FUJITSU
shaping tomorrow with you

匿名加工・再識別コンテスト2015（PWS CUP 2015）振り返り

2016年3月

富士通研究所
山岡 裕司

PWS 2015 とは

国内外においてビッグデータ活用ニーズが急速に高まる中で、データの効果的な活用とプライバシー保護を両立させる技術や規準の発展と確立が強く求められています。

そこで

CSEC研究会主催CSSの併催

PWS CUP 2015 とは

- PWS企画のコンテスト
- 狙い
  - 匿名加工の技術や評価方法の確立
  - 議論や交流の活発化

- コンテスト概要
  1. 匿名加工部門
     - 有用性・安全性が高い匿名加工を競う
  2. 再識別部門
     - 他チームの匿名加工を破ることを競う

世界的に類を見ないコンテスト

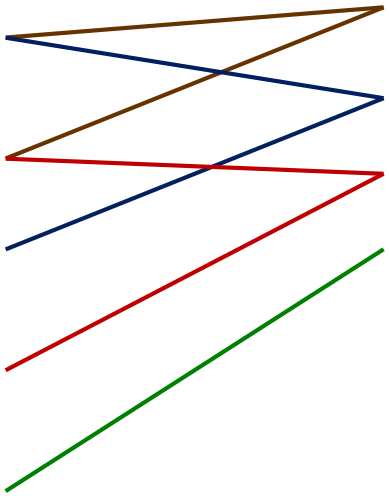
- 以降、ルールを中心に振り返る

ルールの概要

FUJITSU

予備戦・本戦と採点法

内容	最高得点	採点法
予備戦（匿名加工） 8/24～9/24	30	順位
予備戦（再識別） 9/25～10/9	30	順位
本戦（匿名加工） 10/21 in PWS	50	順位
本戦（再識別） 10/21 in PWS	50	順位
最終プレゼン 10/22 in PWS	20	審査点 （審査員3人）



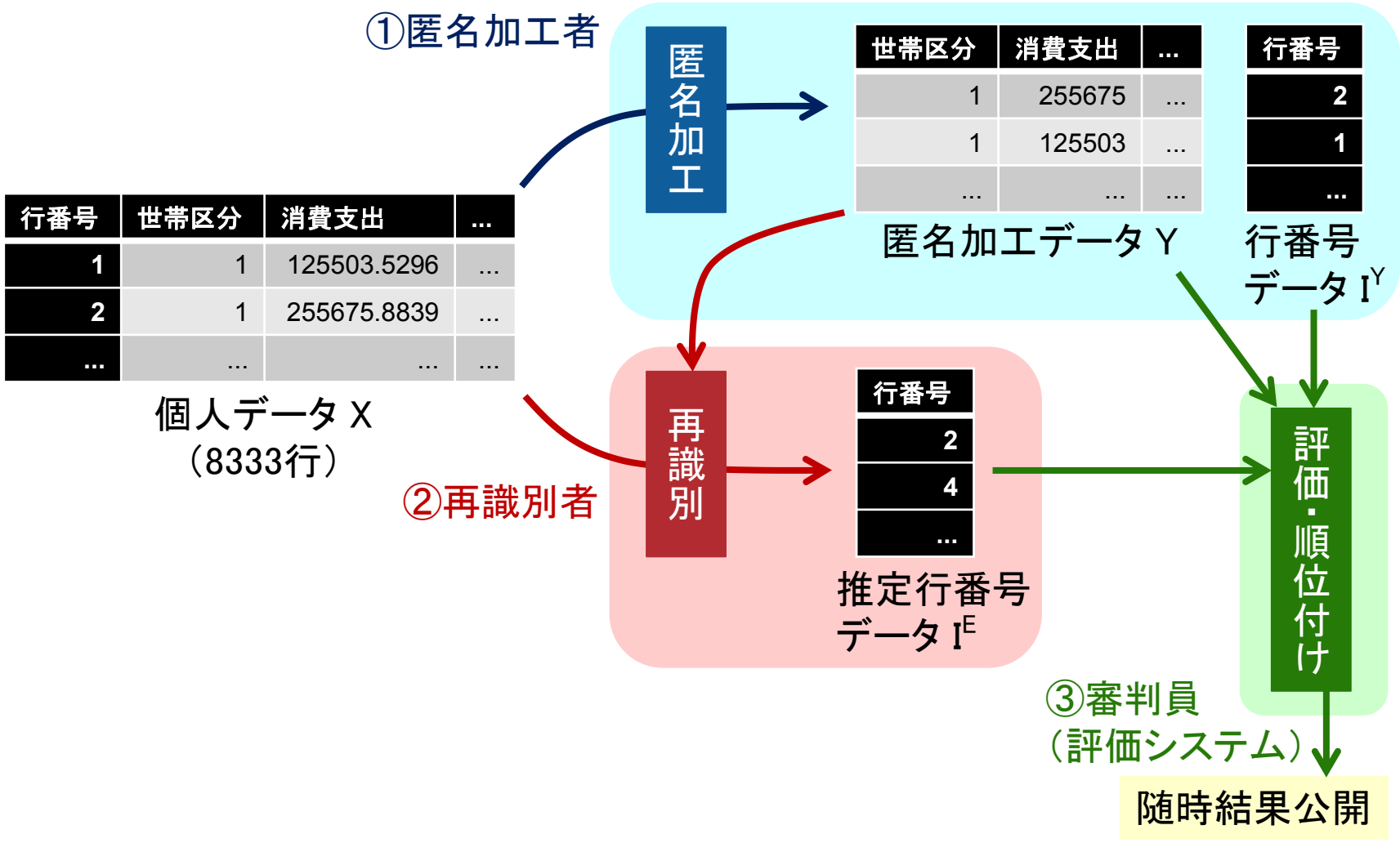
賞
予備戦
匿名加工部門
再識別部門
擬似データ生成部門
総合

1チームが独占！

以降、本戦のルールを説明

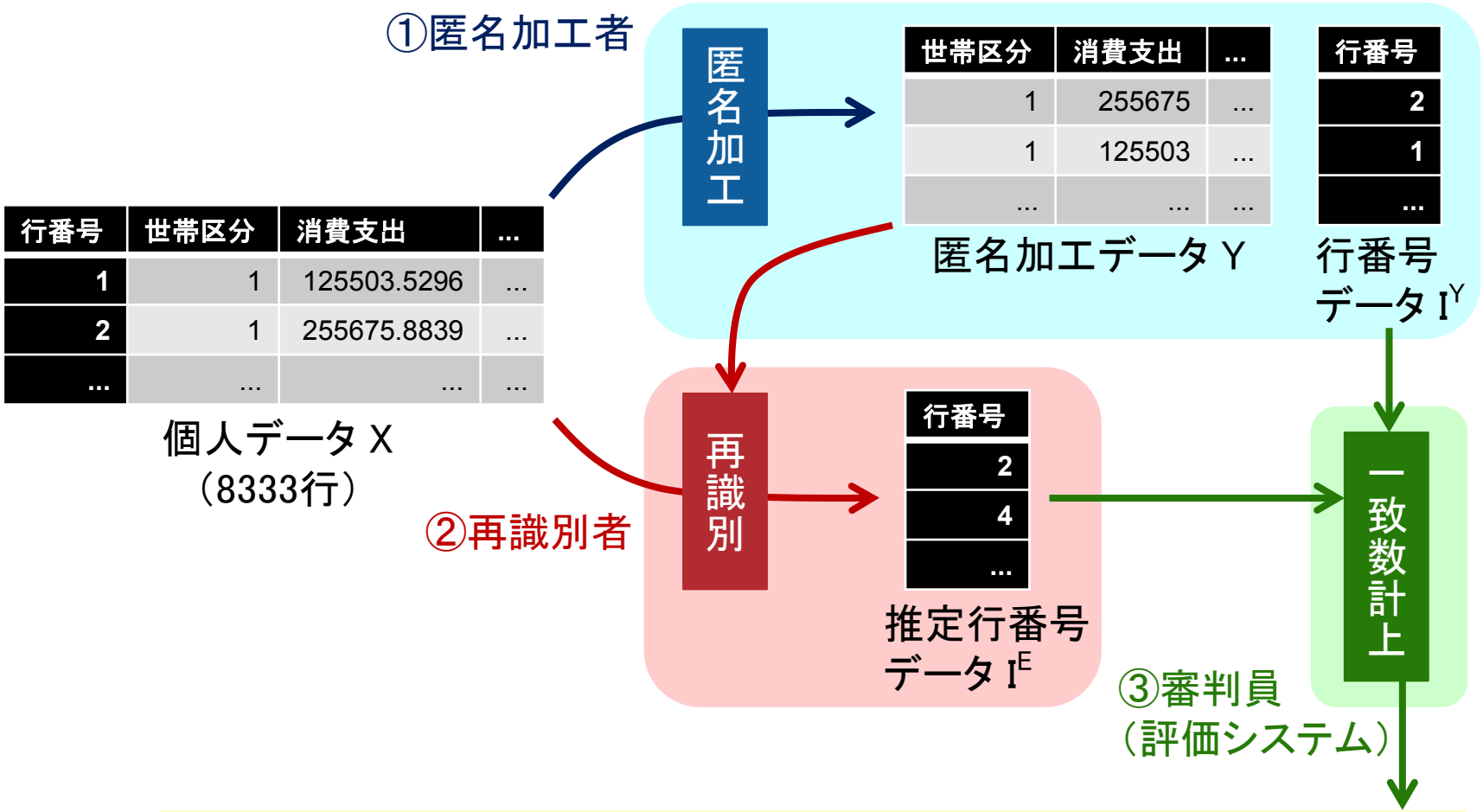
コンテストの流れ

- 匿名加工：再識別されにくく、有用なデータに加工する
- 再識別：匿名加工の行番号を推定・当てる



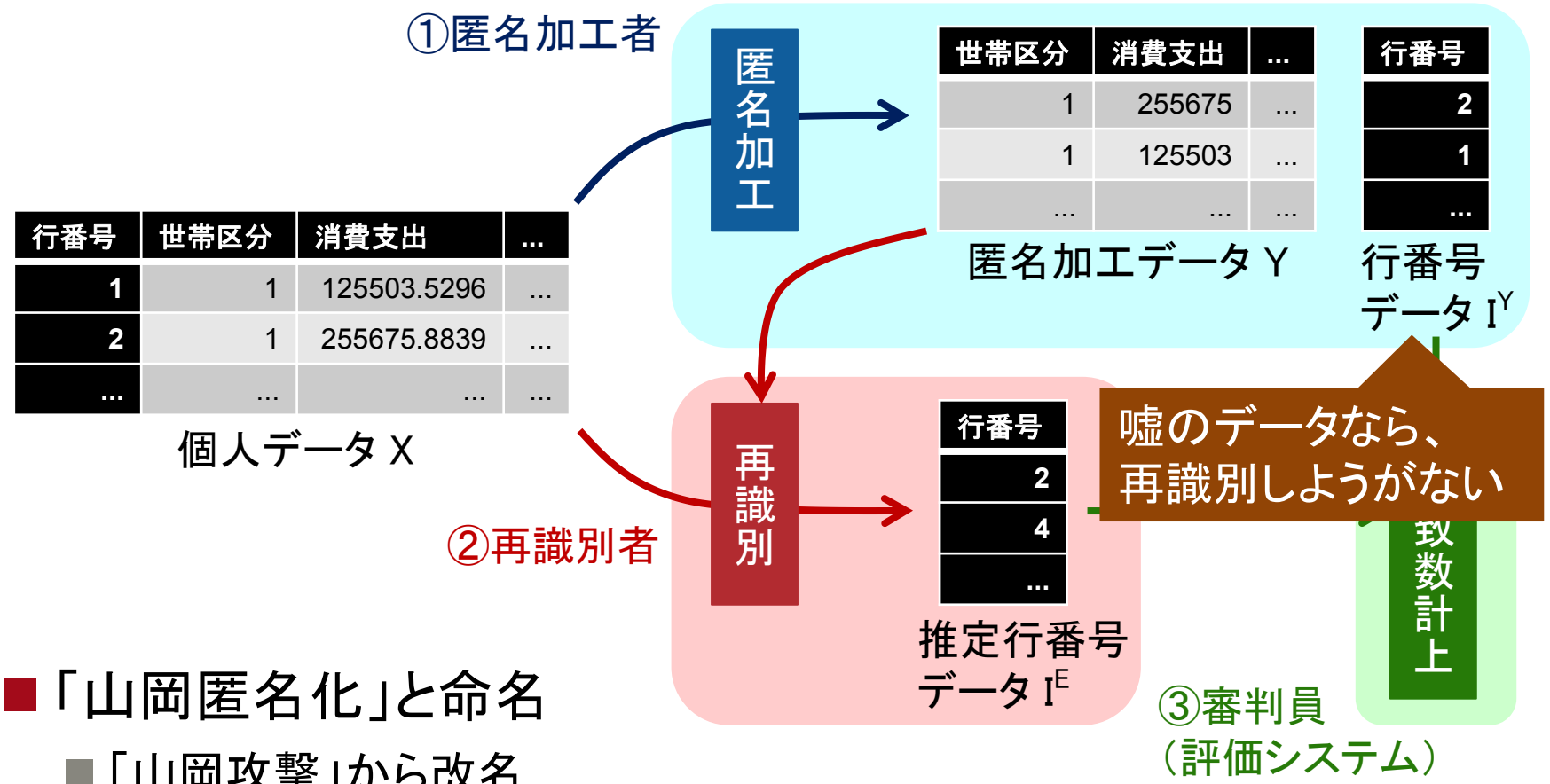
再識別のルール

- 再識別部門の順位：再識別レコード総数（の補正値）の多い順
  - 他チームの各匿名加工データに対する再識別が有効



山岡匿名化

- 問題点：嘘の行番号データ I^Y を申告されると、再識別不可能

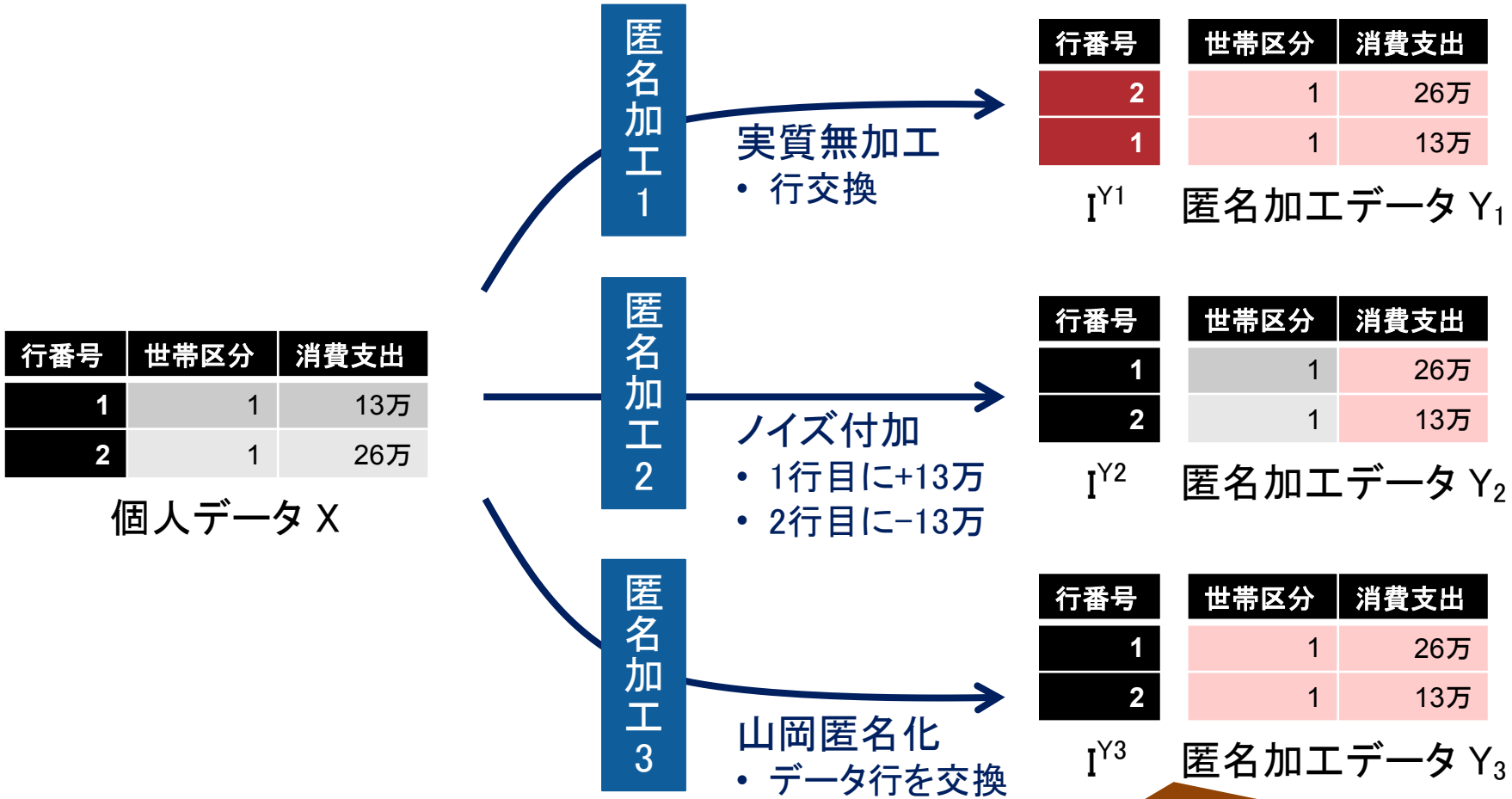


- 「山岡匿名化」と命名
  - 「山岡攻撃」から改名
- 予備戦で猛威を振るった（匿名加工上位10データ全て）
- 本戦でルール変更するなど、山岡匿名化対策に苦心

山岡匿名化対策の難しさ

FUJITSU

■「再識別 ＝ 行番号当て」の宿命



- 有用性が高い（データの変更はない）
- 再識別されない（無加工と見分けられない）
- 不正といえない（ノイズ付加と見分けらない）

参考：法律の「特定の個人を識別」とは

FUJITSU

- 「再識別」コンテストの背景：個人情報保護法の改正
  - 2015年9月に成立、2年以内に施行
  - 「匿名加工情報」を新設
    - 具体的な基準は未公開
      - PWS CUP が一助となる？
  - 「個人情報」の定義（現行法／改正法の一部）

    …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、それによって特定の個人が識別できることとなるものを含む。）
- 疑問：識別とは？山岡匿名化は再識別不可能化？

行番号	氏名	世帯区分	消費支出
1	佐藤 太郎	1	13万
2	鈴木 花子	1	26万

匿名化前データ
（「他の情報」の例）

行番号（CUP）

値

再識別？

行番号	世帯区分	消費支出
1	1	26万
2	1	13万

氏名削除＋山岡匿名化

- 直感的には26万は鈴木の情報と識別可能
- CUPではその識別を防止しているとみなす？

匿名加工のルールの概要

FUJITSU

- 匿名加工部門の順位：有用性と安全性の和の小さい順
  - 有用性（最小値: 1）

$$\frac{1}{6}\sum_{i=1}^{6} Rank(U_i)$$ （U: 有用性指標値）

  - 安全性（最小値: 1）

$$\frac{1}{4}\sum_{i=1}^{2} Rank(S_i) + \frac{1}{2} Rank(\max_j re\text{-}id^{E_j})$$ （S: 安全性指標値, $re\text{-}id^E$: 再識別率）

最大の再識別率が
小さいほど良い

- 事前公開再識別アルゴリズムEの適用結果
- 再識別部門での再識別結果

匿名加工のルール ー 有用性

FUJITSU

- 有用性指標Uは6つ（値が小さいほど良い）

U_1 = MeanMAE,	SA={14,…,25}についての平均絶対誤差
U_2 = crossMeanA, B	A={1, 2, 3, 4, 5, 6}（世帯区分, 世帯人員, 有業人員, 住居の構造, 住居の建て方, 住居の所有関係), B=15 のクロス集計値の平均絶対誤差
U_3 = crossCntA,B	同クロス集計数の平均絶対誤差
U_4 = corMAE	SA={14,…,25}についての全相関係数の平均絶対誤差
U_5 = IL	匿名加工データの各値の平均絶対誤差
U_6 = nrow	個人データと匿名加工データのレコード数の差（の絶対値）

- 例：U_5 $$IL(\boldsymbol{X},\boldsymbol{Y},I^Y)=\frac{1}{m'n'}\sum_{i=1}^{m'}\sum_{j=1}^{n'}\frac{|x^i_{i^Y_j}-y^i_j|}{\max x^i-\min x^i}$$

Xの列25の値の範囲が[0, 5000]の場合、
$|x^{25}_1-y^{25}_1|/\max x^{25}-\min x^{25}$
$=|2000-3000|/5000-0$
$=1/5$

QI　　SA

行番号	1	…	13	14	…	25
1	1	…	10	0	…	2000
2	1	…	16	0	…	3000
…	…	…	…	…	…	…

個人データ X

→

行番号	1	…	13	14	…	25
1	1	…	16	0	…	3000
2	1	…	10	0	…	2000
…	…	…	…	…	…	…

匿名加工データ Y

- Uで唯一、山岡匿名化にペナルティを課す → ∴導入

匿名加工のルール ー 安全性

FUJITSU

■ 安全性指標Sは2つ（値が大きいほど良い）

S_1=k−anony　　k−匿名性のk値（QIのみ）
S_2=k−anonyMean　　その平均値

安全性：

$$\frac{1}{4}\sum_{i=1}^{2} Rank(S_i) + \frac{1}{2} Rank(\max_j re\text{-}id^{E_j})$$

■ 事前公開再識別アルゴリズムEは5つ

- 5つのアルゴリズムE
- 再識別部門

E_1 = IdRand　QI={1,...,13}が同じレコードの中からランダムにレコードを識別

E_2 = IdSA　QI={1,...,13}が同じレコードの中から，SA={15}（消費支出)について識別

E_3 = Sort　SA={14,..., 25}の総和でソートする

E_4 = SA21　SA={21}だけで最小距離を持つレコードを識別

E_5 = AYA　(Anti Yamaoka-~~Attacker~~) Anonymization

本戦で追加

I^E: E_3 で算出した推定行番号データ

$$i_i^{AYA} = \begin{cases} i_i^Y & \text{if } \left|X_{i_i^E} - Y_i\right| < \left|X_{i_i^Y} - Y_i\right| \\ * & \text{otherwise} \end{cases}$$

（距離 $|X_i - Y_j| = \sum_{a \in SA} \frac{|x_i^a - y_j^a|}{\max x^a - \min x^a}$ ）

I^Y: 正答である行番号データ
（AYAでだけ特別に使用）

匿名加工データYの各行iにつき、
E_3で推定される行の方が、I^Yとして正答と申告された行より、「距離」が近い場合に限り、正答を言い当てる

匿名加工のルール ― 安全性(例)

FUJITSU

■ AYAの作用例

行番号	世帯区分	消費支出
1	1	13万
2	1	26万
3	2	8万
4	3	8万

個人データX

匿名加工
（山岡匿名化）

行番号(I^Y)	世帯区分	消費支出
1	1	26万
2	1	13万
3	3	8万
4	2	8万

匿名加工データY
（同色行を交換）

安全性:

$$\frac{1}{4}\sum_{i=1}^{2} Rank(S_i) + \frac{1}{2} Rank(\max_j re\text{-}id^{E_j})$$

- 5つのアルゴリズムE
- 再識別部門

再識別
（普通の再識別者）

I^E
2
1
4
3

$re\text{-}id^E$: 0/4

AYA

$$i_i^{AYA} = \begin{cases} i_i^Y & \text{if } \left|X_{i_i^E} - Y_i\right| < \left|X_{i_i^Y} - Y_i\right| \\ * & \text{otherwise} \end{cases}$$

$I^{E(Sort)}$	$I^{E(Sort)}$はI^Yより近い？	$I^{E(AYA)}$
2	Y	1
1	Y	2
3	N	*
4	N	*

$re\text{-}id^E$: 2/4

まとめ

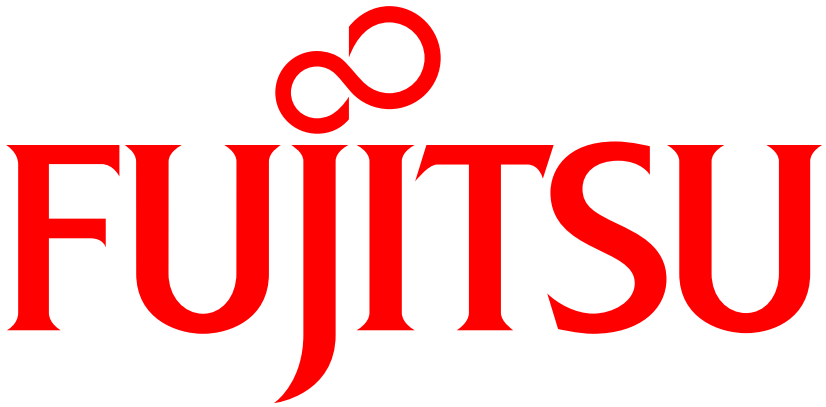
FUJITSU

- PWS CUP 2015 が開催
  - 狙い
    - 匿名加工の技術や評価方法の確立
    - 議論や交流の活発化
  - 世界初
- 本戦のルールを概説
  - 再識別部門の順位：再識別レコード総数（の補正値）の多い順
    - CUP 2015の再「識別」：匿名加工者が申告した行番号を再識別者が当てる
    - 個人情報保護法の「識別」：？
  - 匿名加工部門の順位：有用性と安全性の和の小さい順

$$\frac{1}{6}\sum_{i=1}^{6} Rank(U_i) \quad + \quad \frac{1}{4}\sum_{i=1}^{2} Rank(S_i) + \frac{1}{2} Rank(\max_j re\text{-}id^{E_j})$$

- 山岡匿名化：匿名加工者が嘘の行番号データIYを申告
  - 予備戦で猛威を振るったため、本戦で対策としてAYAを導入

議論の一助になれば幸いです！



shaping tomorrow with you