ソーシャルICTサービスとプライバシー保護

東京大学ソーシャルICT研究センター　山口利恵

東京大学大学院情報理工学系研究科
ソーシャルICT研究センター

今日の目次

* プライバシー保護の必要性
* 匿名化データとは
* 安全な基準と実現
* サービス提供者とユーザとの情報共有

プライバシー保護

東京大学大学院情報理工学系研究科
ソーシャルICT研究センター

個人情報保護法

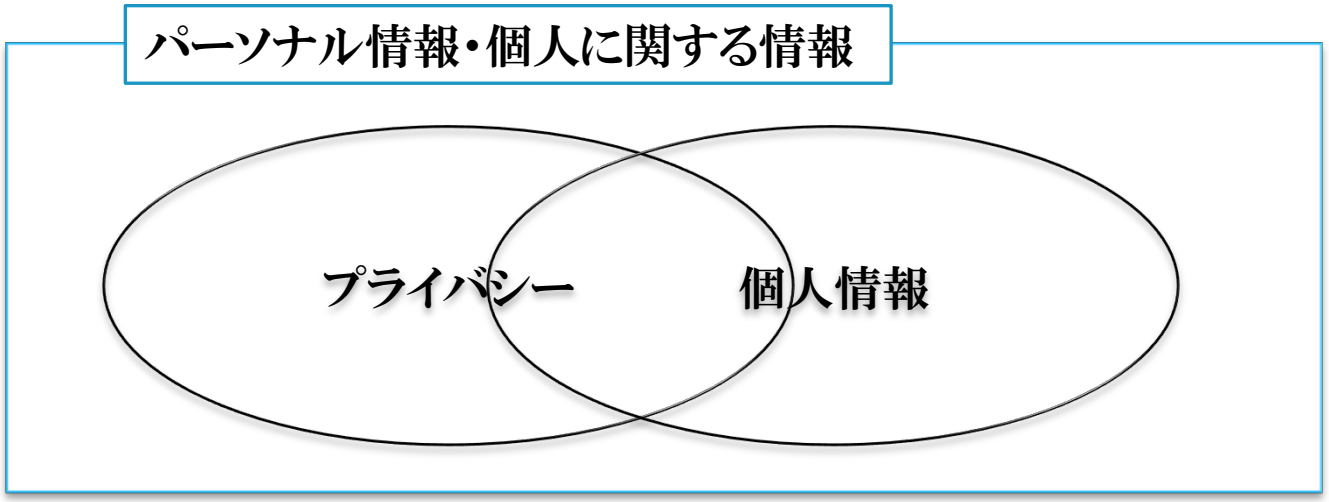
* 個人情報の保護に関する法律
  * 第2条第1項
  * この法律において「個人情報」とは、生存する個人に関する情報であって、当該情報に含まれる氏名、生年月日その他の記述等により特定の個人を識別することができるもの（他の情報と容易に照合することができ、それにより特定の個人を識別することができることとなるものを含む。）をいう。

* 保護法における個人情報が本当に個人情報か？
  * 個人に関する情報（パーソナル情報）のうち、どこまでが「個人情報」なのか
    * 本当に氏名と生年月日が個人情報？
      * もし、仮に名刺情報が漏洩したとしても実害があるのか
      * 個人に紐づく秘匿情報のほうが問題では？

プライバシーとは？

* 名前がないデータであれば匿名データですか？（再掲）
  * 情報そのものの特性だけでなく、データ構造に依存
    * 氏名等、特定の個人を識別できる情報
    * 上記情報と紐付いている情報
    * 上記情報と紐付いていない情報

* どういう情報であれば、プライバシーに関わらない情報？



日本と欧州のプライバシー意識の違い

* IPAセキュリティセンターは、2010年に「eIDに関するリスクの認知と受容の調査」を行い、日本と欧州における個人情報の感覚の違いを調査
* 「オンラインで提供できる情報」を聞いたところ、日欧の違いが読み取れる

「オンラインで提供できる情報」に関する意識

	欧州	日本
名前/姓	86%	37%
年齢	90%	75%
国籍	87%	80%
身分証明書番号(保険証、パスポートなど)	13%	7%
住所	65%	19%
容姿(身長、体重など)	39%	36%
趣味やよくしていること	53%	75%
志向/意見	75%	69%
友人やよく会う人達、同好会の人達など	37%	22%
私が通常行く場所	27%	31%
mixi やフェイスブックのような SNS に提供している情報	50%	41%
自分の写真	58%	7%
財務情報(収益、残高など)	9%	7%
医学情報(健康保険番号など)	7%	4%
銀行情報(銀行カード番号、アカウント番号など)	30%	4%
裁判の情報(前科情報・破産宣告の有無など)	5%	6%
バイオメトリックス情報(指紋、虹彩など)	4%	5%

プライバシー侵害事例:ミログAppLog

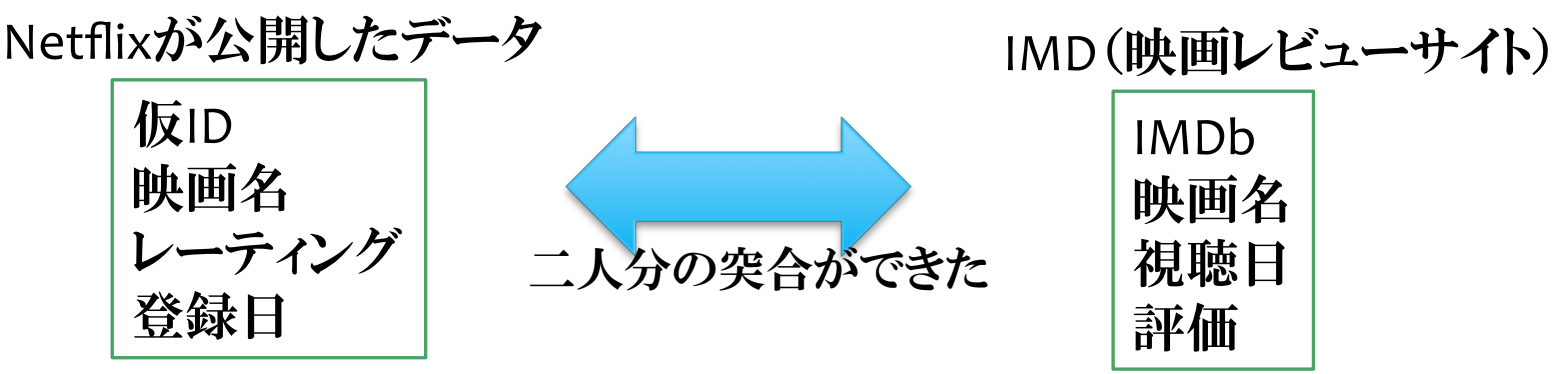
* 2011年7月、日本のベンチャー企業であるミログ社は、Android（アンドロイド）端末にインストールされたアプリケーションのリストや起動履歴を収集、活用する事業を展開していた。
  * 具体的には、ユーザーのアプリケーション情報を基にしたターゲティング広告やリワード広告、統計処理したアプリケーション情報を使ったコンサルティング事業などを手掛けていた。

* だが2011年秋頃から、こうしたアプリ情報の収集が「プライバシーの侵害に当たるのでは」という指摘が相次いでいた。アプリケーション起動履歴などを収集する「app.tv」「AppLog」といった同社が提供するサービスについて、「ユーザーへの十分な説明なく情報を収集している」とし てネットを中心に批判が噴出した。

* ミログは一部のサービスを終了・停止すると共に、内容の全面的な見直しを検討したが今回、「事業環境を総合的に判断した結果」（ミログ）として2012年4月、会社の解散、清算を決定したという。

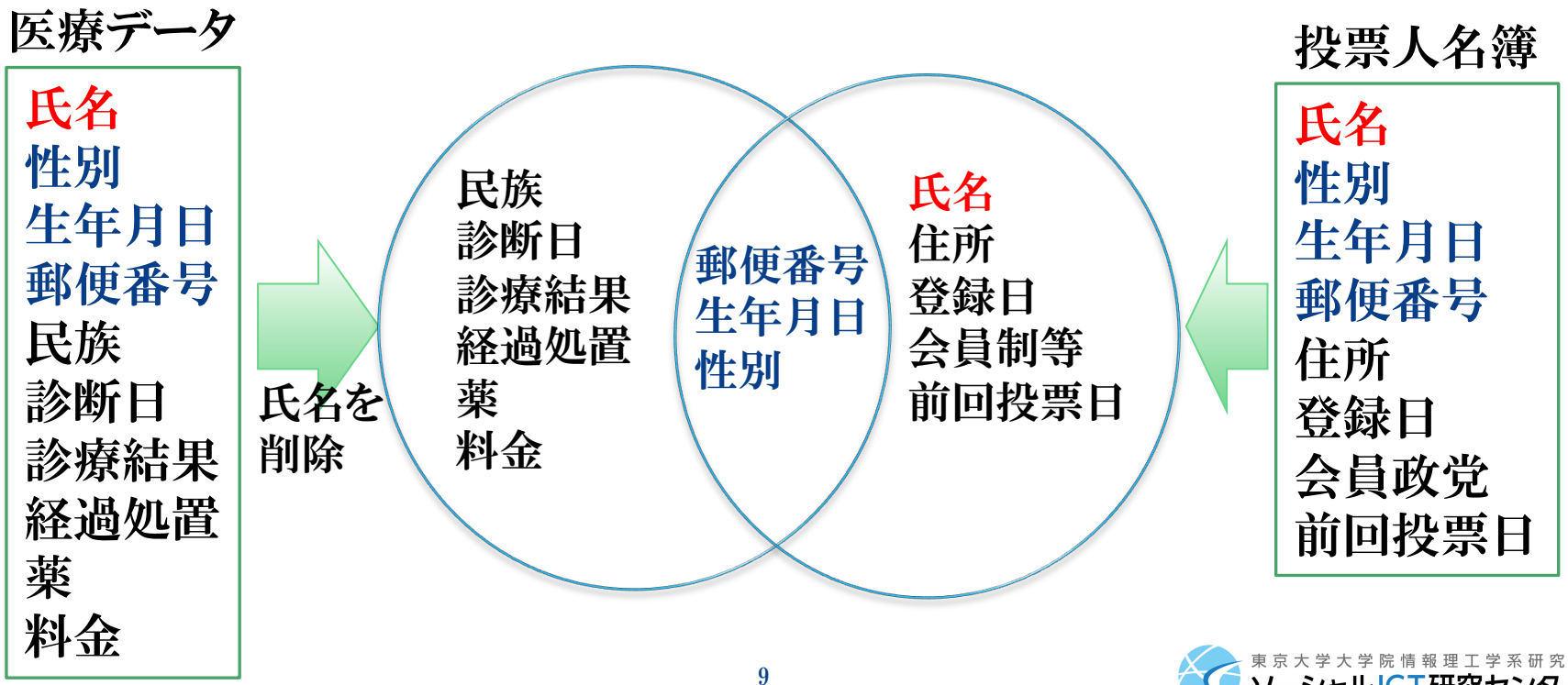
プライバシー侵害事例：Netflix社のDVDレンタル履歴

* 2006年、米国の大手DVDレンタル会社であるNetflix社は、匿名化されたDVDレンタル履歴を公開し、リコメンデーションのためのアルゴリズムを競わせるコンテスト(Netflix Prize)を行った。
  * 約50万ユーザ、1億件分のデータから個人を識別できる情報を削除
* NarayananとShmaikovは、これらの公開データと the Internet Movie Database(映画のレビューサイト)のデータを突き合わせることで、二人の個人が識別できたと発表した。
* このような動きを受け、 Netflixは米国連邦取引委員会(FTC)の調査や法律家による訴訟を受けることになり、計画されていた Netflix Prizeの続編は中止に追い込まれた。



プライバシー侵害事例：マサチューセッツ医療データ

- 2002年Sweeneyの研究
- 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を特定
  - 医療データから氏名を削除して公開
  - 既に公開・販売されている投票者名簿とを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

匿名化データとは
どういうデータなのか？

みなさんのイメージは？

* 匿名データって？　どんなデータですか？
  * 2chのデータ？
    * 最近名前が漏洩したとかありました
  * 名前のないデータであれば全て匿名データですか？
    * 統計情報（例：集計データ、世論調査）
    * おおざっぱに処理されたデータ？
    * 個人情報と連結可能な匿名データ？

匿名性と有用性はトレードオフ

* 匿名性と有用性はそれぞれ評価が可能で、定性的に互いにトレードオフの関係にある。
  * 有用な情報であればあるほど、匿名性が犯されている危険性が高いと一般的にいわれている。
* 有用な情報をどのように扱っていくかについて考えなければならない
  * 匿名性だけを考えるべきではない

匿名化のためのモデル（登場人物）

- ユーザは情報を提供する
- 情報を保有している人はセキュリティを確保する
- 利用する人はプライバシーを保った状態で使いたい

※表の例は位置情報の渋滞情報での活用

同意の取り方

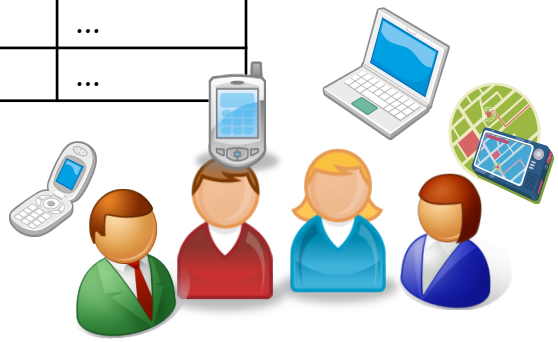
セキュリティの確保、匿名化

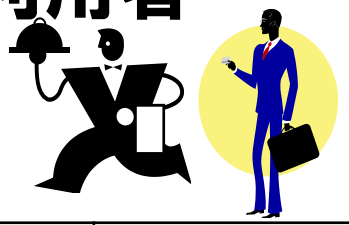
ユーザ → 取得 → パーソナルデータの **取得保有者** → 提供 → パーソナルデータの **利用者**

生のセンサー等履歴データ

匿名化データ

位置情報	GPS
時間情報	年月日時間
…	…
…	…





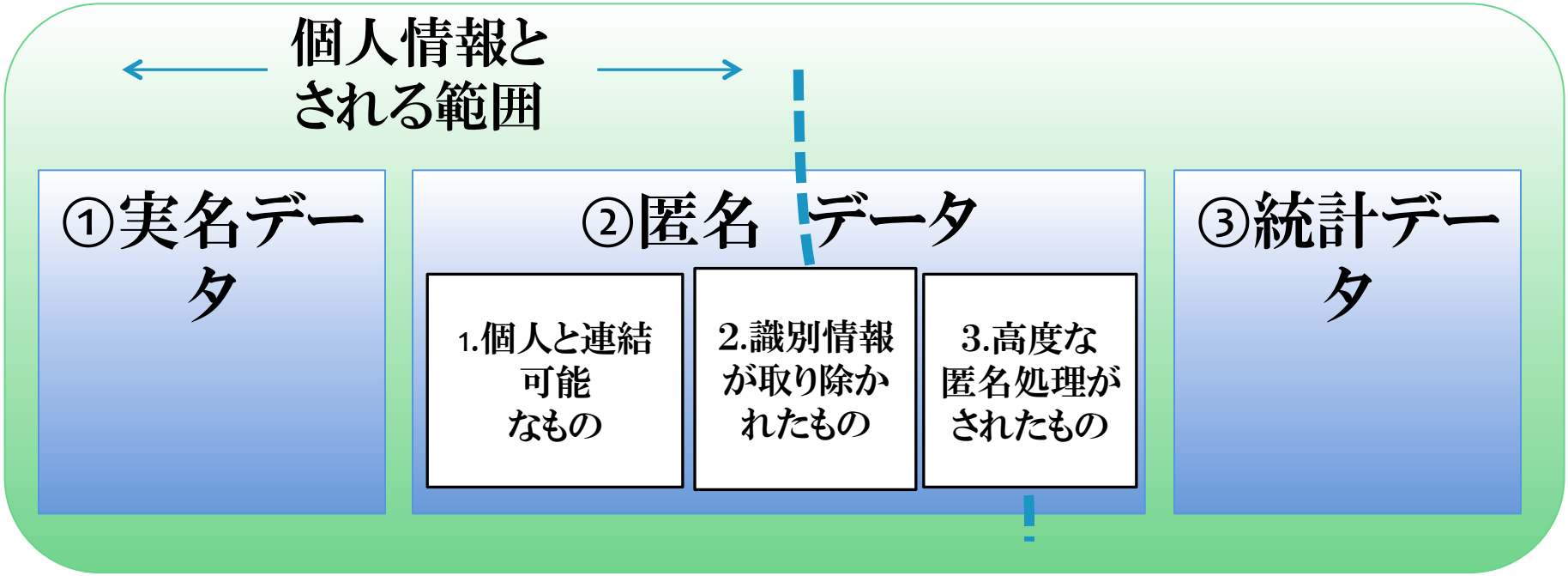
位置情報	＊＊道上り
時間情報	年月日時間は10分単位
…	…
匿名性	＊＊ランクをクリア

東京大学大学院情報理工学系研究科
ソーシャルICT研究センター

「匿名化」と個人情報

* パーソナルデータがどのような処理によって個人情報を削除できるのか
* いわゆる「匿名データ」は技術的に3種類に分類できる

※「実名データ」「匿名データ」「統計データ」という名称は技術用語ではない



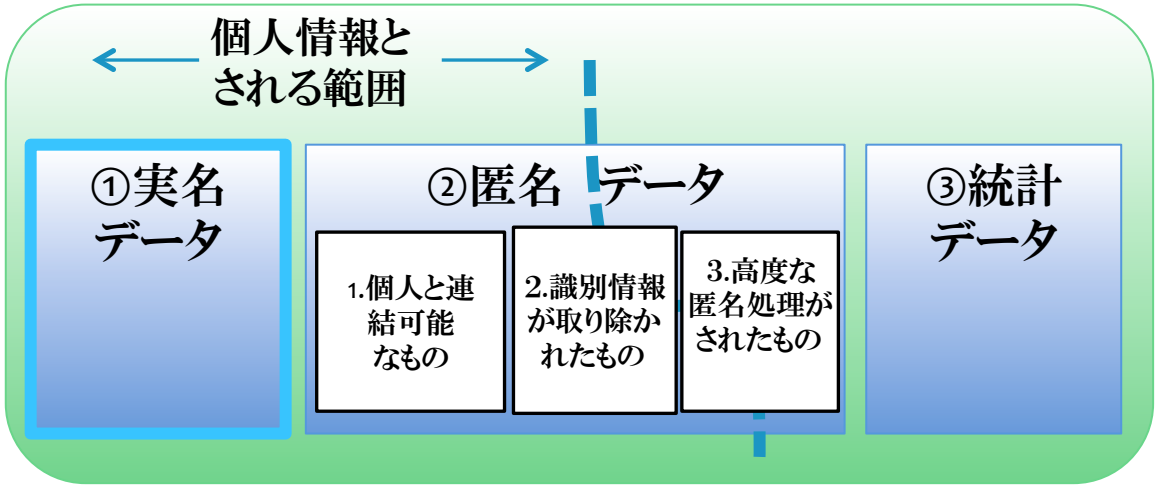
出展：NTT高橋克巳作成の資料

①実名データ(個人情報)

* 氏名等により個人を識別できる情報

①実名データ

氏名	生年月日	位置情報	行動
鈴木よしえ	1978.9.10	34.72, 135.36	野球
松井りんご	1942.10. 7	35.90, 139.71	サッカー



②-1　連結可能匿名データ

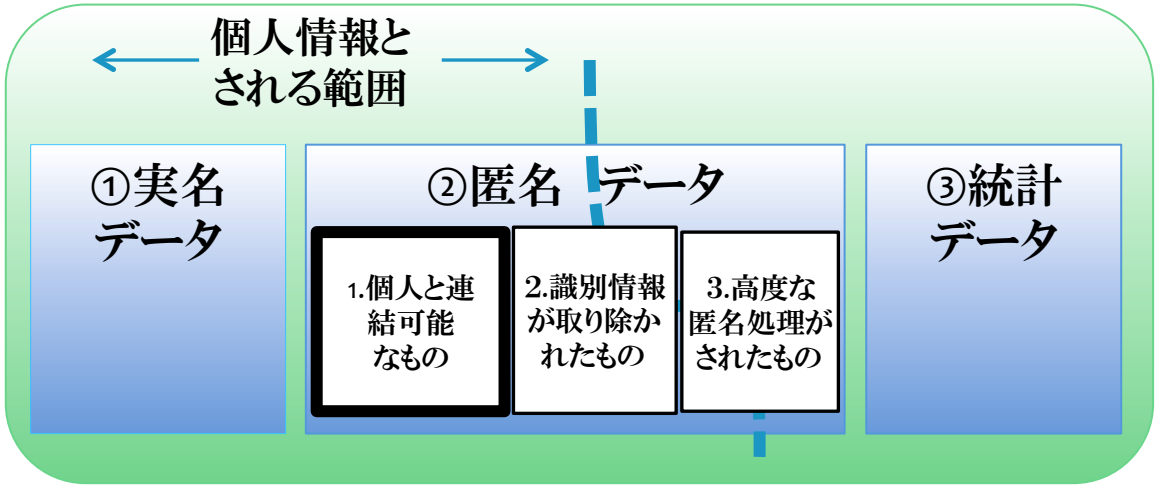
* 他の情報と容易に統合が可能。場合によっては、仮名データとして活用。

①実名データ

氏名	生年月日	位置情報	行動
鈴木よしえ	1978.9.10	34.72, 135.36	野球
松井りんご	1942.10. 7	35.90, 139.71	サッカー

②-1 個人と連結可能な匿名データ

氏名	生年月日		位置情報	趣味
鈴木よしえ	1978.9.10	↔	34.72, 135.36	野球
松井りんご	1942.10. 7	↔	35.90, 139.71	サッカー



②-2 いわゆる匿名データ

* 他の情報と容易に統合が可能。場合によっては、仮名データとして活用。

①実名データ

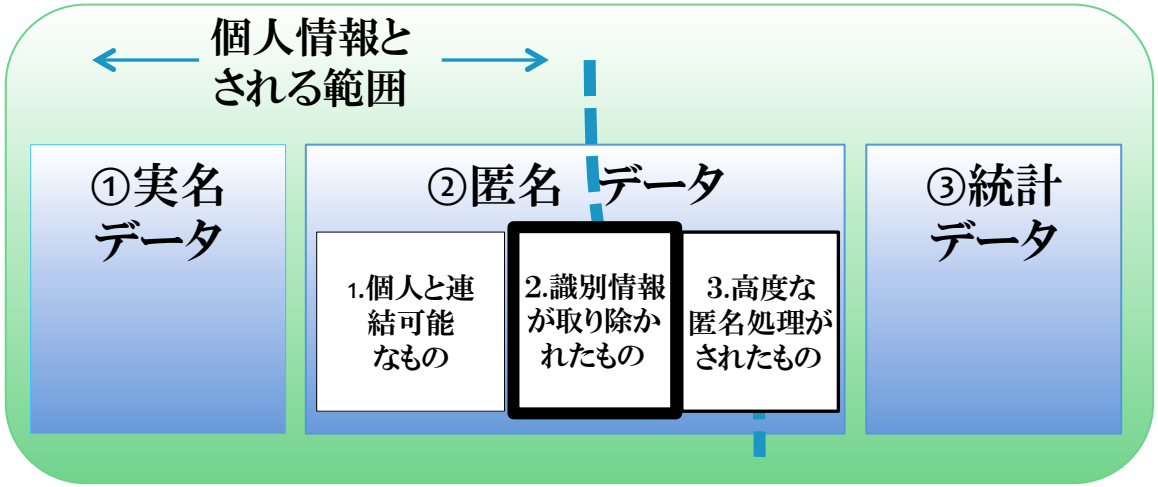
氏名	生年月日	位置情報	行動
鈴木よしえ	1978.9.10	34.72, 135.36	野球
松井りんご	1942.10. 7	35.90, 139.71	サッカー

②-1 個人と連結可能な匿名データ

氏名	生年月日		位置情報	趣味
鈴木よしえ	1978.9.10	↔	34.72, 135.36	野球
松井りんご	1942.10. 7	↔	35.90, 139.71	サッカー

②-2 いわゆる匿名データ

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



②-3　高度な匿名データ

- 特定の個人が識別できないレベルまで匿名化したデータ
- 作成には専門知識が必要

①実名データ

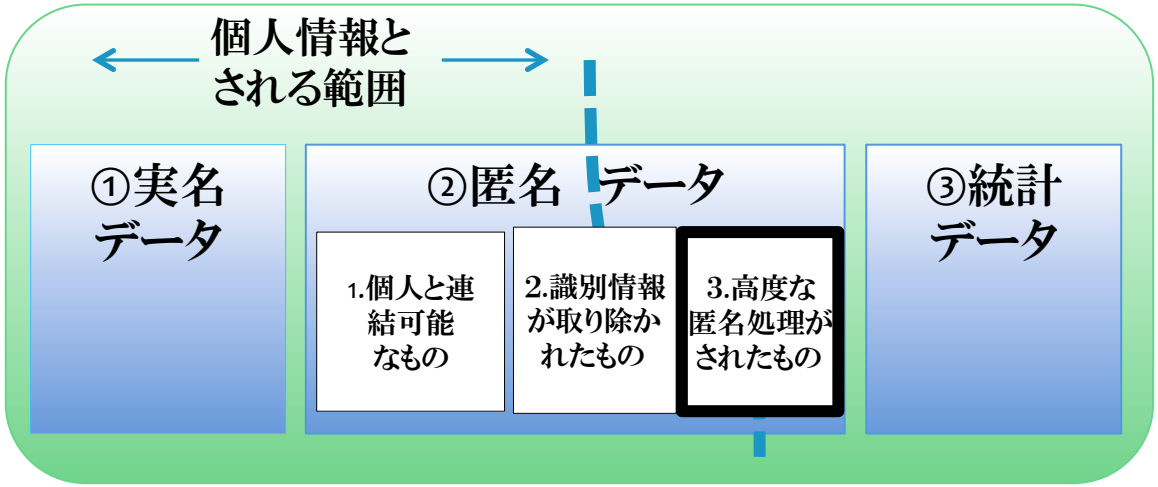
氏名	生年月日	位置情報	行動
鈴木よしえ	1978.9.10	34.72, 135.36	野球
松井りんご	1942.10. 7	35.90, 139.71	サッカー

②-1 個人と連結可能な匿名データ

氏名	生年月日		位置情報	趣味
鈴木よしえ	1978.9.10	↔	34.72, 135.36	野球
松井りんご	1942.10. 7	↔	35.90, 139.71	サッカー

②-2 いわゆる匿名データ

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



②-3 高度な処理の匿名データ

位置情報	趣味
兵庫県	球技
埼玉県	球技

③統計データ

* 統計処理データ
* 作成には専門知識が必要

①実名データ

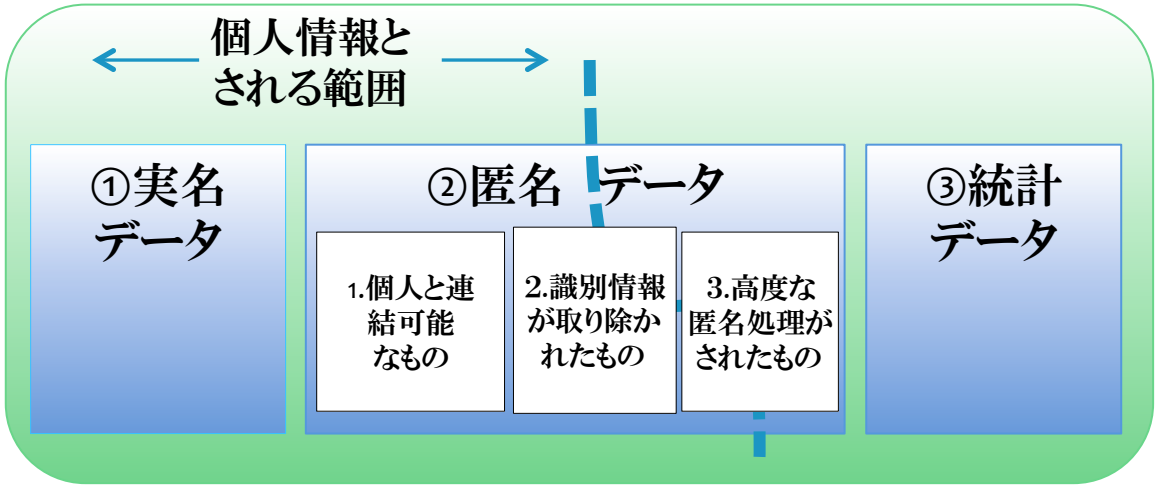
氏名	生年月日	位置情報	行動
鈴木よしえ	1978.9.10	34.72, 135.36	野球
松井りんご	1942.10. 7	35.90, 139.71	サッカー

②-1 個人と連結可能な匿名データ

氏名	生年月日		位置情報	趣味
鈴木よしえ	1978.9.10	↔	34.72, 135.36	野球
松井りんご	1942.10. 7	↔	35.90, 139.71	サッカー

②-2 いわゆる匿名データ

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



②-3 高度な処理の匿名データ

位置情報	趣味
兵庫県	球技
埼玉県	球技

③統計データ

	東京	埼玉
野球	33	8
サッカー	27	32

安全性指標：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
  * [Sweeney 02] k-Anonymity: A Model for Protecting Privacy
  * 準識別情報について、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、共通の準識別情報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

№	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匿名化

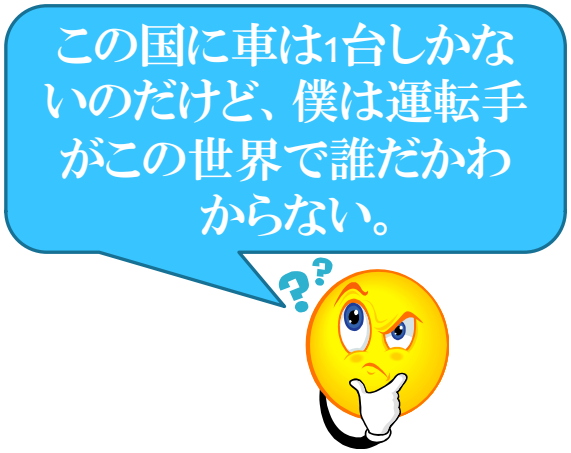
ここでは、郵便番号・性別・年齢に注目

3-匿名性（郵便番号・性別・年齢）

№	郵便番号	性別	年齢	趣味	
1	18000**	男	3*	アニメ	3
2	18000**	男	3*	アニメ	
3	18000**	男	3*	アニメ	
4	18100**	女	4*	映画	3
5	18100**	女	4*	アニメ	
6	18100**	女	4*	ドラマ	
7	18000**	男	50以上	映画	4
8	18000**	男	50以上	ドラマ	
9	18000**	男	50以上	ドラマ	
10	18000**	男	50以上	時代劇	

母集団一意性

* 個人識別の母集団評価
  * 車を持つ人が世界中一人だとしても、世界中の誰であるかを特定できない
  * 例：
    * 世界で一人しかいない新種の病気にかかった病人がいるとして、病院関係者と家族以外は、誰がその病気にかかっているかはわからない。
    * その病気のデータは貴重なデータなので、患者名は公表されないが、病気のデータを学会で発表された。

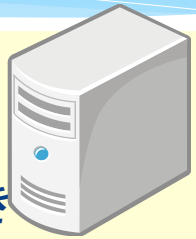


この車は世界に一台

東京大学大学院情報理工学系研究科
ソーシャルICT研究センター

k 匿名性の匿名化レベル

* k 匿名性を満たしている情報にもいろいろなレベルがある.
  1. 匿名化データをどのように準識別子の取り方を考えても, k 匿名性を満たしている状態（統計データとして扱うことができ
  2. ある準識別子に注目すると, k 匿名性を満たしている状態



匿名化データ

1. どの準識別子だとしても大丈夫

3-匿名性（＊）

郵便番号	性別	年齢
18000**	男	3*
18000**	男	3*
18000**	男	3*
18100**	女	4*
18100**	女	4*
18100**	女	4*
18000**	男	50以上
18000**	男	50以上
18000**	男	50以上
18000**	男	50以上

2. ある準識別子に注目すると、k 匿名性を満たしている

3-匿名性（郵便番号・性別・年齢）

郵便番号	性別	年齢	趣味
18000**	男	3*	アニメ
18000**	男	3*	アニメ
18000**	男	3*	アニメ
18100**	女	4*	映画
18100**	女	4*	アニメ
18100**	女	4*	ドラマ
18000**	男	50以上	映画
18000**	男	50以上	ドラマ
18000**	男	50以上	ドラマ
18000**	男	50以上	時代劇

この条件を満たす人は一人しかいないが、3つの準識別子に注目すると、3-匿名性を満たしている

k 匿名性データの多様性

- k 匿名性を満たしたデータを定義したとしても、一意には決まらない
- 準識別子の取り方には複数の種類がある。
  - ものさしの当て方も、一意には決まらない
- 情報利用者にとって、それぞれに必要な情報が違うので、
  有用な匿名化方法を利用者自身が見つけなければならない。

収集データ

情報利用者

どんな風にものさしを
あてようか？

同じデータを利用したとしても、有用性が変わってくるので、
利用者にとって、必要な取り方をしなければならない

3-匿名性（郵便番号・性別・年齢）

郵便番号	性別	年齢	趣味
18000**	男	3*	アニメ
18000**	男	3*	アニメ
18000**	男	3*	アニメ
18100**	女	4*	映画
18100**	女	4*	アニメ
18100**	女	4*	ドラマ
18000**	男	50以上	映画
18000**	男	50以上	ドラマ
18000**	男	50以上	ドラマ
18000**	男	50以上	時代劇

3-匿名性（アニメ）

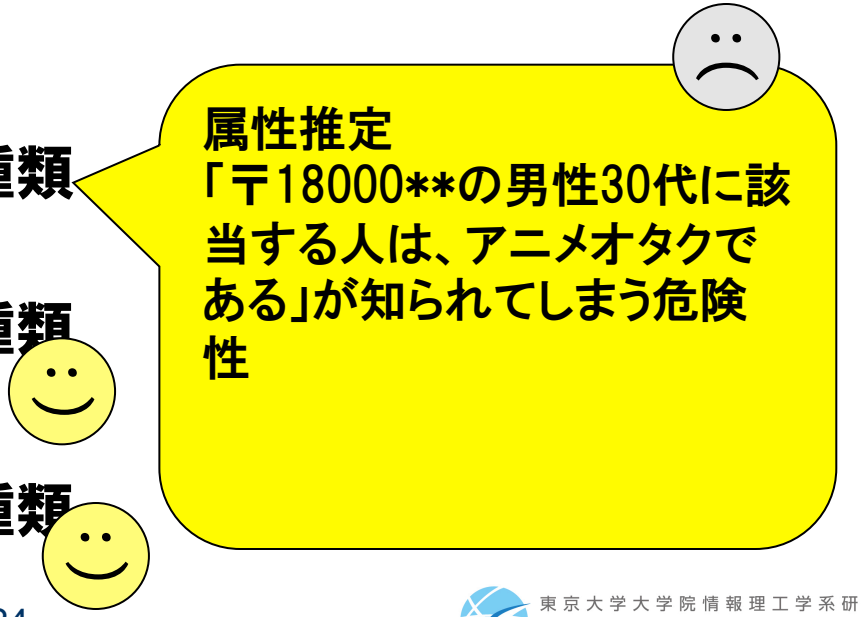
趣味	郵便番号	性別	年齢
アニメ	1800005	男	39
アニメ	1800012	男	32
アニメ	1800003	男	37
アニメ	1810015	女	46
ドラマ	1800021	男	52
ドラマ	1800001	男	60
ドラマ	1810013	女	46
映画	18100**	女	4*
映画	18000**	男	50以上
時代劇	18000**	男	50以上

このデータは利用できない

k-匿名性を補完する:L-多様性

* 開示データからの属性推定を防ぐための匿名化モデル
  * [Machanavajjhala et al.06] l -Diversity: Privacy Beyond k-Anonymity
* 同じ準識別子の組み合わせを持つk個のレコードの中で、関連する属性データがL種の良い多様性を持つこと
* L-多様性には様々な属性推定を防ぐためにバリエーションが提案されている
* T-closeness: L-多様性があったとしても、データに偏り(例:99%、1%)があれば結局傾向としてはわかってしまう。

№	郵便番号	性別	年齢	趣味	
1	18000**	男	3*	アニメ	1種類
2	18000**	男	3*	アニメ	
3	18000**	男	3*	アニメ	
4	18100**	女	4*	映画	3種類
5	18100**	女	4*	アニメ	
6	18100**	女	4*	ドラマ	
7	18000**	男	50以上	映画	3種類
8	18000**	男	50以上	ドラマ	
9	18000**	男	50以上	ドラマ	
	18000**	男	50以上	時代劇	



位置情報とそのほかの履歴の違い

* 同じ履歴情報でも、情報の特性によって、後で統合しやすいものとしにくいものがある、
  * 収集したデータが移動履歴情報であれば、組み合わせの際に、個人特定の可能性がある
* より簡単に個人を特定できる情報となる可能性を持つ

ものの購入 ← 既に曖昧化してある

何時何分何秒 北緯　西経 ← 細かいと必ず特定される

k匿名性の限界

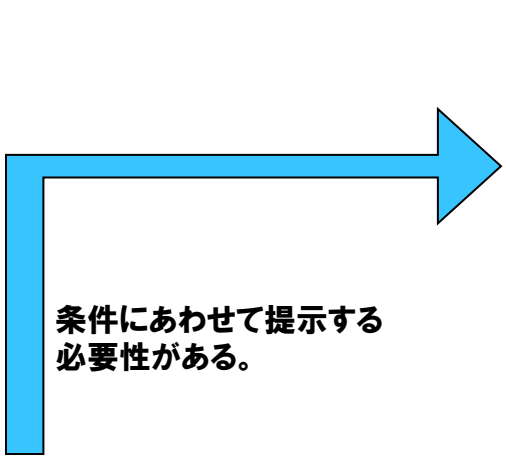
* k匿名性の限界
  * k匿名化(*)については安全
    * k匿名性という概念は、情報を統計情報とするという点で、価値がある。
    * ただし、アプリケーションは限定的となる。
  * k匿名性の持つ危険性
    * 準識別子の選択や組み合わせによって、匿名性が犯される危険性を持つ。
  * k匿名性した情報では、二次利用者にとって、必要な情報すべてが存在しているかどうかがまだわかっていない
    * 匿名化**しすぎ**ている可能性も。
  * 正しく匿名化しているかどうかが、**第三者の検証**が必須
    * 自己評価ではない、方法が必要

* ガイドラインや制度での実現の限界
  * ガイドラインや制度で、匿名化しない準識別子を限定的にすることや情報の組み合わせを行わないことを定めることで実現することも考えられるが、技術的に守られることが保証されていない。

安全な基準ができたとして、誰が実現するのか

責任の所在：誰が実現するのか？

* 誰の責任を持って、前述の匿名性を実現するのか、について定義。
* 認証（本当に実行していることを誰がそれを確認するのか）については、言及しない。
  * 今後は、認証についても議論する必要性がある。



条件にあわせて提示する
必要性がある。

センター（情報収集者）
条件設定（匿名性（1）～（4））にあうようにじ情報の匿名処理をするかどうか。
何らかの匿名処理しなかった場合はどうする？

生データがくる可能性もある。
その場合の条件はどうするか。

情報提供者

プロバイダ（二次収集者）

匿名化基準と実現方法の違い

* 母集団一意性を満たすために技術基準が必要
* 匿名化データをどのように実現するのか
  * 例として、k 匿名化があげられるが、実現のためには、議論は今から
  * 実現するためには、安全性の整理も必須

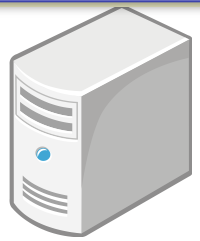
匿名化したあとのデータの性質についての基準

上の基準を満たすためにどのような技術を利用すべきかを考える

情報提供者

収集データ

匿名化



匿名化データ

情報利用者

匿名化基準の必要性

「ものさし」としての基準が必要

実現方法の選定

匿名化データをどのように実現するのか

評価・認証

正しく匿名化データを作成したのか確認

他にもいろいろなプライバシー保護技術が！

* 単純な情報の暗号化
* 様々な分野の匿名化手法
* 匿名認証、署名技術
* 匿名検索

ユーザとの同意

ユーザとの同意（例：スマホアプリ）

- スマホアプリの同意については、適切な方法が難しい
  - 現状では、Androidについてはサービス事業者による過剰なパーミッション要求が散見され、ユーザが判断できない
  - 同意の方法が不適切なため、問題になるアプリも

実行時確認型
（iOS）

インストール時確認型
(Android)

IPA注意喚起より
http://www.ipa.go.jp/security/txt/2012/09outline.html

ユーザ（利用者）への適切な匿名化手法の提示

* 現状では、「適切に処理しています」と、プライバシーポリシーに記載があるだけ
* 今後は、各社がどれだけ責任をもってやっているのかを提示する方が良い

昔の日本って？

* 情報は基本縦割り
  * 日本は、情報がリンクをとれるイメージがない

まとめ

問題解決にむけて

* 各データ種別ごとに、有用な情報の定義を
  * 移動体（携帯電話・プローブ情報システム）、買い物データ、健康データなど、それぞれごとに解析を行い、匿名性と有用性トレードオフの整理が必要。
    * 場合によっては、識別情報の削除だけで、匿名化を満たし、二次利用には優れている可能性がある。

  → 分野ごとに適したプライバシ手法があるのでは？
* 情報の定義ができたとして、誰が実現してそれを保証していくのか
* 安全性検証を負荷した技術の組み合わせ
  * ただし、重くならないよう、ほどほどに実現。