2016.6.12 情報法制研究会

Innovative R&D by NTT

匿名化技術の概要

2016.6.12

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

本日の内容

話すこと

- 匿名化に関する技術情報の提供
  - 匿名化判断の“3要素”

話さないこと

- 匿名加工情報
  - 同情報の解釈に資する技術情報の提供ができることを目標に

匿名化の例

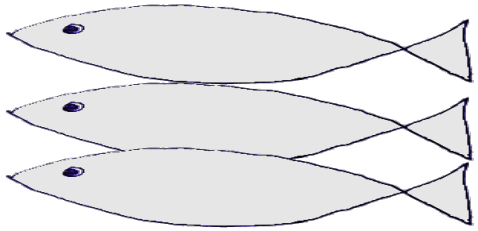
氏名	住所	年齢	性別	年収	体重
△山〇郎	東京都中央区銀座1	33	男	675万円	61.2 kg

氏名	住所	年齢	性別	年収	体重
削除	東京都中央区	30代	男	675万円	61.2 kg

- 一見、何の変哲もない匿名化の例
- しかし影響を正しく判断するには、匿名化の“3要素”の理解が必要
  - 3要素：どれを、どこまで、どうやって

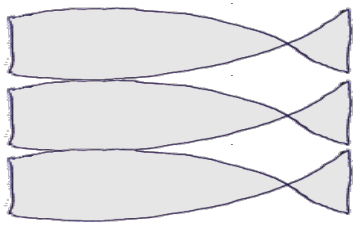
自動的に作ってくれませんか？

この程度なら
できる

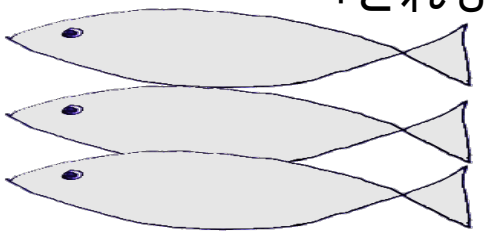


「名前を外す」

匿名化
ぼっくす

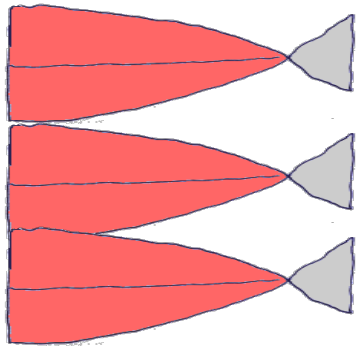


無理

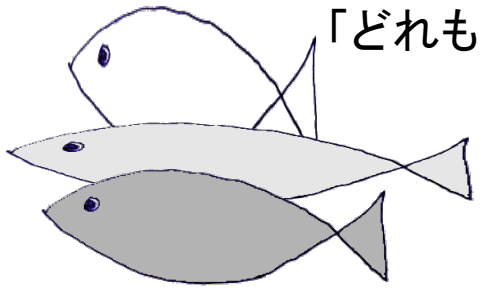


「どれも安全に加工、しかも食べやすく」

匿名化
ぼっくす

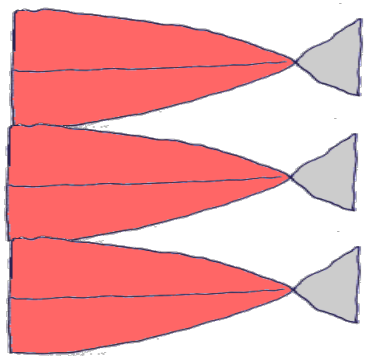


実態



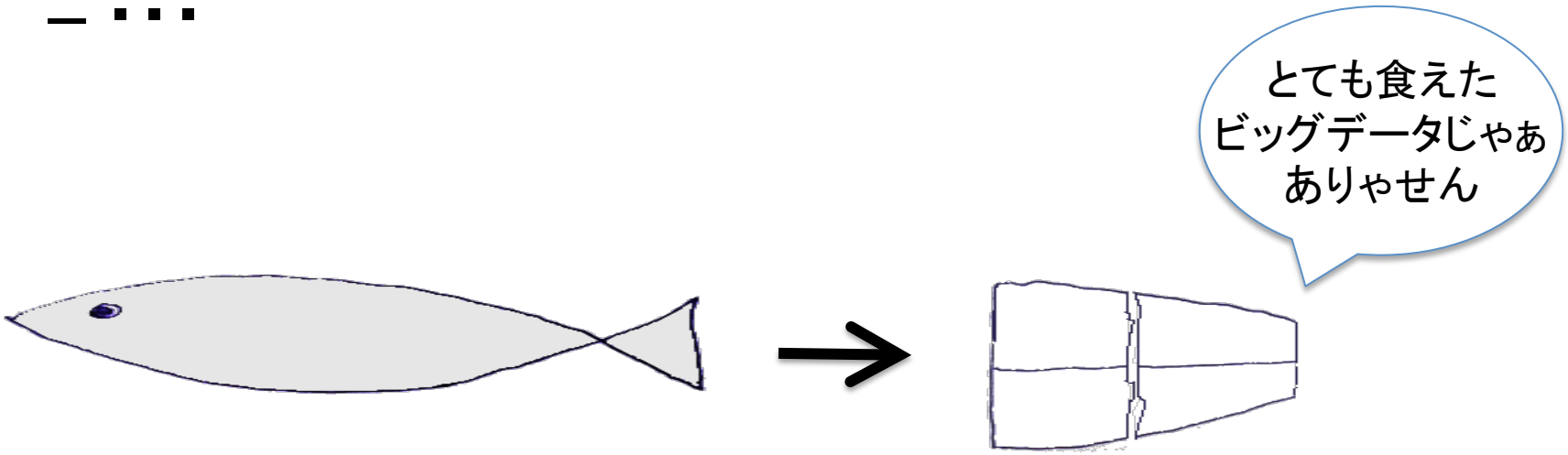
「どれも安全に加工、しかも食べやすく」

職人が
加工



汎用安全基準は ないものか？

- 目黒のさんま
  - 脂は体に悪かろう　→　蒸す
  - 小骨が喉にささっては　→　骨を全部抜く
  - …



匿名化判断の３要素

どれを、どこまで、どうやって

氏名	住所	年齢	性別	年収	体重
削除	東京都中央区	30代	男	675万円	61.2 kg

- どれを
  - 加工は住所と年齢だけでいいの？
- どこまで
  - 住所は市区町村レベルで十分なの？
- どうやって
  - 加工の具体的技法

細かい説明に入る前に

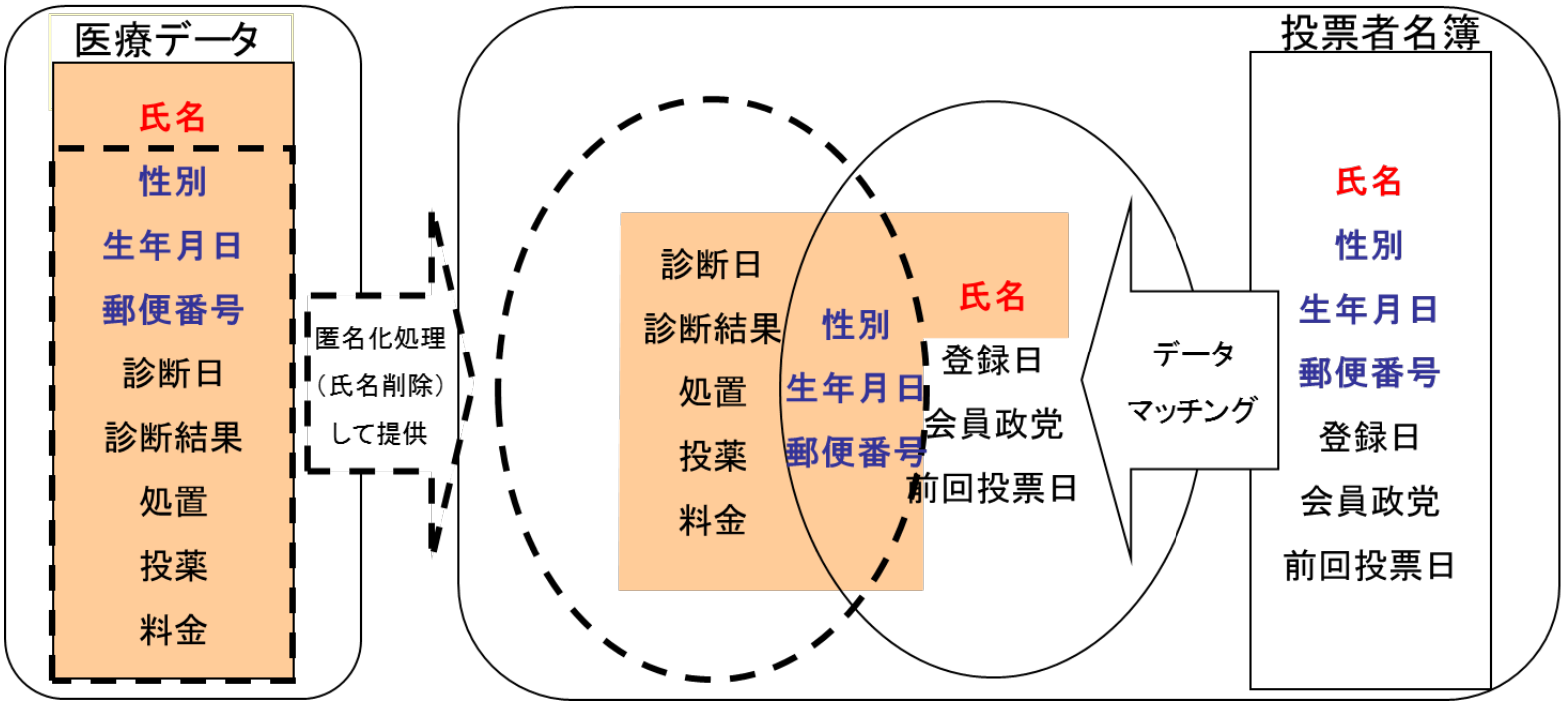
どんなことが起きると困るのか？

代表例

- 匿名化したつもりのデータから特定の個人が識別されてしまう（再識別事故）

再識別事故の例その１

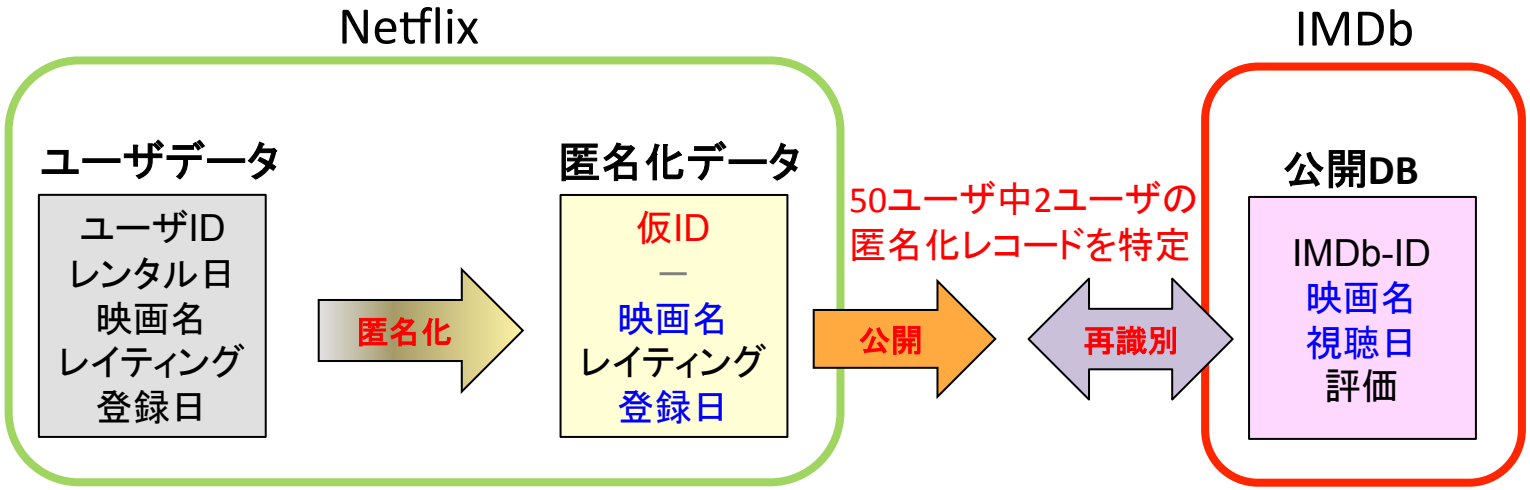
- マサチューセッツ州が公開した匿名化処理した医療データから州知事の情報が特定された
  - 医療データから氏名を削除して公開
  - 公開されている投票者名簿とマッチングしたところ、知事と同じ生年月日のレコードが6人、うち3人が男で、郵便番号から1人に特定可能



出典）L.Sweeney, k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5), 2002 を基に作成

再識別事故の例その２

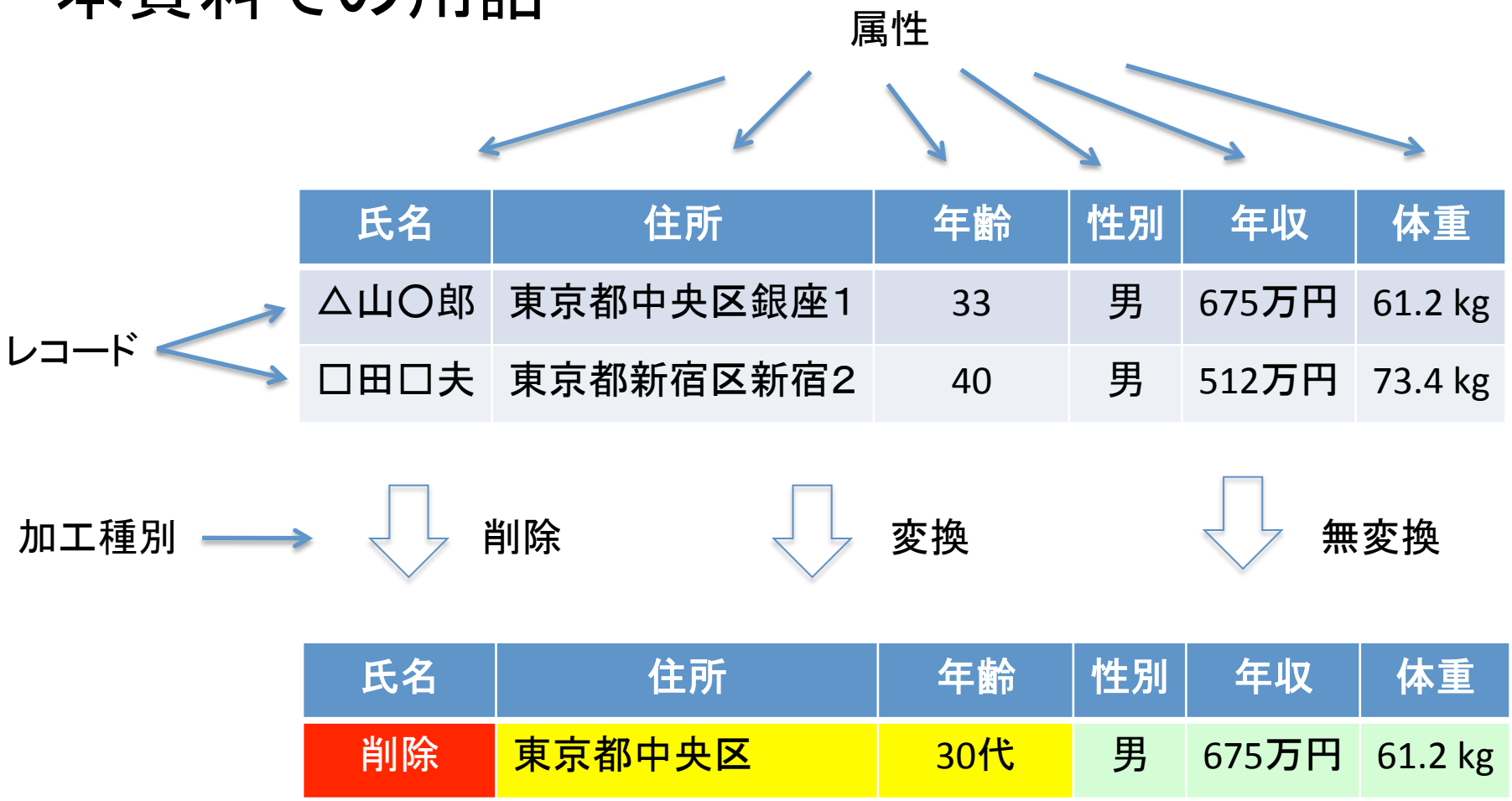
- 匿名化して公開されたDVD視聴履歴データから、ユーザが特定された
  - DVDレンタルのネットフリックス社（米国）が、映画推薦エンジンコンテストのため、匿名データを公開
  - 匿名化：約50万ユーザ、1億件分のレイティングデータから個人を識別できる情報を削除
  - 公開DBである Internet Movie Database（IMDb）の50ユーザの視聴映画と視聴日を使って再識別
  - 再識別：50ユーザ中2ユーザの公開された匿名化データが特定可能



出典）Arvind Narayanan and Vitaly Shmatikov, Robust De-anonymization of Large Sparse Datasets, 2008 IEEE Symposium on Security and Privacy を基に作成

細かい説明に入る前に

本資料での用語



要素１ どれを

- 匿名化対象の属性を「削除」「変換」「無変換」のどれかに仕分けすること

- 技術的観点からの仕分け
  - 個人が特定できる属性（識別属性）→ 削除
    - 「センシティブな属性（値）」かどうかという側面もある
  - 組み合わせで個人が特定できる属性（準識別属性）→ 変換
  - 個人の特定に関与しない属性（非識別属性）→ 無変換

要素１ どれを：識別属性

- 識別属性の技術的定義（と考えられるもの）
  - 特定個人の多量または多様な情報収集ができる
  - 本人密接性、一意性、共用性、不変性
    - 上記を全て満たすもの
- 例
  - 氏名
  - 運転免許証番号、パスポート番号
  - （スマートフォンの）MACアドレス
  - 指紋、顔認識データ

※ パーソナルデータ検討会技術検討WG報告書（2014.5）における「（仮称）準個人情報」の議論より

要素１ どれを：準識別属性

- 準識別属性の技術的定義
  - 組み合わせて個人特定が可能になる属性
  - 例えば、住所、年齢、生年月日の組み合わせ
  - 判断基準は確立していないと言わざるをえない
- 準識別属性と非識別属性の線引きは非常に困難
  - 「先週の日曜日に銀座に行った」は準識別属性か？
    - 従来は、（親しい間柄のみ知る事柄ゆえ）非識別扱い
    - ビッグデータ・IoTの時代では識別性が生ずる場合がある
      - 個人の行動を記録する手段が増えた
      - 個人の行動記録を入手する手段が増えた

要素１ どれを：履歴

- 履歴をどの期間（回数）使うかは、慎重に検討する必要がある（← そもそも識別性の有無に関わらず）
- 履歴が個人識別性を持つかどうかは、慎重に検討する必要がある
  - 種別：車の購入 vs. コンビニでの購入
  - 期間（回数）：
    - １回の購入（バスケット） vs. 10年分の買物歴
  - 精度：
    - どの日に何県に居た vs. cm単位秒単位の位置情報
  - 組み合わせ：
    - 音楽CDを購入した
    - サッカー観戦に行った
    - 法律の勉強会に行った
  - あらゆる属性・履歴に個人識別性が生じる可能性

要素２ どこまで

- 匿名化対象の属性の尺度と変換のレベルを決めること
  - 属性の値を尺度に応じて上位の値や概念に置き換えることが一般的
    - ３３歳→３０代、キュウリ→野菜
- 技術的性質
  - 尺度：対象によって異なる
  - レベル：属性ごとに固定的に定めるのは困難
    - 「年齢ならば１０歳刻みでよい」といった相場はあまりない
    - あったとしても、組み合わせると意味を持たない場合あり

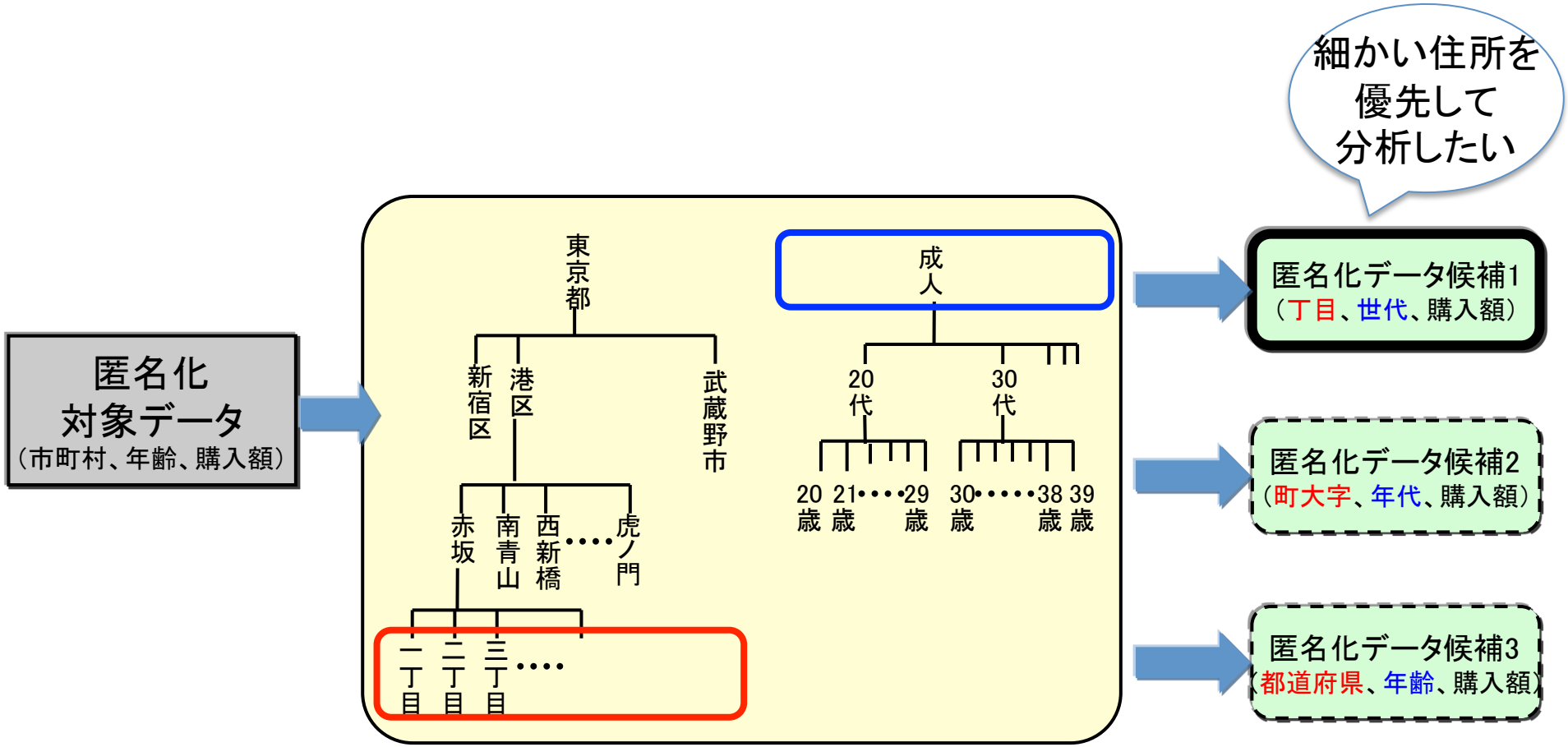
要素２ どこまで：属性尺度の決定（例）

- 年齢：
  - 数値尺度は比較的自明 → ５歳刻み＜１０歳刻み…
  - 意味的尺度（成年／未成年など）を入れると「知識」とその「共通化」が必要
- 位置：
  - 意味的尺度は比較的自明
    → 丁目番地号＜町字＜市区町村（行政区画）
  - 数値尺度は未確立ではないか
    - 100m四方、1km四方、10km四方の意味は地域（人口密度）によって異なる
- 商品
  - 意味的尺度を決める「知識」とその「共通化」が必要
    → 商品分類シソーラス（概念木）

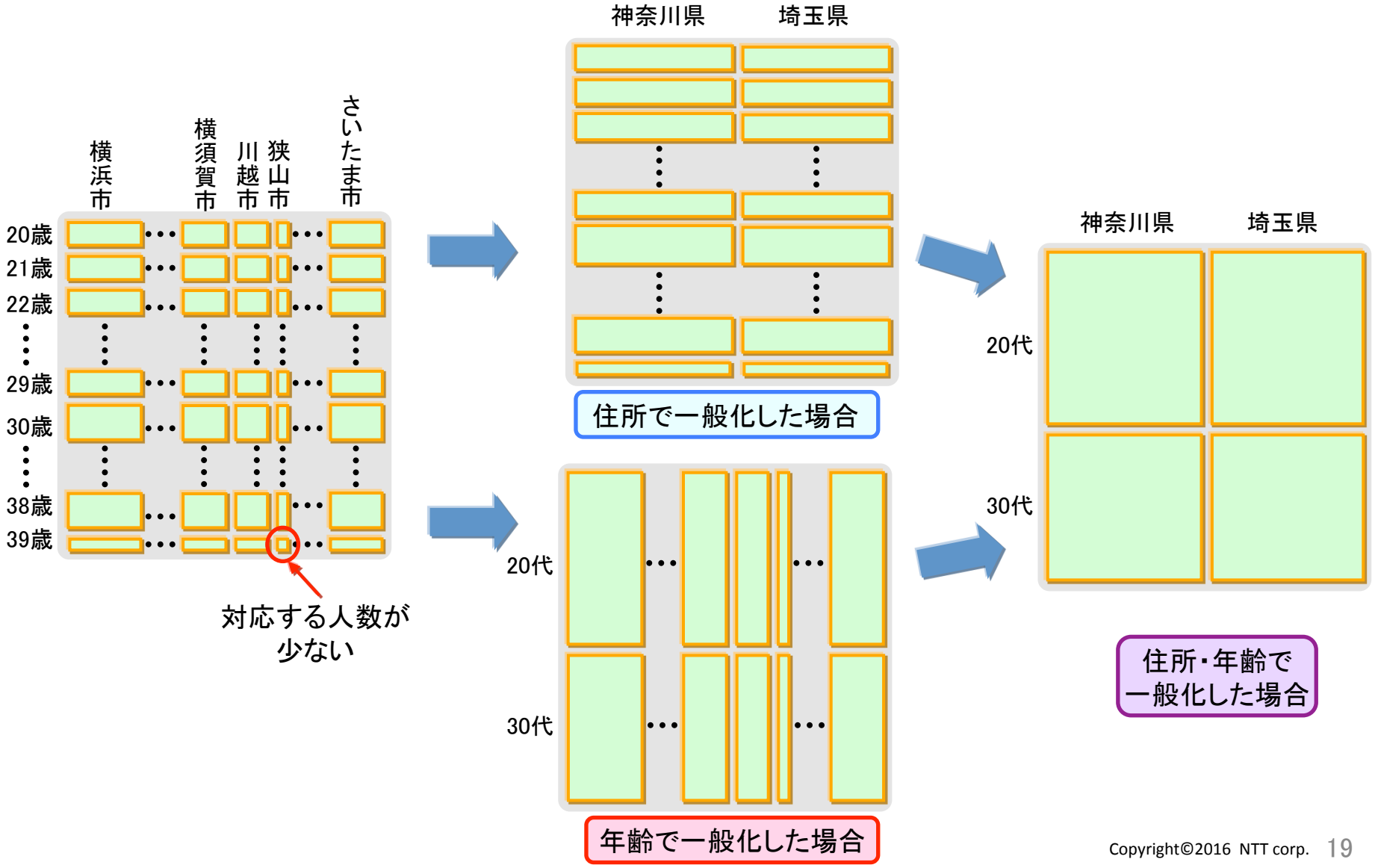
要素２ どこまで：
属性変換レベルの決定（例）

- 個別属性ごとに固定的に定めるのは困難
- その匿名化データの準識別属性の組み合わせた結果のレベル（解像度）で個人特定がないようになっているように、個別属性の変換レベルを決める
- 上記を実現するには？
  - 準識別情報の組み合わせの確定し、
  - どの程度のプライバシーを達成したいのか（どの程度のリスクを許容するのか）を意思決定し、
  - 以下の選択肢を実行する
    - 利用目的で決める（どの属性を優先的に分析したいか）
    - 数理的に最適なものを決める
  - 上記がいわゆる「k-匿名性」の概念

要素２ どこまで：
属性変換レベルの決定（利用目的で）



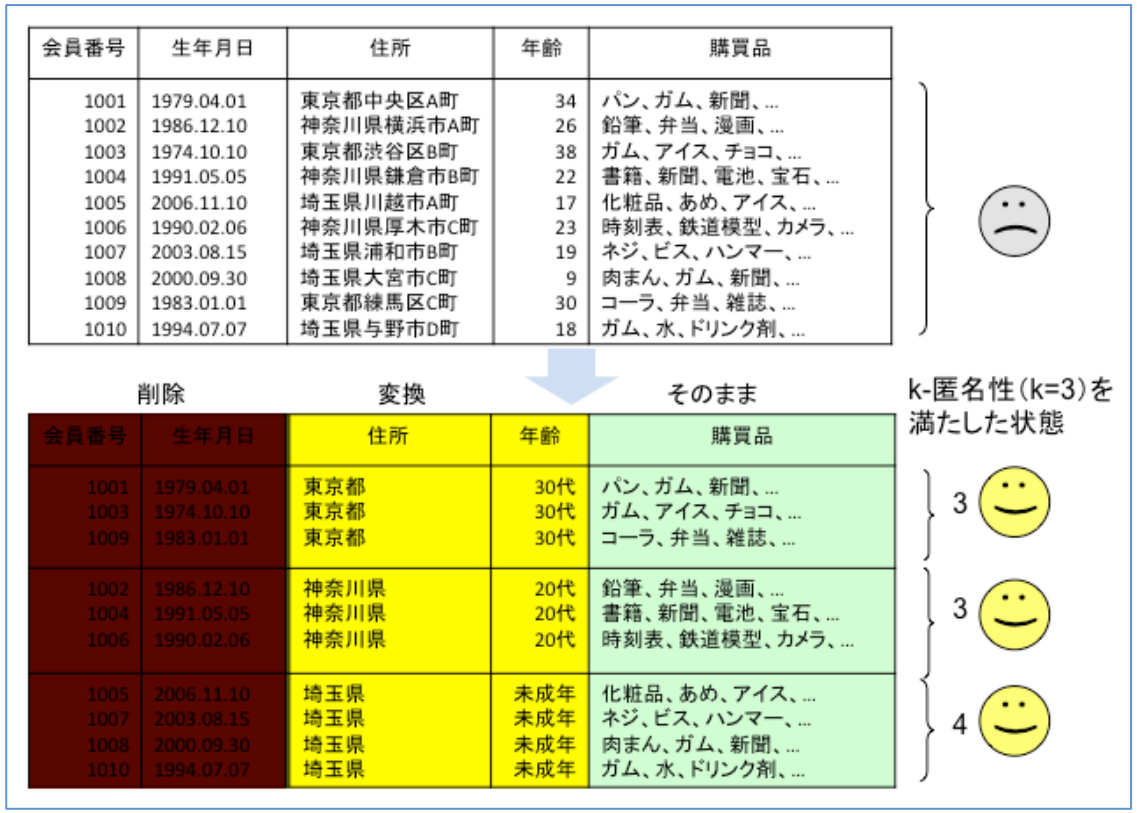
要素２ どこまで：
属性変換レベルの決定（数理的に）



（参考）要素２ どこまで： k-匿名性

同じ属性の組み合わせの人が「k人」以上いるようにする

- k-匿名性とは
  - 変換対象の属性を仕分けし（要素１）
  - 準識別属性に対して「個人特定が起きない」ということを「同じ準識別属性の組み合わせの人が必ず複数人いる」こととして担保する方法



会員番号	生年月日	住所	年齢	購買品
1001	1979.04.01	東京都中央区A町	34	パン、ガム、新聞、…
1002	1986.12.10	神奈川県横浜市A町	26	鉛筆、弁当、漫画、…
1003	1974.10.10	東京都渋谷区B町	38	ガム、アイス、チョコ、…
1004	1991.05.05	神奈川県鎌倉市B町	22	書籍、新聞、電池、宝石、…
1005	2006.11.10	埼玉県川越市A町	17	化粧品、あめ、アイス、…
1006	1990.02.06	神奈川県厚木市C町	23	時刻表、鉄道模型、カメラ、…
1007	2003.08.15	埼玉県浦和市B町	19	ネジ、ビス、ハンマー、…
1008	2000.09.30	埼玉県大宮市C町	9	肉まん、ガム、新聞、…
1009	1983.01.01	東京都練馬区C町	30	コーラ、弁当、雑誌、…
1010	1994.07.07	埼玉県与野市D町	18	ガム、水、ドリンク剤、…

削除　変換　そのまま

k-匿名性（k=3）を満たした状態

会員番号	生年月日	住所	年齢	購買品	
1001	1979.04.01	東京都	30代	パン、ガム、新聞、…	3
1003	1974.10.10	東京都	30代	ガム、アイス、チョコ、…	
1009	1983.01.01	東京都	30代	コーラ、弁当、雑誌、…	
1002	1986.12.10	神奈川県	20代	鉛筆、弁当、漫画、…	3
1004	1991.05.05	神奈川県	20代	書籍、新聞、電池、宝石、…	
1006	1990.02.06	神奈川県	20代	時刻表、鉄道模型、カメラ、…	
1005	2006.11.10	埼玉県	未成年	化粧品、あめ、アイス、…	4
1007	2003.08.15	埼玉県	未成年	ネジ、ビス、ハンマー、…	
1008	2000.09.30	埼玉県	未成年	肉まん、ガム、新聞、…	
1010	1994.07.07	埼玉県	未成年	ガム、水、ドリンク剤、…	

要素３ どうやって

- 「変換する」と仕分けされた匿名化対象の属性の値を、指定の尺度・レベルに従って（どれを・どこまで）変換すること

- 一般化と削除がよく用いられる技法の代表
  - 具体的技法の一覧（後述）
- 履歴データの扱い
  - 後述

要素３ どうやって：技法 (1/3)

- 属性情報の削除
  - 属性（列）削除
    - 直接個人を特定可能な属性（氏名等）を削除すること
  - 仮名化（仮ID）
    - 直接個人を特定可能な属性またはその組み合わせ（氏名・生年月日）を符号や番号等に置き換えること。例えば、ハッシュ関数
- 属性情報の一般化
  - 一般化
    - 属性の値を上位の値や概念に置き換えること。例えば、10歳刻み、キュウリ→野菜
    - データ全体に行うものをGlobal Recoding、局所的に行うものをLocal Recodingと呼ぶ
    - 四捨五入や二捨三入などを丸め法（Rounding）と呼ぶ
  - トップコーディング
    - 数値属性に対して、特に大きい、もしくは小さい属性値をまとめる。例えば、100歳以上の人は「100歳以上」とする

※ パーソナルデータ検討会技術検討WG報告書（2013.12）の議論より

要素３ どうやって：技法 (2/3)

- 属性情報の可能技法
  - ミクロアグリゲーション
    - 元データをグループ化した後、同じグループのレコードの各属性値を、グループの代表値に置き換えること
  - ノイズ（誤差）の付加
    - 数値属性に対して、一定の分布に従った乱数的なノイズを加えること
  - データ交換
    - カテゴリー属性に対して、レコード間で属性値を（確率的に）入れ替えること
  - 疑似データ作成
    - 元のデータと統計的に疑似させる人工的な合成データを作成すること

※ パーソナルデータ検討会技術検討WG報告書（2013.12）の議論より

要素３ どうやって：技法 (3/3)

- その他技法
  - レコード（行）削除
    - 特に大きい等、特殊な属性（値）を持つレコードを削除する。例えば、120歳以上のレコードは削除する
  - セル削除
    - センシティブな属性値等、分析に用いるべきでない属性値を削除する
  - ソート
    - レコードの並び順をランダムにする
  - リサンプリング
    - 元データ全体から一定の割合・個数でランダムに抽出

※ パーソナルデータ検討会技術検討WG報告書（2013.12）の議論より

要素３ どうやって：履歴データ

- 履歴データを扱う場合、その匿名化データに履歴の「どれ」を含めるかは、要素1を考える段階で考慮すべき問題である
- しかし履歴の「長さ（期間や回数）」はデータ処理の実務と技法に依存することが容易に想像される
  - 履歴を含んだ匿名化データを作成する場合、人単位で履歴データを束ねたレコードを作成する
  - 履歴データを束ねる期間が履歴データの長さを律する

まとめ

- データに対して匿名化判断の３要素（どれを、どこまで、どうやって）を検討したデータ処理結果が匿名化データである
- 作成した匿名化データは、下記を評価可能と考える
  - 何が含まれていて、
  - 何をどこまで変換したか（保護したか）、
  - 保護しない属性には何があるか
- 技術で担保できない部分がある場合は、それをよく理解した上で、残存するリスクの対応を制度と運用で行う必要がある

おわりに

- 匿名化技術に関して、その「できること」をできるだけ明らかにすることを目的に、従来と異なったアプローチでの説明を試みた
- 本資料で網羅できていない視点が多々あることをお許しいただきたい

- ありがとうございました

