2019年3月5日

第84回CSEC研究発表会

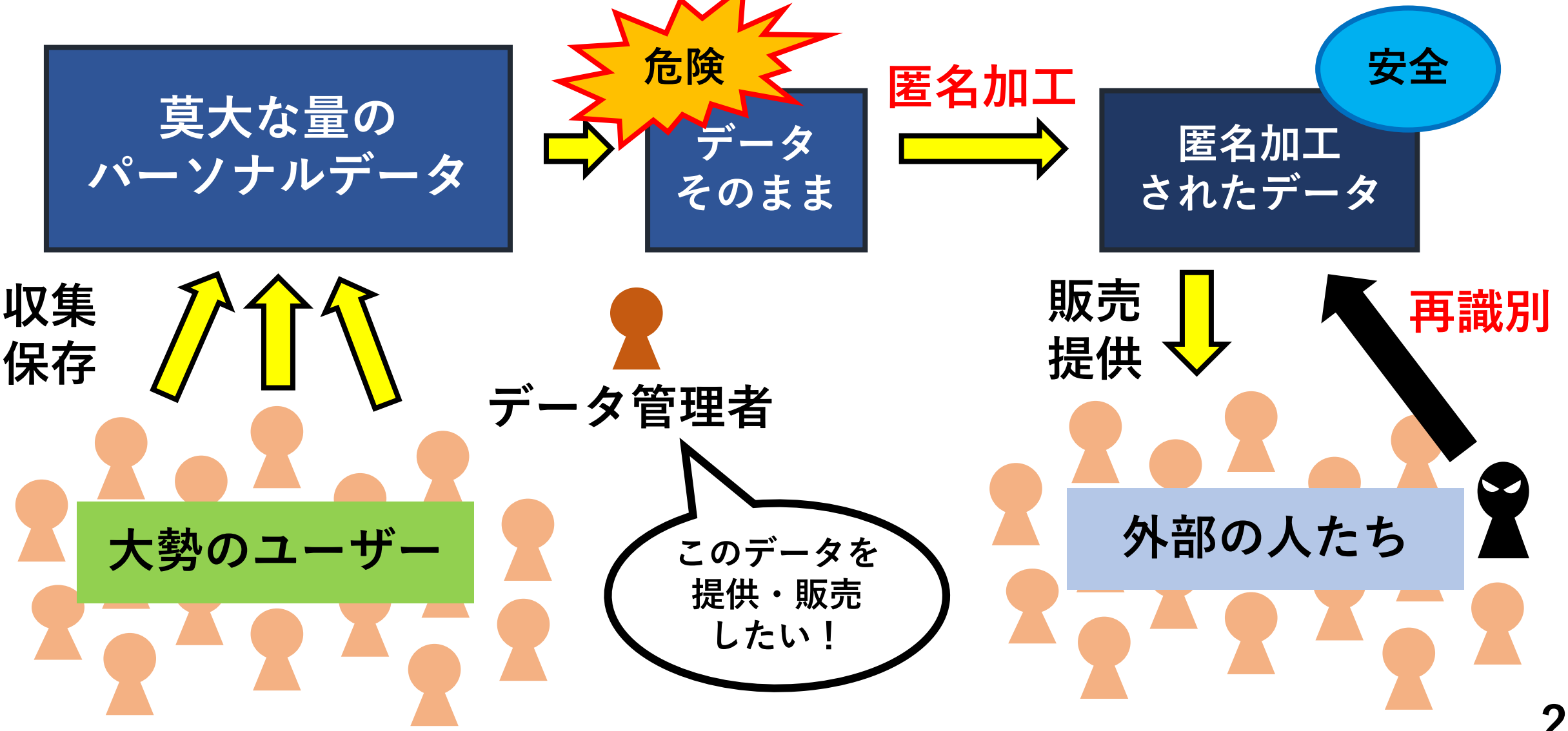
攻撃者の平均識別確率を用いた匿名加工情報の再識別リスク評価モデルの提案と評価

伊藤 聡志（明治大学大学院先端数理科学研究科）

菊池 浩明（明治大学総合数理学部）

匿名加工とは？

匿名加工の研究において
攻撃者の想定は大きな課題である



攻撃者と背景知識

明治大学生の試験結果

ID	数学	英語	物理
A	90	50	70
B	90	50	60
C	90	70	70
D	50	70	60
E	50	50	80
F	50	50	10
G	30	70	80
H	30	70	10

どれかが伊藤

このデータから
伊藤の試験結果を
知りたい！

攻撃者

攻撃者が
このデータから
伊藤を識別
できる確率

$= \frac{1}{8} (12.5\%)$

背景知識

攻撃者と背景知識

明治大学生の試験結果

ID	数学	英語	物理
A	90	50	70
B	90	50	60
C	90	70	70
D	50	70	60
E	50	50	80
F	50	50	10
G	30	70	80
H	30	70	10

攻撃者の持つ背景知識によってデータの危険度は変わる

攻撃者X

伊藤の英語の点数は50点である

伊藤を識別できる確率

$= \frac{1}{4}(25\%)$

攻撃者Y

伊藤の物理の点数は10点である

伊藤を識別できる確率

$= \frac{1}{2}(50\%)$

本研究について

研究目的

- **どんな背景知識を持つ攻撃者が危険なのか？**
- **データ中のどの属性が危険であるのか？**
- **匿名加工の際にどの属性を加工したらよいのか？**

これらを明らかにする

解決手法

- **データのリスク評価を行う理論的なモデルを提案する**
- **実際のデータを用いて評価実験を行う**

想定するケース

元データ

真名	属性X
伊藤	2019/2/1
伊藤	2019/2/1
山田	2019/2/2
岡本	2019/2/2
岡本	2019/2/3

確率$Pr(x)$で
ある顧客の
属性Xについての
背景知識xを得る

攻撃者

確率$Pr(\mathrm{idf}|x)$で
再識別に成功する

仮名化データ

仮名	属性X
A	2019/2/1
A	2019/2/1
B	2019/2/2
C	2019/2/2
C	2019/2/3

例：x =“2019/2/2”の場合

元データ

真名	購買日
伊藤	2019/2/1
伊藤	2019/2/1
山田	2019/2/2
岡本	2019/2/2
岡本	2019/2/3

$Pr(x) = 2/5$の確率で
「あるユーザが2019/2/2に
買い物をした」
という背景知識を得る

攻撃者

仮名化データ

仮名	購買日
A	2019/2/1
A	2019/2/1
B	2019/2/2
C	2019/2/2
C	2019/2/3

$Pr(\mathrm{idf}|x) = 1/2$の確率で
あるユーザの再識別に成功する

例：x =“2019/2/2”の場合

背景知識 x の危険度

$$Pr(\mathrm{idf}, x) = Pr(x)Pr(\mathrm{idf}|x)$$

属性Xの危険度(平均識別確率)

$$Pr(\mathrm{idf}, X) = \sum Pr(\mathrm{idf}, x)$$

$Pr(x) = 2/5$の確率で
あるデータが2019/2/2に
買い物をしたという背景知識を得る

元データ

真名	購買日
伊藤	2019/2/1
伊藤	2019/2/1
山田	2019/2/2
岡本	2019/2/2
岡本	2019/2/3

攻撃者

仮名化データ

仮名	購買日
A	2019/2/1
A	2019/2/1
B	2019/2/2
C	2019/2/2
C	2019/2/3

$Pr(\mathrm{idf}|x) = 1/2$の確率で
あるユーザの再識別に成功する

平均識別確率と平均レコード数 α_x (1)

α_x：xについての平均レコード数

m：データセットのレコード数

$$\alpha_x = \frac{x\text{を満たすレコードの数}}{x\text{を満たすユーザの数}}$$

例：購買履歴データ

真名	購買日
伊藤	2019/2/1
伊藤	2019/2/1
山田	2019/2/2
岡本	2019/2/2
岡本	2019/2/3

$x =$“2019/2/1”のとき $\alpha_x = \frac{2}{1} = \mathbf{2}$

$x =$“2019/2/2”のとき $\alpha_x = \frac{2}{2} = \mathbf{1}$

$x =$“2019/2/3”のとき $\alpha_x = \frac{1}{1} = \mathbf{1}$

$m = \mathbf{5}$

平均識別確率と平均レコード数 α_x (2)

このとき，平均識別確率は $Pr(\mathbf{idf}, X) = \sum \frac{\alpha_x}{m}$ と表せる．

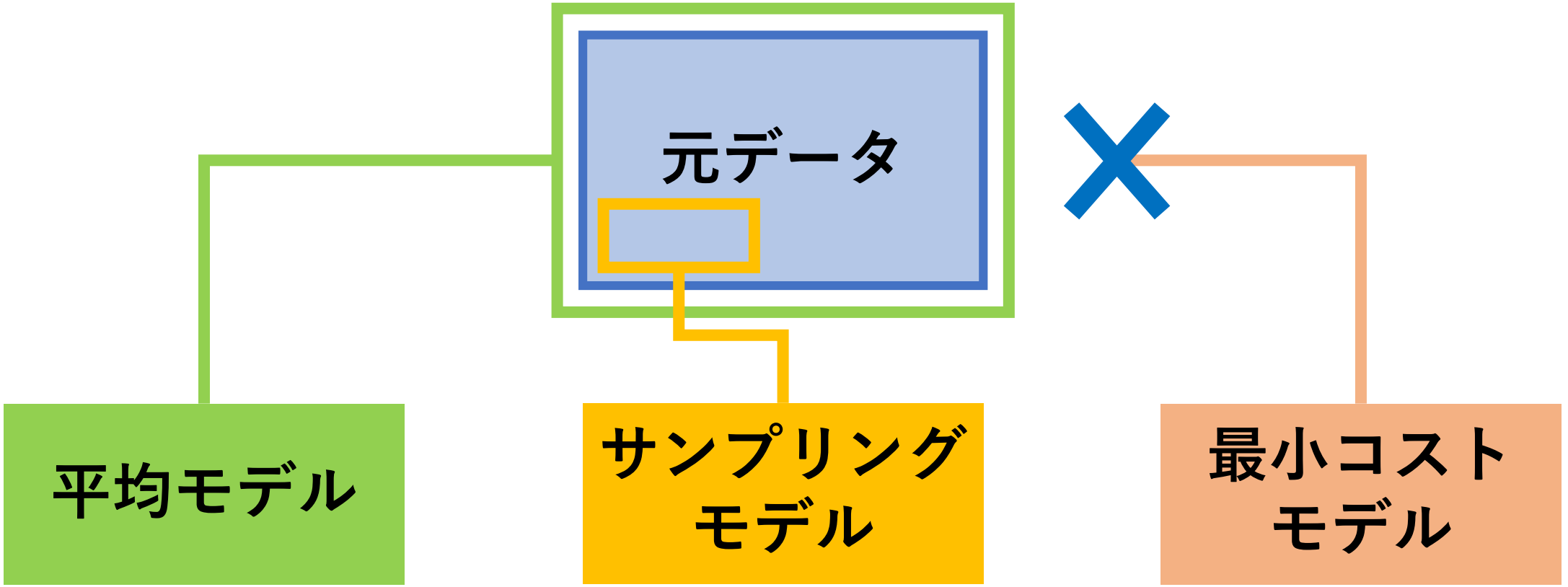
例：購買履歴データ

真名	購買日
伊藤	2019/2/1
伊藤	2019/2/1
山田	2019/2/2
岡本	2019/2/2
岡本	2019/2/3

$$Pr(\mathbf{idf}, X) = \sum \frac{\alpha_x}{m} = \frac{2}{5} + \frac{1}{5} + \frac{1}{5} = \frac{4}{5}$$

しかし，ビッグデータについて全ての
α_xを計算するのは困難なので，
これを近似してリスク評価を行う
モデルを提案する．

平均識別確率を近似する３つのモデル



データの全体から
α_xの平均値α_Xを求め
リスクを計算する

データの一部から
α_xの平均値$\alpha_{x'}$を求め
リスクを計算する

全てのxについて
$\alpha_x = 1$として
リスクを計算する

各モデルの式

1. 平均識別確率 $$Pr(\mathbf{idf}, X) = \sum \frac{\alpha_x}{m}$$

2. 平均モデル $$\mathbf{R_{mean}(X)} = \sum \frac{\alpha_X}{m}$$

3. 最小コストモデル $$\mathbf{R_{cost}(X)} = \sum \frac{1}{m}$$

4. サンプリングモデル $$\mathbf{R_{sample}(X)} = \sum \frac{\alpha_{x'}}{m}$$

評価実験

提案した３つのモデルで以下の３データのリスク評価をした

T_1：Online Retail Dataset
英国の１年間の小売データ

T_2：Diabetes Dataset
糖尿病患者の入院履歴データ

T_3：Adult Dataset
国勢調査による世帯収入データ

各データの大きさ

	レコード数	ユーザ数	属性数
T_1	38,087	400	7
T_2	101,766	71,518	50
T_3	32,561	32,561	16

実験には
一部の属性を
用いる

各モデルの評価結果

※サンプリングサイズは10
※90%信頼区間

データ	属性	真値	平均モデル	最小コストモデル	サンプリングモデル
T_1	購買時刻	0.3217	0.3217	0.0145	[0.1411, 0.5998]
	購買日	0.1860	0.1860	0.0076	[0.1267, 0.2786]
	購買商品	0.0965	0.0965	0.0730	[0.0718, 0.0982]
	価格	0.0121	0.0121	0.0048	[0.0036, 0.0132]
	個数	0.0080	0.0080	0.0025	[0.0017, 0.0152]
T_2	入院日数	1.45E-04	1.45E-04	1.38E-04	[1.46E-04, 1.52E-04]
	年齢	1.33E-04	1.33E-04	9.83E-05	[1.21E-04, 1.42E-04]
	人種	7.73E-05	7.73E-05	5.90E-05	[6.92E-05, 8.31E-05]
	性別	3.78E-05	3.78E-05	2.95E-05	[3.08E-05, 4.30E-05]
T_3	年齢	2.24E-03	2.24E-03	2.24E-03	[2.24E-03, 2.24E-03]
	職業	4.61E-04	4.61E-04	4.61E-04	[4.61E-04, 4.61E-04]
	婚姻状況	2.15E-04	2.15E-04	2.15E-04	[2.15E-04, 2.15E-04]
	人種	1.54E-04	1.54E-04	1.54E-04	[1.54E-04, 1.54E-04]

各モデルの評価結果

※サンプリングサイズは10
※90%信頼区間

データ	属性	真値	平均モデル	最小コストモデル	サンプリングモデル
T_1	購買時刻	0.3217	0.3217	0.0145	[0.1411, 0.5998]
	購買日	0.1860	0.1860	0.0076	[0.1267, 0.2786]
	購買商品	0.0965	0.0965	0.0730	[0.0748, 0.0982]
	価格	0.0121	0.0121	0.0048	[0.0030, 0.0152]
	個数	0.0080	0.0080	0.0025	[0.0017, 0.0152]
T_2	入院日数	1.45E-04	1.45E-04	1.45E-04	[1.45E-04, 1.45E-04]
	年齢	1.33E-04	1.33E-04	9.83E-05	[1.21E-04, 1.42E-04]
	人種	7.73E-05	7.73E-05	7.73E-05	[7.15E-05, 8.31E-05]
	性別	3.78E-05	3.78E-05	3.78E-05	[3.78E-05, 3.78E-05]
T_3	年齢	2.24E-03	2.24E-03	2.24E-03	[2.24E-03, 2.24E-03]
	職業	4.61E-04	4.61E-04	4.61E-04	[4.61E-04, 4.61E-04]
	婚姻状況	2.15E-04	2.15E-04	2.15E-04	[2.15E-04, 2.15E-04]
	人種	1.54E-04	1.54E-04	1.54E-04	[1.54E-04, 1.54E-04]

平均識別確率を用いることにより
どの属性が危険であるかを判断できる

T_1: 購買時刻, T_2: 入院日数, T_3: 年齢

匿名加工をする際に
どの属性を優先的に加工するか
決めることができる

各モデルの評価結果

※サンプリングサイズは10
※90%信頼区間

データ	属性	真値	平均モデル	最小コストモデル	サンプリングモデル
T_1	購買時刻	0.3217	0.3217	0.0145	[0.1411, 0.5998]
	購買日	0.1860	0.1860	0.0076	[0.1267, 0.2786]
	購買商品	0.0965	0.0965	0.0730	[0.0718, 0.0982]
	価格	0.0121	0.0121	0.0048	[0.0036, 0.0152]
	個数	0.0080	0.0080	0.0025	[0.0017, 0.0152]
T_2	入院日数	1.45E-04	1.45E-04	1.38E-04	[1.46E-04, 1.52E-04]
	年齢	1.33E-04	1.33E-04	9.83E-05	[1.21E-04, 1.42E-04]
	人種	7.73E-05	7.73E-05	5.90E-05	[6.92E-05, 8.31E-05]
	性別	3.78E-05	3.78E-05	2.95E-05	[3.08E-05, 4.30E-05]
T_3	年齢	2.24E-03	2.24E-03	2.24E-03	[2.24E-03, 2.24E-03]
	職業	4.61E-04	4.61E-04	4.61E-04	[4.61E-04, 4.61E-04]
	婚姻状況	2.15E-04	2.15E-04	2.15E-04	[2.15E-04, 2.15E-04]
	人種	1.54E-04	1.54E-04	1.54E-04	[1.54E-04, 1.54E-04]

平均モデルの結果は
平均識別確率の真値と一致する

$$\sum \frac{\alpha_x}{m} = \sum \frac{\alpha_X}{m}$$

各モデルの評価結果

※サンプリングサイズは10
※90%信頼区間

データ	属性	真値	平均モデル	最小コストモデル	サンプリングモデル
	購買時刻	0.3217	0.3217	0.0145	[0.1411, 0.5998]
T_1	購買日	0.1860	0.1860	0.0076	[0.1267, 0.2786]
	購買商品	0.0965	0.0965	0.0730	[0.0718, 0.0982]
	価格	0.0048	0.0048	0.0048	[0.0036, 0.0132]
	個数	0.0080	0.0080	0.0025	[0.0017, 0.0152]
T_2	入院日数	1.45E-04	1.45E-04	1.38E-04	[1.46E-04, 1.52E-04]
	年齢	1.33E-04	1.33E-04	9.83E-05	[1.21E-04, 1.42E-04]
	人種	7.73E-05	7.73E-05	5.90E-05	[6.92E-05, 8.31E-05]
	性別	3.78E-05	3.78E-05	2.95E-05	[3.08E-05, 4.30E-05]
T_3	年齢	2.24E-03	2.24E-03	2.24E-03	[2.24E-03, 2.24E-03]
	職業	4.61E-04	4.61E-04	4.61E-04	[4.61E-04, 4.61E-04]
	婚姻状況	2.15E-04	2.15E-04	2.15E-04	[2.15E-04, 2.15E-04]
	人種	1.54E-04	1.54E-04	1.54E-04	[1.54E-04, 1.54E-04]

評価値はサンプリングの結果によって変化する

ここではサンプリングサイズ10のときの90%信頼区間を示している（10レコードではなく，10種類のx）

データの一部しか用いていないが属性の危険度を精度よく評価できる

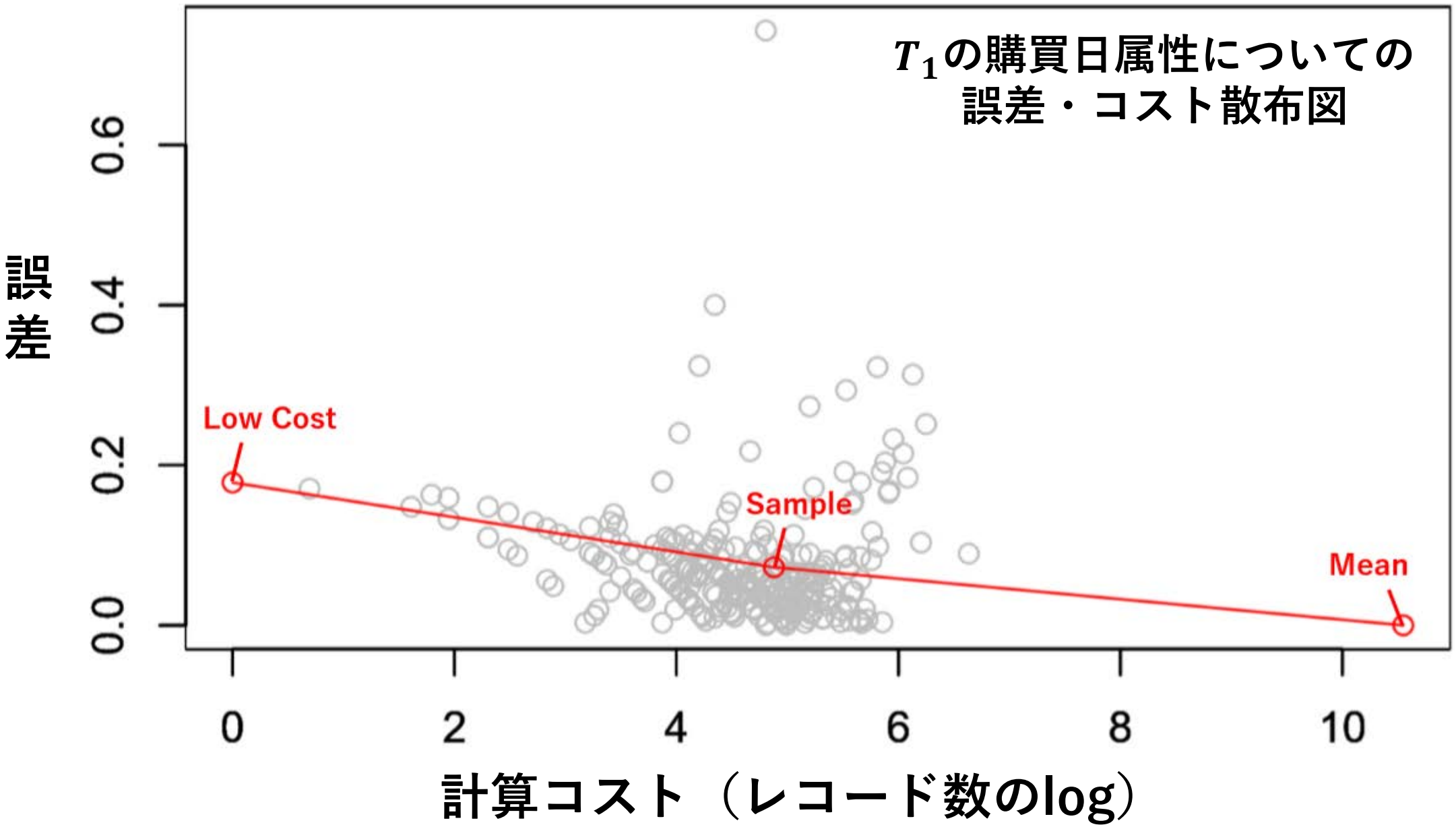
各モデルの評価結果

※サンプリングサイズは10
※90%信頼区間

データ	属性	真値	平均モデル	最小コストモデル	サンプリングモデル
T_1	購買時刻	0.3217	0.3217	0.0145	[0.1411, 0.5998]
	購買日	0.1860	0.1860	0.0076	[0.1257, 0.2780]
	購買商品	0.0965	0.0965	0.0730	[0.0718, 0.0982]
	価格	0.0121	0.0121	0.0048	[0.0096, 0.0146]
	個数	0.0080	0.0080	0.0025	[0.0015, 0.0035]
T_2	入院日数	1.45E-04	1.45E-04	1.38E-04	[1.46E-04, 1.52E-04]
	年齢	1.33E-04	1.33E-04	9.83E-05	[1.21E-04, 1.42E-04]
	人種	7.73E-05	7.73E-05	5.90E-05	[6.92E-05, 8.31E-05]
	性別	3.78E-05	3.78E-05	2.95E-05	[3.08E-05, 4.30E-05]
T_3	年齢	2.24E-03	2.24E-03	2.24E-03	[2.24E-03, 2.24E-03]
	職業	4.61E-04	4.61E-04	4.61E-04	[4.61E-04, 4.61E-04]
	婚姻状況	2.15E-04	2.15E-04	2.15E-04	[2.15E-04, 2.15E-04]
	人種	1.54E-04	1.54E-04	1.54E-04	[1.54E-04, 1.54E-04]

**データによっては
最小コストモデルでも
属性の危険度を
精度よく評価できる**

各モデルの比較 (誤差・計算コスト)



まとめ

- 匿名加工の研究には攻撃者の想定が不可欠である
- 元データのある属性から背景知識を得る攻撃者を想定し，それらの危険度の期待値(平均識別確率)を用いたリスク評価モデルを提案した
- 提案したモデルを用いて3つの実データを評価し，データ中の危険な属性を明らかにした
- 平均識別確率を近似する3つのモデルを提案し，それらの精度とコストを評価した

質疑応答用スライド

平均識別確率とα_x

R_x：xを満たすレコードの集合，U_x：xを満たすユーザの集合

m：データのレコード数，D_X：属性Xの値の集合，$\omega_X = |D_X|$

属性 X の背景知識 x について，$\frac{|R_x|}{|U_x|} = \alpha_x$とおくと，平均識別確率は

$$Pr(\mathbf{idf}, X) = \sum_{x \in D_X} Pr(x) Pr(\mathbf{idf}|x) = \sum_{x \in D_X} \frac{|R_x|}{m} \frac{1}{|U_x|} = \sum_{x \in D_X} \frac{\alpha_x}{m}$$

と求めることができる．

しかし，ビッグデータについて全てのα_xを計算するのは困難なので，

これを近似してリスク評価を行うモデルを提案する．

3つの近似モデル

1. 平均モデル

$$\mathbf{R_{mean}(X)} = \sum_{x \in D_X} \frac{\alpha_X}{m} = \frac{\alpha_X \omega_X}{m}$$

2. 最小コストモデル

$$\mathbf{R_{cost}(X)} = \sum_{x \in D_X} \frac{1}{m} = \frac{\omega_X}{m}$$

3. サンプリングモデル

$$\mathbf{R_{sample}(X)} = \sum_{x \in D_X} \frac{\alpha_{x'}}{m} = \frac{\alpha_{x'} \omega_X}{m}$$

各データの概要・$Pr(\mathrm{idf}, X)$

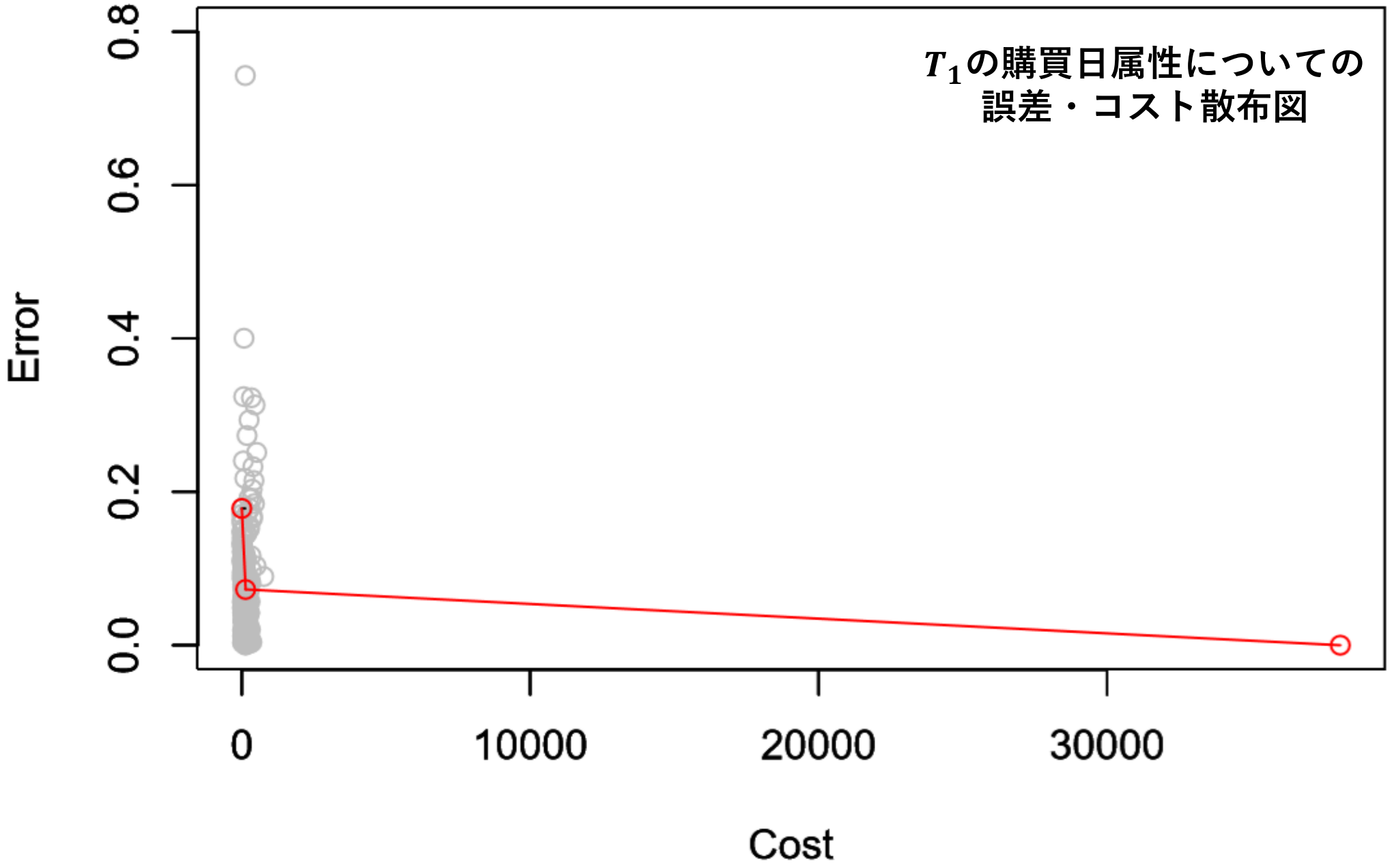
各データの大きさ

	m	n	#Attribute
T_1	38,087	400	7
T_2	101,766	71,518	50
T_3	32,561	32,561	16

データ・属性によって
α_Xの値は大きく異なる

各属性の分析

T	X	α_X	ω_X	$Pr(\mathrm{idf}, X)$	σ
T_1	Date	24.42	290	0.186	0.140
	Time	22.23	551	0.322	0.228
	Goods	1.32	2781	0.097	0.151
	Price	2.49	184	0.012	0.066
	Number	3.15	97	0.008	0.043
T_2	Race	1.31	6	$7.73 \cdot 10^{-5}$	$2.08 \cdot 10^{-4}$
	Gender	1.28	3	$3.78 \cdot 10^{-5}$	$1.81 \cdot 10^{-3}$
	Age	1.35	10	$1.33 \cdot 10^{-4}$	$3.20 \cdot 10^{-4}$
	Time	1.05	14	$1.45 \cdot 10^{-4}$	$1.66 \cdot 10^{-4}$
T_3	Age	1	73	$2.24 \cdot 10^{-3}$	$1.01 \cdot 10^{-2}$
	Martial	1	7	$2.15 \cdot 10^{-4}$	$1.20 \cdot 10^{-3}$
	Occupation	1	15	$4.61 \cdot 10^{-4}$	$1.21 \cdot 10^{-3}$
	Race	1	5	$1.54 \cdot 10^{-4}$	$4.79 \cdot 10^{-4}$


T_1の購買日属性についての
誤差・コスト散布図
Error
0.0
0.2
0.4
0.6
0.8
0
10000
20000
30000
Cost