データの
「匿名性」と「公共性」

データの「匿名性」と「公共性」

- 「匿名性」
  - データから「漏れてはいけない」情報が特定されない度合い
  - データの「匿名化」
    - いわゆる**「お客様の情報は統計的に処理されるので，個人が特定されることはありません」**を立証する行為・手続き

- 「公共性」
  - データが誰でも使える度合い
    - データ大好きな我々「もっと自由にデータを使えるようになれ！」
    - **共通言語Rを持っている我々はバベルの塔を建築し始める……**
  - データの公共化
    - 誰も困った思いをすることなく，分析する価値のあるデータを構築し公開すること

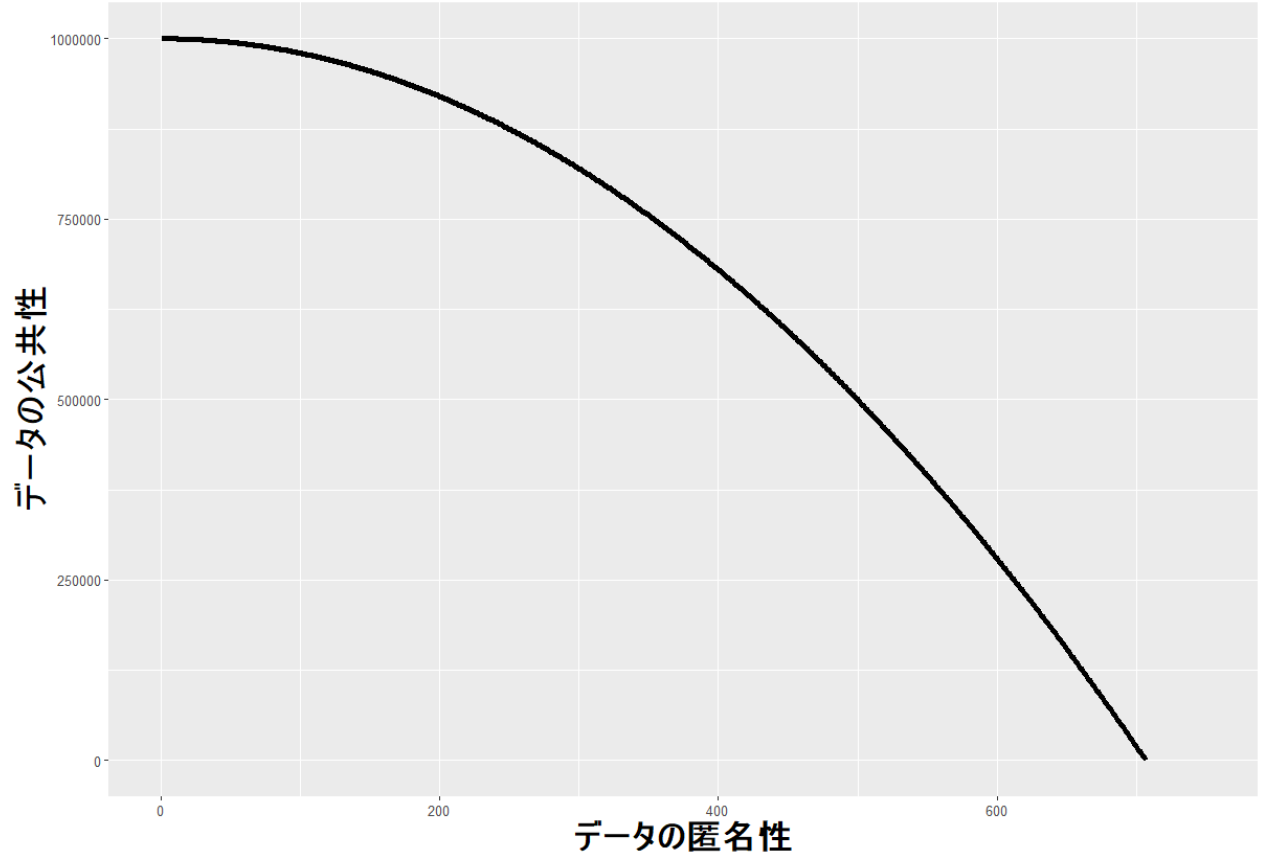
理想は

- 公共性も匿名性も高いデータがあれば最強
  - 公共性は高すぎるとプライバシーを無視
  - 匿名性が高すぎると公共性の高さを担保できない
  - どっちも低いデータは……

- そうはいかんざき
  - 個人情報保護のための圧力
    - お茶の間を騒がせがちな官公庁の情報漏れ(耳が痛い)

つまり

- 匿名性と公共性はトレードオフ
  - ggplot2を使っているのでTokyoRのLTです



ルール的には

- データが漏洩しないように処理しなさーい！
- データが漏洩したとしても**個人が特定されないように処理**しなさーい！

- 「個人が特定されないような処理」って何？
  - どんな処理をすればどの程度「安全」なのか？

本題
データの匿名化手法

知られたくない情報と知りたがり(3)

- こんなデータが入っていました
- ちょっとAという変数観てみよう

id	A	B	C	D
1	44	124	M	26
2	5	735	F	20
3	30	173	F	29
4	5	1005	F	18
5	43	1380	F	19
6	16	1273	M	28
7	47	485	M	30

```
> dat <- readr::read_csv(“Survey1.csv”)
> summary(dat$A)
>  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   13.00   26.00   24.75   37.00   47.00
```

特定してみる

- 「最小値が1で最大値が47のA，な〜んだ？」
  - ……「アレ」か
- 「FとMで記号化されてるCといえば」
  - ……「アレ」かもな

- 官公庁の集計済みデータとかSNSとか使って
  - みいつけた

BAD END

つまるところ

- 企業・組織が持つ個人がユニークとなっているデータの漏洩
  - 公開されている集計済み官公庁データ
  - 個人のSNSや写真・映像
  - これらを組み合わせると割と簡単に個人は特定できてしまう

- 企業・組織: **「漏洩しない」努力**は必須
  - 一方で**「漏洩したときの被害を最小限にする」**こともも重要
  - こちらの話

k-anonymity

- 似た情報を持つ人達を『一人』として扱う
  - クラスタは特定できても，その中にいる「個人」を特定することは難しいことを根拠にした匿名化.
    - 計算複雑性理論ベースでは，k-匿名化されたデータから個人を特定するという問題はNP困難な問題らしい

id	A	B	C	D
classA	Var_A	*	M	21-30
classA	Var_A	*	F	21-30
classA	Var_A	*	F	21-30
classB	Var_B	*	F	10-20
classB	Var_B	*	F	10-20
classA	Var_A	*	M	21-30
classC	Var_C	*	M	31-40

k-anonymity

- 「知識による攻撃」を受けづらい記号化を行う(抑制)
  - 例: 変数Aは順序の明らかでないカテゴリに変換する
    - var_a,var_bなどに変換するときは順序に規則性がないようにランダム化する
    - あえて「使わない」手段(B列のアスタリスク)

- IDが「集団」に対してユニーク
  - 「個人」に対してユニークではない
  - 「計算機」にとっては特定しづらい
  - でも「**人間**」にとっては……？

k-anonymity

- 実装コストが低い(メリット)
  - 例: 1クラスタあたりk人が含まれるようにクラスタリング
    - 本人が特定される確率は単純に考えれば1/k
    - kが大きいほど特定リスクは下がる.

- データの粒度が粗くなる(デメリット)
  - 解析した結果の評価指標が悪く計算されることがある
    - p-valueや相関係数は若干出づらい(経験則)
    - 予測精度の指標も落ちる

k-anonymity

- 匿名化されない情報もある(デメリット)
  - 「似た情報」を持つことが特定につながる可能性がある**(同質性攻撃)**

    例: C列の情報は抑制も一般化もされていない

    →特定したい個人が属性Fを持っていることを知っていると，

    データから情報を取得できる危険性は排除しきれていない.

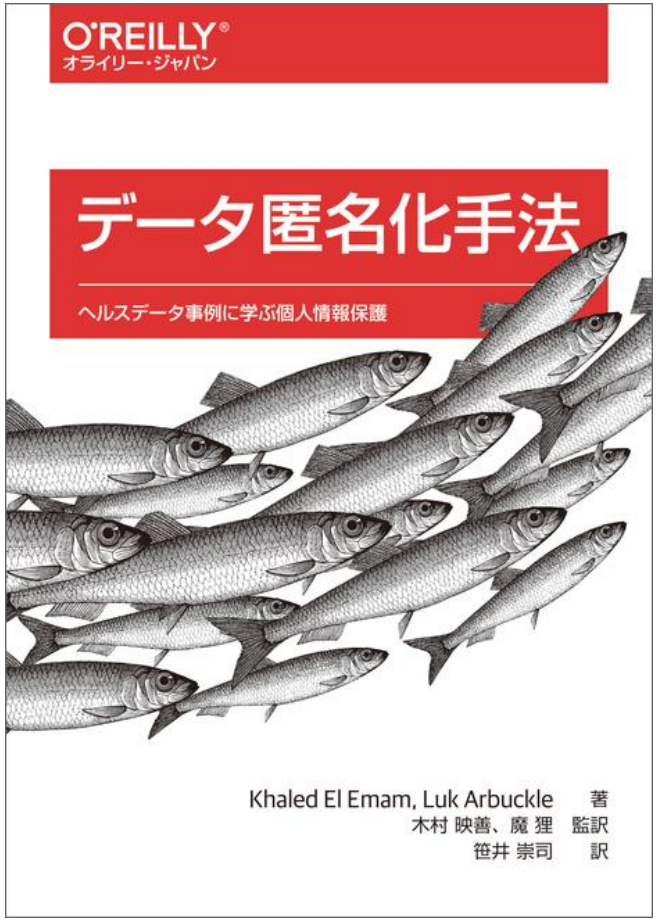
- 「どのクラスタに属しているか」がわかると特定される可能性がある
  - 漏洩したデータに20代女性が含まれることを悪用**(背景知識攻撃)**
    - 計算機にとって特定が難しくても，人間にとって特定が難しいとは限らない
    - **AI(笑)よりまだ人間の推論力のほうがこわい**

まとめ

- for 情報を提供する人
  - 情報を求めている組織が信用できるか見極めよう
    - プライバシーマークとか参考にはなるかもしれない
  - 安易にSNSに情報を流すのは控えよう
    - 昨今話題になっている炎上動画より普通の画像や発言のほうがよほど危険

- for 情報を扱う人
  - 安易な記号化だけでは個人のプライバシーを守れるとはいえない
  - 第三者とデータを共有する際はデータの匿名化も検討しない？

参考



その他各種論文とか

enjoy!