを可能とする（連結可能匿名データ）

②-2 個人を識別できるデータを削除する（連結不可能匿名データ）

②-3 ②-2の属性値をさらに保護し個人の識別などを難しくする（高度な匿名データ）

③ 生データから同じ属性を持つ人数を数え上げる（統計表） ／ ④ 分析の結果

①生データ（実名データ）

氏名	生年月日	勤務地	趣味
鈴木二郎	1973.10.23	東京都千代田区○町1	野球
三浦数良	1967.02.27	神奈川県横浜市□町2	サッカー

②-1 連結可能匿名データ

識別番号	勤務地	趣味
1	東京都千代田区○町1	野球
2	神奈川県横浜市□町2	サッカー

識別番号	氏名	生年月日
1	鈴木二郎	1973.10.23
2	三浦数良	1967.02.27

②-2 連結不可能匿名データ

氏名	生年月日	勤務地	趣味
鈴木二郎	1973.10.23	東京都千代田区○町1	野球
三浦数良	1967.02.27	神奈川県横浜市□町2	サッカー

②-3 高度な匿名データ

氏名	生年月日	勤務地	趣味
鈴木二郎	1973.10.23	東京都	球技
三浦数良	1967.02.27	神奈川県	球技

③統計表

	全体	東京	埼玉
全体	100	60	40
野球	41	33	8
サッカー	59	27	32

④分析結果

「埼玉ではサッカーが盛ん」

基本概念の整理

3. 守るべき原則の分類

再度カブキアン博士のプライバシー原則を振り返ってみる

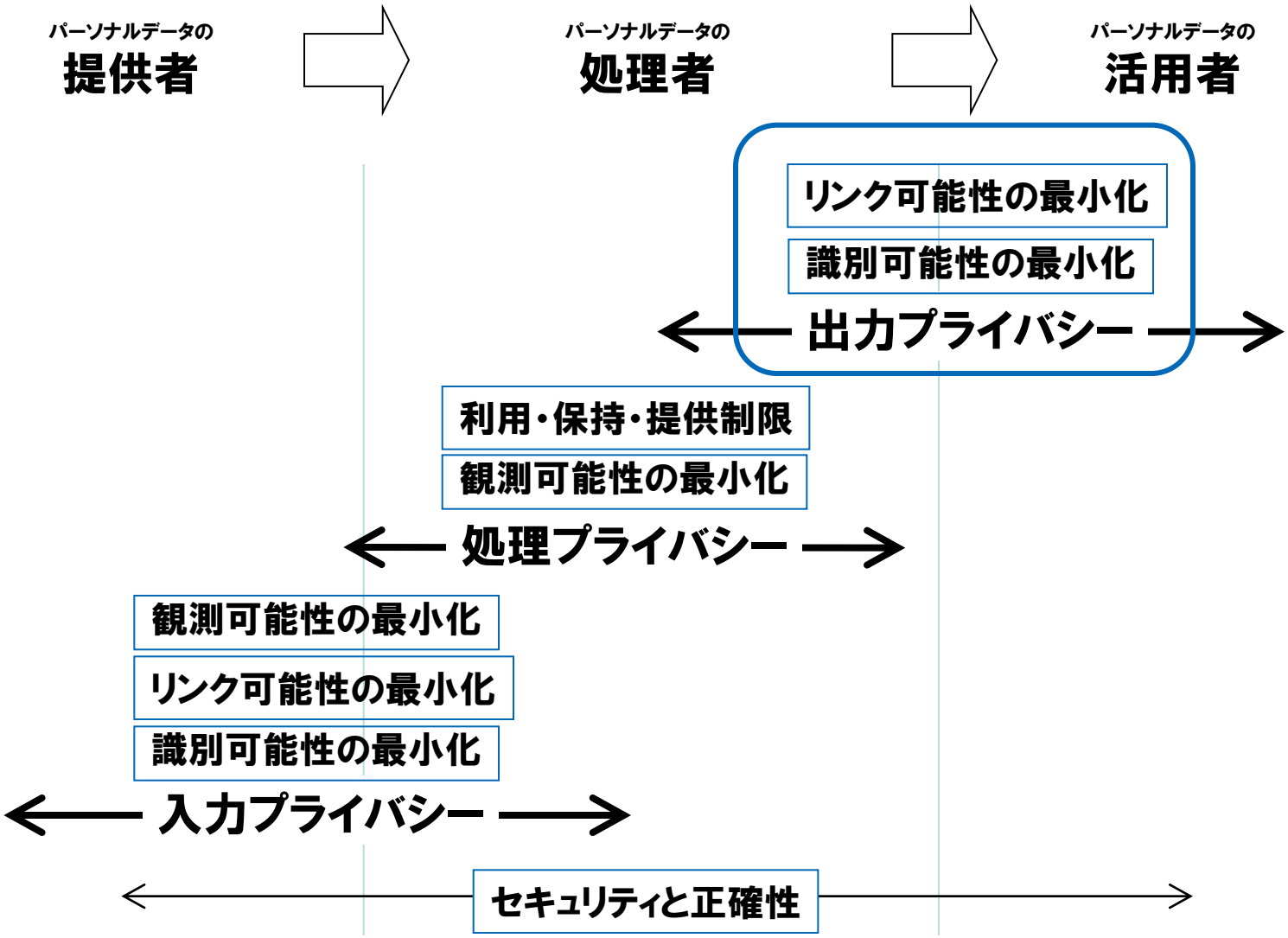
4. Collection Limitation
   The collection of personal information must be fair, lawful and limited to that which is necessary for the specified purposes.
   - Data Minimization − The collection of personal information should be kept to a strict minimum. The design of programs, information technologies, and systems should begin with non-identifiable interactions and transactions as the default. Wherever possible, identifiability, observability, and linkability of personal information should be minimized.
5. Use, Retention, and Disclosure Limitation
   Organizations shall limit the use, retention, and disclosure of personal information to the relevant purposes identified to the individual, except where otherwise required by law. Personal information shall be retained only as long as necessary to fulfill the stated purposes, and then securely destroyed.

∗ Privacy Principles #4 and #5 from “Creation of a Global Privacy Standard” Ann Cavoukian (2006)

技術的解決が期待されるプライバシー原則

- **識別可能性** (identifiability) **の最小化**
  - **データからできるだけ個人が識別できないようにする**
- **観測可能性** (observability) **の最小化**
  - **データをできるだけ閲覧できないようにする**
- **リンク可能性** (linkability) **の最小化**
  - **データをできるだけ他のデータと結び付けられないようにする**
  - Privacy Principles #4
- **利用・保持・提供制限** (Use, Retention, and Disclosure Limitation)
  - **データが正しい目的以外で使われないようにする**
  - Privacy Principles #5
- **セキュリティ** (Security) **と正確性** (Accuracy) **の確保**
  - Privacy Principles #7 and #6

プライバシー原則のプロセスモデルへのマッピング



識別可能性とリンク可能性の考察

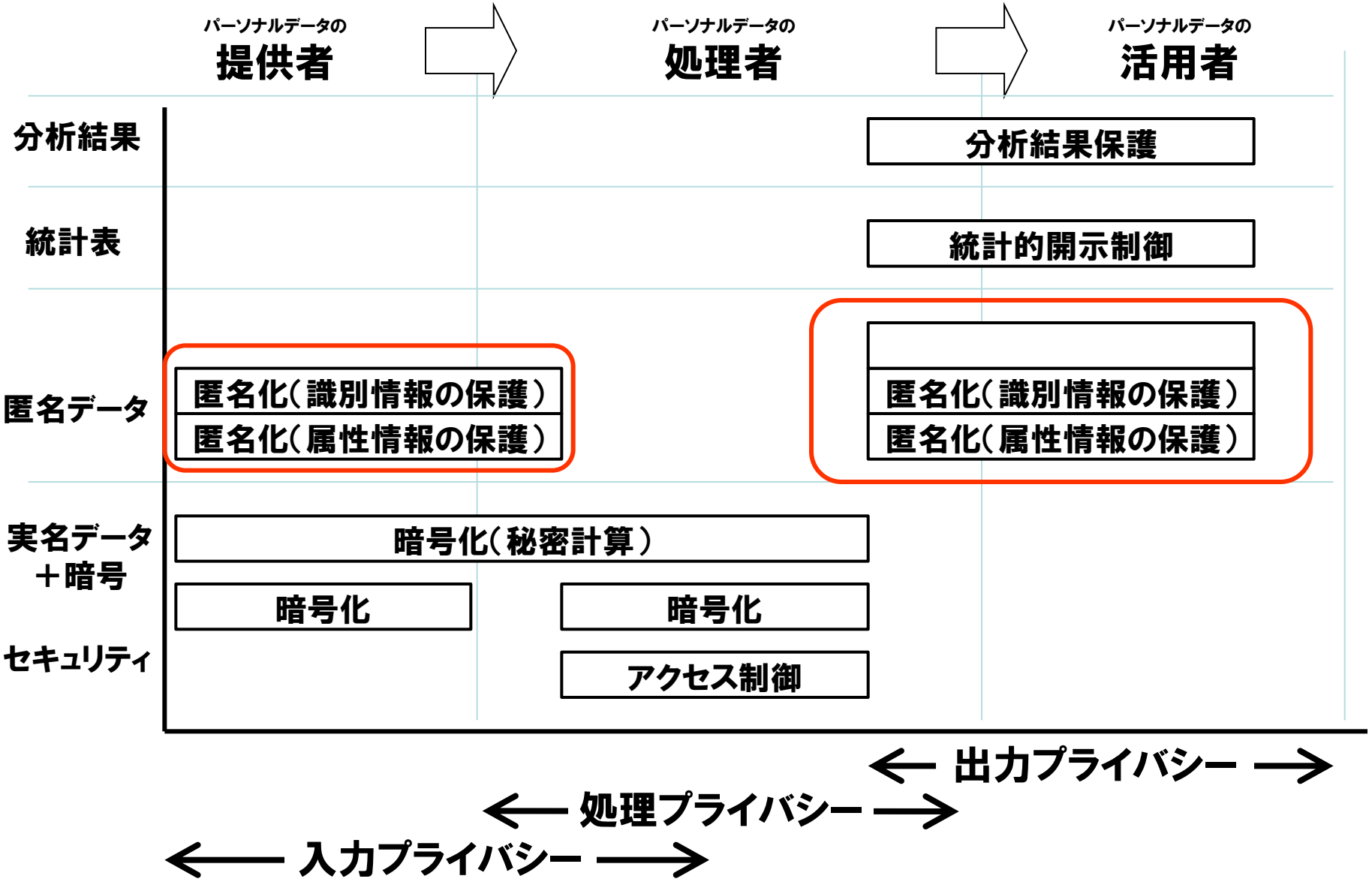
- 識別可能性
  - 匿名データの一部が、そのデータに含まれる年齢、住所等の属性情報を組み合わせることにより、本人が誰であるか認識できでしまうこと
- リンク可能性
  - 匿名のデータが他のデータと結び付けられることによって、識別可能性が高くなったり、知られていなかった属性の値が推定されてしまうこと

代表的なリスク因子

- 開示データ受領者の持つ背景知識
  - 受領者が開示データの母集団や構成される個人に関する知識を有する場合、識別可能性・リンク可能性のリスクが高くなる
- 開示データの形式
  - 匿名データ、統計表、分析表から直接個人識別が起こることはないが、それぞれのデータにおける匿名状態の作り方に応じて、識別可能性やリンク可能性が生じる
- 開示データの性質
  - 匿名データの個人あたりの属性が多い場合（いわゆる、長いログデータ）リスクが高まる
  - 匿名データに希少な属性値が含まれる場合リスクが高まる
  - 同じあるいは類似した母集団のデータを繰り返して開示を受ける場合リスクが高まる
- 開示データの機微度
  - リスクの大きさはデータの「価値」に依存し、属性の機微度が高ければリスクも大きくなる
- 開示データの管理
  - 開示したデータに対して、行ってよいデータ操作が多い場合リスクが増大する
  - データを公開した場合リスクは最大になり、開示先とデータ操作を限定する場合リスクは小さくなる
- ID管理
  - 匿名データで用いられるIDの利用期間が長いほど、利用範囲が広いほどリスクが増大
- 開示データの有用性
  - プライバシー保護とデータ有用性のトレードオフの関係にあり、より開示データに有益と思われるデータを多く残す場合、プライバシーリスクは高くなる

プライバシーを制御する技術

プライバシーを制御する要素技術の分類



プライバシー制御の要素技術

出力プライバシー／入力プライバシーのための匿名化

- 匿名化（識別情報の保護）
  - 【概要】提供者の識別情報を保護するために、識別情報を符号に変換する、あるいは削除する
  - 【プロセスモデル】入力プライバシー／出力プライバシー
  - 【データ形式】個別あるいは集合の生データから匿名データへの変換
  - 【プライバシー原則】識別可能性の最小化
  - 【成熟度】実用レベル
- 匿名化（属性情報の保護）
  - 【概要】提供者の属性情報を保護するために、グルーピング（数値属性やカテゴリ属性を上位の値に変換）、トップコーディング（特に高い値や低い値を「○○以上」などとまとめる）、誤差の混入（確率的にノイズを入れる）、スワッピング（確率的に別な値と入れ替える）を行う
  - 【プロセスモデル】入力プライバシー／出力プライバシー
  - 【データ形式】個別あるいは集合の生データから匿名データへの変換
  - 【プライバシー原則】識別可能性の最小化、リンク可能性の最小化
  - 【成熟度】実用レベル

出力プライバシー／入力プライバシーのための匿名化

- 集合データの高度な匿名化（k-匿名化）
  - 【概要】集合データの識別等のリスクを減じるため、同じ属性の組み合わせを持つ個人が複数存在するようになることを基本に、属性情報の保護を行う
  - 【プロセスモデル】出力プライバシー
  - 【データ形式】集合生データから匿名データへの変換
  - 【プライバシー原則】識別可能性の最小化、リンク可能性の最小化
  - 【成熟度】実用レベル（様々な推定リスクは研究が行われている）

（参考）k-匿名性

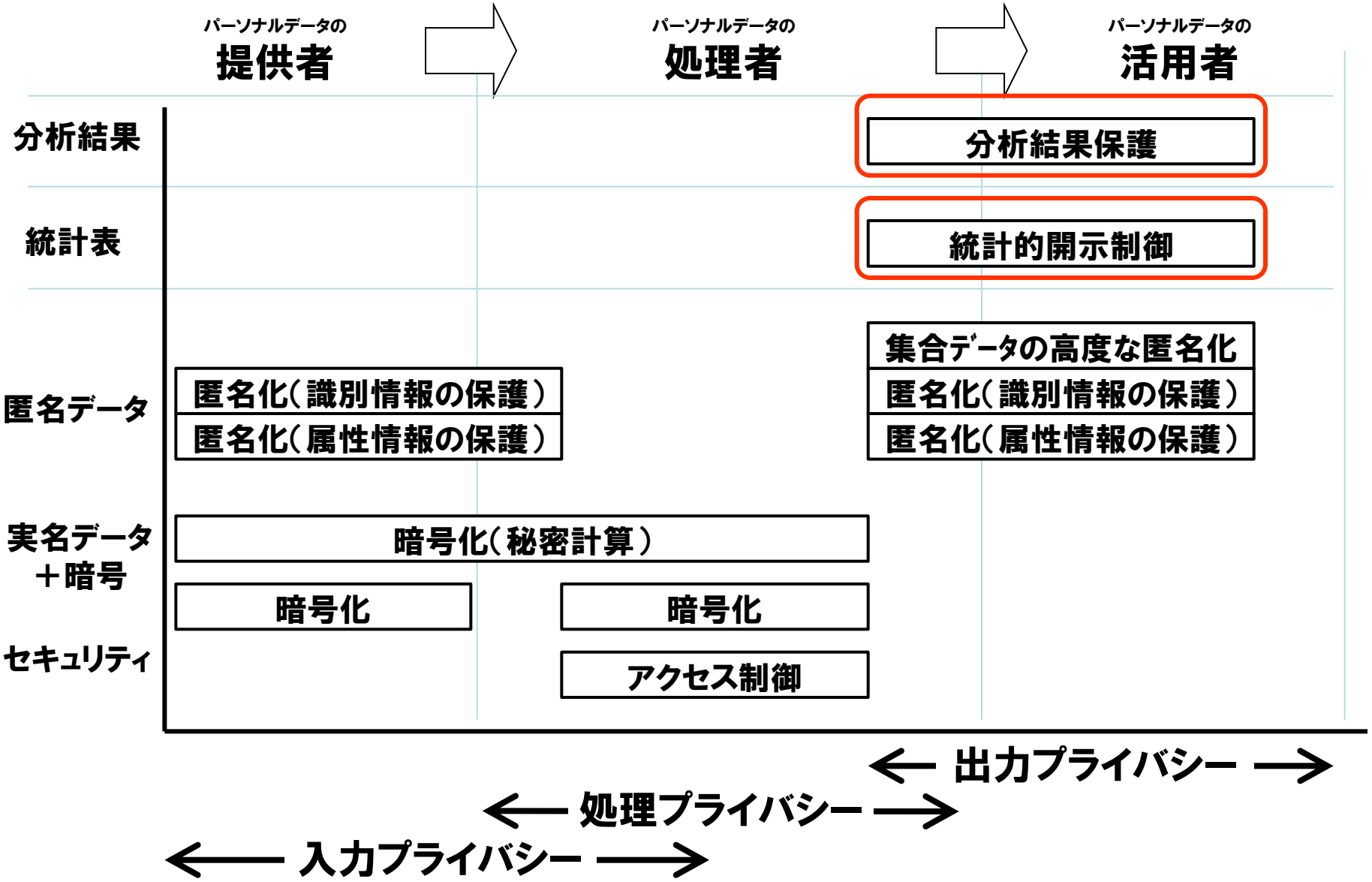
- 開示データからの個人識別を防ぐための匿名化モデル
  - 保護する属性について、共通の組み合わせを持つレコードが少なくともk個以上存在する時、開示データはk-匿名性をみたすと言う
- k-匿名化
  - 属性の一般化や秘匿などにより、k-匿名性をみたすように、共通の属性の組み合わせを持つ複数のレコード集合を構成すること

No.	郵便番号	性別	年齢	趣味
1	1800005	男	39	アニメ
2	1800012	男	32	アニメ
3	1800003	男	37	アニメ
4	1810015	女	40	映画
5	1810015	女	46	アニメ
6	1810013	女	43	ドラマ
7	1800003	男	50	映画
8	1800021	男	52	ドラマ
9	1800001	男	60	ドラマ
10	1800099	男	66	時代劇

3-匿名性

No.	郵便番号	性別	年齢	趣味	
	18000**	男	3*	アニメ	3
	18000**	男	3*	アニメ	
	18000**	男	3*	アニメ	
	18100**	女	4*	映画	3
	18100**	女	4*	アニメ	
	18100**	女	4*	ドラマ	
	18000**	男	50以上	映画	4
	18000**	男	50以上	ドラマ	
	18000**	男	50以上	ドラマ	
	18000**	男	50以上	時代劇	

プライバシーを制御する要素技術の分類

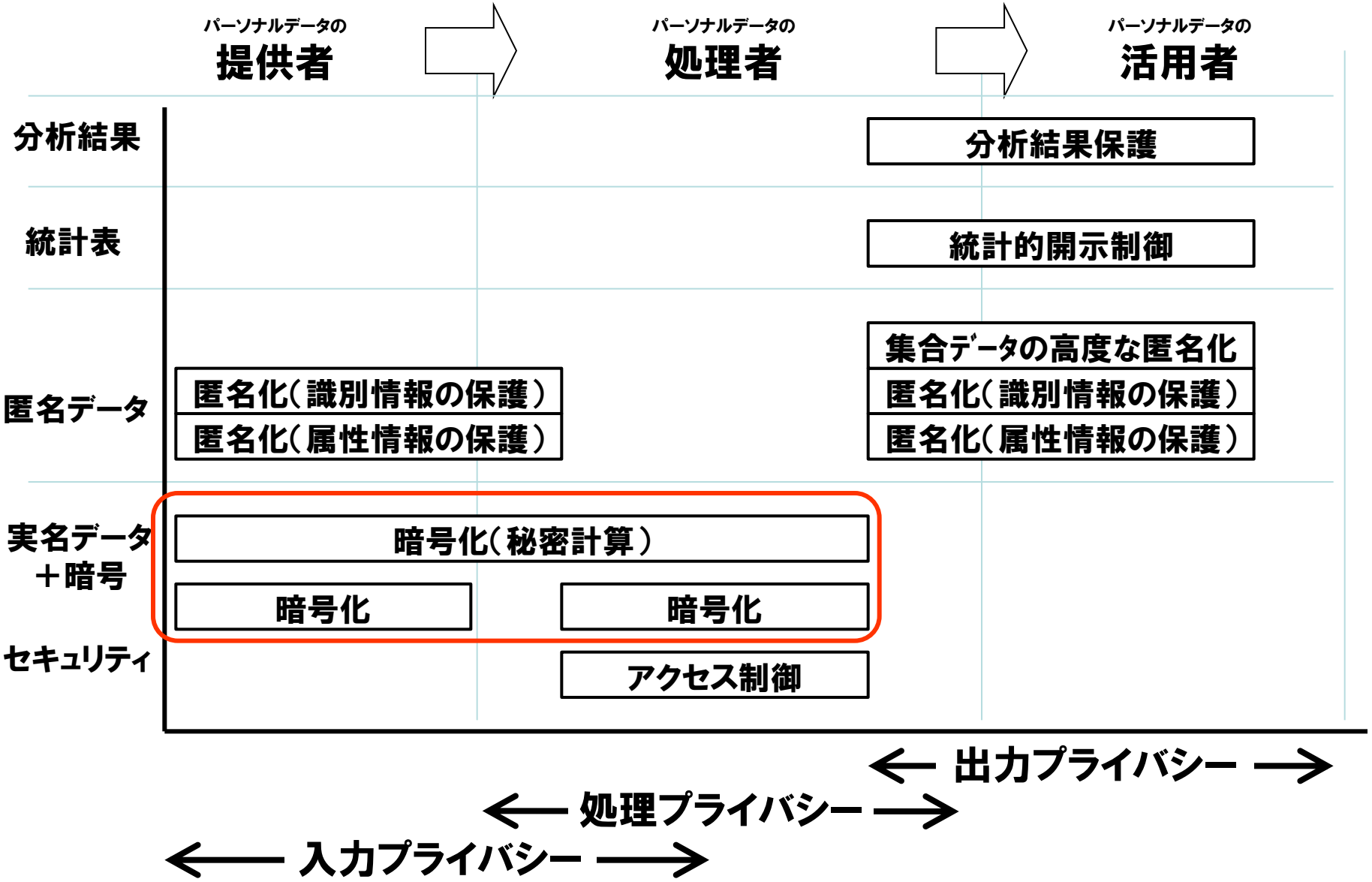


プライバシー制御の要素技術

統計表や分析結果のプライバシー

- 統計的開示制御
  - 【概要】統計表上のリスクがあるセルを削除するために、セルの数値が少ないもの、セルの値が一部少数の個人に依存するものなどを削除する
  - 【プロセスモデル】出力プライバシー
  - 【データ形式】集合データから統計表への変換
  - 【プライバシー原則】識別可能性の最小化、リンク可能性の最小化
  - 【成熟度】実用レベル
- 分析結果保護(差分プライバシー)
  - 【概要】分析結果のリスクを低減するために、分析結果にノイズを混入する
  - 【プロセスモデル】出力プライバシー
  - 【データ形式】分析結果
  - 【プライバシー原則】識別可能性の最小化、リンク可能性の最小化
  - 【成熟度】研究レベル

プライバシーを制御する要素技術の分類



プライバシー制御の要素技術

暗号化等

- 暗号化（秘密計算）
  - 【概要】意図しない閲覧や利用を制限するために、データを暗号化して入力し、暗号化の状態を保ったまま（統計）計算を行い、結果の暗号を出力する
  - 【プロセスモデル】入力プライバシーと処理プライバシー
  - 【データ形式】暗号
  - 【プライバシー原則】観測可能性の最小化、利用・保持・提供制限 、セキュリティ
  - 【成熟度】実証実験レベル（大規模化の研究が行われている）
- 暗号化
  - 【概要】権限（鍵）を持つもの以外に閲覧できなくするために、暗号で保護する
  - 【プロセスモデル】入力プライバシー、処理プライバシー
  - 【データ形式】暗号
  - 【プライバシー原則】観測可能性の最小化、セキュリティと正確性の確保
  - 【成熟度】実用レベル
- アクセス制御
  - 【概要】データ処理者において権限を持つもの以外に閲覧できなくする
  - 【プロセスモデル】処理プライバシー
  - 【データ形式】任意のパーソナルデータ
  - 【プライバシー原則】観測可能性の最小化
  - 【成熟度】実用レベル

まとめ

- パーソナルデータを利用するためのプライバシー保護の技術情報の共有を以下の観点から行った
  - 利用プロセスモデル、データ形式、守るべき原則、技術分類

- 匿名化処理の方法を決めるときに、現実にどのような危険があるかについても考えておく必要がある。統計情報の場合、住所、氏名が流出することはあり得ない。
  （中略）しかし、もし対象を特定するような試みが実際に行われたら、それはミクロデータ提供の危険性、ひいては統計調査の危険性を指摘するものとして利用されてしまうであろう。ところが、絶対的な匿名性を担保しようとすると、ドイツでの経験のように提供できる情報が極めて限られてしまう。したがって、この問題は匿名化処理だけで対策を考えるべきものではなく、そのような試みを行うこと自体を制限しておくことが必要となる。このため、データを提供するときには、利用目的を限定し、データの管理を適正に行わせることを義務付けておかなくてはならない。
  （匿名データの作成・提供の係るガイドライン 改正 平成23年3月28日 総務省政策統括官（統計基準担当）決定 別紙１匿名化処理の考え方）

参考文献

- 匿名化に関するもの
  - L. Sweeney. k-Anonymity: A Model for Protecting Privacy, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol.10, No.5 (2002), pp. 557-570
  - Charu C. Aggarwal and Philip S. Yu, Privacy-Preserving Data Mining: Models and Algorithms, Springer, 2008.
- 統計に関するもの
  - C. Skinner, Statistical Disclosure Control for Survey Data, in Sample Surveys: Design, Methods and Applications, Handbook of Statistics 29A, pp. 381-396, 2009.
- 暗号に関するもの
  - Oded Goldreich, Foundations of Cryptography, Cambridge University Press, 2000.
  - 医療統計処理における秘密計算技術を世界で初めて実証, NTT, http://www.ntt.co.jp/news2012/1202/120214a.html
- プライバシー原則
  - Ann Cavoukian, Privacy by Design, 2011.
- 解説
  - 佐久間淳 高橋克巳, クラウドを支えるデータストレージ技術：7. クラウドストレージにおける個人情報の利活用とプライバシー保護, 情報処理, Vol.52, No.6, 2011