

ユークリッド距離を用いた再識別手法と PWSCup2015の匿名加工データを用いた評価

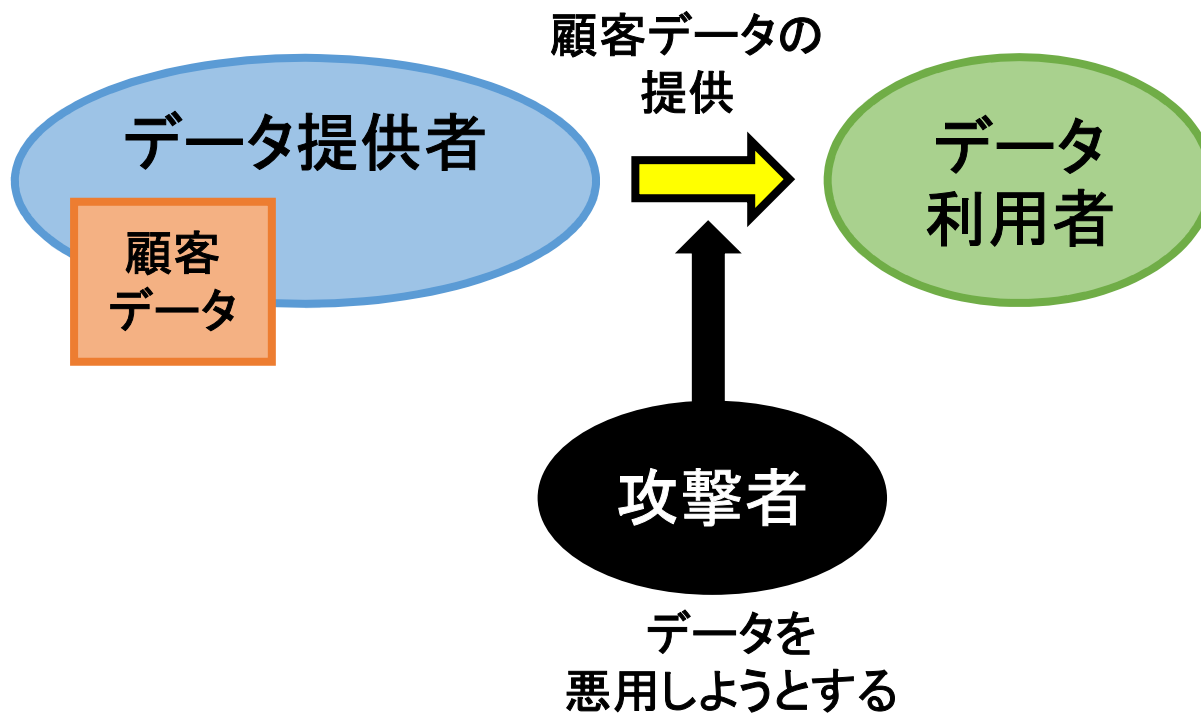
伊藤聡志 菊池浩明

明治大学 総合数理学部 先端メディアサイエンス学科

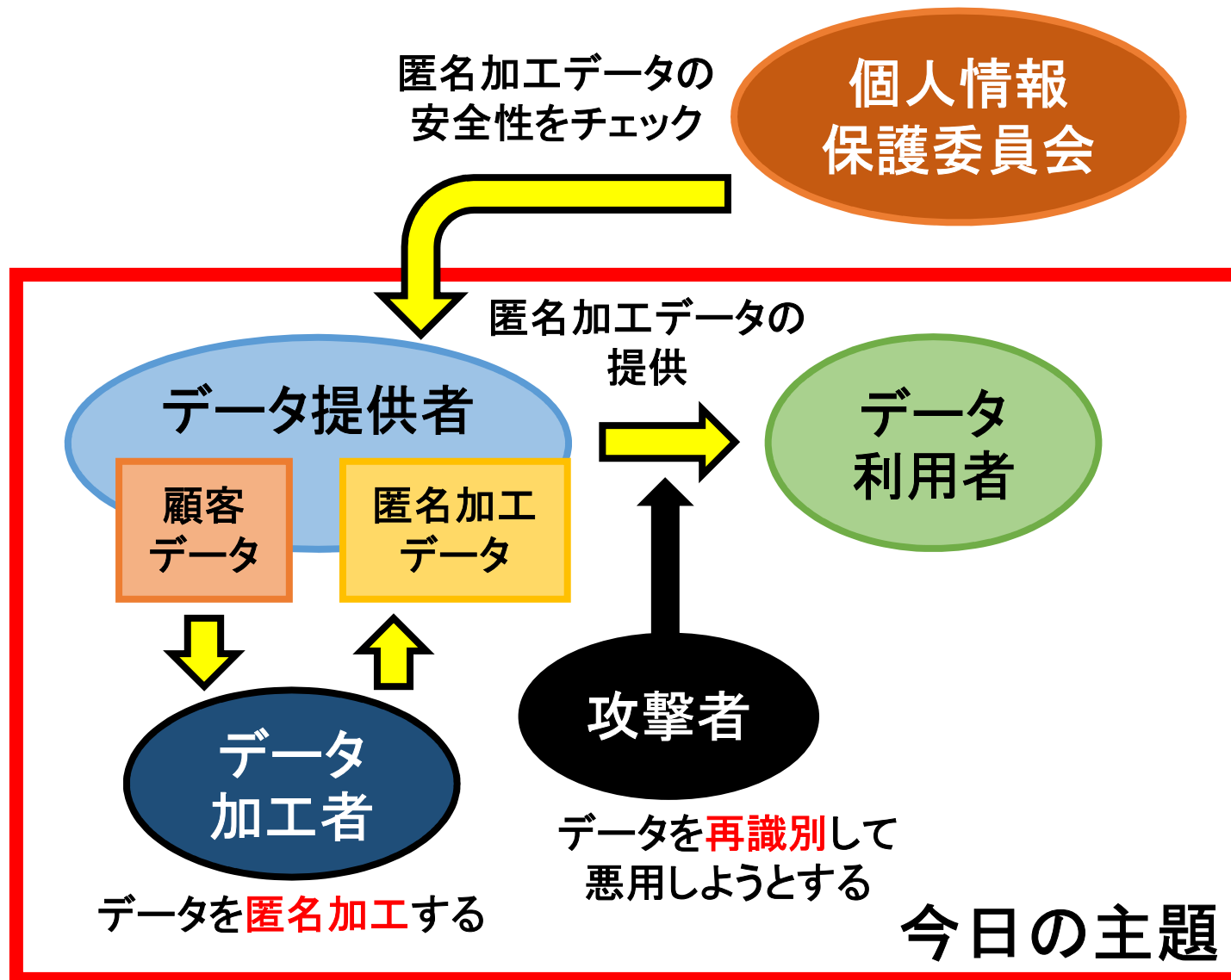
発表概要

- 新たな再識別手法の提案
- PWSCUP2015に提出された匿名加工データの解析

匿名加工とは？



匿名加工とは？



保護法改正とPWSCUP

2015年9月に個人情報保護法が改正され、「**匿名加工情報**」が定義された。

2015年10月に第1回「**PWSCUP 匿名加工・再識別コンテスト**」が長崎で開催された。
第2回は秋田で開催される。



The poster for the PWS CUP Ice And Fire contest features a laptop displaying a mission interface. The interface includes four main sections: 'Micro-data Anonymization Department' (ミクロデータの匿名加工部門), 'Anonymized Data Re-identification Department' (匿名加工データの再識別部門), 'Synthetic Data Generation Department' (疑似データの生成部門), and a 'Mission' (Mission) section. The background of the poster is dark with blue and orange flames, and the text 'Ice And Fire' is prominently displayed in a stylized font. Below the laptop, the event details are provided: '長崎ブリックホール' (Nagasaki Brick Hall) on '10/21水~10/23金' (October 21st Wednesday to October 23rd Friday). The poster also mentions '第1回プライバシーワークショップ(PWS2015)' and provides a URL: 'http://www.iwsec.org/pws/2015/'. At the bottom, it states 'PWS Cup 参加エントリー再募集期間 8/18火~9/11金' (PWS Cup participation entry re-recruitment period August 18th Tuesday to September 11th Friday) and '申し込み先はこちら' (Apply here) with 'PWS CUP 2015 実行委員会事務局' (PWS CUP 2015 Executive Committee Secretariat).

PWS CUP匿名加工・再識別コンテスト
アイスアンドファイヤー
Ice And Fire
Mission
会場 **長崎ブリックホール** 日時 **10/21水~10/23金**
第1回プライバシーワークショップ(PWS2015) <http://www.iwsec.org/pws/2015/>
PWS Cup 参加エントリー再募集期間 **8/18火~9/11金** 申し込み先はこちら **PWS CUP 2015 実行委員会事務局**

問題点と研究目的

1. 既存再識別手法には問題点がある.

→既存手法の弱点を改善した新たな再識別手法を提案し,
PWSCUP2015の本戦に提出された匿名加工データを用いて
既存手法との比較を行う

2. PWSCUP2015の本戦に提出された匿名加工データがどのような 手法で加工されているか不明である.

→本戦に提出された匿名加工データの解析を
小規模データを用いて行う

問題点1: 既存再識別手法

「PWSCUP 2015 匿名加工・再識別コンテスト」で匿名加工データの安全性評価に用いられた4つの再識別手法

- identify.rand** : ランダムに再識別を行う
- identify.sa** : ある1つのSAを用いて再識別を行う
- identify.sort** : SAの合計値をソートして再識別を行う
- identify.sa21** : 特定のSAを用いて再識別を行う

問題点

再識別に用いる属性の数が少ないため、加工に弱い

問題点2: 提出された匿名加工データ

PWSCUP2015本戦には様々な企業や大学から13チームが参加し, 24の匿名加工データが提出された.

しかし, どのデータがどのような手法で匿名加工されているかは不明である.

今回は5チーム(自チーム含む)が提出した12個のデータを研究に用いる.

データ名	作成チーム	成績
D_1, D_2	T_A (明治大学)	
D_3, D_4	T_B	2位
D_5, D_6	T_C	
D_7, D_8, D_9	T_D	1位
D_{10}, D_{11}, D_{12}	T_E	3位

提案手法 identify.euc

identify.euc

匿名加工データのレコードと同じQIのベクトルを持つレコードの中からSAのユークリッド距離 $D(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i \in S} (b_i - a_i)^2}$ で再識別を行う。

元データ							匿名加工データ				
○ ×	QI1	QI2	QI3	SA1	SA2		QI1	QI2	QI3	SA1	SA2
	2	1	1	100	100	14.142	2	1	1	110	90
	2	1	1	200	400	322.8	2	1	1	220	390
	1	1	2	300	200		1	1	2	280	210
	1	1	2	400	500		1	1	2	390	520

既存手法との違い

既存手法 identify.sa

元データ

	QI1	QI2	QI3	SA1	SA2
×	2	1	1	100	100
○	2	1	1	110	300
	1	1	2	300	200
	1	1	2	400	500

匿名加工データ

	QI1	QI2	QI3	SA1	SA2
	2	1	1	150	100
	2	1	1	160	300
	1	1	2	350	200
	1	1	2	450	500

50

40

提案手法 identify.euc

元データ

	QI1	QI2	QI3	SA1	SA2
○	2	1	1	100	100
×	2	1	1	110	300
	1	1	2	300	200
	1	1	2	400	500

匿名加工データ

	QI1	QI2	QI3	SA1	SA2
	2	1	1	150	100
	2	1	1	160	300
	1	1	2	350	200
	1	1	2	450	500

50

203.96

提案手法 EUC1, EUC2

EUC1

元データ

QI1	QI2	QI3	SA1	SA2
2	1	1	100	100
2	1	1	200	400
1	1	2	300	200
1	1	2	400	500

匿名加工データ

QI1	QI2	QI3	SA1	SA2
2	1	1	110	90
2	1	1	220	390
1	1	1	280	210
1	1	1	390	520



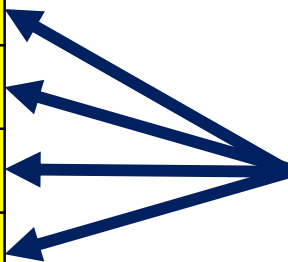
EUC2

元データ

QI1	QI2	QI3	SA1	SA2
2	1	1	100	100
2	1	1	200	400
1	1	2	300	200
1	1	2	400	500

匿名加工データ

QI1	QI2	QI3	SA1	SA2
2	1	1	110	90
2	1	1	220	390
1	1	1	280	210
1	1	1	390	520



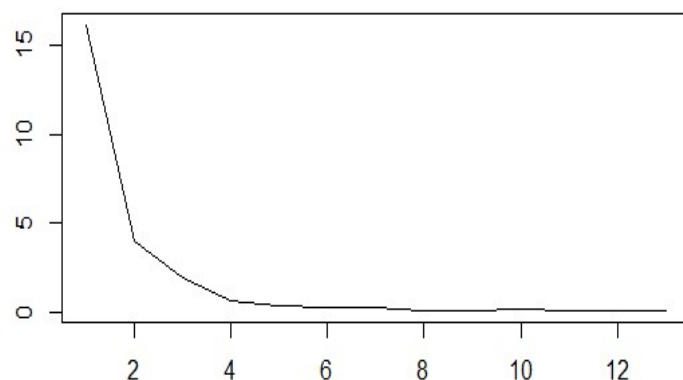
実験結果1: 既存手法との比較 (EUC1の精度)

匿名加工 データ	既存方式				提案方式
	Id-rand	Id-sa	Id-sort	Id-sa21	EUC1
D ₁	0.0326	0.8238	*1.0000	0.1858	0.3010
D ₂	0.6485	*0.6507	0.0012	0.0022	0.4780
D ₃	0.1990	0.2412	*0.2482	0.0511	0.2070
D ₄	0.1894	0.2401	*0.2526	0.0455	0.2110
D ₅	0.0000	0.0223	0.0004	0.0002	*0.0743
D ₆	0.0000	0.0223	0.0004	0.0002	*0.0743
D ₇	0.0023	0.0223	0.0001	0.0014	*0.8762
D ₈	0.0000	0.0000	0.0004	0.0002	*0.0011
D ₉	0.0001	0.0002	0.0004	0.0000	*0.0024
D ₁₀	0.0060	*0.0066	0.0001	0.0005	0.0043
D ₁₁	*0.0180	0.0164	0.0001	0.0001	0.0080
D ₁₂	*0.0214	*0.0214	0.0004	0.0001	0.0080
平均	0.0931	0.1723	0.1261	0.0240	*0.1871
標準偏差	0.1741	0.2578	0.2681	0.0499	0.2426
最適数	2	3	3	0	5

最適数と平均再識別率が最大である

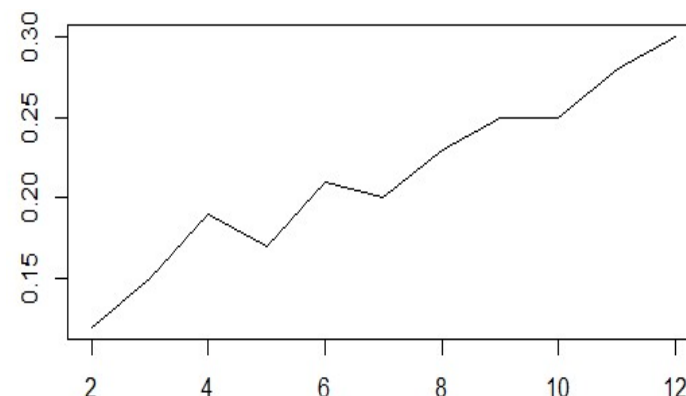
EUC1の処理性能評価

処理時間 (s)



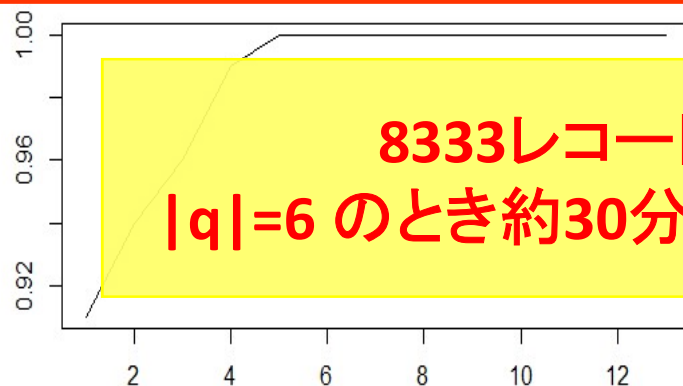
QIのサイズ |q|

処理時間 (s)

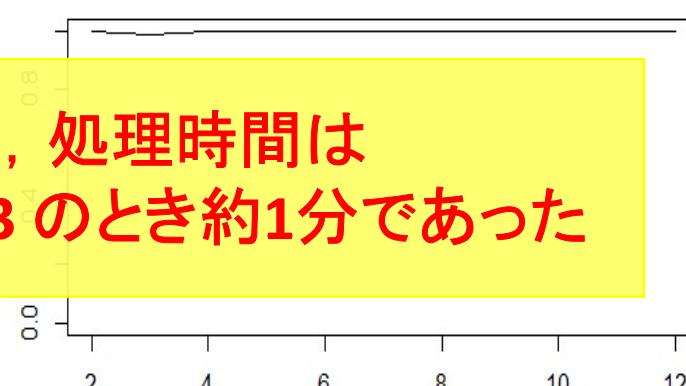


SAのサイズ |s|

再識別率



QIのサイズ |q|



SAのサイズ |s|

8333レコードの場合、処理時間は
 $|q|=6$ のとき約30分, $|q|=13$ のとき約1分であった

処理性能評価は小規模データ(100レコード, 25属性)を用いて行った

実験2: 匿名加工データの解析

単一の手法を用いて加工した小規模匿名加工データ $D_A, \dots, D_H$ を用いて $D_1, \dots, D_{12}$ の加工手法を予測する.

データ名	匿名加工手法
D_1	?
D_2	?
D_3	?
D_4	?
D_5	?
D_6	?
D_7	?
D_8	?
D_9	?
D_{10}	k-匿名化+SA平均化
D_{11}	?
D_{12}	?

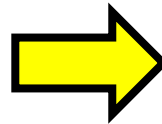
データ名	匿名加工手法	加工対象
D_A	k-匿名化	QI
D_B	SAノイズ付加	SA
D_C	山岡匿名化	ID
D_D	QI統一 (対象外)	QI
D_E	QI統一 (対象内)	QI
D_F	SA平均化	SA
D_G	QI内スワップ	SA
D_H	レコード削除	レコード

匿名加工手法の例

・k-匿名化

QI1	QI2	QI3	SA1	SA2
2	1	1	100	100
2	1	2	200	400
1	1	1	300	200
1	1	2	400	500

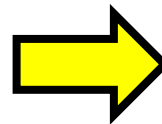
2-匿名化



QI1	QI2	QI3	SA1	SA2
2	1	1	100	100
2	1	1	200	400
1	1	1	300	200
1	1	1	400	500

・SA平均化 (マイクロアグリゲーション)

QI1	QI2	QI3	SA1	SA2
2	1	1	100	100
2	1	1	200	400
1	1	1	300	200
1	1	1	400	500



QI1	QI2	QI3	SA1	SA2
2	1	1	150	250
2	1	1	150	250
1	1	1	350	350
1	1	1	350	350

期待される効果

		対象	匿名加工手法							
			k-匿名化	ノイズ付加	YA	QI統一 (対象外)	QI統一 (対象内)	SA 平均化	QI内 スワップ	レコード 削除
有用性	U1	SA	—	△	—	—	—	—	—	×
	U2	QI,SA	×	△	—	—	×	—	—	×
	U3	QI	×	△	—	—	×	—	—	×
	U4	SA	—	△	—	—	—	×	×	×
	U5	SA	—	△	×	—	—	×	×	×
	U6	行数	—	—	—	—	—	—	—	×
安全性	E1	SA	×	△	○	×	×	×	×	×
	E2	QI,SA	△	×	○	△	△	×	△	×
	E3	SA	×	△	○	×	×	○	○	×
	E4	SA	×	△	○	×	×	○	△	×
	EUC1	QI,SA	△	×	○	△	△	×	○	×

SAを加工する手法であるため
SAが対象の有用性を損ない、
SAが対象の攻撃に対して強い

手法の組み合わせによる効果

	DA	DF	D10
加工手法	k-匿名化	SA平均化	k-匿名化+SA平均化
U1	－	－	－
U2	×	－	×
U3	×	－	×
U4	－	×	×
U5	－	×	△
U6	－	－	－
S1	○	×	○
S2	○	×	○
E1	△	×	△
E2	△	×	△
E3	×	○	○
E4	×	○	○
EUC1	△	×	○

実験2結果：匿名加工データの予測結果

[illegible]

実験2結果：匿名加工データの予測結果

		匿名加工データ											
		D1	D2	D3	D4	D7	D5	D6	D8	D9	D10	D11	D12
有用性	U1	－	－	－	－	－	－	－	－	－	－	－	－
	U2	×	×	×	×	×	－	－	－	－	×	×	×
	U3	×	×	×	×	△	－	－	－	－	×	×	△
	U4	－	△	△	△	△	△	△	△	△	×	×	×
	U5	－	△	△	△	△	△	△	△	△	△	△	△
	U6	－	－	－	－	－	－	－	－	－	－	－	－
安全性	S1	×	×	△	△	△	×	×	×	×	○	○	△
	S2	×	×	△	△	○	△	○	○	○	○	○	○
	E1	△	×	×	×	○	△	○	○	○	△	△	△
	E2	×	×	×	×	△	△	○	○	○	△	△	△
	E3	×	○	×	×	○	△	○	○	○	○	○	○
	E4	×	○	△	△	○	△	○	○	○	○	○	○
	EUC1	×	×	×	×	×	△	△	○	○	○	○	○
匿名加工手法	DA	－	－	○	○	○	－	－	－	－	○	○	○
	DB	－	－	－	－	－	－	－	－	－	－	－	－
	DC	－	－	－	－	－	○	○	○	○	○	○	○
	DD	－	－	－	－	○	○	○	－	－	○	○	○
	DE	○	－	－	－	－	－	－	○	○	○	○	○
	DF	－	○	－	－	－	－	－	－	－	○	○	○
	DG	－	－	○	○	○	－	－	○	○	－	－	－
	DH	－	－	－	－	－	－	－	－	－	－	－	－

グループ1
EUC1が有効

グループ2
山岡匿名化+他手法

グループ3
k-匿名化+SA平均化

実験2結果：匿名加工データの予測結果

		匿名加工データ											
		D1	D2	D3	D4	D7	D5	D6	D8	D9	D10	D11	D12
有用性	U1	—	—	—	—	—	—	—	—	—	—	—	—
	U2	×	—	×	×	×	—	—	—	—	×	×	×
	U3	×	—	△	△	△	—	—	—	—	×	×	△
	U4	—	△	—	△	△	△	△	△	△	×	×	×
	U5	—	△	△	△	△	△	△	△	△	△	△	△
	U6	—	—	—	—	—	—	—	—	—	—	—	—
匿名加工手法	S1	×	×	△	△	△	×	×	×	×	○	○	△
	S2	×	×	△	△	○	△	○	○	○	○	○	○
	EUC1	×	×	×	×	×	△	△	○	○	○	○	○
	DA	—	—	○	○	○	—	—	—	—	○	○	○
	DB	—	—	—	—	—	—	—	—	—	—	—	—
	DC	—	—	—	—	—	○	○	○	○	—	—	—
	DD	—	—	—	—	○	○	○	—	—	—	—	—
	DE	○	—	—	—	—	—	—	○	○	—	—	—
	DF	—	○	—	—	—	—	—	—	—	○	○	○
	DG	—	—	○	○	○	—	—	○	○	—	—	—
	DH	—	—	—	—	—	—	—	—	—	—	—	—

**D_8 は本戦1位の匿名加工データ
山岡匿名化と他手法を組み合わせた
データ(グループ2)が上位に多く見られた**

まとめと今後の課題

- まとめ

- 既存手法の問題点を改善した, ユークリッド距離を用いる手法を提案した. 12個の匿名加工データのうち, 5個で最高の再識別率である.
- 提出された12個のデータで用いられていた匿名加工手法を明らかにした. 山岡匿名化と他手法を組み合わせたデータが上位に多い.

- 今後の課題

- identify.eucのさらなる改善
- 新たな再識別, 匿名加工手法の開発

- 謝辞

- データを提供していただいた企業と大学の皆様に感謝いたします.

山岡匿名化について

山岡匿名化

データのIDを入れ替える匿名加工手法

元データ

ID	QI1	QI2	QI3	SA1	SA2
1	2	1	1	100	100
2	2	1	1	200	400
3	1	1	2	300	200
4	1	1	2	400	500

匿名加工データ

ID	QI1	QI2	QI3	SA1	SA2
2	2	1	1	100	100
3	2	1	1	200	400
4	1	1	2	300	200
1	1	1	2	400	500