乗降と物販履歴データの識別リスク分析と匿名加工の検討

伊藤聡志　原田玲央　菊池浩明
明治大学

研究背景

- 2015年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.

- 履歴データから個人が識別されるリスクを明らかにする必要がある.

先行研究

乗降履歴データの識別リスク

- “乗降履歴データの安全な匿名化は可能か？”（菊池，高橋，2014）

購買履歴データの識別リスク

- “PWSCUP:履歴データを安全に加工せよ”（菊池，小栗，野島，濱田，村上，山岡，山口，渡辺，2016）

しかし，その相関や，組み合わせから生じるリスクについては不明だった.

問題点

1. 何件の履歴から個人を識別できるか？

2. 交通履歴と物販履歴で，識別リスクが高いのはどちらか？

3. 交通履歴と物販履歴を組み合わせるとリスクは増えるか？

本研究で注目するデータ

本研究では5種類の用途の履歴からなる
交通ICカード総合履歴データに注目した.

顧客ID	日付	回数	乗車駅	降車駅	乗車路線	降車路線	用途	使用場所	料金
1	2016/10/30	2	上野	高田馬場	JR東北本線	JR山手線	交通	NA	-194
1	2016/10/30	1	高田馬場	上野	JR山手線	JR東北本線	交通	NA	-194
2	2016/10/8	1	NA	NA	NA	NA	チャージ	券売機	2000
2	2016/10/1	1	NA	NA	NA	NA	物販	自販機	-120

研究目的

1. 交通ICカード総合履歴データの分析

2. 同データの識別リスク分析

3. 同データの匿名加工・評価指標の検討

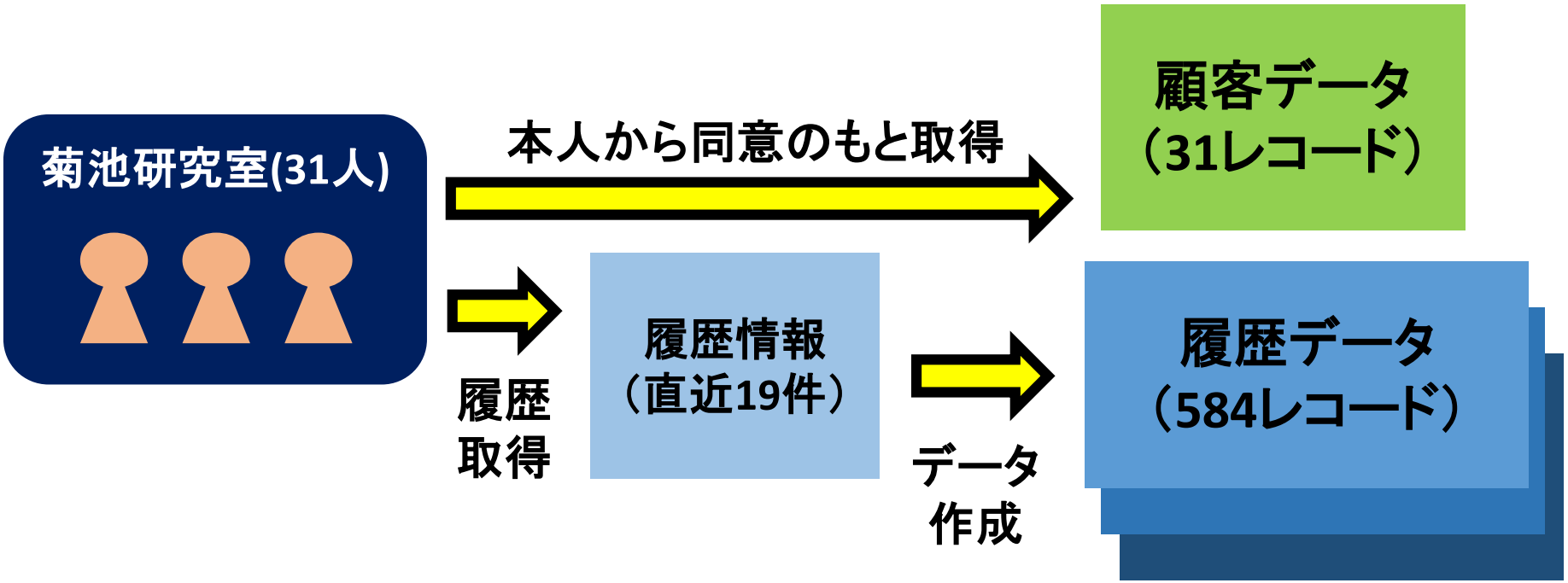
1. 交通ICカード総合履歴データの分析

2. 同データの識別リスク分析

3. 同データの匿名加工・評価指標の検討

交通ICカード総合履歴データ

明治大学菊池研究室に所属する31人（n=31）の情報や
交通ICカードから,顧客データと履歴データを取得した



データの例

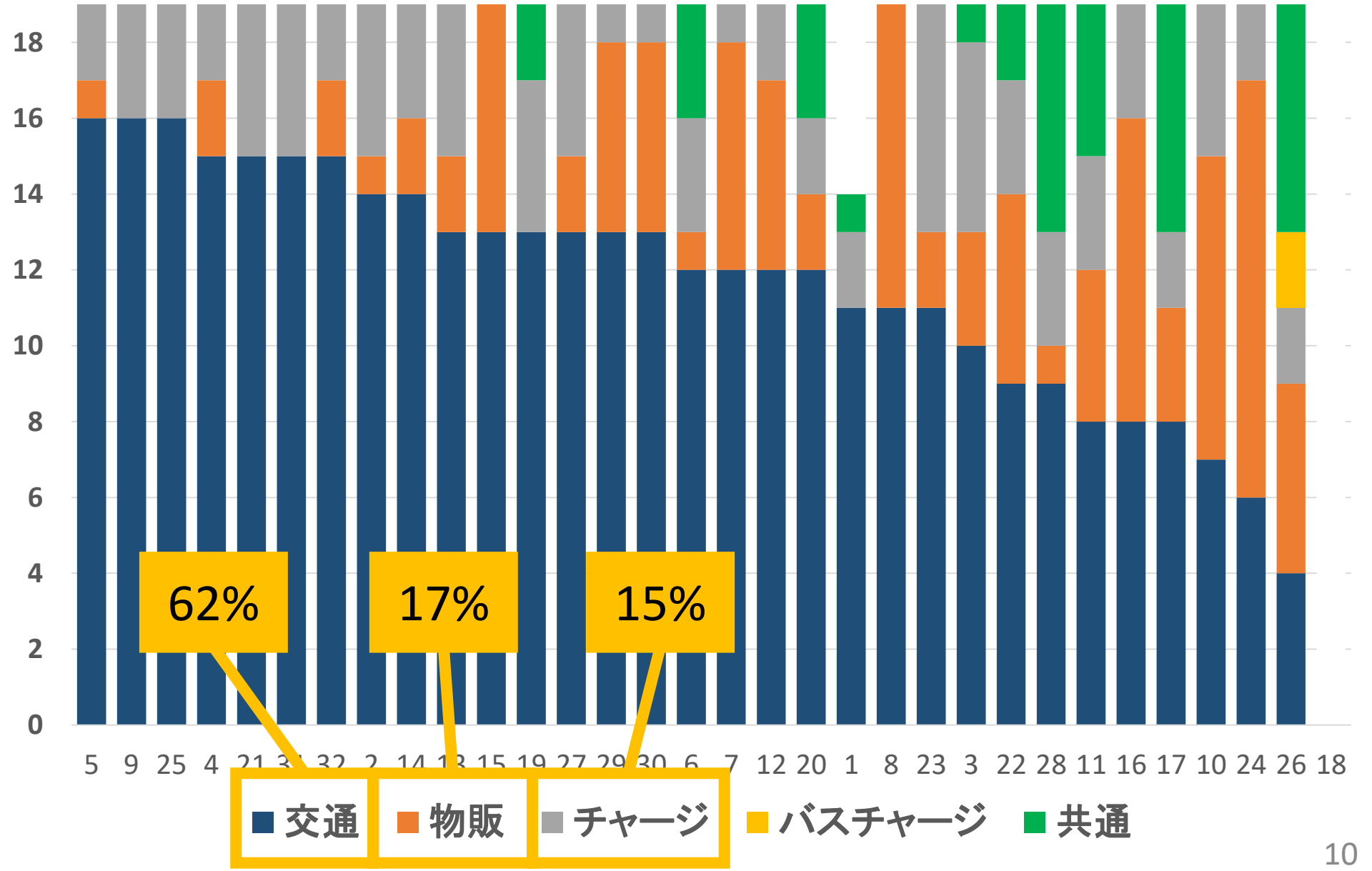
顧客データ例 M

顧客ID	性別	学年	住所	定期券範囲1	定期券範囲2
1	男	1	千葉県	NA	NA
2	女	3	東京都	中野	新宿

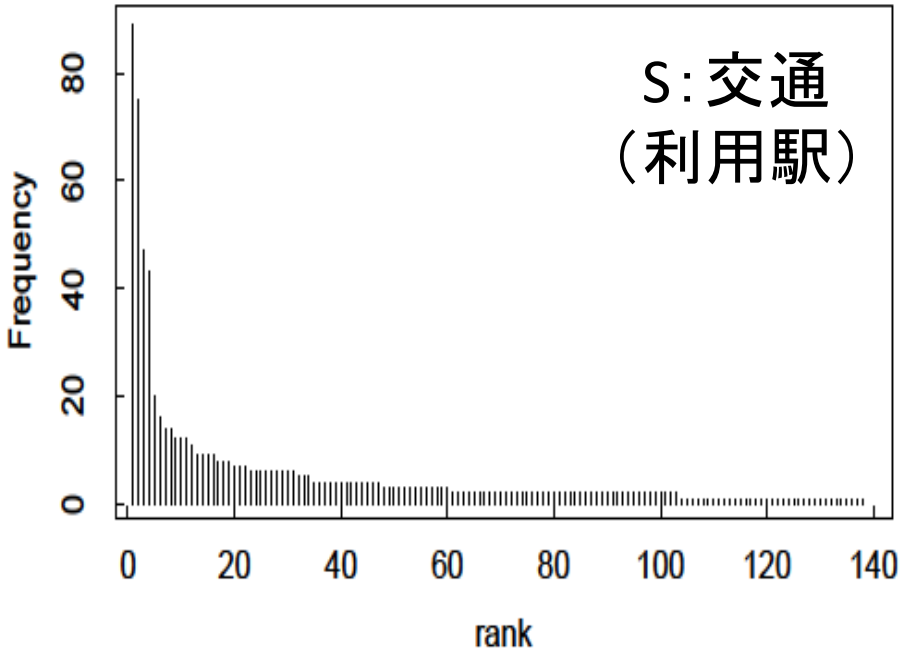
履歴データ例 T

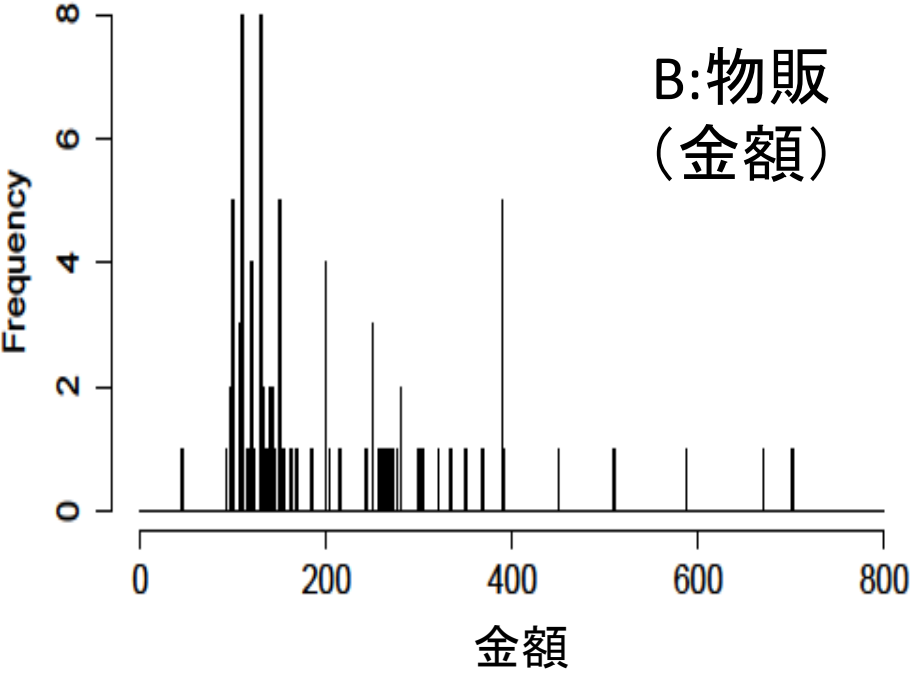
顧客ID	日付	回数	乗車駅	降車駅	乗車路線	降車路線	用途	使用場所	料金
1	2016/10/30	2	上野	高田馬場	JR東北本線	JR山手線	交通	NA	-194
1	2016/10/30	1	高田馬場	上野	JR山手線	JR東北本線	交通	NA	-194
2	2016/10/8	1	NA	NA	NA	NA	チャージ	券売機	2000

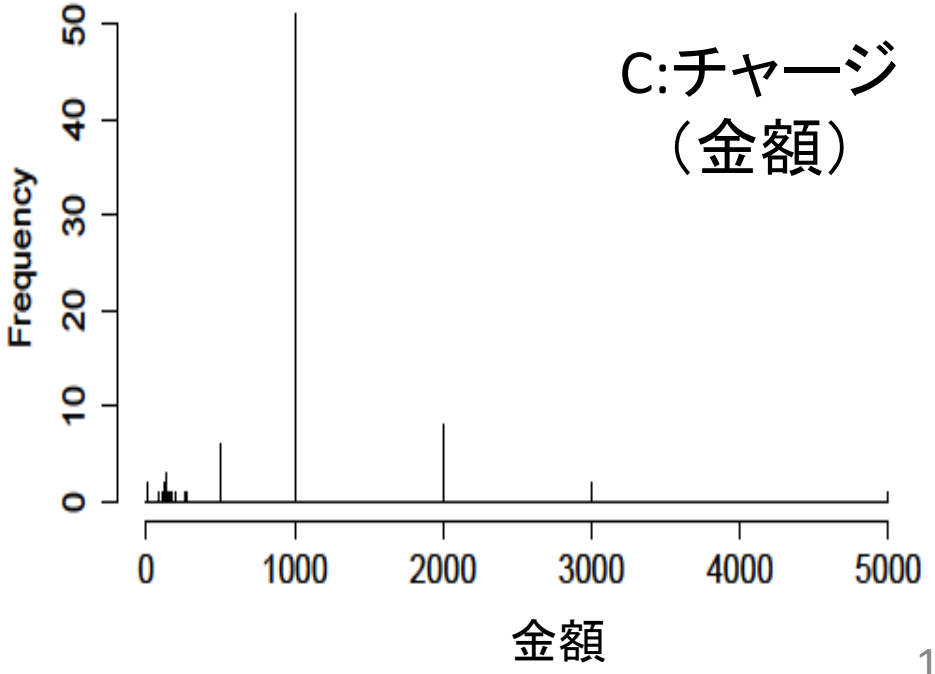
顧客ごとの用途の内訳


18
16
14
12
10
8
6
4
2
0
62%
17%
15%
5 9 25 4 21 32 2 14 15 19 27 29 30 6 7 12 20 1 8 23 3 22 28 11 16 17 10 24 26 18
交通
物販
チャージ
バスチャージ
共通

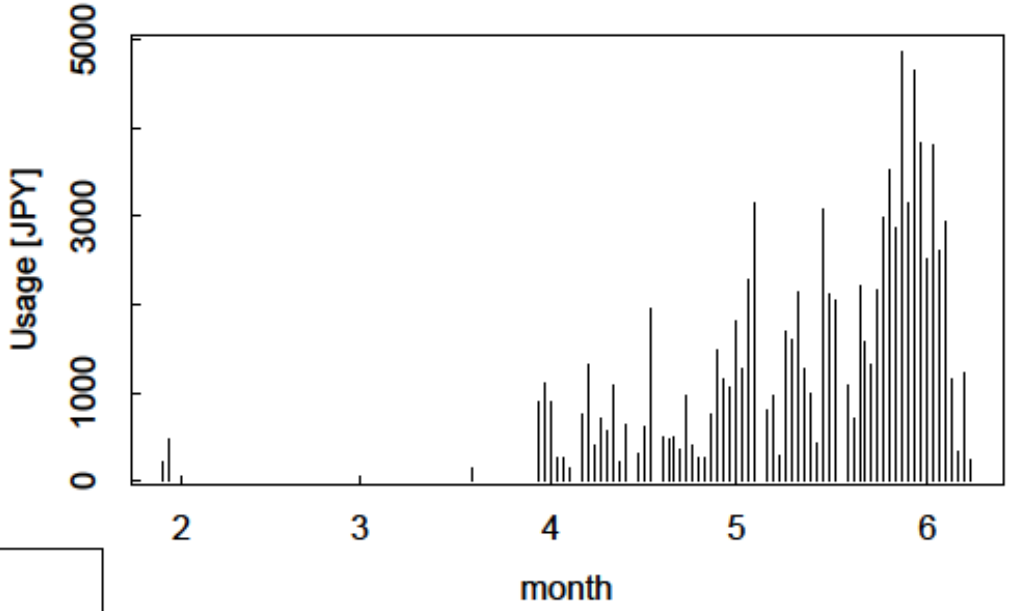
各値の出現頻度


S:交通
（利用駅）
Frequency
0
20
40
60
80
0
20
40
60
80
100
120
140
rank


B:物販
（金額）
Frequency
0
2
4
6
8
0
200
400
600
800
金額

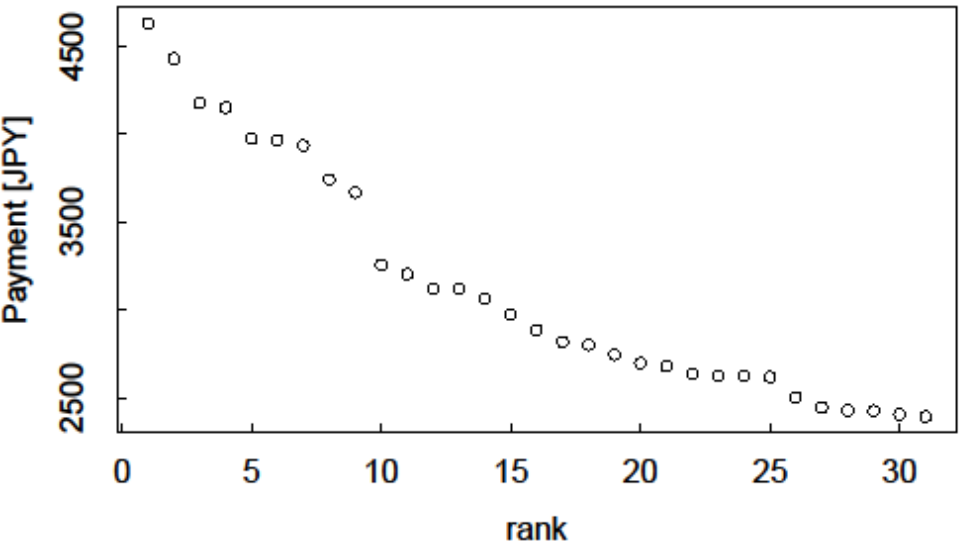

C:チャージ
（金額）
Frequency
0
10
20
30
40
50
0
1000
2000
3000
4000
5000
金額

使用料金についての分析



月ごとの使用料金

各顧客の総使用料金



データ取得日
5月下旬～6月上旬
（顧客によって異なる）

1. 交通ICカード総合履歴データの分析

2. 同データの識別リスク分析

3. 同データの匿名加工・評価指標の検討

識別リスク

A

顧客ID	利用駅1	利用駅2
1	新宿	中野
2	新宿	中野
3	新宿	中野
4	新宿	中野
5	新宿	中野

各顧客が同じような駅を
利用している
↓
個人は識別されにくい

B

顧客ID	利用駅1	利用駅2
1	新宿	中野
2	静岡	浜松
3	岐阜	大垣
4	熱海	品川
5	島田	藤枝

各顧客が利用している
駅が似ていない
↓
個人が識別されやすい

駅の識別しにくさ

交通履歴についての集計表例

ユーザ/駅	中野	品川	熱海
u_1	2	1	0
u_2	4	0	4
u_3	4	4	0

全員利用している
識別リスク小さい

１人しか利用していない
識別リスク大きい

駅の識別しにくさ　　中野 ＞ 品川 ＞ 熱海

条件付きエントロピー

ユーザ / 駅	中野	品川	熱海	計	$P(U=u_i)$
u_1	2	1	0	3	3/19
u_2	4	0	4	8	8/19
u_3	4	4	0	8	8/19
$H(U\|S=s_i)$	1.52	0.72	0		
$P(S=s_i)$	10/19	5/19	4/19		

条件付きエントロピー$[bit/履歴]$

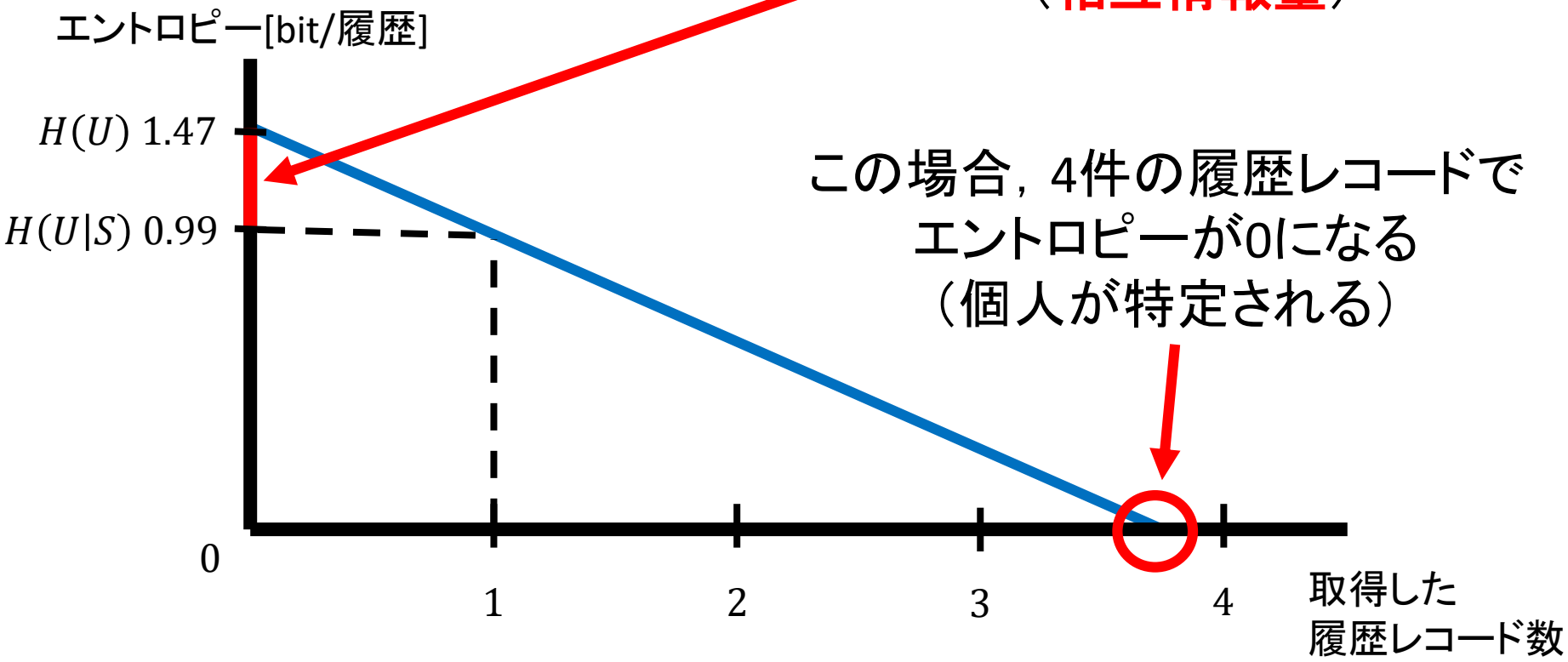
$$1.52 = H(U|S=\text{中野}) > (U|S=\text{品川}) > H(U|S=\text{熱海}) = 0$$

$$H(U) = -\sum_{i=1}^{n} P(U=u_i)\, log_2 P(U=u_i) = 1.47$$

$$H(U|S) = -\sum_{i=1}^{m} P(S=s_i)\, H(U|S=s_i) = 0.99 \qquad I(U;S) = H(U) - H(U|S) = 0.48$$

相互情報量

1履歴レコードから得られる
情報量の期待値
（**相互情報量**）



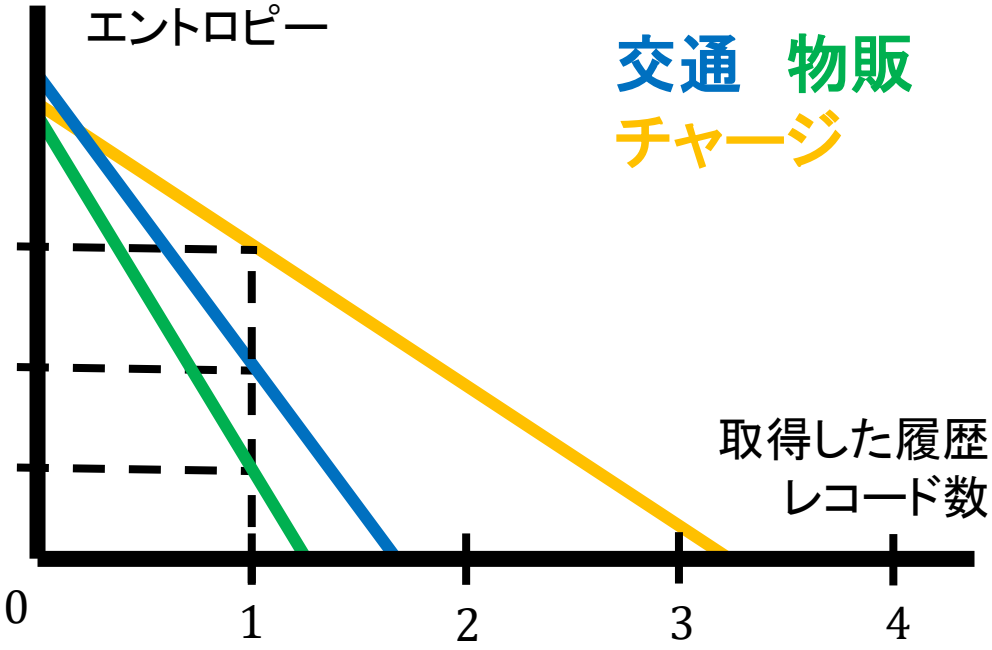
この場合，4件の履歴レコードで
エントロピーが0になる
（個人が特定される）

個人が識別される平均確率は$1/2^{\text{エントロピーの値}}$と等しい

実データのエントロピー等の値

交通ICカード結合履歴データを用途ごとに分け，識別リスクを分析した

	交通 (S)	物販 (B)	チャージ (C)
$H(U)$	4.900	4.338	4.736
$H(U\|x)$	1.814	0.948	3.256
$I(U;x)$	3.085	**3.389**	1.479
$P(U\|x)$	0.284	0.518	0.105



- 交通・物販の履歴は2件判明するとほぼ個人を特定できてしまう
- 物販履歴の情報量が最も大きく，リスクが高い

複数の用途の組み合わせ

交通履歴についての集計表例

ユーザ/駅	s_1	s_2	s_3
u_1	2	1	0
u_2	4	0	4
u_3	4	4	0

物販履歴についての集計表例

ユーザ/料金	b_1	b_2
u_1	2	0
u_2	1	3
u_3	0	1

交通･物販履歴についての集計表例

	s_1, b_1	s_1, b_2	s_2, b_1	s_2, b_2	s_3, b_1	s_3, b_2
u_1	4	0	2	0	0	0
u_2	4	12	0	0	4	12
u_3	0	4	0	4	0	0

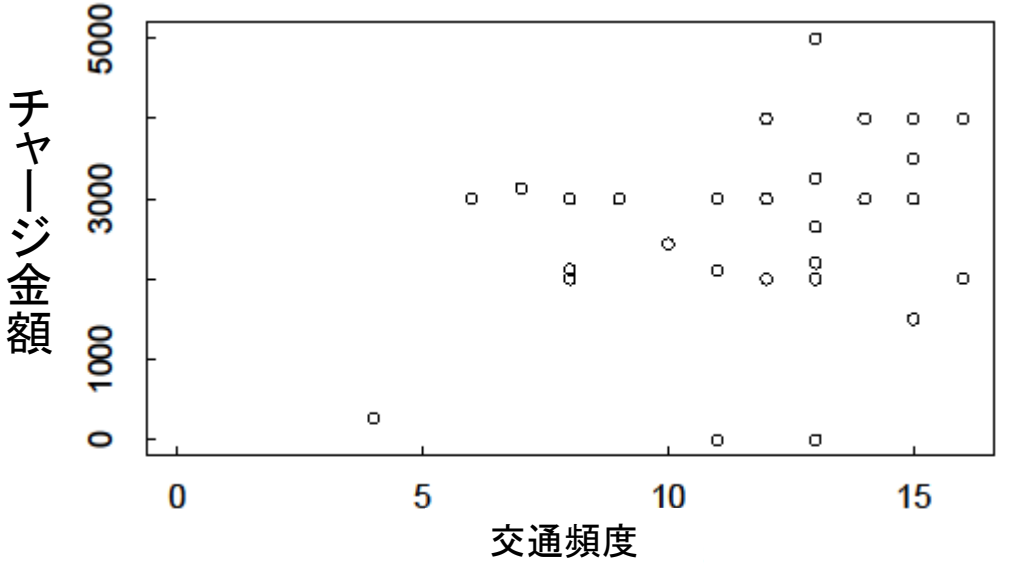
交通と物販を組み合わせたリスク

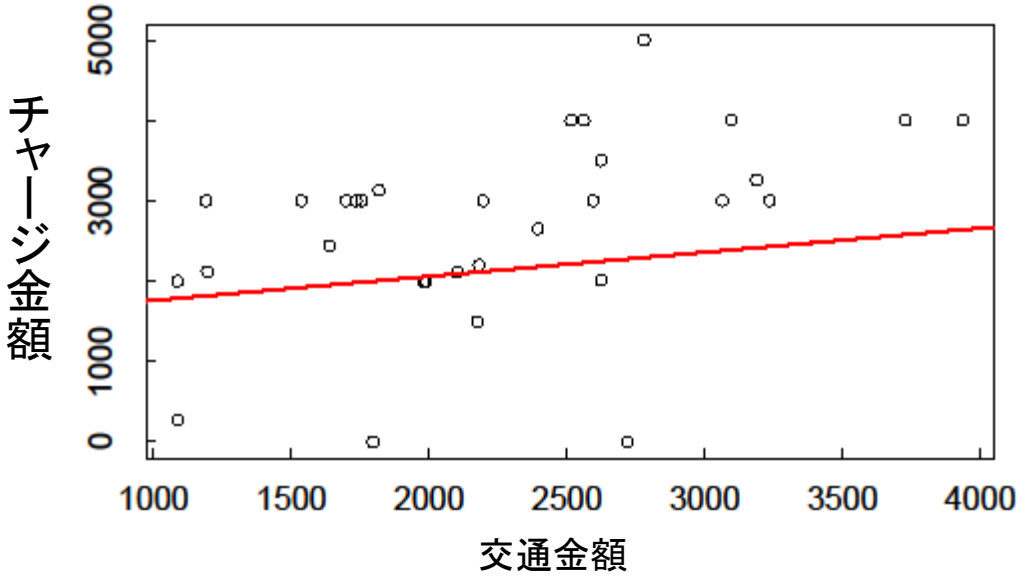
「交通」「物販」用途から1履歴ずつ与えられた場合，
個人が識別されるリスクは88.1%まで増加

	交通・物販(S,B)
$H(U)$	4.412
$H(U\|x)$	0.182
$I(U;x)$	4.230
$P(U\|x)$	0.881

$I(U;S)+I(U;B)=6.474>4.230=I(U;S,B)$であるため，
交通と物販は独立ではない

交通とチャージの相関





- チャージ金額と交通頻度・交通金額の間には弱い相関があった（相関係数は順に0.469，0.315）
- 交通用途の情報からチャージ用途の情報が推測されるリスクも考えられる

分析のまとめ

1. 何件の履歴から個人を識別できるか？

   →交通・物販履歴は2件，チャージ履歴は4件

2. 交通履歴と物販履歴で，識別リスクが高いのはどちらか？

   →物販履歴の情報量が最も大きい

3. 交通と物販を組み合わせるとリスクは増えるか？

   →88.1%まで増加した

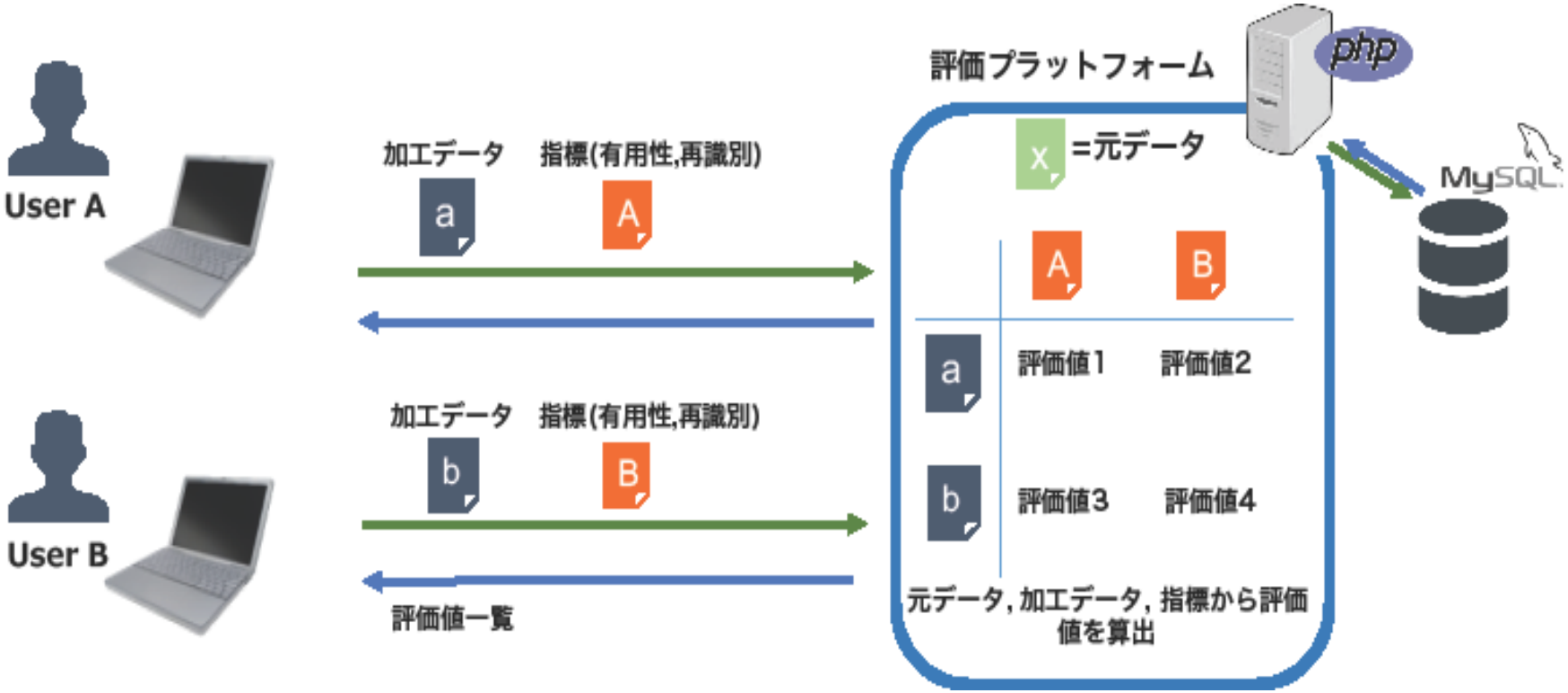
1. 交通ICカード総合履歴データの分析

2. 同データの識別リスク分析

3. 同データの匿名加工・評価指標の検討

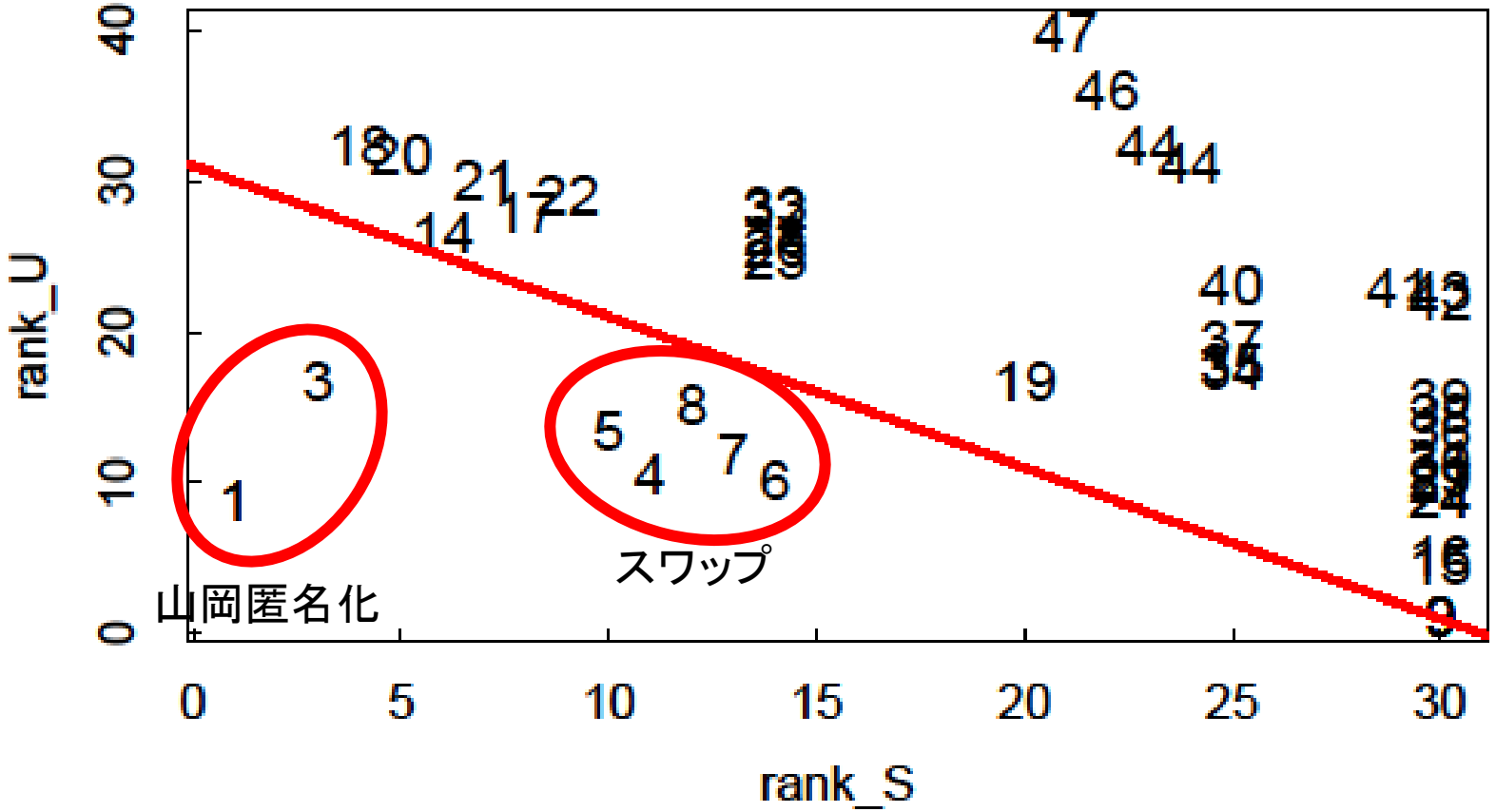
プチPWSCUP　1

交通ICカード総合履歴データに対する匿名加工手法と評価指標を検討するために，コンテスト形式（プチPWSCUP）の実験を行った



プチPWSCUP　2

プチPWSCUP（参加者4名・提出データ47個）の結果，山岡匿名化や属性スワップ等の手法で加工されたデータの評価が高かった



まとめ

1. 31人の交通ICカードから履歴データを取得し，分析を行った結果，交通ICカードは交通以外に物販やチャージ用途にも多く利用されていることがわかった.
（交通：62%，物販：17%，チャージ：15%）

2. エントロピー等を用いて識別リスクを分析した結果，物販と交通の履歴は2件判明すると個人を特定でき，物販用途の履歴の情報量が最も多く，また，交通と物販の履歴を組み合わせると識別リスクが88.1%まで増加するという結果が得られた.

3. 匿名加工と評価指標を検討するために，コンテスト形式の実験を行った

質疑応答用スライド

	交通(S)	物販(B)	チャージ(C)
$H(U)$	4.900	4.338	4.736
$H(U\|x)$	1.814	0.948	3.256
$I(U;x)$	3.085	3.389	1.479
$P(U\|x)$	0.284	0.518	0.105
n_x	31	25	29
m_x	138	58	17

	交通・物販(S,B)	交通・チャージ(S,C)	物販・チャージ(B,C)
$H(U)$	4.412	4.677	4.149
$H(U\|x)$	0.182	1.065	0.529
$I(U;x)$	4.230	3.612	3.620
$P(U\|x)$	0.881	0.478	0.692
n_x	31	31	31
m_x	8004	2346	986