

匿名化リスク評価コンテスト PWS Cupの8年間

菊池浩明
明治大学



菊池 浩明Hiroaki Kikuchi



- 明治大学総合数理学部 教授
 - 大学院教務主任 / 情報基盤本部 副本部長
 - 理化学研究所革新知能統合研究センター客員研究員
 - JPCERT/CC代表理事
- 略歴
 - 明治大学卒, 富士通研究所, 東海大学情報通信学部教授、カーネギーメロン大学計算機科学科訪問研究員などを経て, 2013年より明治大学
 - 電子情報通信学会 ICSS研究会 専門委員会顧問
 - 情報処理学会 CSEC研究会顧問
 - 日本知能情報ファジィ学会理事

- 外部委員
 - JISAプライバシーマーク審査委員会・委員長
 - NICT映像センサー使用大規模実証実験検討委員会・委員長
 - 内閣官房IT戦略室 パーソナルデータに関する検討会技術検討WG・構成員
 - 内閣官房 健康医療戦略室・参事官

Agenda

- 背景
- コンテスト基本方針
- PWS Cup 2015-2022
- 我々はこれから何を目指すか

個人情報保護法

個人情報の保護に関する法律(平成十五年法律第五十七号)

- 2005年4月施行
- 2017年5月30日改正
 - 個人情報保護委員会(プライバシーコミッショナー)
 - 匿名加工情報導入
- 2020年(令和2年)改正・(令和3年改正)
 - 仮名加工情報
 - 官民統合

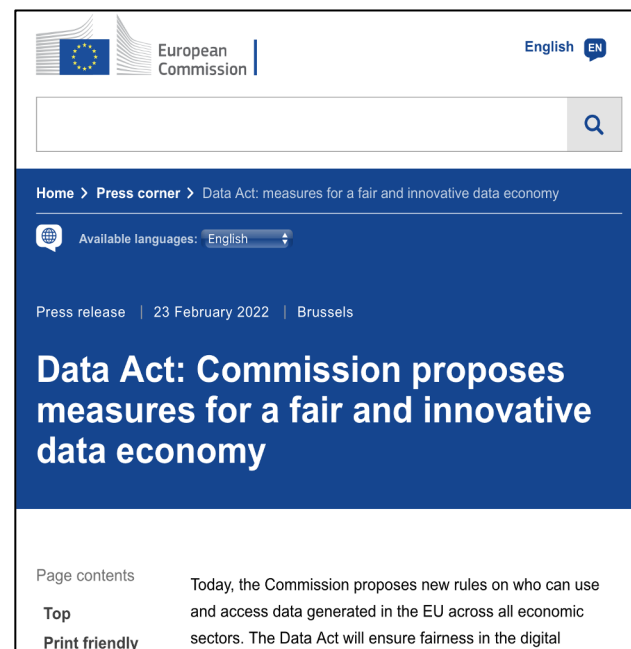
マイナンバーカード
PRキャラ「マイナちゃん」



GDPR (General Data Protection Regulation)

欧州一般個人情報保護規則

- 2018年5月25日施行
 - クッキーを含む個人情報のEEA域外への転移は違反
 - 個人の権利が中心
 - 「忘れられる権利」
 - 「プロファイリングを距離する権利」
- 2020年2月
 - データ法. (B2C/B2B/B2Gでのアクセスを可能とする)
 - データガバナンス法
 - データスペース法案(健康医療, EHDS)



背景：保護法の改正（平成29年）

- 1. 個人情報定義の明確化
 - 個人識別符号(第2条2項), 要配慮個人情報(第2条3項)
- 2. 個人情報の有用性確保
 - 匿名加工情報(第2条9項)
 - 認定個人情報保護団体による個人情報保護指針(第53条)
- 3. 個人情報の保護を強化
 - 個人情報データベース等提供罪(第83条)
- 4. 個人情報保護委員会の新設と権限
 - 新設(第5章)
- 5. 個人情報の取り扱いのグローバル化
 - 国境を越えた適用(第75条), 外国執行局への情報提供(第24条)
- 6. オプトアウトなど
 - 届出厳格化(第23条2項), 利用目的の変更禁止(第15条2項)

問題点

- 匿名加工の方法

- ☐ 個人情報保護指針？k-匿名が必要か？kはいくらか？ l -多様性は？ t -近似性は？差分プライバシー
- ☐ 有用性と安全性のトレードオフ

- リスクと安全性の評価

- ☐ 誰が評価するのか？個人情報保護委員会？認定個人情報保護団体？

- 加工方法やリスクのデータ依存

- ☐ 個人情報データを提供してくれない
- ☐ 研究者ごとに都合のいいデータ利用



問題3. 安全性の曖昧な基準

規則	対象	例
(1)	個人情報 ^① の全部又は一部を削除すること	氏名を削除
(2)	個人識別符号 ^② の全部を削除すること	マイナンバーを削除
(3)	個人情報と当該個人情報に措置を講じて得られる情報とを連結する符号 ^③ を削除する	管理用 ID を削除
(4)	特異な記述 ^④ 等を削除	116歳を90歳以上に置換
(5)	個人情報データベース等を構成する他の個人情報に含まれる記述等との差異その他の当該個人情報データベース等の性質を勘案し、その結果を踏まえて適切な措置 ^⑤ を講ずること	・自宅住所が推定できる位置情報削除 ・購入者が限定される特殊商品を一般カテゴリーに置換

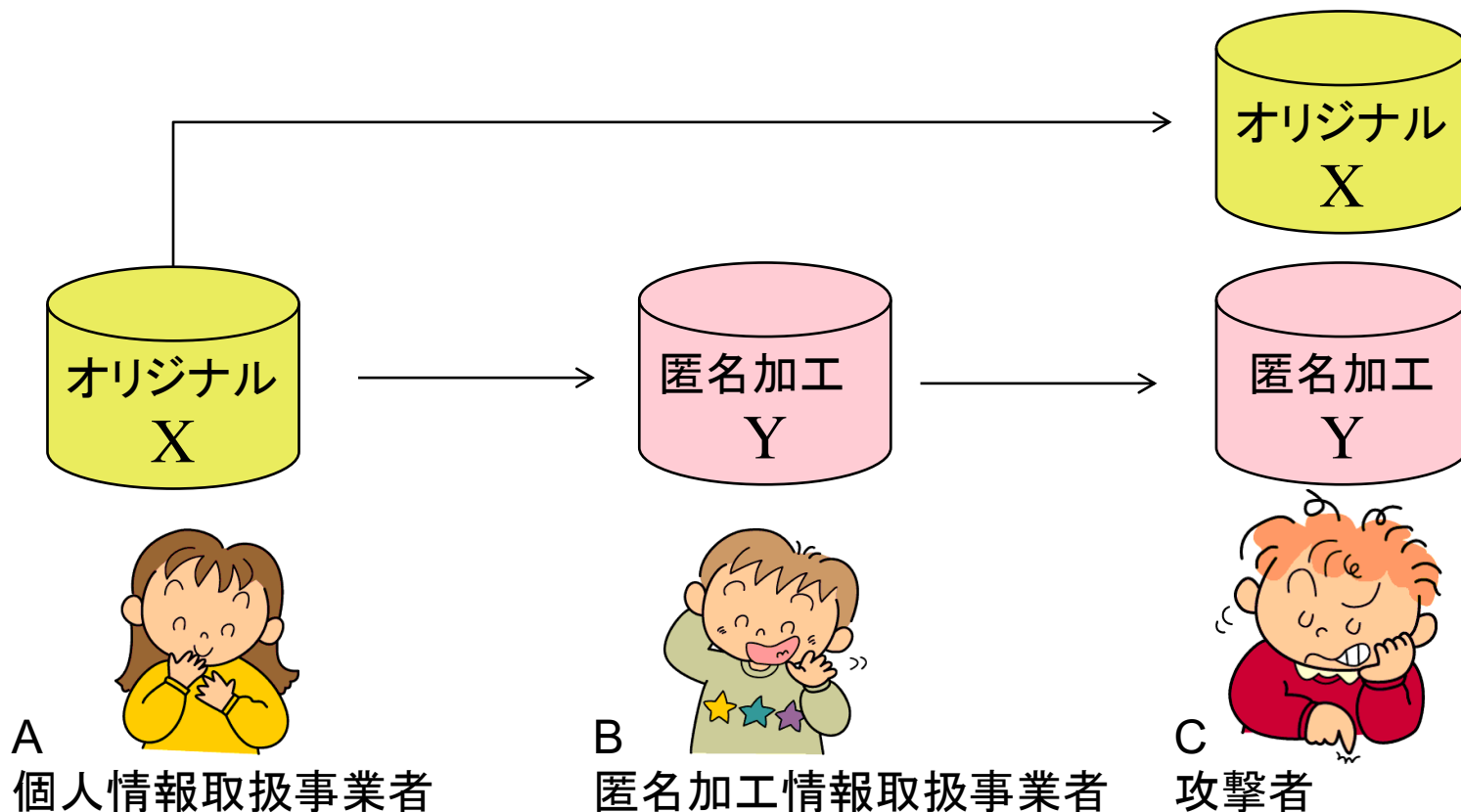
「個人情報保護ガイドライン(匿名加工情報編)」より抜粋

CUPの基本原則

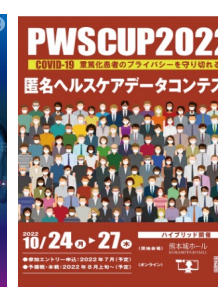
- 1. 共通のデータを使うこと
 - チームごとに合成データを出題する場合でも、合成の元になるデータは 共通とすること。原則的にオープンデータとする
- 2. 公平なルールで、客観的に勝ち負けを決めること
 - プレゼンテーション賞などの一部に主観評価は認めるが、総合優勝などは定量的に勝負を決めるものとする
- 3. 実践的なユースケースを想定すること
 - 学者の遊びになってしまわない様に、法制度を考慮して、実社会で有益なシナリオとする。
 - 法制度は国内を中心とするが、世界的な視野で考え、将来的な修正もスコープとする

加工者と攻撃者モデル

- 例) 世帯別消費動向をコンビニ出店計画に応用する業者とそれを悪用して再識別してゆする不正者



PWS Cup editions



2015	2016	2017	2018	2019	2020	2021	2022
10/21-22 長崎	10/11-12 秋田	10/23-24 山形	10/23-24 長野	10/21-24 長崎	10/26-29 オンライン	10/26-29 オンライン	10/24-27 熊本
13チーム (20名)	15チーム (42名)	14チーム (43名)	14チーム	21チーム	17	14	15
疑似マイクロデータ(世帯消費額)	UCI Dataset "Online Retail" (購買履歴)			ナイトレイ「疑似人流データ」合成位置	UCI Adult dataset	NHEANS dataset (ヘルスケア医療データ)	
25	11 (顧客4+履歴7)			11			

匿名加工・再識別コンテスト Ice & Fire の設計 2015

2015「擬似マイクロデータ」

■ (独)統計センター

- 申請書を提出することで誰でも無料で利用
- 平成16年度全国消費実態調査に基づく合成データ

データ セット	レコード 数	QI 数	SA 数	
	n		支出項目	収入項目
大規模データ	32,027	14	149	34
簡易データ	8,333	14	11	N/A

The screenshot shows the official website of the National Statistical Center (NSTAC) in Japan. The header includes the NSTAC logo and navigation links. The main content area is titled '業務案内' (Business Guide) and features a section for '擬似マイクロデータの利用' (Synthetic Microdata Utilization). This section explains that the center provides synthetic microdata for research and education, derived from the 2004 National Survey of Consumer Expenditure. It also mentions that users can apply for the data through an online system. A sidebar on the left lists various services, including '統計作成実務' (Statistical Production), '公的統計基盤サービス' (Public Statistical Basic Service), and '統計利用向けサービス及び公的基盤サービス' (Services for Statistical Utilization and Public Basic Service). The '公的統計のマイクロデータ利用' (Microdata Utilization of Public Statistics) link is highlighted.

<http://www.nstac.go.jp/services/giji-microdata.html#P2>

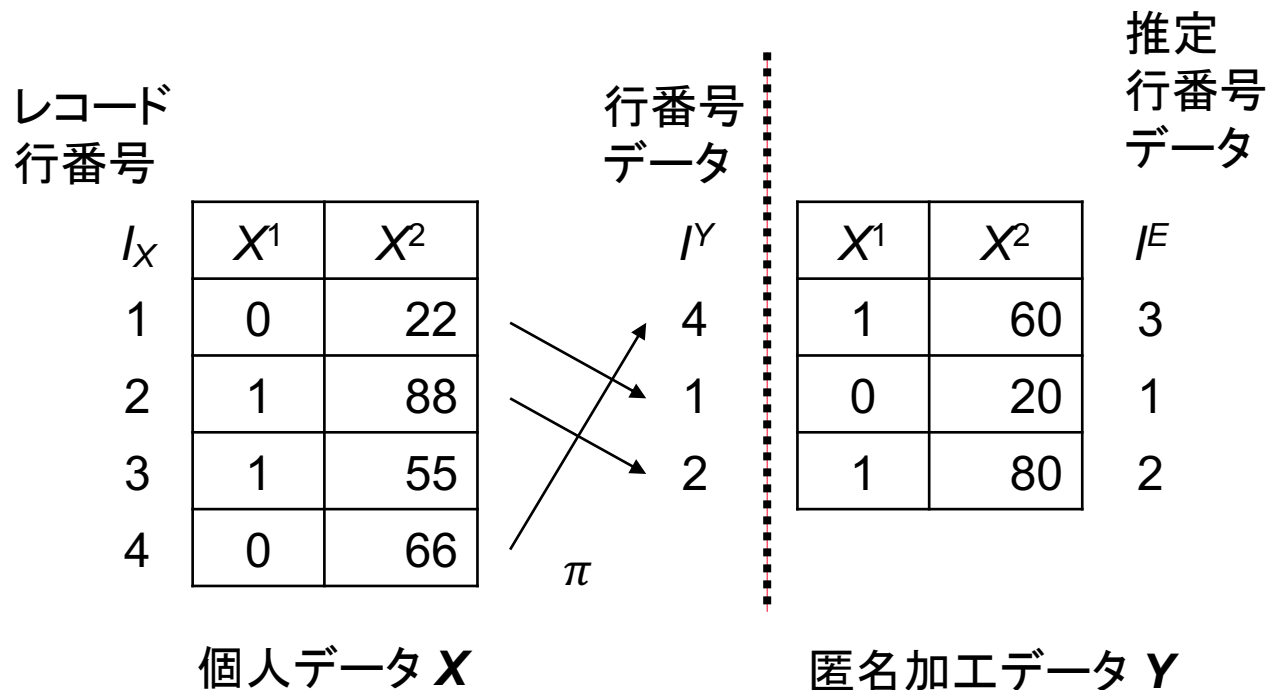
QAとSA（簡易データ）

	SetaiKubun	1=勤労	世帯区分
	SetaiJinin	名	人員
	ShuugyouJinin	名	有業人員
	Kouzou	1=木造	住居構造
	Tatekata	1=一戸建	建て方
	Shoyuu	1=持家	所有
	S1_Sex	1=男	世帯主性別
	S1_Age	2=25~19歳	年齢
	S1_Shuugyou	1=就業	就業
	S1_KigyokuKubun	1=民営	企業区分
	S1_KigyokuKibo	2=5~29人、3=30~	企業規模
	S1_Sangyou	5=建設業、6=製造業	産業符号
	S1_Shokugyou	1=常設作業者	職業
	Weight		集計用乗率
	Youto037		消費支出
	Youto038		食糧
	Youto079		住居
	Youto084		光熱
	Youto089		家具
	Youto099		衣類
	Youto117		医療
	Youto122		交通通信
	Youto129		教育
	Youto133		娯楽
	Youto142		その他

標準																								
#DIV/0! #DIV/0! #DIV/0! #DIV/0! #DIV/0! #DIV/0! #DIV/0! 1.7E+00																								
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25

int int int f f f int f f f f f f n n n n n n n n n n n
1 1 2 5 7 8 2 11 3 3 7 15 6

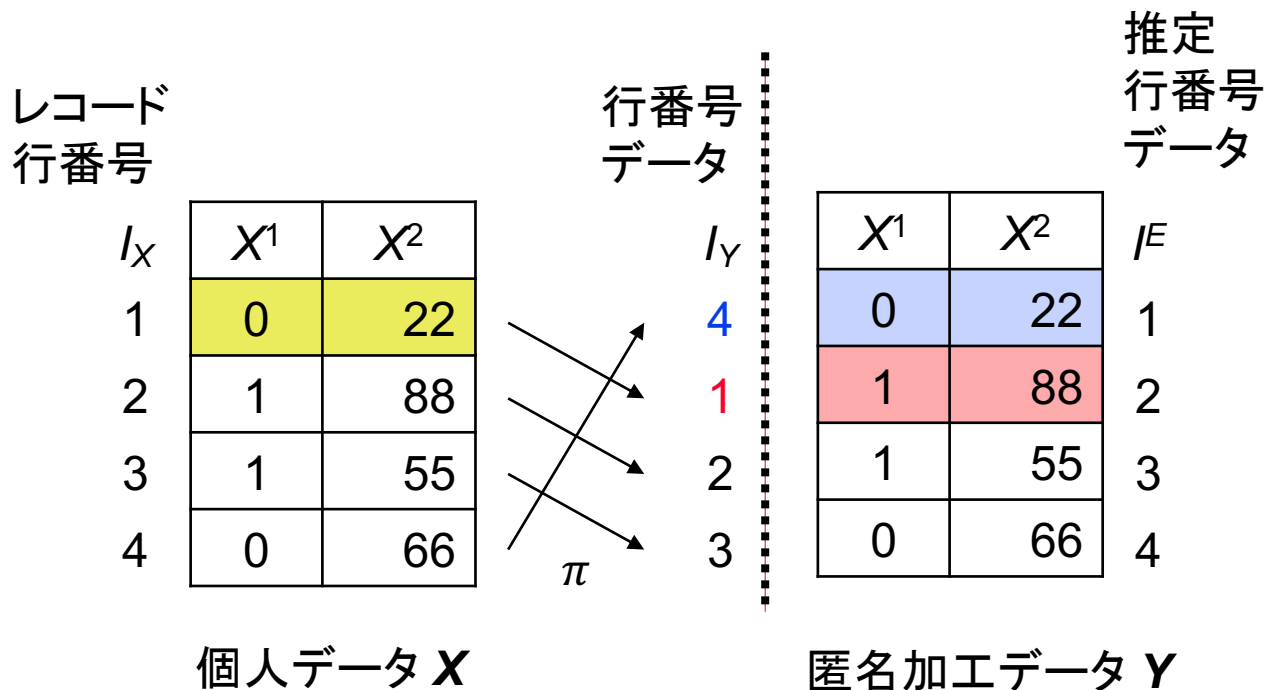
レコード再識別問題



再識別率

$$\text{Re-id } I^E(I^Y, I^E) = |\{j \text{ in } \{1, \dots, n'\} \mid i_j^Y = i_j^E\}| / n'$$

再識別問題の課題「山岡攻撃」



有用性定義

■ 匿名加工データYの有用性

$$\frac{1}{6} \sum_{i=1}^6 \text{Rank}(U_i)$$

U1	MeanMAE	SA={14,...,25}についての平均絶対誤差	$\text{meanMAE}(X, Y) = \frac{1}{m} \sum_{i \in SA} \mu(X^i) - \mu(Y^i) $
U2	crossMean A,B	A={7, 8, 9}, B=15のクロス集計値の平均絶対誤差	$\text{crossMean}^{A,B}(X, Y) = \frac{1}{ R(A) }$
U3	crossCnt A,B	同クロス集計数の平均絶対誤差	$\sum_{a \in R(A)} \mu(x^B x^A = a) - \mu(y^B y^A = a) $
U4	corMAE	SA={14,...,25}についての全相関係数の平均絶対誤差	$\text{corMAE}(X, Y) = \frac{1}{ SA ^2} \sum_{i,j \in SA} \text{cor}(X^i, X^j) - \text{cor}(Y^i, Y^j) $
U5	IL	匿名加工データの各値の平均絶対誤差	$IL(X, Y, I^Y) = \frac{1}{m'n'} \sum_{i=1}^{m'} \sum_{j=1}^{n'} \frac{ x_{ij}^i - y_j^i }{\max x^i - \min x^i}$
U6	nrow	個人データと匿名加工データのレコード数の差	$\text{nrow}(Y) = n - Y = n - n' $

安全性定義

- 匿名加工データの安全性

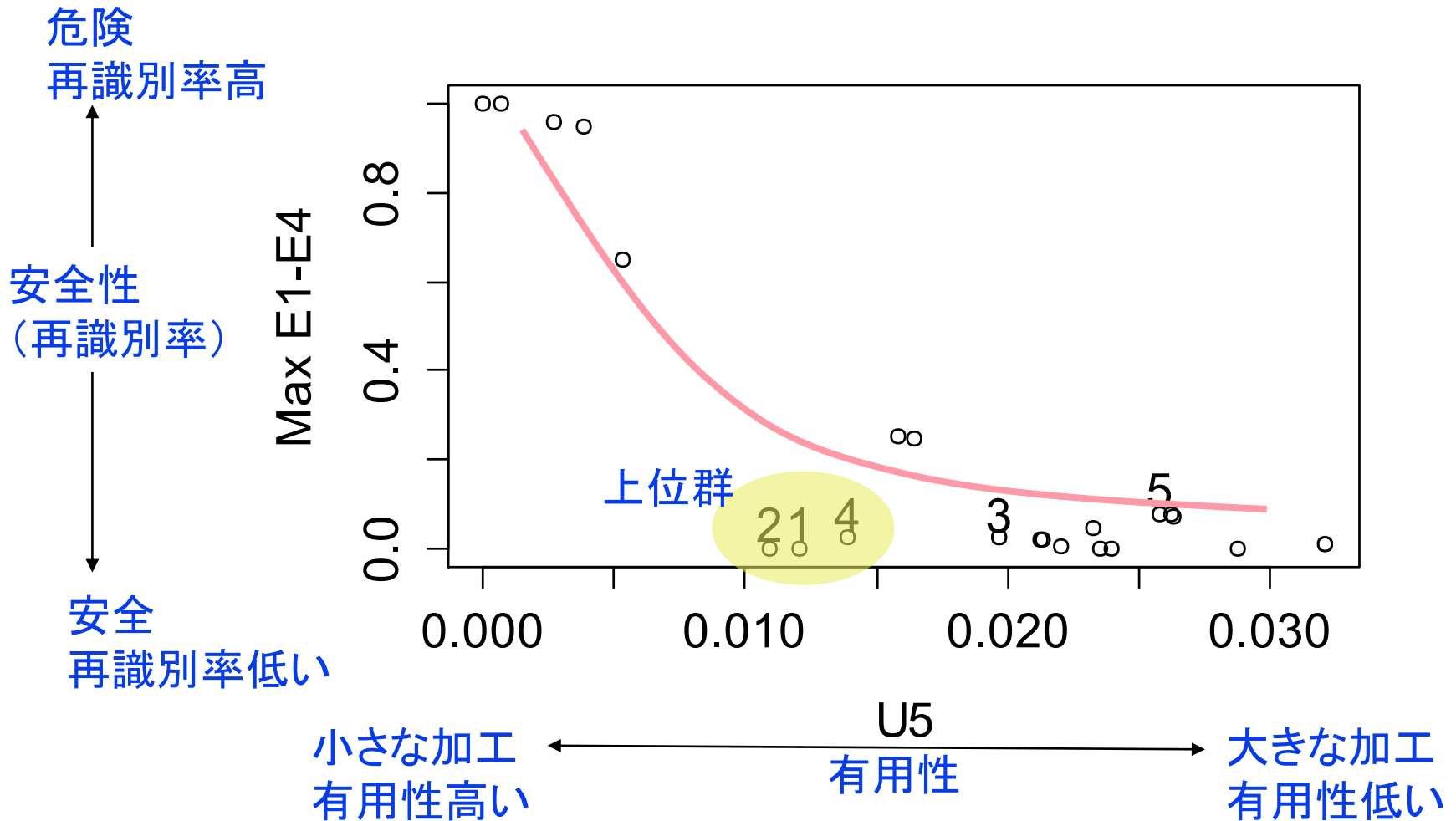
- $\frac{1}{4} \sum_{i=1}^2 Rank(S_i) + \frac{1}{2} Rank(\max_j re-id^{E_j})$

- 安全性指標

- S1=k-匿名性の最小値(いわゆるk)

- S2=k-匿名性の平均値

有用性Uと安全性Eのトレードオフ



PWSCUP 2016 匿名加工・再 識別コンテスト

菊池浩明（明治大学）

2016年度対象データセット

- UCI で公開されている Online Retail データセット
 - <https://archive.ics.uci.edu/ml/datasets/Online+Retail>
- 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
 - 約54万行
 - 7属性 + 品名属性
 - » 顧ID (顧客ID)
 - » 国 (顧客の国)
 - » 伝ID (伝票ID)
 - » 日時 (伝票の日時)
 - » 品ID (品目ID)
 - » 単価 (当該伝IDでの品IDの単価)
 - » 数量 (当該伝IDでの品IDの数量)
 - 主な品目: 贈り物
 - 主な顧客: 卸売り業者



2016年度PWSCUP概要

マスター M

顧客ID	性別	生年月日	国
12346	f	1960/12/25	UK
12347	f	1957/5/15	Iceland
12348	m	1947/2/19	Finland

トランザクション T

顧客ID	年月日	製品ID
12347	2010/12/7	85116
12347	2010/12/7	22375
12346	2011/1/18	23166

加工 (仮名化, 攪乱, シャッフル, 削除)

行番号

匿名加工マスター M'

匿名加工トランザクション T'

Q	P	仮ID	性別	生年月日	国
3	3	10	m	1947/01/01	UK
2	1	20	f	1960/01/01	UK
2	2	30	f	1960/01/01	UK

仮ID	年月日	製品ID
10	2010/12/1	85123A
30	2010/12/1	85123A
30	2010/12/7	20000
20	2011/1/18	20000



再識別者が推定した行番号Q

$$\text{再識別率 } \text{Re-ID}(P, Q) = \frac{\text{正解行数}}{n'} = 2/3$$

2種類のデータテーブル

■ マスター M

□ $n = 400$ 名

□ 36か国

■ トランザクション (履歴) T

□ $m = 38,087$ 行

□ 2,781 商品

表 4 顧客マスターデータ M の例

c^1 顧客 ID	c^2 性別	c^3 生年月日	c^4 国籍
12360	m	1976/2/24	Austria
12361	f	1954/2/14	Belgium
12362	f	1963/12/26	Belgium
12364	f	1960/9/16	Belgium

表 5 購買履歴データ T の例

t^1 顧 ID	t^2 伝票	t^3 購買日	t^4 時刻	t^5 商品	t^6 単価	t^7 数
12362	544203	2011/2/17	10:30	21913	3.75	4
12362	544203	2011/2/17	10:30	22431	1.95	6
12361	545017	2011/2/25	13:51	22630	1.95	12
12361	545017	2011/2/25	13:51	22555	1.65	12
12362	551346	2011/4/28	9:12	21866	1.25	12
12362	551346	2011/4/28	9:12	20750	7.95	2
12362	551346	2011/4/28	9:12	22908	0.85	12
12360	554132	2011/5/23	9:43	21094	0.85	12
12360	554132	2011/5/23	9:43	23007	14.95	6

有用性と3つのユースケース

■ 1. RFM分析

- Recency(最終購買日), Frequency(購買頻度), Monetary(累計購買額)で優良顧客を選別

U3: ut-rfm

■ 2. バスケット分析

- 相関ルール(同時に購買している商品の間の広く観測される性質)

U4: ut-top_item

■ 3. クロス集計

- 性別, 年齢別での購買パターン(商品, 公倍数, 額)を集計

U1: ut-cmae

U2: ut-cmae2

サンプル再識別 re-tnum-bi(1位)



■ 再識別アルゴリズム

- Step 1: 各顧客IDに対して、対応するトランザクションの行数を求める
- Step 2: (トランザクション数, 生年月日)で顧客ID/仮IDをソートする
- Step 3: 顧客ID－仮IDの対応付けを行い、行番号Pを推定する

トランザクション数	マスター M			トランザクション T		
	顧客ID	生年月日	...	顧客ID	...	
	12346	1960/12/25	...	12346	...	
	12347	1957/5/15	...	12346	...	
2				12347	...	
1						
0						

匿名加工マスター M'	匿名加工マスター M'			匿名加工トランザクション T'		
	仮ID	生年月日	...	仮ID	...	
	10	1947/01/01	...	10	...	
	20	1960/01/01	...	20	...	
1				20	...	
2						
0						

過加工の除外

■ 山岡匿名加工

	M		M' (=M)	P		再識別 Q
$X = 1$	12346	f	GE	3	$\neq$	1
	12347	f	UK	1	$\neq$	2
$P(X) = 3$	12348	m	UK	2	$\neq$	3

■ 次の条件を「過加工」とみなす.

□ Y1 (subset) > 50,000

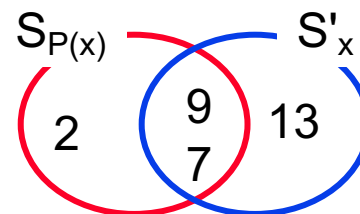
□ Y2 (Jaccard) > 0.7

Y2: 顧客Xの購買した全商品集合

Y1:

$\mu_{P(X)} = P(X)$ の30日間総購買額 = 305

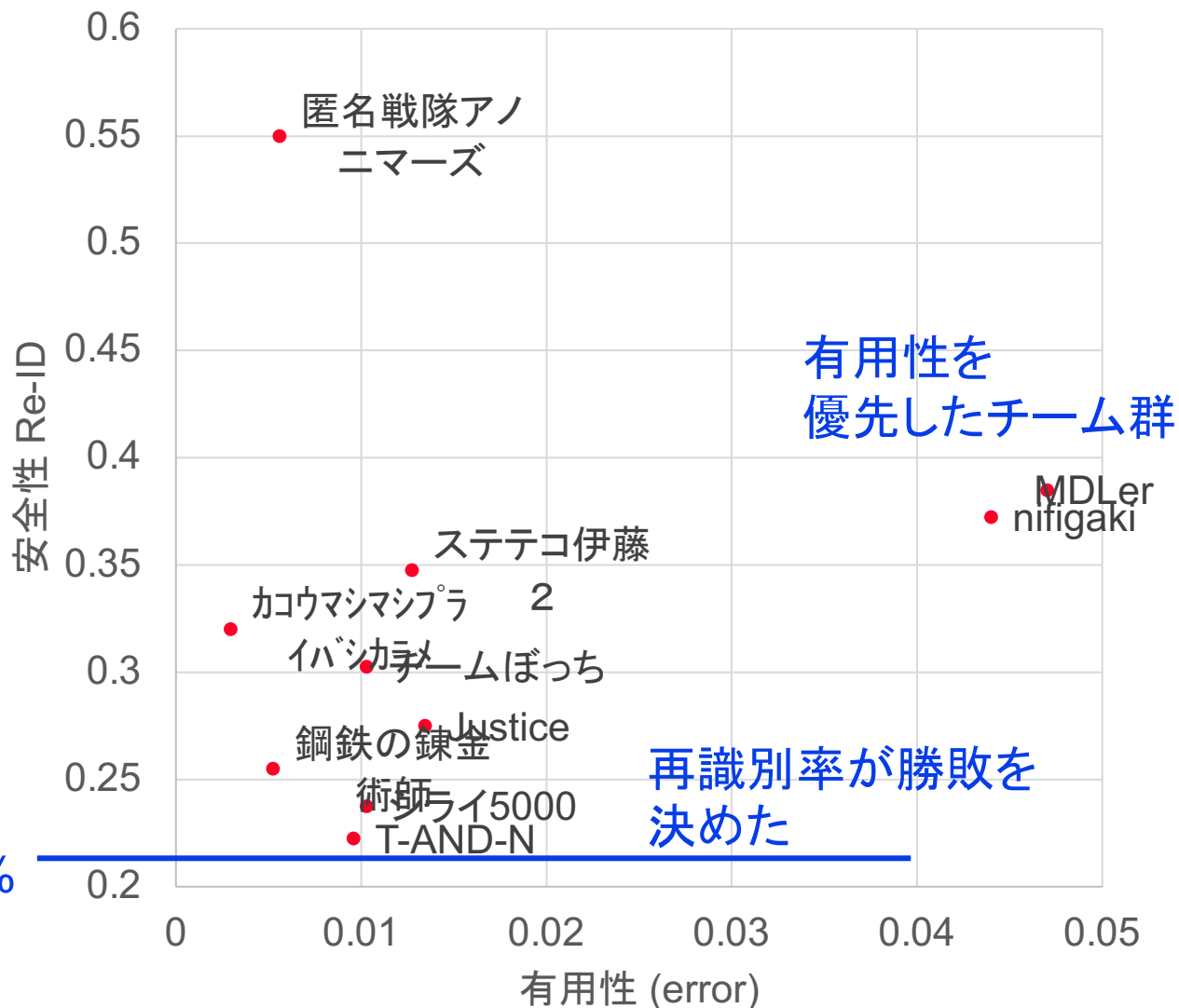
$\mu_{P(X)} = X$ の30日間購買平均単値 = 405



コンテストの結果(上位10チーム)

No	チーム名
1	T-AND-N
2	シライ5000
3	鋼鉄の錬金術師
4	Justice
5	チームぼっち
6	カウマシマシプライバシラメ
7	ステテコ伊藤 2
8	nifigaki
9	MDLer
10	匿名戦隊アノニマーズ

最低再識別率 = 22.25%



個人データに対応する匿名化と 再識別コンテスト PWSCUP 2017

菊池浩明(明大)

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。

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

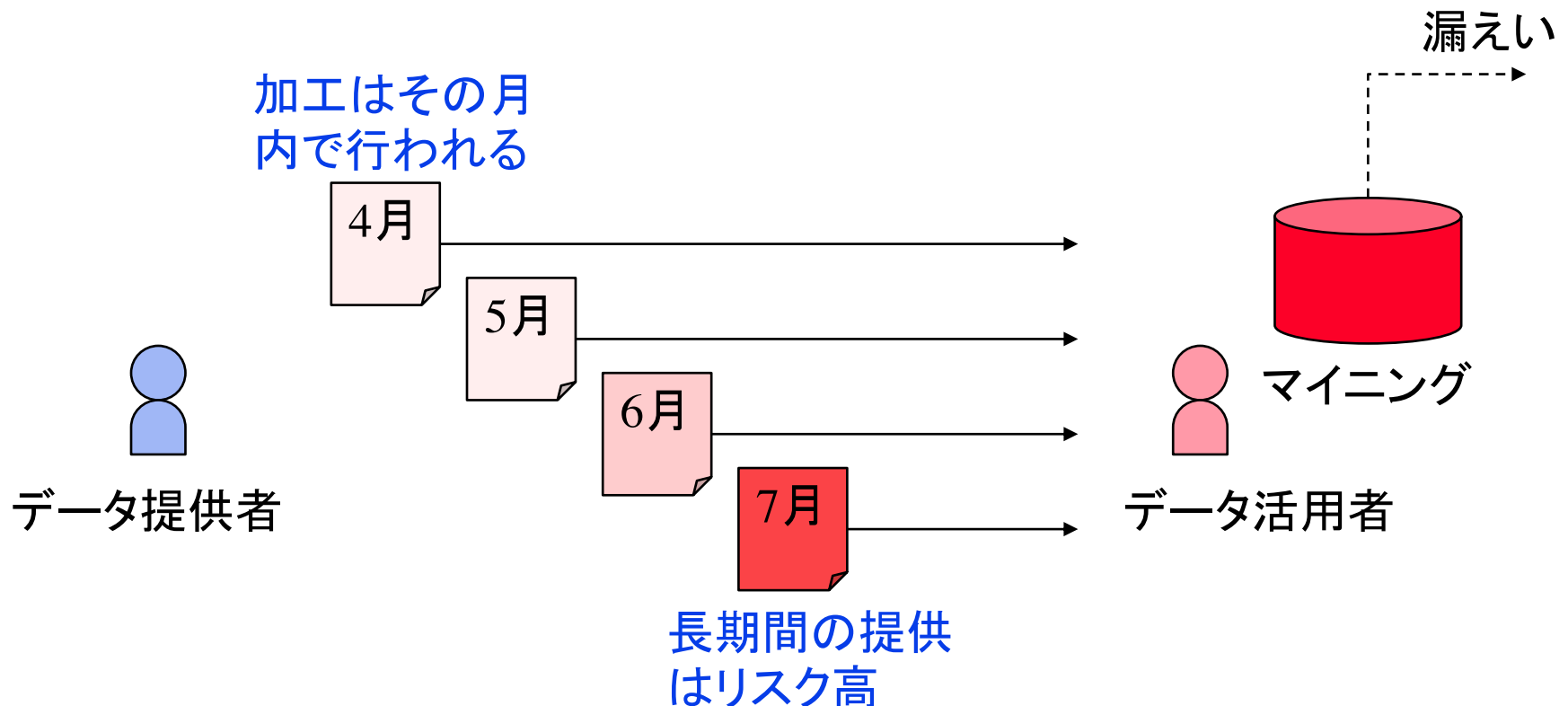
□ 2,781 商品

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

ユースケースシナリオ

■ 匿名加工された購買履歴からの商品推薦システム



(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

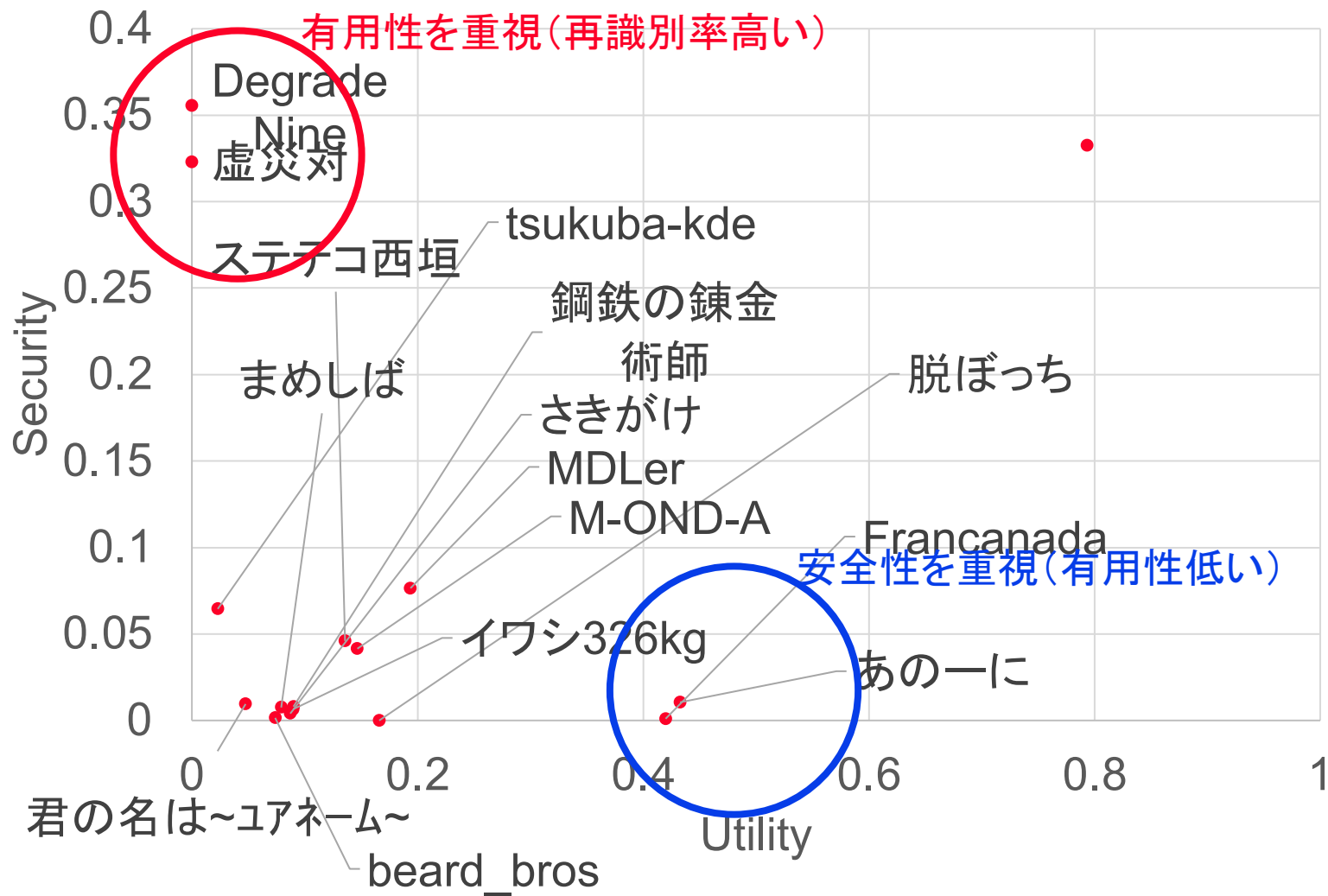
仮名表 F

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	

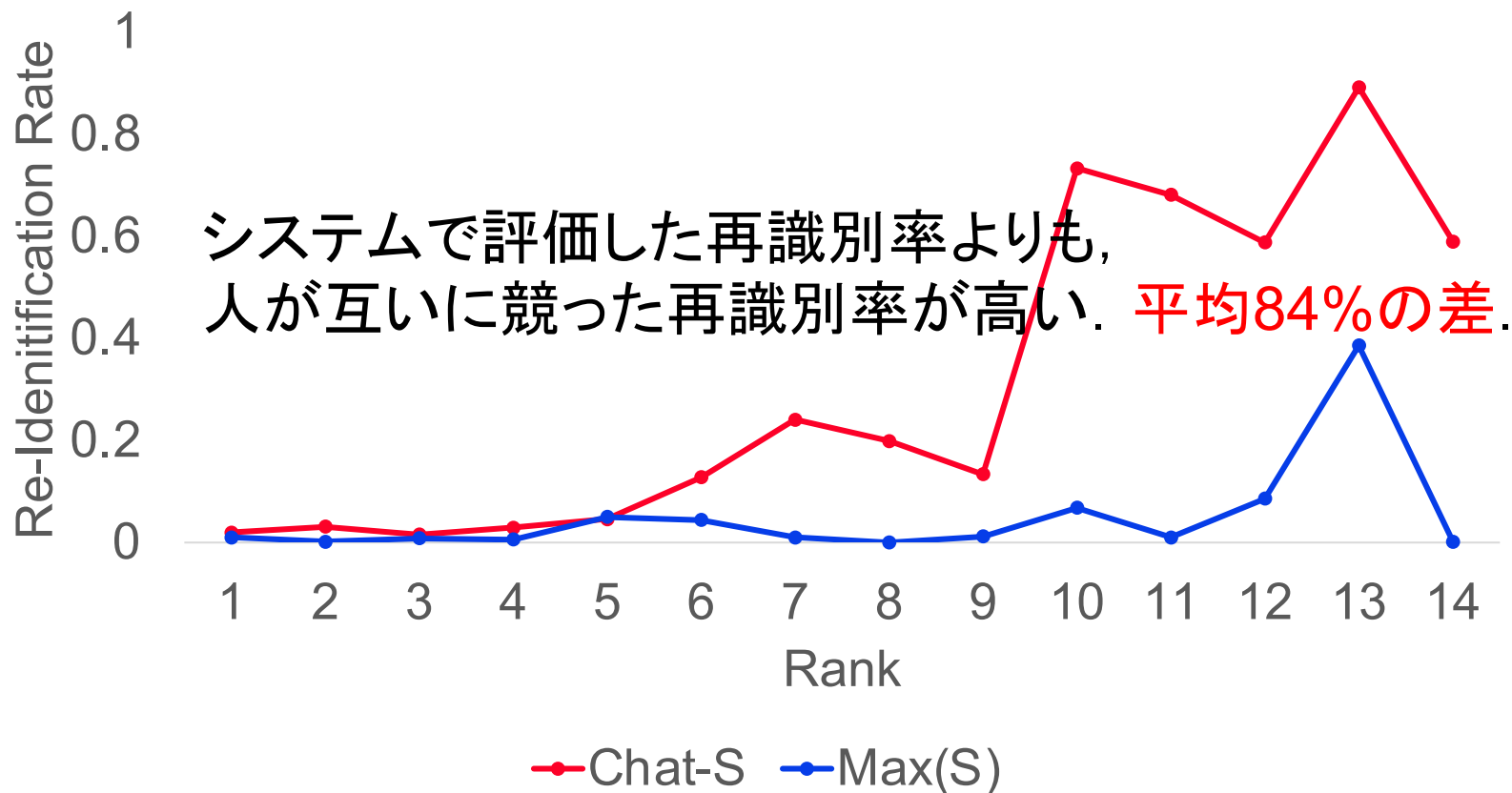
推測 $\hat{F}$

$$\text{Reid}(F, \hat{F}) = \text{正しい推測顧客数}/n = 2/5$$

有用性と安全性の関係



再識別率



PWS Cup 2021 – 糖尿病罹患リスクを 予測するヘルスケアデータの匿名化コ ンテスト

NHANES概要



- National Health and Nutrition Examination Survey
 - ❑ CDC (米国疾病対策センター)の国民健康栄養調査プログラム
 - ❑ 1960年代から行われている調査. 全米15箇所で, 年5,000人を調査している.
 - ❑ 疫学研究, 健全な公共健康政策やサービスの施策に活用.
 - ❑ 被験者世帯は, NCHS所長からのレターを受け取る. 報酬と診断結果を得る. プライバシーは法律で守られている (privacy is protected by public laws)

CDC Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™

National Center for Health Statistics

CDC > NCHS > National Health and Nutrition Examination Survey

National Health and Nutrition Examination Survey

About NHANES +

What's New +

Questionnaires, Datasets, and Related Documentation -

Survey Methods and Analytic Guidelines

Search Variables

Frequently Asked Questions

All Continuous NHANES +

NHANES 2019-2020 +

NHANES 2017-2018 +

NHANES 2015-2016 +

NHANES 2013-2014 +

NHANES 2011-2012 +

NHANES 2009-2010 +

NHANES Questionnaires, Dataset, and Documentation

Survey Methods
Plan & Operations, Sample Design, Estimation & Weighting Procedures, Analytic Guidelines, etc.

Continuous NHANES

NHANES 2019-2020

NHANES 2017-2018

NHANES 2011-2012

NHANES 2009-2010

NHANES 2003-2004

NHANES 2001-2002

ゲーム概要

加工者
 i

元データ
 B

1			
2			
3			
4			

top, bottom
kanony等

第1匿名化

削除行
 $X^{(i)}$

1
3

$C^{(i)}$

rr, dp等

第2匿名化

匿名化データ
 $D^{(i)}$

1			
2			

checkDX

提出

$D^{(i)}, X^{(i)}$

審判



安全性1

$\text{Uniqrt}(B, C^{(i)})$

安全性2

$\text{Iloss}(C^{(i)}, D^{(i)})$

有用性

$\text{Umark}(B, D^{(i)})$

匿名化フェーズ

攻撃フェーズ

攻撃者
 j

元データ
 B

1			
2			
3			
4			

pick等

サンプリング

テストデータ
 $C_T^{(i,j)}$

1			
4			

rlink等



匿名化データ
 $D^{(i)}$

1			
2			

推定

推定
 $E^{(i,j)}$

-1
3

正解
 $E_a^{(i,j)}$

-1
2

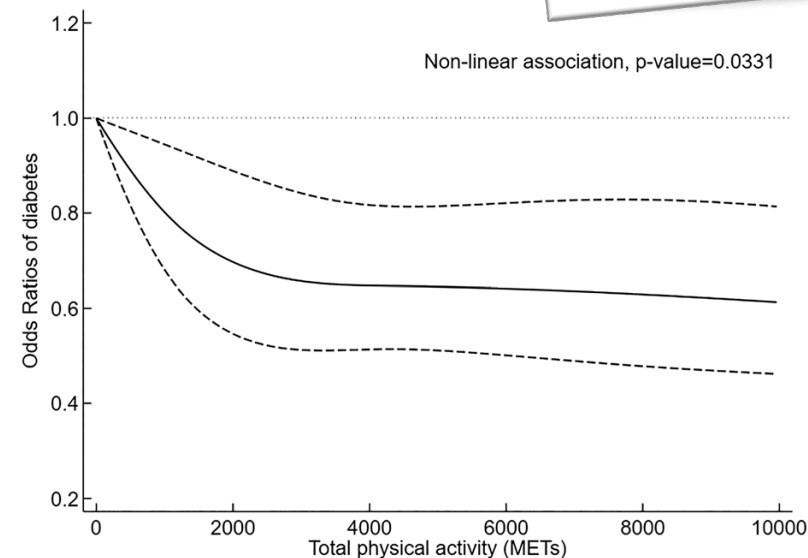
安全性3

$\text{Lmark}(E^{(i,j)}, E_a^{(i,j)})$

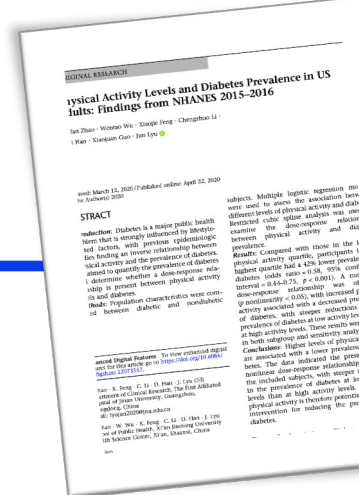
利活用例「糖尿病罹患リスク」

■ ORs (Model III, 因子調整済み)

活動量 METs	オッズ 比 OR	95% 信頼区間	p値
Q1	1		
Q2	0.71	0.56-0.89	0.003
Q3	0.66	0.52-0.84	0.001
Q4	0.58	0.44-0.75	< 0.001



Fanfan Zhao, et. al (The First Affiliated Hospital of Jinan, Xi'an Jiaotong Univ.),
 “Physical Activity Levels and Diabetes Prevalence in US Adults: Findings from
 NHANES 2015–2016”, Diabetics Ther 2020.



Umark (Utility benchmark)有用性指標

rate

罹患率クロス集計
(33x2)

Table 1 [Zhao 2020]

	cnt		rate	
	0	1	0	1
diabetes				
Female	1710	407	0.408	0.097
Male	1607	466	0.384	0.111
(19, 44]	1574	110	0.376	0.026
(44, 64]	1033	379	0.247	0.090
(64, 80]	710	384	0.169	0.092
Black	647	208	0.154	0.050
Hispanic	443	127	0.106	0.030
Mexican	518	200	0.124	0.048
Other	537	112	0.128	0.027
White	1172	226	0.280	0.054
11th	373	120	0.089	0.029
9th	386	168	0.092	0.040
College	975	239	0.233	0.057
Graduate	851	156	0.203	0.037
HighSchool	732	190	0.175	0.045
Divorced	342	103	0.082	0.025
Married	1669	502	0.398	0.120
Never	652	87	0.156	0.021
Parther	331	57	0.079	0.014
Separated	101	32	0.024	0.008
Widowed	222	92	0.053	0.022
(15.0, 18.5]	64	4	0.015	0.001
(18.5, 25.0]	1027	114	0.245	0.027
(25.0, 30.0]	1149	241	0.274	0.058
(30.0, 70.0]	1075	514	0.257	0.123
dep_0	2666	647	0.636	0.154
dep_1	651	226	0.155	0.054
pir 0	2656	650	0.634	0.155

cor

共分散行列
(30x30)

	0_Male	0_Female	1	2_White	2_Black	2
0_Male	0.00	0.00	0.00	0.00	0.00	
0_Female	-1.00	0.00	0.00	0.00	0.00	
1	0.01	-0.01	0.00	0.00	0.00	
2_White	0.02	-0.02	0.12	0.00	0.00	
2_Black	-0.02	0.02	-0.05	-0.37	0.00	
2_Mexican	-0.02	0.02	-0.03	-0.34	-0.23	
2_Other	0.04	-0.04	-0.11	-0.30	-0.20	
2_Hispanic	-0.03	0.03	0.02	-0.29	-0.19	
3_Graduate	0.01	-0.01	-0.06	0.09	-0.07	
3_HighSchool	0.03	-0.03	-0.01	0.02	0.07	
3_College	-0.06	0.06	-0.05	0.11	0.07	
3_11th	0.04	-0.04	0.01	-0.10	0.02	
3_9th	-0.01	0.01	0.16	-0.21	-0.10	
3_nan	-0.02	0.02	0.02	-0.01	-0.01	
4_Married	0.10	-0.10	0.14	0.04	-0.15	
4_Divorced	-0.06	0.06	0.14	0.03	0.04	
4_Parther	0.02	-0.02	-0.21	-0.03	0.00	
4_Separated	-0.03	0.03	0.02	-0.06	0.04	
4_Never	0.00	0.00	-0.37	-0.08	0.16	
4_Widowed	-0.14	0.14	0.34	0.07	-0.01	
5	-0.08	0.08	0.06	-0.04	0.09	
6	-0.08	0.08	-0.03	0.03	-0.04	
7	-0.03	0.03	0.00	-0.21	0.04	
10_Q2	-0.06	0.06	0.09	0.00	0.01	
10_Q1	-0.11	0.11	0.21	-0.08	-0.03	
10_Q3	0.02	-0.02	-0.07	0.07	-0.04	

OR

オッズ比
(21x2)

Table 2 [Zhao 2020]

	Coef	OR	pvalue
Intercept	-7.319	0.001	0.00
gen[T.Male]	0.380	1.463	0.00
race[T.Hispanic]	-0.302	0.740	0.00
race[T.Mexican]	0.084	1.088	0.51
race[T.Other]	0.020	1.020	0.90
race[T.White]	-0.844	0.430	0.00
edu[T.9th]	-0.151	0.860	0.40
edu[T.College]	-0.073	0.930	0.63
edu[T.Graduate]	-0.150	0.861	0.35
edu[T.HighSchool]	-0.192	0.825	0.24
mar[T.Married]	0.278	1.320	0.00
mar[T.Never]	0.326	1.386	0.10
mar[T.Parther]	0.316	1.372	0.10
mar[T.Separated]	0.081	1.084	0.71
mar[T.Widowed]	-0.134	0.875	0.51
qm[T.Q2]	-0.228	0.796	0.00
qm[T.Q3]	-0.328	0.720	0.00
qm[T.Q4]	-0.401	0.670	0.00
age	0.057	1.058	0.00
bmi	0.097	1.102	0.00
dep	0.440	1.552	0.00
pir	0.147	1.158	0.17

Lmark (Link benchmark) 安全性指標



審判

■ 推測の正しさの評価

□ 正解行番号 E_a と推測行番号 E について,

□ 所属推定: $recall = \frac{|E_a^X \cap E^X|}{|E_a^X|}, prec = \frac{|E_a^X \cap E^X|}{|E^X|}$

再識別率: $top_k = \frac{|\{\ell \in E_a^X | e_{a\ell} \in E_\ell\}|}{|E_a^X|}$

□ 例)

» $E^X = \{1, 3, 4, 5\}$, $E_a^X = \{1, 3, 4\}$

» $recall = 3/3 = 1.0$

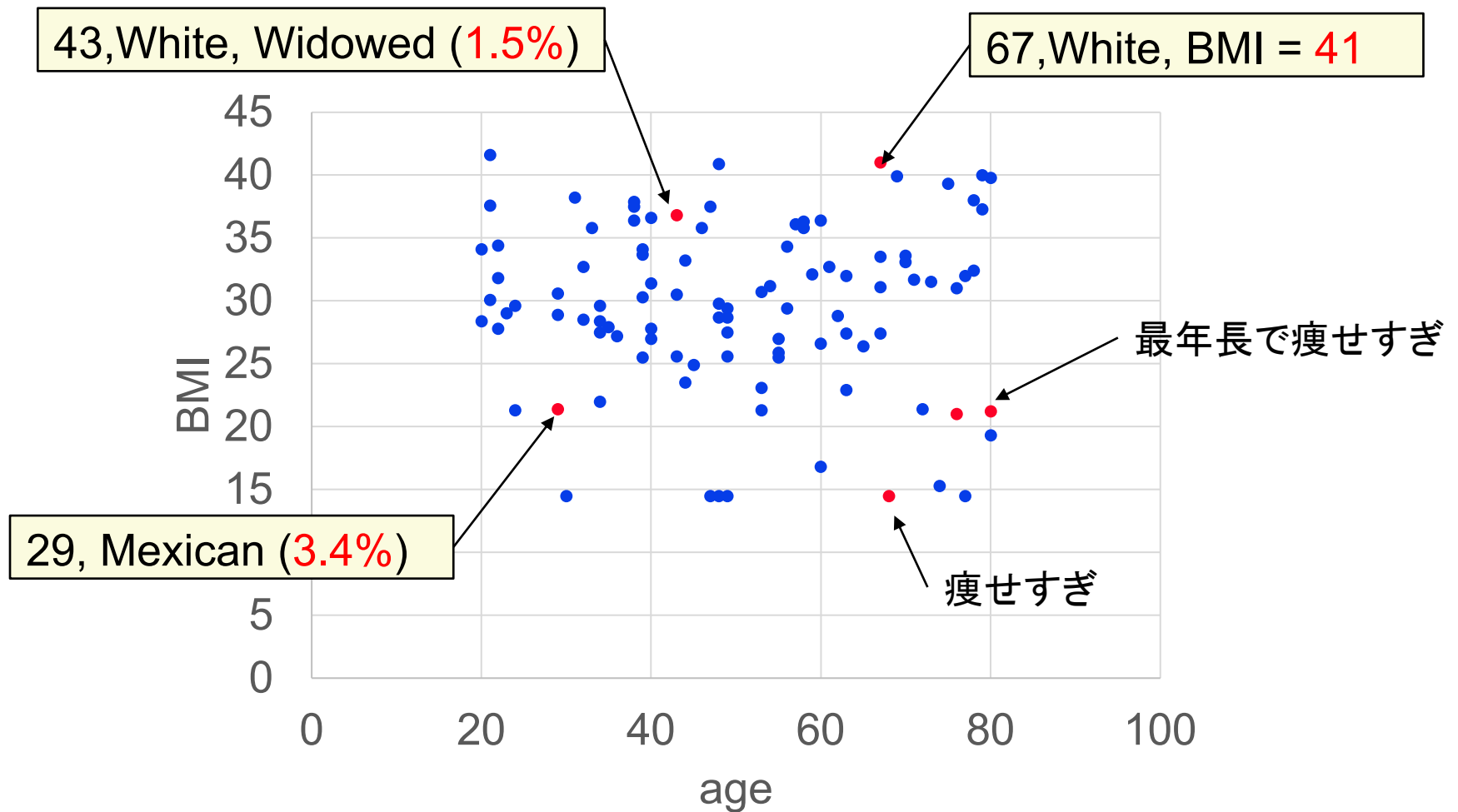
» $prec = 3/4 = 0.75$

» $top_k = 2/3 = 0.6$

	$E_{\ell,1}$	$E_{\ell,2}$	$E_{\ell,3}$	E_a (正解)	判定
1	29	847	2599	29	✓
2	-1	-1	-1	-1	✓
3	2038	2345	2336	2345	✓
4	2702	1378	2331	80	NG
5	134	1820	2580	-1	NG

» $Lmark = (1.0 - \delta)(0.75 - \delta)(0.6 - \delta) = 0.43$

識別された6つのレコード



攻撃: 危険度Rの平均Aを大きくできるほど高評価

本戦結果

$$A^{(j)} = \frac{1}{3} \sum_{i \in T_3} R^{(i,j)}$$

例: Team01がTeam02を攻撃した結果(危険度R)

加工者 攻撃者	01	02	03	04	05	06	07	08	09	10	11	12	13	14	A 平均	攻撃
01		0.520	0.005	0.016	0.010	0.005	0.002	0.025	0.006	0.045	0.486	0.192	0.035	0.026	0.022	6
02	0.068		0.003	0.003	0.002	0.003	0.006	0.005	0.004	0.042	0.004	0.002	0.003	0.007	0.016	9
03	0.093	0.970		0.004	0.049	0.001	0.000	0.004	0.006	0.034	0.462	0.003	0.008	0.119	0.013	12
04	0.010	0.007	0.011		0.002	0.024	0.002	0.035	0.026	0.030	0.018	0.002	0.020	0.056	0.027	5
05	0.005	0.005	0.005	0.005		0.067	0.015	0.283	0.257	0.057	0.005	0.005	0.046	0.036	0.043	2
06	0.022	0.015	0.005	0.015	0.026		0.008	0.045	0.017	0.067	0.037	0.005	0.057	0.057	0.041	3
07	0.054	0.721	0.003	0.005	0.028	0.001		0.004	0.005	0.036	0.236	0.013	0.009	0.009	0.014	11
08	0.027	0.026	0.005	0.015	0.057	0.015	0.005		0.109	0.026	0.005	0.005	0.036	0.015	0.019	7
09	0.081	0.046	0.005	0.067	0.165	0.036	0.011	0.067		0.057	0.048	0.057	0.077	0.036	0.053	1
10	0.027	0.026	0.005	0.026	0.015	0.046	0.026	0.221	0.166		0.027	0.036	0.036	0.057	0.036	4
11	0.059	0.857	0.003	0.008	0.018	0.002	0.002	0.003	0.005	0.024		0.002	0.002	0.007	0.011	13
12	0.042	0.008	0.006	0.016	0.048	0.008	0.008	0.002	0.014	0.007	0.008		0.010	0.012	0.010	14
13	0.011	0.007	0.003	0.004	0.021	0.031	0.008	0.039	0.016	0.012	0.005	0.007		0.000	0.016	10
14	0.023	0.203	0.003	0.011	0.007	0.001	0.000	0.002	0.018	0.043	0.084	0.002	0.009		0.018	8
R参考	0.093	0.970	0.011	0.067	0.165	0.067	0.026	0.283	0.257	0.067	0.486	0.192	0.077	0.119	0.024	
R(Max)	0.093	0.970		0.067	0.165	0.067		0.283	0.257	0.067	0.486	0.192	0.077	0.119	1.000	
lmark	0.045	0.585	0.003	0.003	0.019	0.001	0.000	0.004	0.007	0.034	0.148	0.003	0.002	0.031		
加工	5	12		1	7	1		10	9	1	11	8	4	6		

$R^{(i)} = \max_{j \neq i} R^{(i,j)}$ 失格

3つのチームが同列1位

加工: 危険度Rの

最大値が小さいほど高評価

コンテストで達成できたこと

- 匿名化の技巧によって安全性の差があることを確認
- コンテストにすることで, k -匿名よりも強い山岡匿名化が存在する
- データセットの特性(離散, 数値, 履歴)に応じて, 適した評価関数がある.
- レコード再識別, 属性推定, メンバーシップ推定の攻撃モデルが有る

出来なかったこと

- 有用性指標, 安全性指標の標準化(ベンチマーク), オープンソース, オープンデータ
- コンテストから得られた知見を法制度や実践にフィードバック
- 適切な攻撃者モデルの識別

これから我々は何を目指か

- PWS Cupの原則（共通のデータ，明確な勝ち負け，ユースケース想定）を他の問題（データ合成，差分プライバシー，仮名加工）へ適用していく
- 技術者，科学者の立場から，法制度や規制に対して妥当性を検討していく

結論

- 匿名化されたヘルスケアデータを用いたコンテストPWS Cup 2021を開催した.
- 活動量が糖尿病の罹患リスクをどれだけ下げるかを正確に, かつ, 再識別されないように安全にする匿名加工データを作成した.