NTT

個人識別できない匿名データは作成できるか？

2013.10.17

NTTセキュアプラットフォーム研究所
高橋克巳

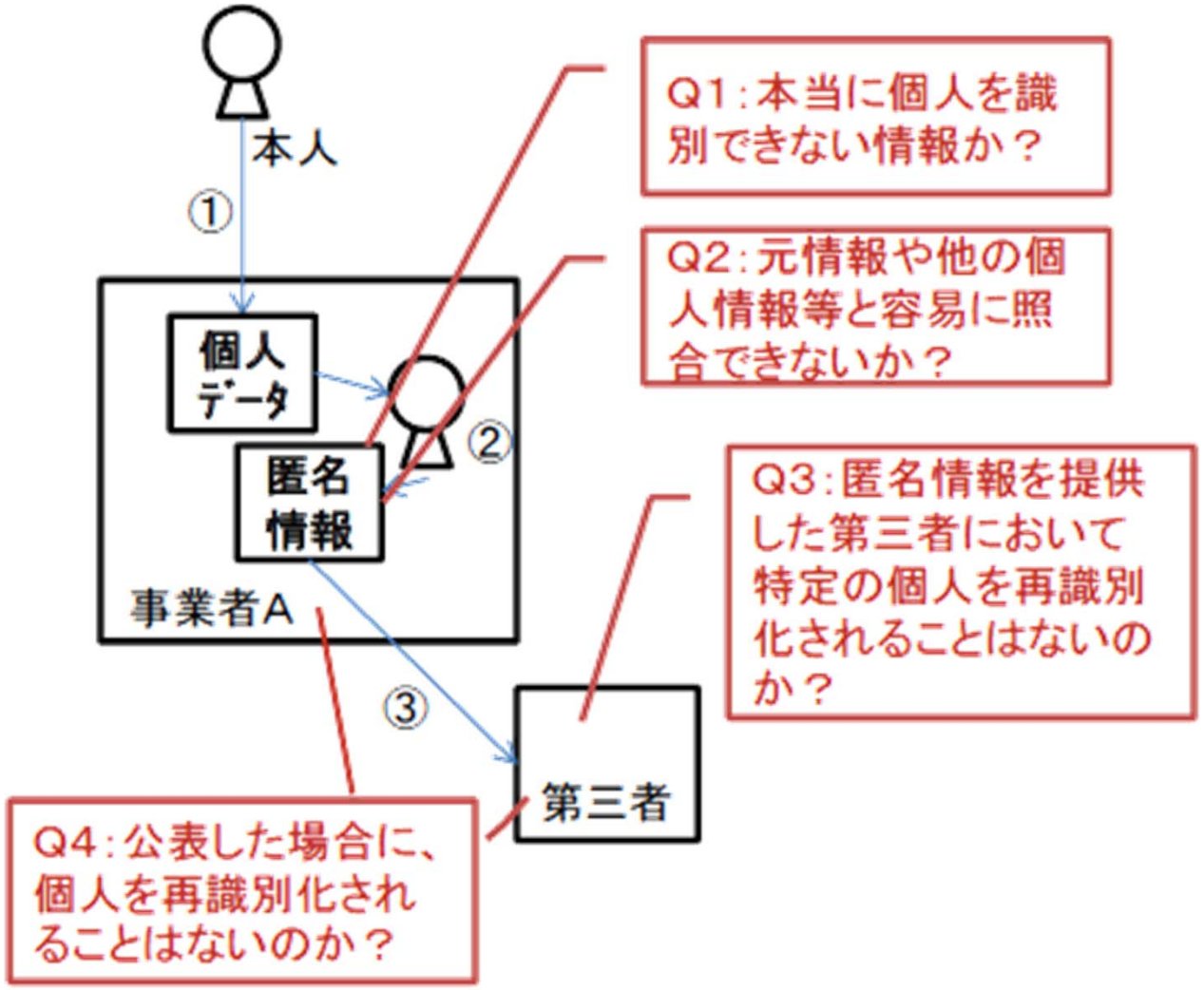
本資料のねらい

- 絶対的な匿名化を一律に定義することは困難*との認識に立ち、匿名情報の個人識別性が事実上低いと考えられる匿名化の技術的な例（十分条件）をボトムアップ的に考察してみる。

  ＊第１回技術検討ワーキングでの議論の方向性

- 以下の観点から分析を行った
  - 匿名情報の第三者提供の類型
  - データ分析の目的類型
  - その他

匿名情報の第三者提供の類型及び課題



第１回技術検討ワーキング【資料３】より

匿名情報の第三者提供の類型とは、技術的にどのような問題設定になるか

- Q1 本当に個人を識別できない情報か？
  - 「匿名情報」に対しそれ単体で個人識別を試みる問題
- Q2 元情報や他の個人情報等と容易に照合できないか？
  - 「匿名情報」に対し「元情報」や「他の個人情報」を用いて個人識別を試みる問題
- Q3 匿名情報を提供した第三者において特定の個人を再識別化されることはないのか？
  - 「匿名情報」に対し「ある限定された情報」を用いて個人識別を試みる問題
- Q4 公表した場合に、個人を再識別化されることはないのか？
  - 「匿名情報」に対し「全ての入手可能な情報」を用いて個人識別を試みる問題

Q1「匿名情報」に対しそれ単体で個人識別を試みる

- 十分条件の例
  - ① 名前等の直接識別できる情報を削除
    - 社会通念上識別が容易にできるとされる基本4情報等
- 留意点
  - 「直接識別できる情報」以外からの個人識別の可能性は否定できないが、これを考慮に入れると、Q3またはQ4になる
    - (例)「位置情報が 34.72, 135.36 で野球をしていた人」を識別するためには、なんらかの情報(知識)が必要

氏名	生年月日	位置情報	行動
A木イチロー	1973.10.23	34.72, 135.36	野球
B浦カズヨシ	1967.02.27	35.90, 139.71	サッカー

[redacted]

位置情報	行動
34.72, 135.36	野球
35.90, 139.71	サッカー

Q2「元情報」や「他の個人情報」を用いて個人識別を試みる

- 十分条件の例
  - ① 直接識別情報を削除して、かつ残り全ての属性が非可逆に変換されている
    - 一般化（k匿名化）、かく乱
  - ② 直接識別情報を削除して、かつ残り全ての属性の組み合わせが希少になるレコードが全て削除されている
    - （一意にならないレコードだけを残す）
  - ③ 直接識別情報（元情報）が完全に捨てられていることが確認できる
    - 確認できるのであれば、Q1と同等の問題に帰着
- 留意点
  - ①②は行った操作に対応した安全性を持つ
    - 例えば、k匿名化の場合は「k人未満に絞り込まれない」

※他の個人情報を用いた照合を試みる場合も基本的には元情報の場合と同様

氏名	生年月日	位置情報	行動
A木イチロー	1973.10.23	34.72, 135.36	野球
B浦カズヨシ	1967.02.27	35.90, 139.71	サッカー

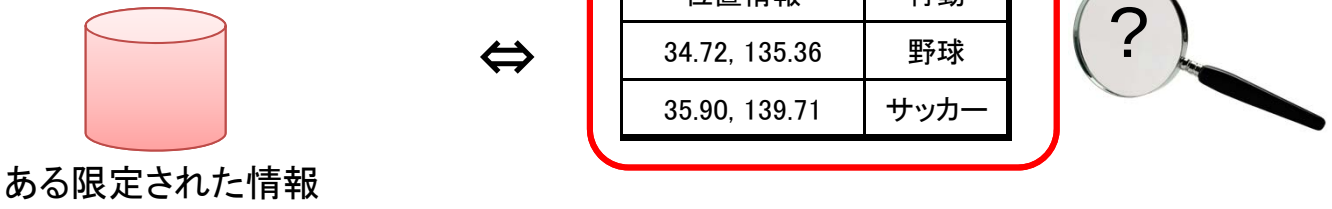
元情報

⇔

位置情報	行動
34.72, 135.36	野球
35.90, 139.71	サッカー

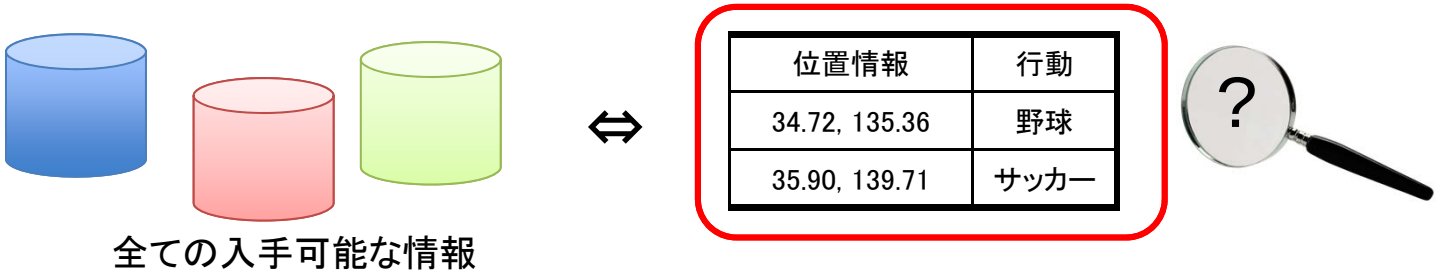
Q3「ある限定された情報」を用いて個人識別を試みる

- 十分条件の例
  - ① 直接識別情報を削除して、かつ全ての属性が「k匿名化」されている
    - 希少なレコードの全削除によるk匿名化を含む
  - ② 直接識別情報を削除して、かつ「提供された匿名情報に含まれる属性を用いない」ことが確認できる
    - 確認できるのであれば、Q1やQ2③と同等の問題に帰着
- 留意点
  - ①は行った操作に対応した安全性を持つ
    - 例えば、k匿名化の場合は「k人未満に絞り込まれない」
  - ②はこのような問題設定に意味があるのかの議論の必要がある
    - 第三者に匿名情報が提供されたときに、第三者が用いる情報や行う行為に制限を課す技術的手段は存在しない
      - 高度な暗号プロトコル等で技術的手段が構成できる可能性はある（研究課題）



Q4「全ての入手可能な情報」を用いて個人識別を試みる

- 十分条件の例
  - ① 直接識別情報を削除して、かつ全ての属性が「k匿名化」されている
    - 希少なレコードの全削除によるk匿名化を含む
- 留意点
  - ①は行った操作に対応した安全性を持つ
    - k匿名化の場合は「k人未満に絞り込まれない」



ここまでの議論のまとめ

- 匿名情報を第三者提供する場合に有効な手だては、識別情報以外の全ての属性を十分に加工して保護（k匿名化）に限られる（下の図）
  - データの有用性には制限がある
- 一方、「万能」の目的に使えるデータはQ1-①に準ずる形式（上の図）と考えられる
  - ビッグデータ（属性数が多い）場合、Q3・Q4の状況で匿名情報と呼ぶことは事実上困難
- 分析目的を類型化して、その目的を達成するための中間処理的なデータで、事実上個人識別性が低いものが定義できれば好ましいのではないか？

会員番号、生年月日、住所、年齢、購買品1、購買品2、購買品3、．．．．．

削除　　加工（保護）　　そのまま（非保護）

会員番号、生年月日、住所、年齢、購買品1、購買品2、購買品3、．．．．．

削除　　加工（k匿名化して保護）

データ分析の目的の例

- バスケット分析
  - 目立つ属性間の関係を数え上げてルールを作成すること
    - 全体に対する評価はしなくてもよい
  - →希少なレコードの削除(“ロングデールのしっぽ切り”)がそれなりに有効
    - 属性の種類が多いと、ほとんどのレコードが削除されてしまうようなことも
- 数え上げ
  - 代表的な属性や属性間の関係の数え上げ(クロス集計)をすること

  →希少なレコード削除の有効性は？

  →属性を選択したうえ(全属性は使わない)希少な組み合わせを慎重に排除したデータの作成
    - k匿名化や統計が行っていることの基本的思想

  ※「任意の組み合わせの数え上げ」ができる匿名データの作成は実質上困難ではなかろうか

データ分析の目的の例（続き）

- 数え上げ（続）
  - →「長い」履歴情報から、「短い」関係情報の取り出して分析に利用の有効性は？
    - 例：{ a, b, c, d, e} という履歴から、ユーザ属性を排除して、{a, b} {b, c} {c, d} {d, e} という二項関係のみを使う
- データの分布の調査・検定
  - データ全体の分布やデータ全体から事実を導きだす

※データ全体の傾向を保ち、かつ意味のある粒度の匿名情報を作成することは困難

- →疑似データの有用性の検討

その他整理が必要と考えられる事項

- 個人識別とは？
  - 匿名情報が実在する誰と結びつくのかわかること
- 個人識別とデータ一意（k=1）の関係
  - k=1は必ずしも個人識別を意味しない
- 属性をそのまま非保護で開示してよい条件
  - 基本的には加工保護していない属性は将来的に組み合わされて識別情報となる可能性がある
  - 属性種による違いはあるか？
- 個人識別が起きなくても生じる問題の分析
  - 特定の属性の値の組み合わせの人に対して、知られたくない属性の値が確定的に割り当てられる（属性推定／「不都合な事実」の公表）