PWSCUP 2017
コンテスト振り返り

菊池浩明(明大)

2017年5月30日

- 改正個人情報保護法の全面施行

朝日新聞　２０１７年５月２９日　朝刊　２ページ　東京本社

ビッグデータの活用 後押し

改正個人情報保護法 あす施行

改正個人情報保護法の主なポイント

特定できないよう個人情報を加工し、復元もできなくする

新設 **匿名加工情報**
外部提供の本人同意は不要

情報を自由に流通させられるようになる

名前 ○○○○	識別できる記述	全部または一部を削除
生年月日 XXXX年○月○日		
年齢 (例)116歳	推定できる特異な記述	削除
病歴 症例数が極めて少ない		
マイナンバー XXXX XXXX XXXX	個人識別符号	すべて削除
運転免許証の番号 XXXX XXXX XXXX		

新設 **要配慮個人情報**
原則、本人同意を義務づけ
人種や信条、病歴、犯歴 → 差別や偏見につながるため **区別**

新設 不正な利益を図る目的での個人情報の盗用や流用に罰則

新設 「個人情報保護委員会」の設置。監視・監督などの権限を一元化

適用対象に 取り扱う個人情報が5千人分以下の中小企業や自治会など

スマホ用アプリ「TRAVEL JAPAN Wi-Fi」の画面。担当者は「改正法の施行で新たなビジネスが生まれる」と期待する＝東京都中央区

情報匿名化し提供 可能に

個人特定 懸念なお

改正個人情報保護法が30日、全面施行される。ビッグデータの利活用を後押しするのが狙いの一つで、個人を特定できないように加工した情報であれば自由に流通させられるようになる。一方、保護規制が強化され、小規模事業者も法の適用対象となることに戸惑いが広がっている。

スマートフォンのアプリを立ち上げると、近くの百貨店や観光情報が出てくる。KDDIのグループ会社「ワイヤ・アンド・ワイヤレス」が訪日外国人向けに提供する「TRAVEL JAPAN Wi-Fi」だ。全国20万カ所以上の公衆無線LANが無使え、観光地や店舗情どの広告も表示する。

位置情報の取得への利用の前提で、累計ンロード数は約200個人の行動履歴ではな端末IDから得られる訪問先、滞在時間なビッグデータを「統計で丸めて」自治体など売している。

責任者の川名義輝マジャーは改正法で新た場が生まれるのを期待る。本人を特定できなう情報を「匿名加工」ば、より細かいデータてコンサルティング会どからの需要があるか

JIPDEC匿名加工情報事例集

- 2017年7月7日．日本情報経済社会推進協会
- 認定個人情報保護団体として，会員の個々の事例について，整理したもの．参考とする（保護指針に沿っているが独立）

事例	事業者	データ例	備考
1. 所有車データ提供	整備工場が，自動車販売店に対して提供	顧客（数万） 車両（数万） 整備（数百万）	
2. 顧客データ提供	質屋が調査会社に提供	顧客（数千） 取引（数十万）	
3. 購買履歴の提供	商店街が，新規出店事業者に	顧客（数千） 購買（数十万）	13カ月
4. 移動履歴	経路サービス事業者が，自治体の委託により，駐輪場事業者	顧客（数千） レコード（数千万）	5人以上の通行者がいる部分を可視化

事例3: 匿名加工の方法

■ 提供した形式（イメージ）

データ		処理
氏名	甲野太郎	削除
会員番号	1000001	削除
誕生日	1963/1/1	50代
性別	男	男
住所	S区AA町４丁目19番地	S区AA町４丁目

■ 安全サイドへマージンの大きい加工.

- 13ヵ月の履歴だが，会員番号は削除されていて，個人の識別は困難と考える.
- 商品推薦や相関ルールマイニングなどの応用は不能で，有用性は低い.

会員番号	0000001
購買店	D魚店
日時	2016/12/5　11:40
購買総額	300円
付与ポイント	3

加工

購買日時	火曜日　14-15時
購買日時	月曜日　14-15時
購買日時	月曜日　11-12時
購買店	食品
購買総額	201～400円

仮IDの制御

- 規則第19号（1号）

規則第19 条（第1 号）
(1) 個人情報に含まれる特定の個人を識別することができる記述等の全部又は一部を削除すること（当該全部又は一部の記述等を復元することのできる規則性を有しない方法により他の記述等に置き換えることを含む。）。

＝仮名化

仮ID を付す場合には、元の記述を復元することのできる規則性を有しない方法でなければならない。
（中略） なお、同じ乱数等の他の記述等を加えた上でハッシュ関数等を用いるなどの手法を用いる場合には、乱数等の他の記述等を通じて復元することができる規則性を有することとならないように、提供事業者ごとに組み合わせる記述等を変更し、定期的に変更するなどの措置を講ずることが望ましい。

研究目的

- 有用性高く安全な匿名加工方式の追求

(1) 最大知識攻撃者

- 部分知識攻撃者モデル

(2) 現実との乖離

- 加工方法明示
- 加工規則19条対応
- アイテムベース協調フィルタリング

(3) 長期間の履歴データ

- データ分割と仮名の制御

PWSCUP匿名加工コンテスト

	2015	2016	2017
開催	10/21-22 長崎ブリックホール	10/11-12 秋田キャッスルホテル	10/23-24 山形国際ホテル
参加者数	13チーム(20名）	15チーム(42名）	14チーム(43名）
データセット	疑似ミクロデータ （世帯消費額）	UCI Dataset “Online Retail” (購買履歴）	
属性数	25	11 （顧客4属性＋履歴7属性）	
顧客数	8,333	400	500
履歴数	なし	18,524	44,917
履歴期間	1年間	1年間	12カ月

本戦 (2017年10月23日山形国際ホテル

公開データセット Online Retail

- UCI Machine Learning Repository で公開されているデータ
- 英国のオンライン店舗での、2010年12月から約1年分の購買履歴
  - 主な製品：贈り物
  - 主な顧客：卸売り業者
    - » 競技では個人とみなす
  - 約54万行

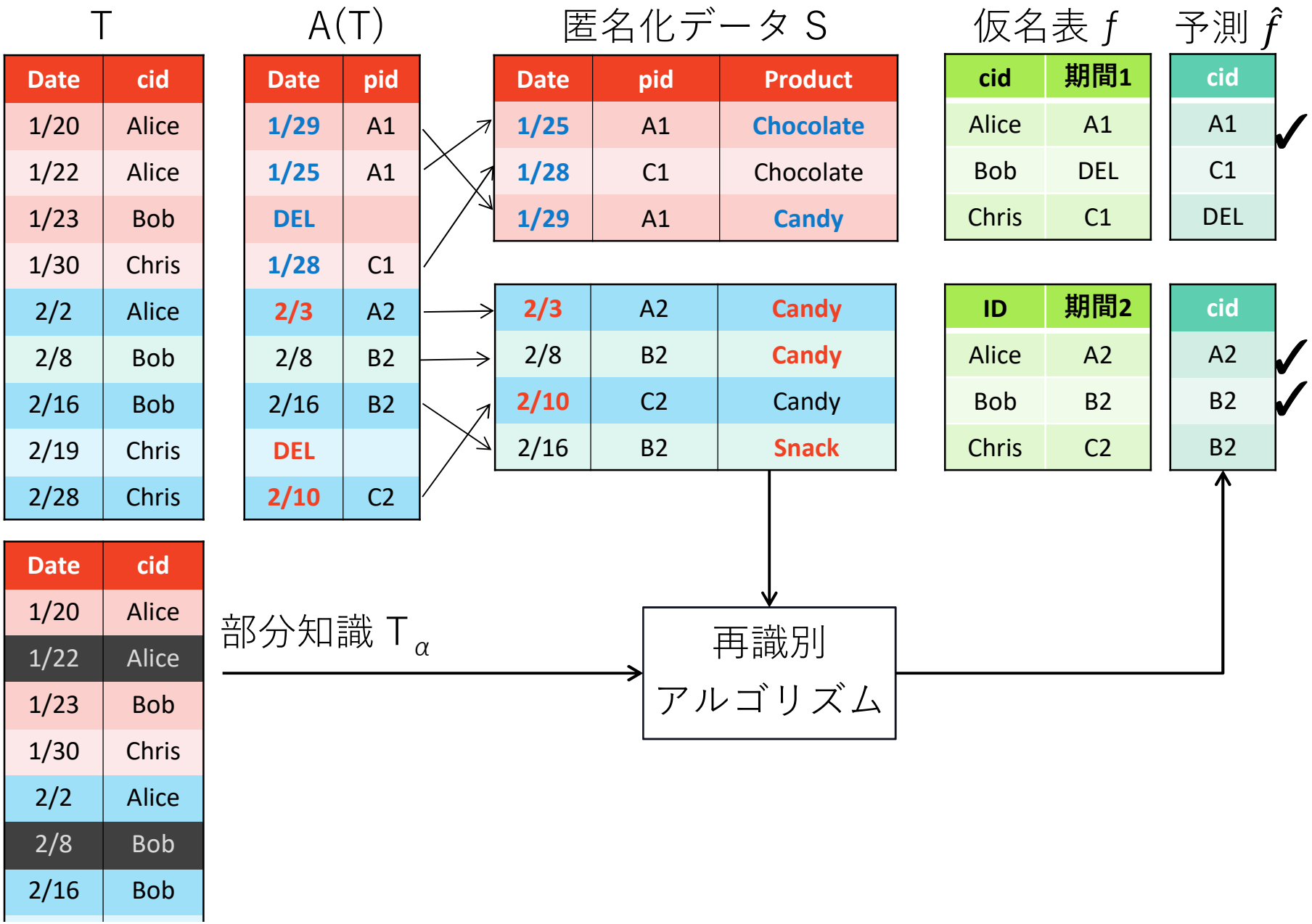
Dataset ‘Online Retail’

- 顧客表（マスター）*M*
  - n = 500 人
  - From 4 国
- 購買履歴（トランザクション） *T*
  - m = 38,087 レコード
  - 2,781 商品

顧客 ID	Sex	誕生日	国籍
Online retail	合成		Online retail
12360	M	1876/2/24	UK
12361	F	1954/2/14	Germany
12362	F	1963/12/2	France
12364	F	1960/9/16	Others

顧客 ID	伝票ID	日付	時刻	商品番号 Stock Code	単価 Unit Price	個数
12362	0	2011/2/17	10:30	21913	3.75	4
12362	0	2011/2/17	10:30	22431	1.95	6
12361	0	2011/2/25	13:51	22630	1.95	12
12361	0	2011/2/25	13:51	22326	2.95	6

再識別者


T
A(T)
匿名化データ S
仮名表 f
予測 $\hat{f}$
Date cid
1/20 Alice
1/22 Alice
1/23 Bob
1/30 Chris
2/2 Alice
2/8 Bob
2/16 Bob
2/19 Chris
2/28 Chris
Date pid
1/29 A1
1/25 A1
DEL
1/28 C1
2/3 A2
2/8 B2
2/16 B2
DEL
2/10 C2
Date pid Product
1/25 A1 Chocolate
1/28 C1 Chocolate
1/29 A1 Candy
2/3 A2 Candy
2/8 B2 Candy
2/10 C2 Candy
2/16 B2 Snack
cid 期間1
Alice A1
Bob DEL
Chris C1
cid
A1
C1
DEL
ID 期間2
Alice A2
Bob B2
Chris C2
cid
A2
B2
B2
Date cid
1/20 Alice
1/22 Alice
1/23 Bob
1/30 Chris
2/2 Alice
2/8 Bob
2/16 Bob
部分知識 T_α
再識別
アルゴリズム

(1) 部分知識攻撃者モデル

C ID	Date	S Code	Price	Qt
12362	2/17	21913	3.75	4
12362	2/17	22431	1.95	6
12361	2/25	22630	1.95	12
12361	2/25	22326	2.95	6
15100	12/1	21258	10.95	32
15100	12/1	22197	0.85	6
12431	12/1	22941	8.5	6
12431	12/1	21622	4.95	8

T

最大知識モデル

C ID	Date	S Code	Price	Qt
12362	2/17	21___	3.75	4
12362	2/17	22___	1.95	6
15100	12/1	21___	10.95	32
12431	12/1	22___	8.5	6

T_{50}

C ID	Date	S Code	Price	Qt
12362	2/17	21___	3.75	4
15100	12/1	21___	10.95	32

T_{25}

部分知識モデル

(2.1) 加工方法の明示

顧客ID	Date	S Code	Price	Qt
12362	2/17	21913	3.75	4
12362	2/17	22431	1.95	6
12362	3/25	22630	1.95	12
12362	3/25	22326	2.95	6
15100	12/1	21258	10.95	32
15100	12/1	22197	0.85	6
12431	12/1	22941	8.5	6
12431	12/1	21622	4.95	8

システムでスワップし，
山岡匿名化（レコードスワップ）は無効に

仮名ID	Date	S Code	Price	Qt
62	2/17	21913	3.75	4
62	2/17	22431	1.95	8
63	3/25	22630	1.95	12
63	3/3	22326	3.00	6
DEL				
15	12/1	**22326**	0.85	6
DEL				
31	12/1	21622	4.95	**0**

オリジナル履歴データ
T

$|T| = |A(T)|$

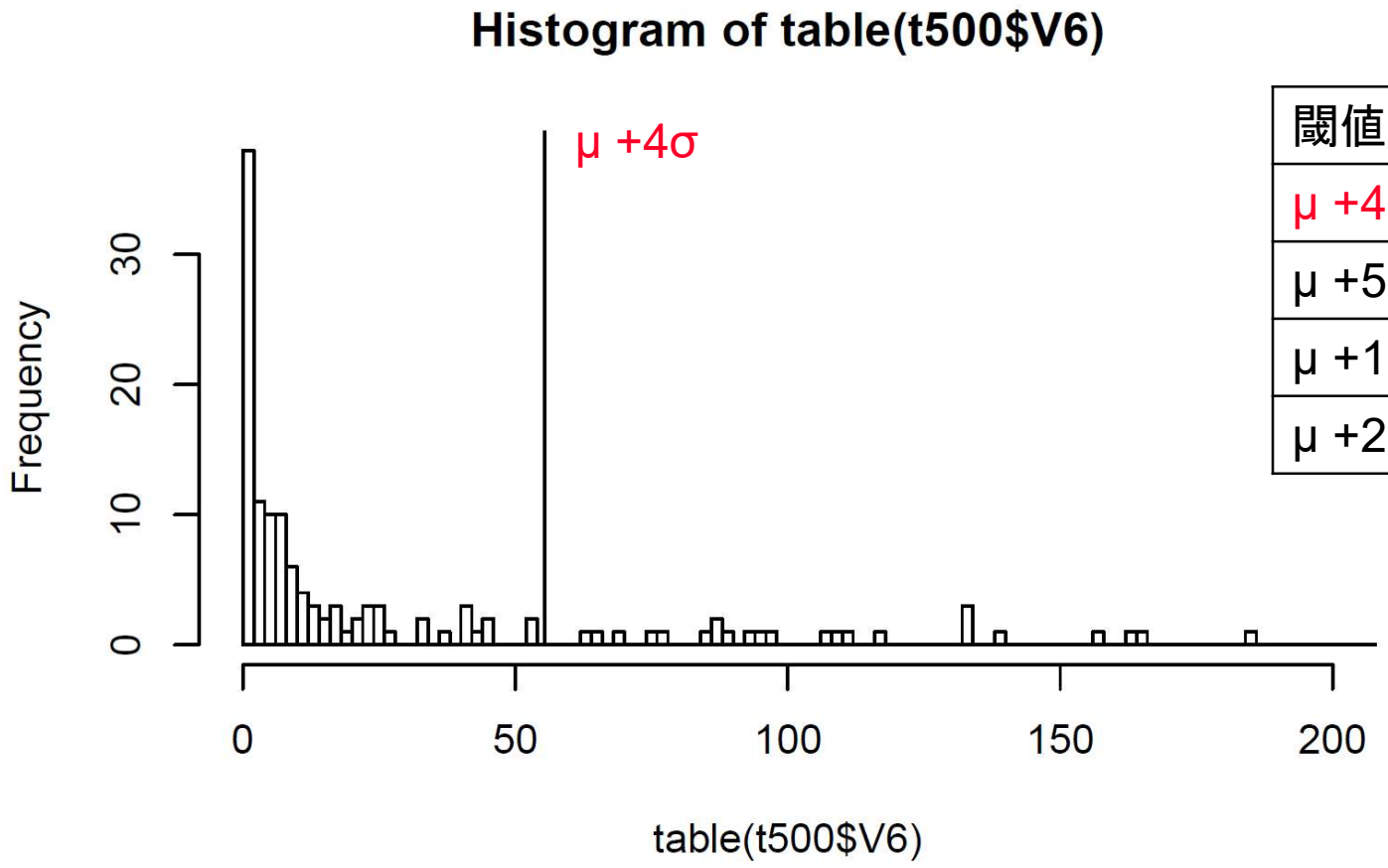
匿名加工データ
A(T)

(2.2) 加工規則19条対応

規則	対象	解釈	措置
(1)	個人情報	各チームで検討	
(2)	個人識別符号	該当なし	
(3)	連結する符号	顧客ID	削除（ルール6) Mのいかなる顧客とも一致するIDが存在しないようにする
(4)	特異な記述等を削除	各チームで検討	（トップコーディングか？）
(5)	他の記述等との差異と性質を勘案し，適切な措置	各チームで検討	（k-匿名化か?）

「特異な記述」の解釈

- 商品の価格分布



閾値	個数
μ +4σ	31
μ +5σ	30
μ +10σ	16
μ +20σ	10

(3) 長期間の仮名化

- 仮名表: 匿名加工で月ごとに割り当てた仮名

仮名ID	Date	S Code	Price	Qt
62	1/17	21913	3.75	4
62	1/17	22431	1.95	8
63	2/25	22630	1.95	12
63	2/**3**	22326	3.00	6
15	4/1	**22326**	0.85	6
DEL				
62	6/5	21622	4.95	**0**

顧客 12361の加工

C ID	1	2	3	4	5
12361	62	63		15	DEL

仮名表 F

$$F = \begin{pmatrix} c_{1,1} & f^{(1)}(c_{1,1}) & \cdots & f^{(12)}(c_{1,1}) \\ \vdots & & \ddots & \vdots \\ c_{n,1} & f^{(1)}(c_{n,1}) & \cdots & f^{(12)}(c_{n,1}) \end{pmatrix}$$

再識別の定義

- 「再識別」
  - 全ての区間で仮名の顧客ID(個人情報)を正しく推測すること

C ID	1	2	3	4
12361	61	61	61	63
12362	62	62	DEL	DEL
12431	31	DEL		31
12628		28		
15100	10	20		40

仮名表 F

C ID	1	2	3	4	
12361	**61**	**61**	**61**	**63**	✓
12362	62	**77**	**77**	DEL	
12431	**31**	**DEL**	**DEL**	**31**	✓
12628	**DEL**	**28**	**28**	**DEL**	
15100	**10**	**20**	**DEL**	**99**	

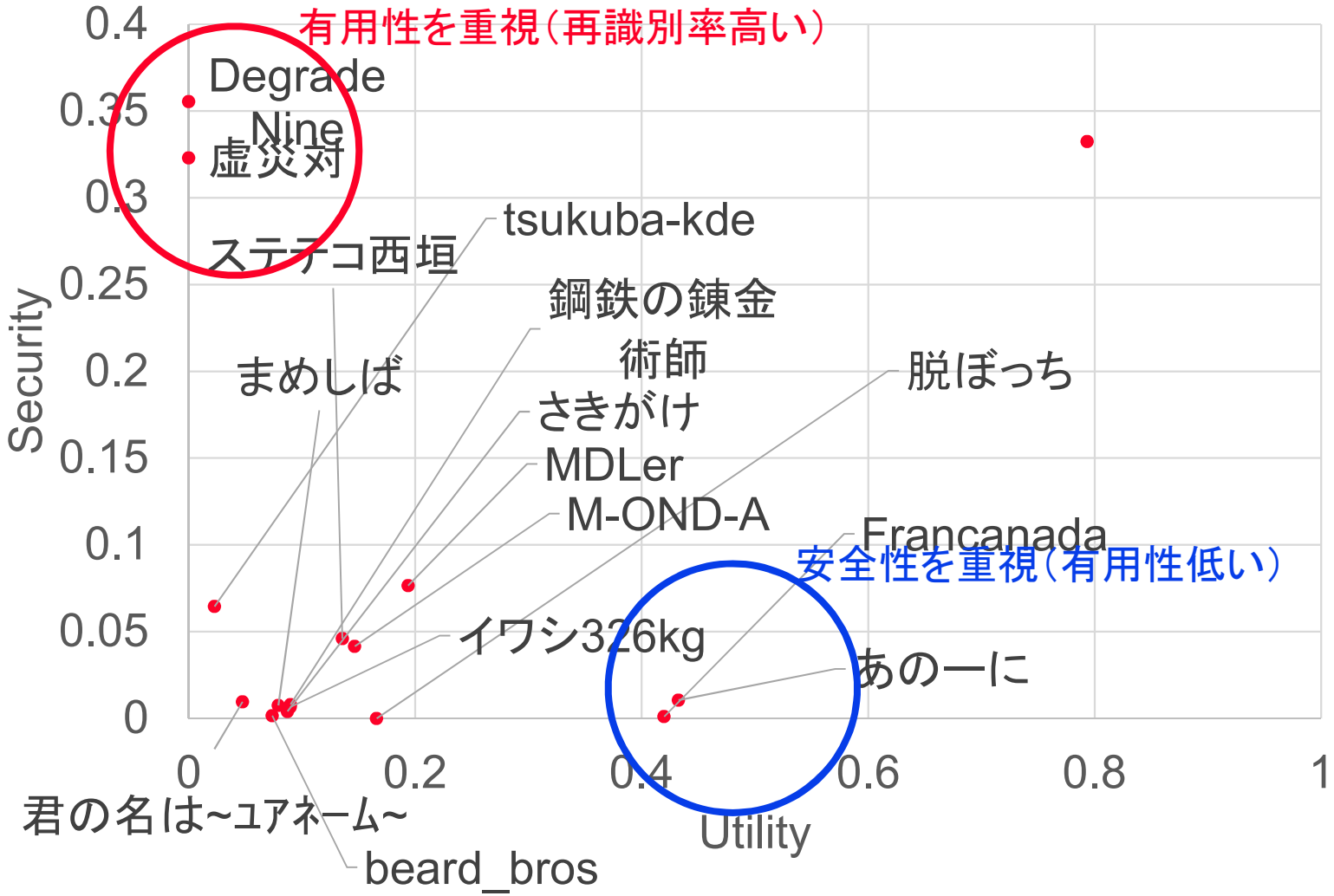
推測 ^F

Reid(F,^F) = 正しい推測顧客数/n = 2/5

PWSCUP2017本戦結果

No	Team	Utility	Security	Average
1	君の名は~ユアネーム~	0.048	0.016	0.032
2	beard_bros	0.074	0.031	0.052
3	イワシ326kg	0.090	0.016	0.053
4	さきがけ	0.087	0.029	0.058
5	ステテコ西垣	0.136	0.046	0.091
6	M-OND-A	0.147	0.127	0.137
7	鋼鉄の錬金術師	0.090	0.241	0.165
8	脱ぼっち	0.166	0.199	0.183
9	あの一に	0.433	0.062	0.247
10	tsukuba-kde	0.023	0.733	0.378
15	Francanada	0.420	0.590	0.505

有用性と安全性の関係


有用性を重視（再識別率高い）
安全性を重視（有用性低い）
Degrade Nine
虚災対
ステテコ西垣
tsukuba-kde
鋼鉄の錬金術師
まめしば
さきがけ
MDLer
M-OND-A
脱ぼっち
Francanada
イワシ326kg
あのーに
君の名は~ユアネーム~
beard_bros
Security
Utility
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1

Questions

- 機械と人力による再識別率には差があるか？
- 仮名の割り当て期間は長いほど再識別されやすくなるか？
- 顧客当たりの仮名割り当て数と再識別リスクの関係は？

再識別率

システムで評価した再識別率よりも，
人が互いに競った再識別率が高い．平均84%の差．

平均仮名長

- 仮名表

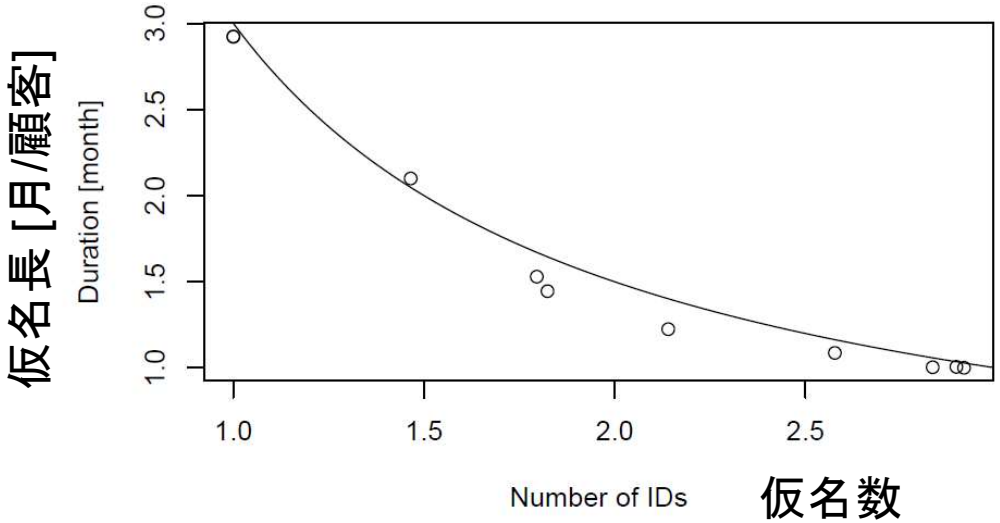
表 5: 仮名長の算出例

a	483	DEL	540	DEL	DEL	540	960	483
b	483		540			540	960	483
c	1		2				1	1

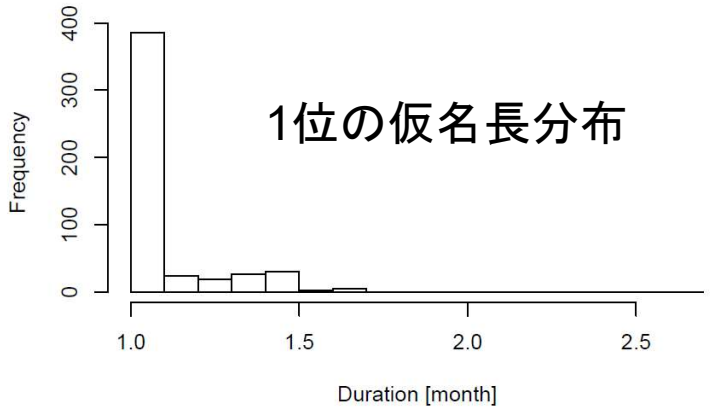
  - P = (483, 540, 960, 483')
- 仮名数
  - |P| = 4
- 平均仮名長 [月]

$$\mu(P) = \frac{1+2+1+1}{4} = 1.25$$

仮名長と仮名数

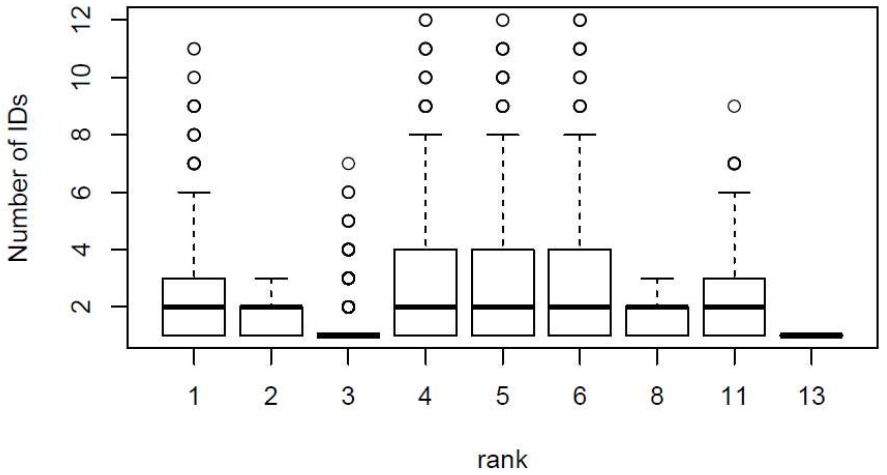


仮名数

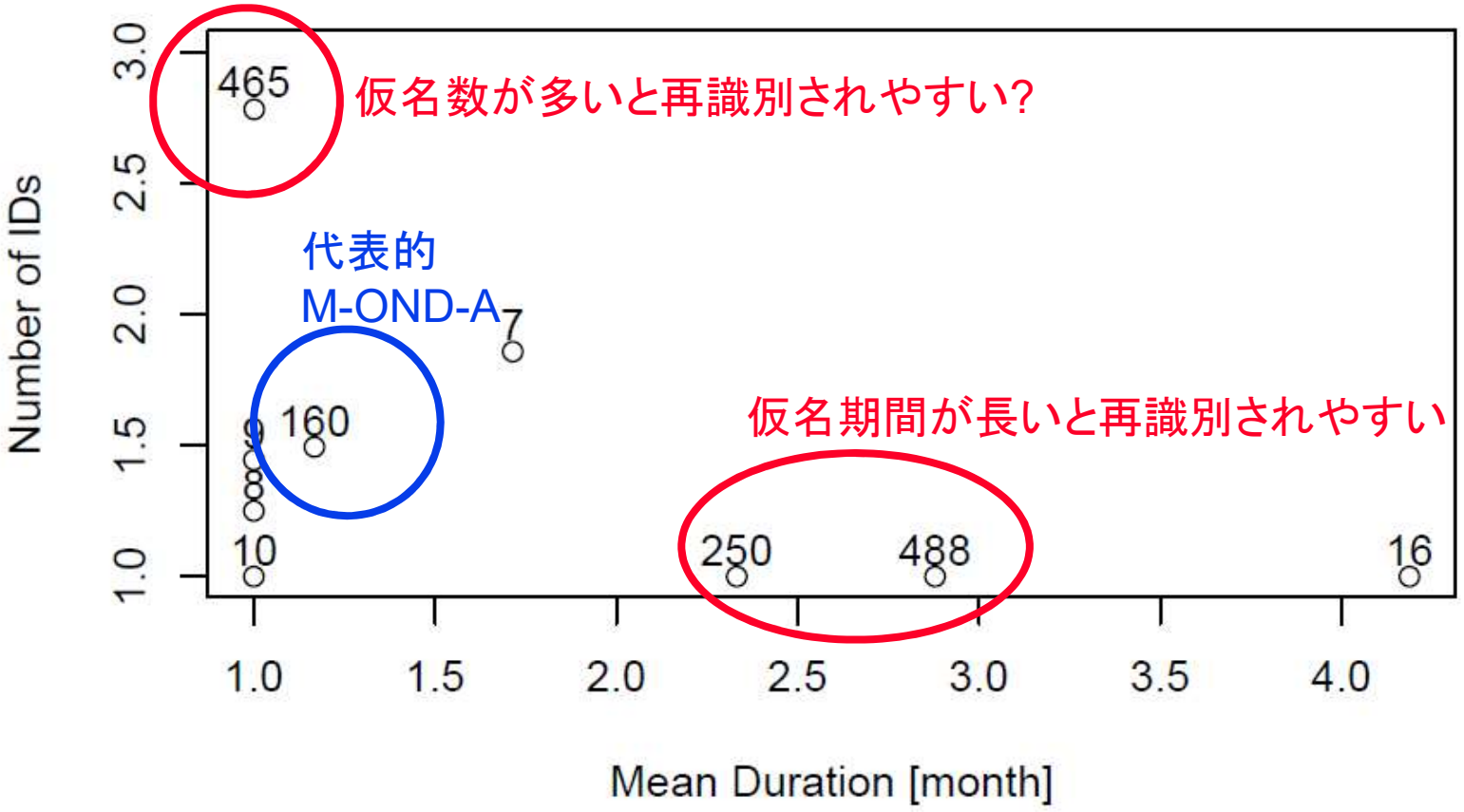


1位の仮名長分布

チームごとの仮名数

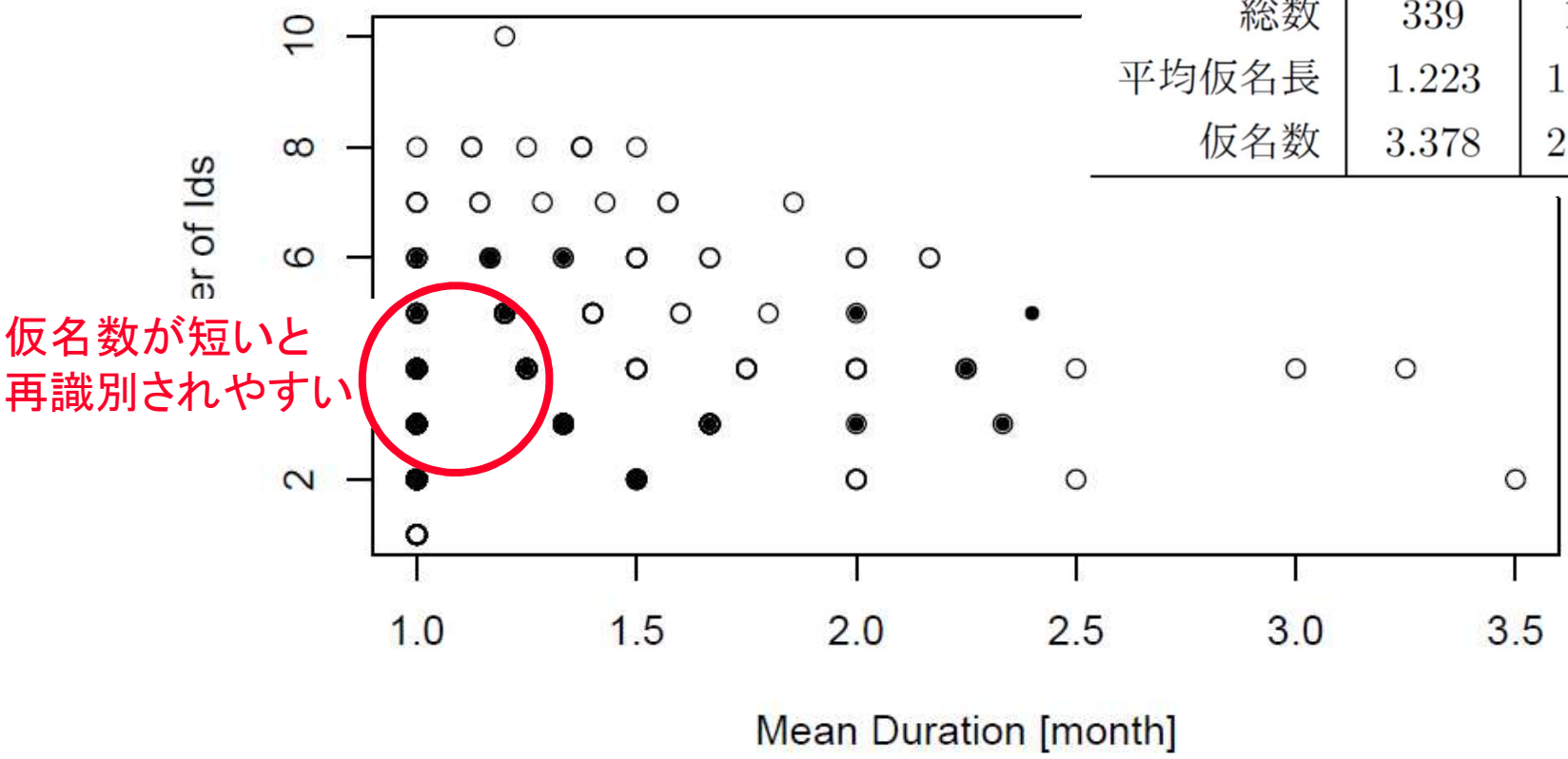


再識別された顧客数(チーム毎)



匿名加工データ M-OND-A

項目	非識別	識別
総数	339	161
平均仮名長	1.223	1.104
仮名数	3.378	2.491



ロジスティック回帰

- 回帰結果

表 8: ロジスティック回帰 (M-OND-A)

変数	係数	標準エラー	統計量 z	$Pr(>\|z\|)$
定数	1.32668	0.43034	3.083	0.00205^{**}
len	−0.46118	0.09212	−5.006	$5.56e-07^{***}$ 水準0.01で有意
mean	−0.64639	0.37055	−1.744	0.08109

- オッズ比
  - $OR = \frac{Pr(識別|len=1)}{Pr(識別|len=0)} = e^{-0.46} = 0.63$

仮名数を1増やすと
再識別リスクが63%に下がる

結論

- 法改正により匿名加工情報による個人情報活用が始まった
- 有用性と安全性の間にトレードオフがある
- システムと手動で再識別率に83%の差がある. リスク評価の困難さを意味する.
- 長期間の履歴データに対して，データ分割して異なる仮名を割り当てることにより再識別リスクが下がる. 仮名数を1増やすごとにリスクが63に下がる.