[招待講演]
複数用途からなる交通IC カードデータの再識別リスク分析
(from AINA 2018)

伊藤聡志，原田玲央，菊池浩明
明治大学

これまでの研究

研究テーマ: 匿名加工と再識別

2015年

- **ユークリッド距離を用いた再識別手法の研究**
  (CSEC-73, NBiS-2017)

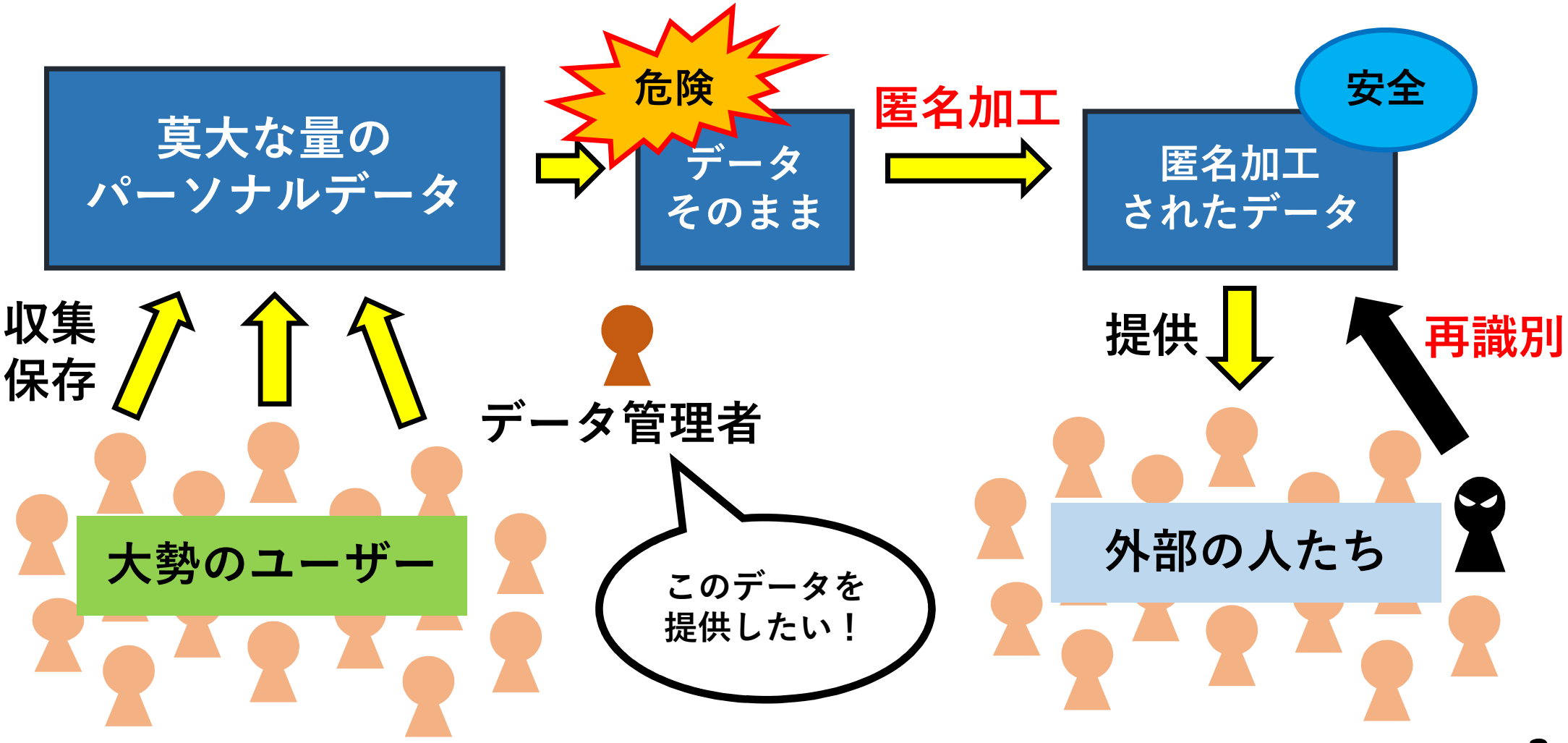
2016年

- **複数用途の履歴を組み合わせた際の識別リスクの研究(本発表)**
  (CSEC-76, AINA-2018)
- **Jaccard距離を用いた再識別手法とそれに対する匿名加工手法の研究**
  (SCIS-2017, MDAI-2019)

2017-2018年

- **攻撃者の持つ背景知識に注目した識別リスク分析の研究**
  (CSS-2017, MDAI-2018, CSEC-84)

匿名加工と再識別とは？

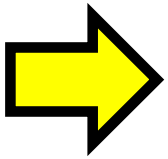

莫大な量の
パーソナルデータ
危険
データ
そのまま
匿名加工
匿名加工
されたデータ
安全
収集
保存
提供
再識別
データ管理者
大勢のユーザー
このデータを
提供したい！
外部の人たち

匿名加工と再識別　例

学生の試験結果

ID	数学	英語	物理
A	40	90	30
B	60	80	40
C	90	0	60
D	80	100	70

加工



加工された学生の試験結果

ID	数学	英語	物理
A	40-60	80-90	35
B	40-60	80-90	35
C	80-90	-	65
D	80-90	-	65

攻撃者

伊藤の数学は
40点だ！

攻撃者は伊藤を
一意に識別できる
伊藤＝A

攻撃者は伊藤を
一意に識別できない
伊藤＝A or B

研究背景

- **国内では改正個人情報保護法で，国外ではGDPRやISOで匿名加工情報が定義された（個人情報を本人が特定できないように加工をしたもので、当該個人情報を復元できないようにした情報）**

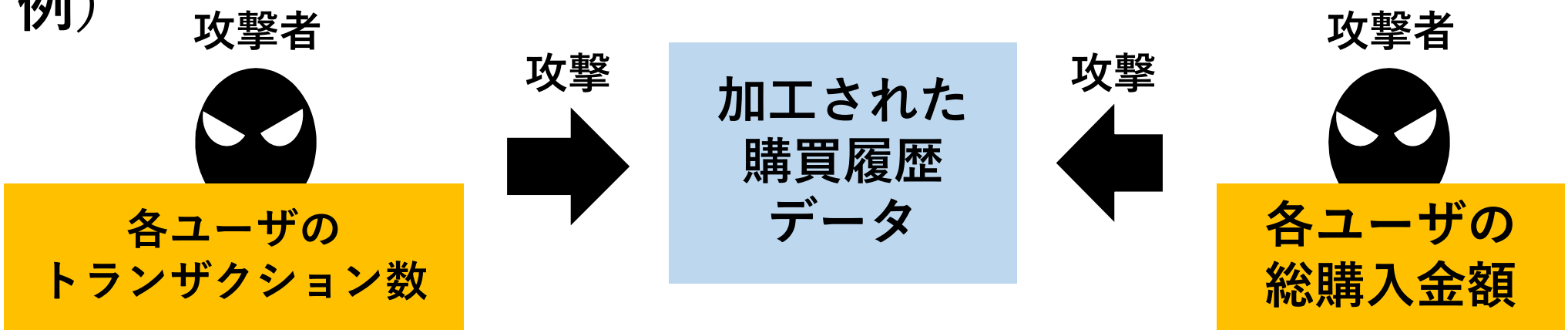
- **個人情報データを匿名加工するためには，そのデータから個人が識別されるリスクの評価を行う必要がある**

- **購買履歴や乗降履歴のような様々なデータに対するリスク評価指標が提案されている**

既存研究と問題点

2015年から開催されている匿名加工コンテストPWS Cupでは購買履歴データのリスク評価指標が数多く提案されている

例）



異なる用途のデータを組み合わせたとき，
識別リスクはどう変化するか？が不明であった

本研究で注目するデータ

本研究では，交通ICカードデータに注目する
交通ICカードは，5種類の用途についての使用履歴を保存している
(交通，物販，チャージ，バスチャージ，共通)

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山手線	JR山手線	交通	NA	-194
1	2016/10/30	1	高田馬場	上野	JR山手線	JR山手線	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

研究目的とリサーチクエスチョン

研究目的

- **異なる用途のデータを組み合わせたときの識別リスク分析**

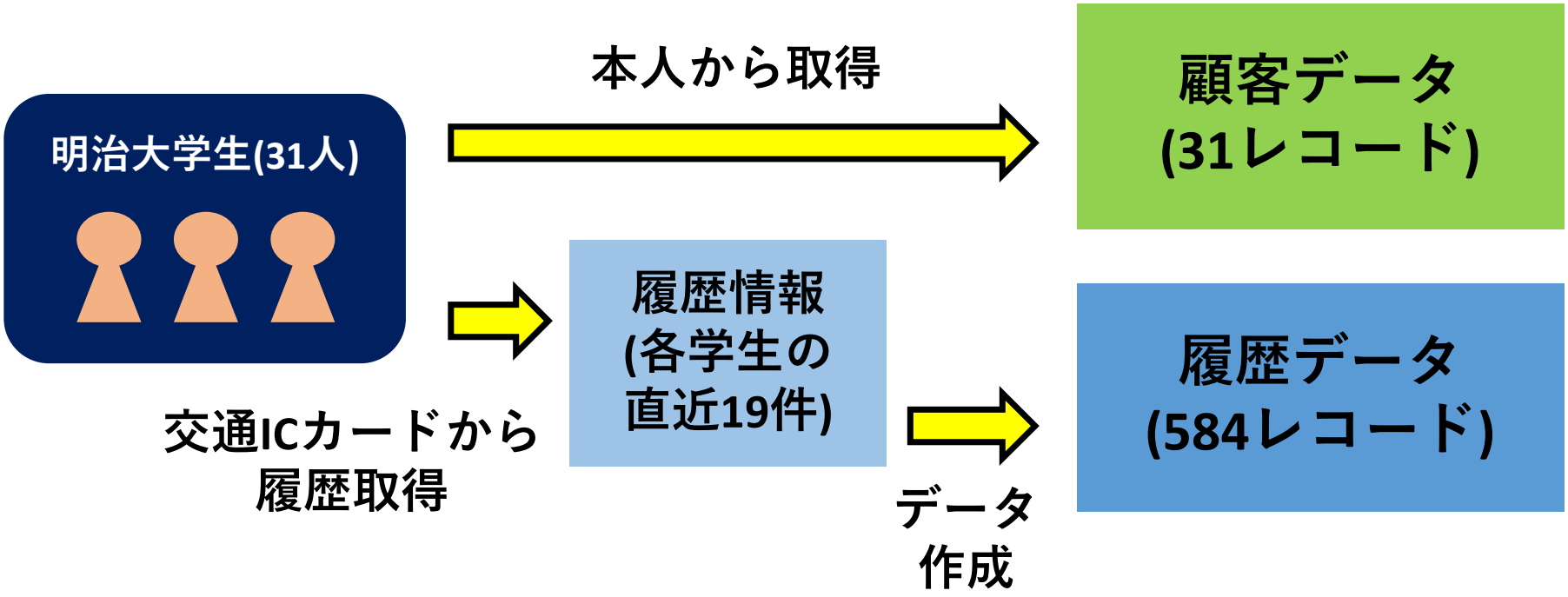
研究目的を達成するために，以下の4点に注目して分析を行う

リサーチクエスチョン

1. **交通ICカードデータはどのような特性を持っているか？**
2. **個人を一意に識別するために，どれだけのレコードが必要か？**
3. **交通データと購買データ，どちらの方がリスクが高いか？**
4. **複数用途のデータが組み合わされたとき，識別リスクはどう変化するか？**

交通ICカードデータ

本研究に用いる交通ICカードデータは，31名の明治大学生から同意のもと収集した584レコードのデータである



交通ICカードデータの例

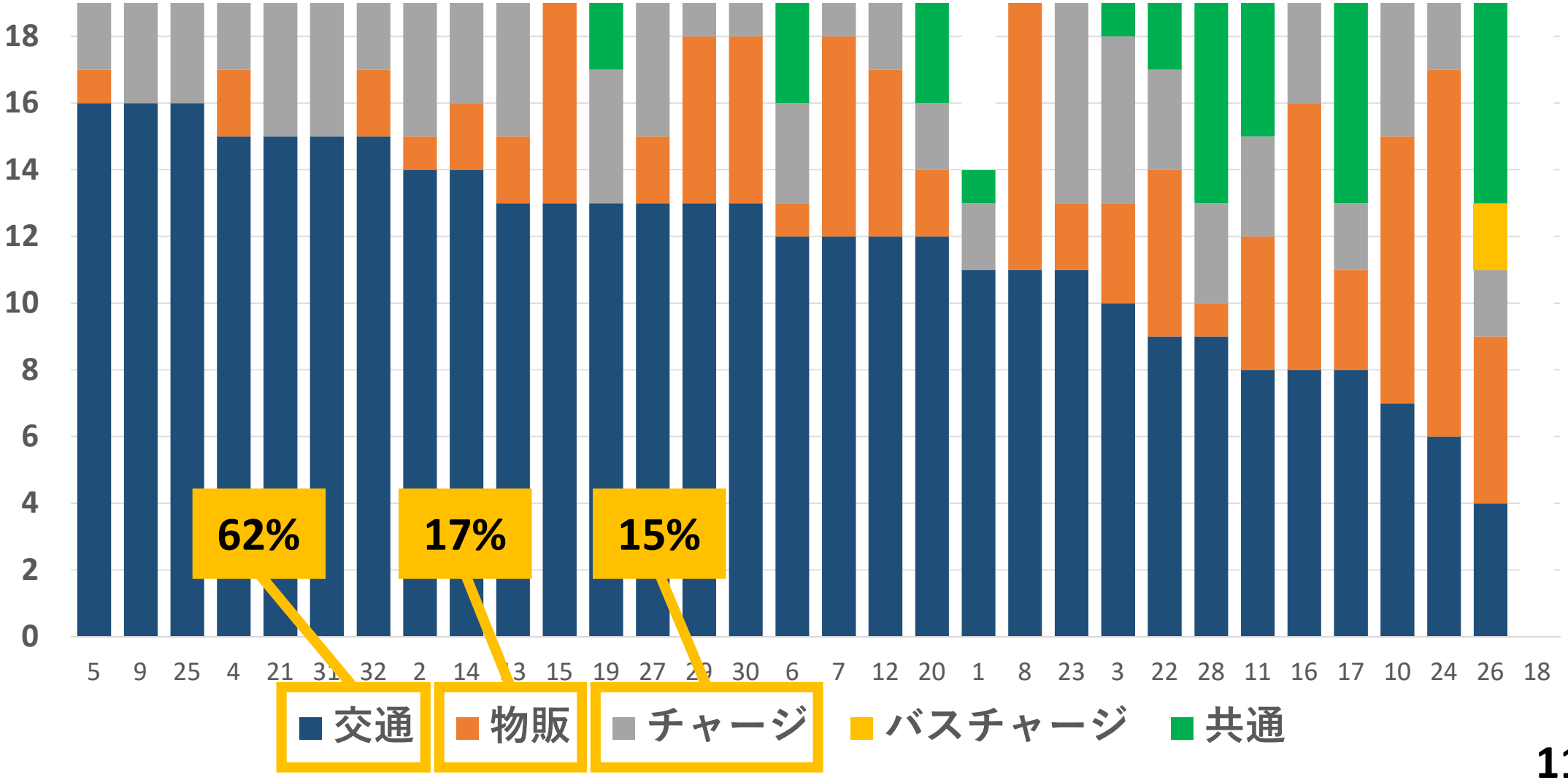
顧客データ例 M

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

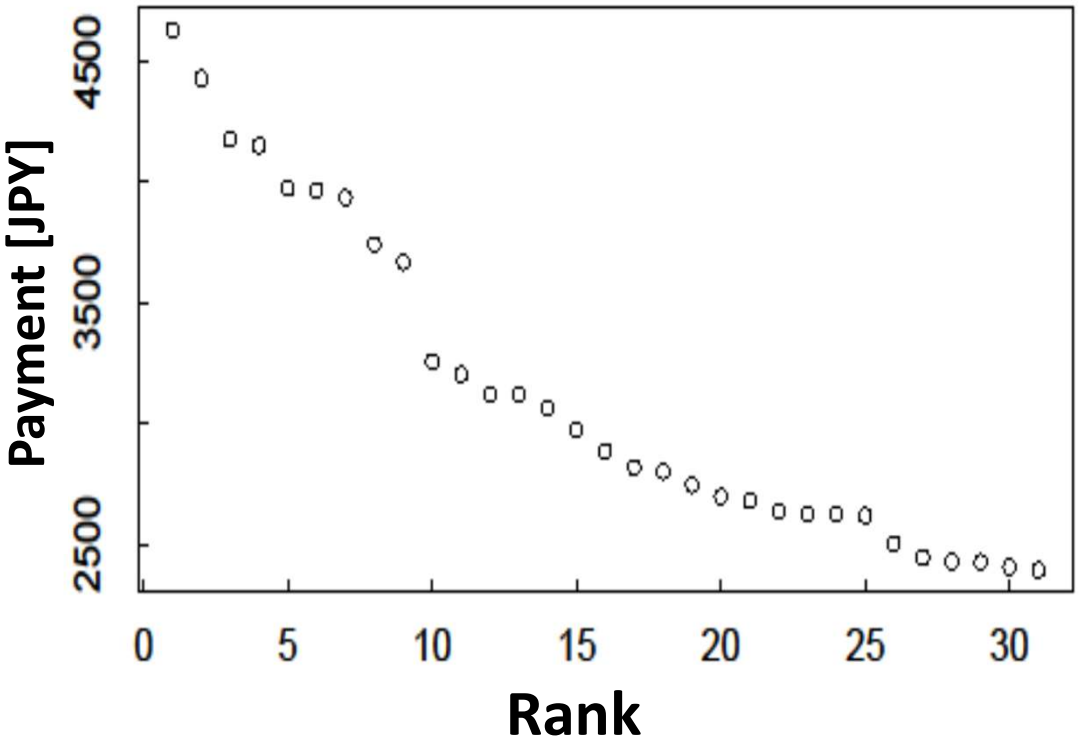
顧客ID	日付	回数	乗車駅	降車駅	乗車路線	降車路線	用途	使用場所	料金
1	2016/10/30	2	上野	高田馬場	JR山手線	JR山手線	交通	NA	-194
1	2016/10/30	1	高田馬場	上野	JR山手線	JR山手線	交通	NA	-194
2	2016/10/8	1	NA	NA	NA	NA	チャージ	券売機	2000

顧客ごとの使用用途の内訳

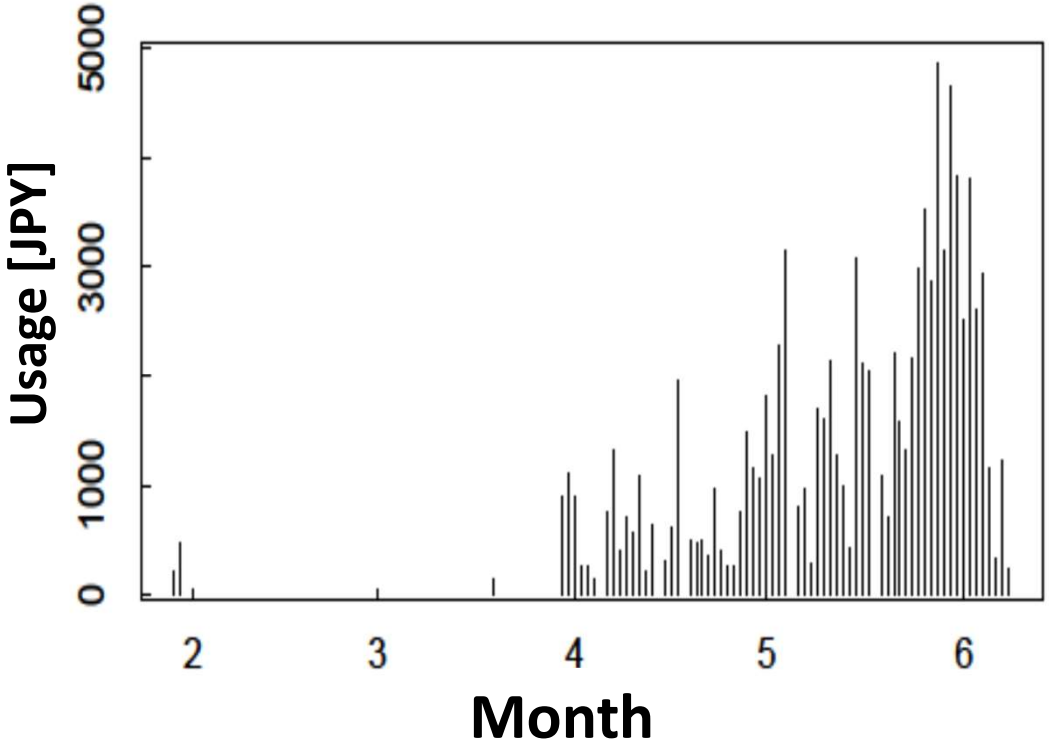


利用金額の分析

全顧客についての利用金額散布図

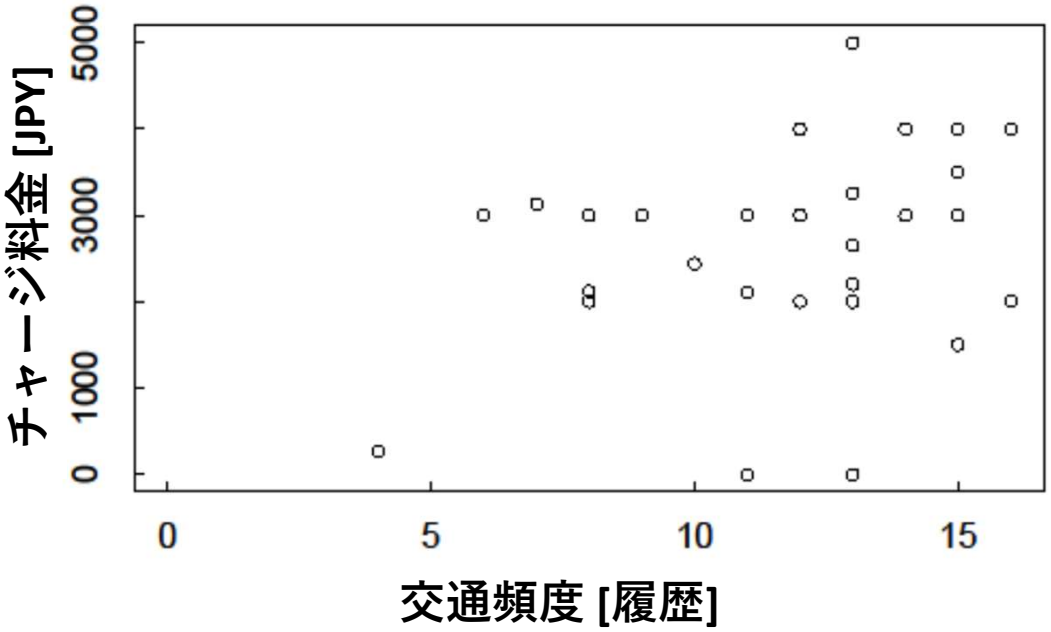


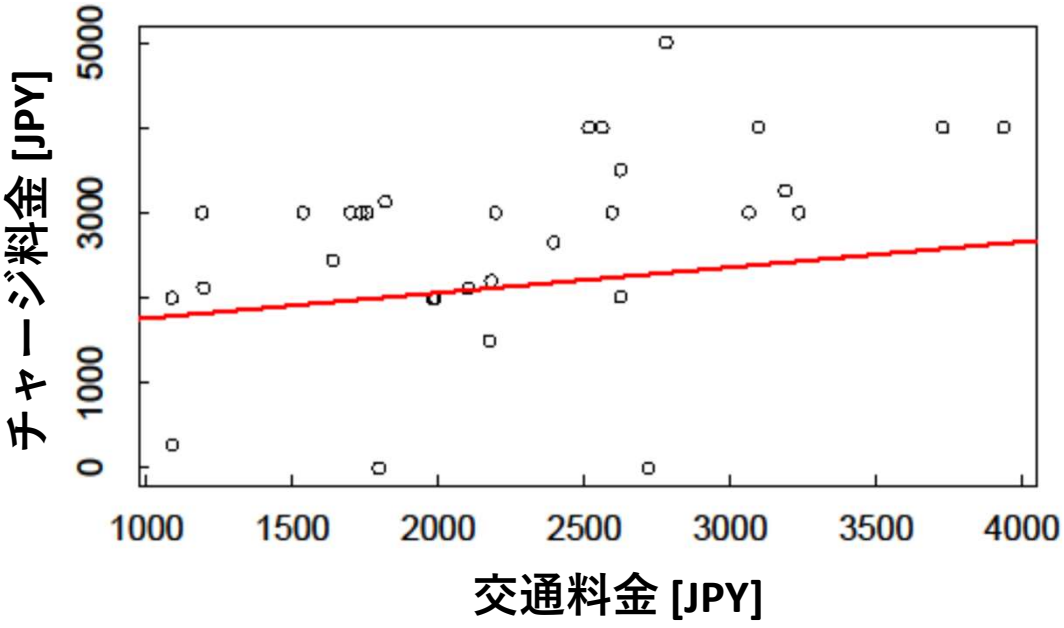
月ごとの利用金額分布



- **顧客によって利用金額や利用時期が大きく異なり，ここから個人が識別されるリスクも考えられる**

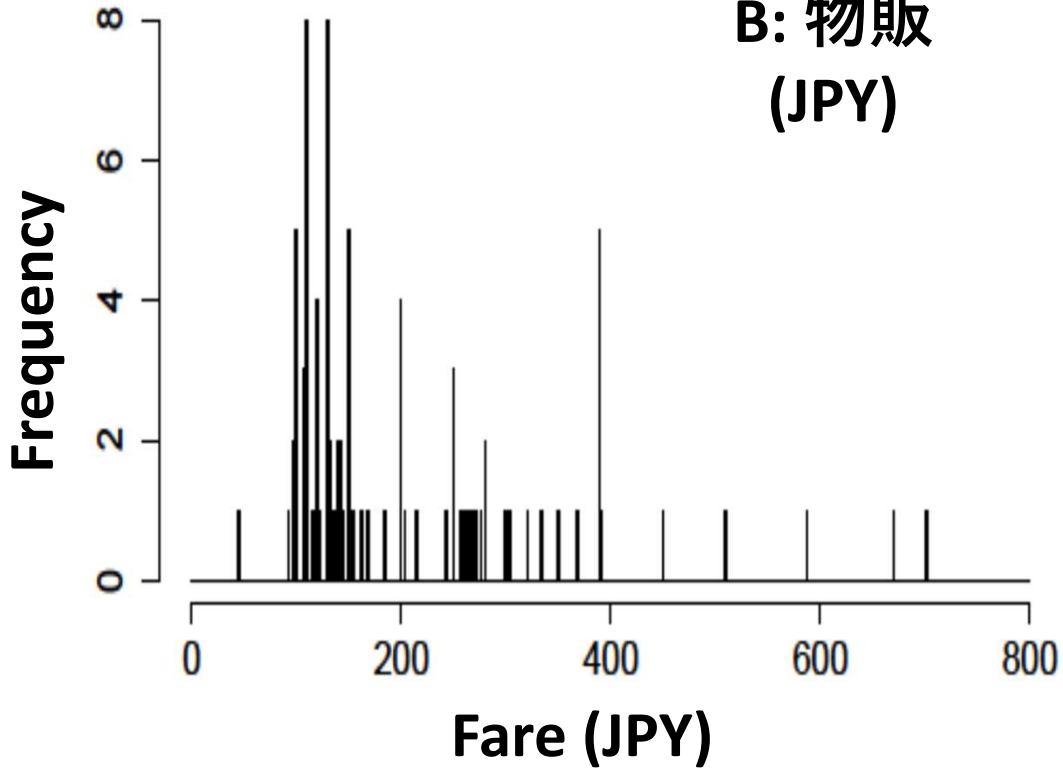
用途間の相関

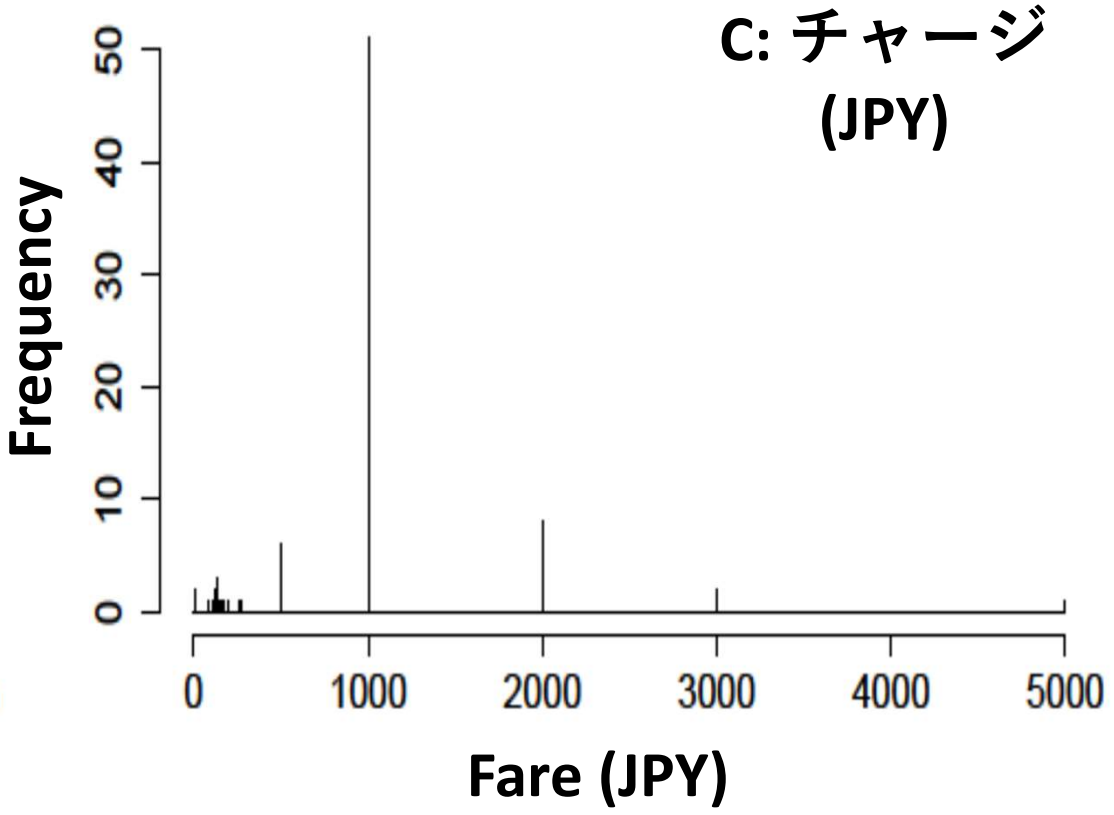




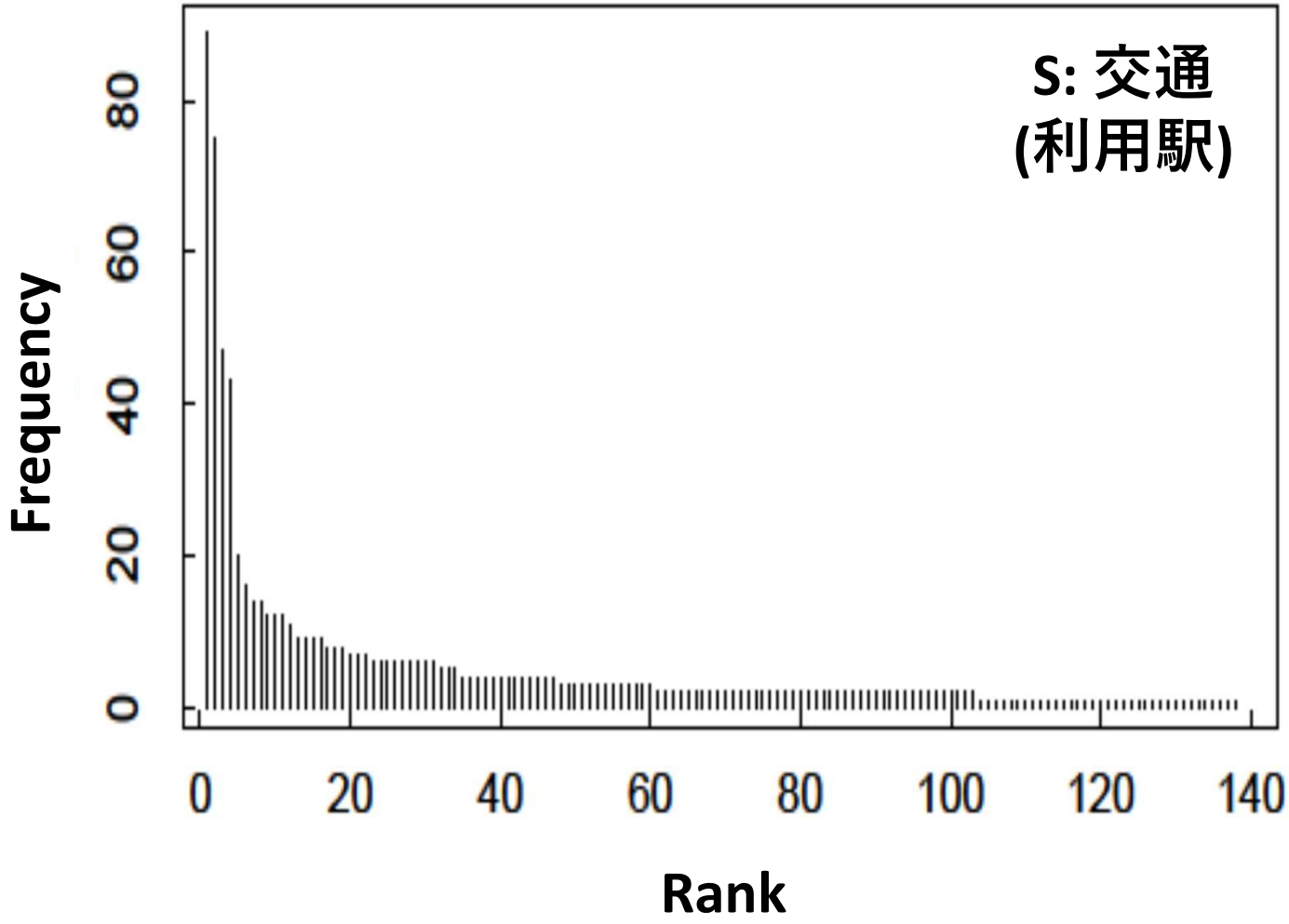
- チャージ金額と交通頻度・交通金額の間には弱い相関があった（相関係数は順に0.469，0.315）
- 交通用途の情報からチャージ用途の情報が推測されるリスクも考えられる

物販・チャージ用途についての金額の頻度





交通用途についての利用駅の頻度



駅名	利用回数
新宿	89
中野	75
渋谷	47
高田馬場	43
明大前	20
…	…

個人の識別されやすさ

交通履歴データ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

交通履歴データ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

利用駅から個人を識別されるリスク

顧客ごとの駅利用回数

ユーザ/利用駅	東京	大阪	京都
u_1	2	1	0
u_2	4	0	4
u_3	4	4	0

東京駅は3人とも
利用しているため
識別リスクが小さい

京都駅を利用しているのは
u_2のみであるため
識別リスクが大きい

識別リスク: 東京<大阪<京都

識別リスクを顧客についてのエントロピーを用いて評価する

条件付きエントロピー

User/Station	東京	大阪	京都	合計	$P(U=u_i)$
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		

元のエントロピー
$H(U) = 1.47$

条件付エントロピー [bit/履歴]

$H(U|S=$東京$) > H(U|S=$大阪$) > H(U|S=$京都$)$

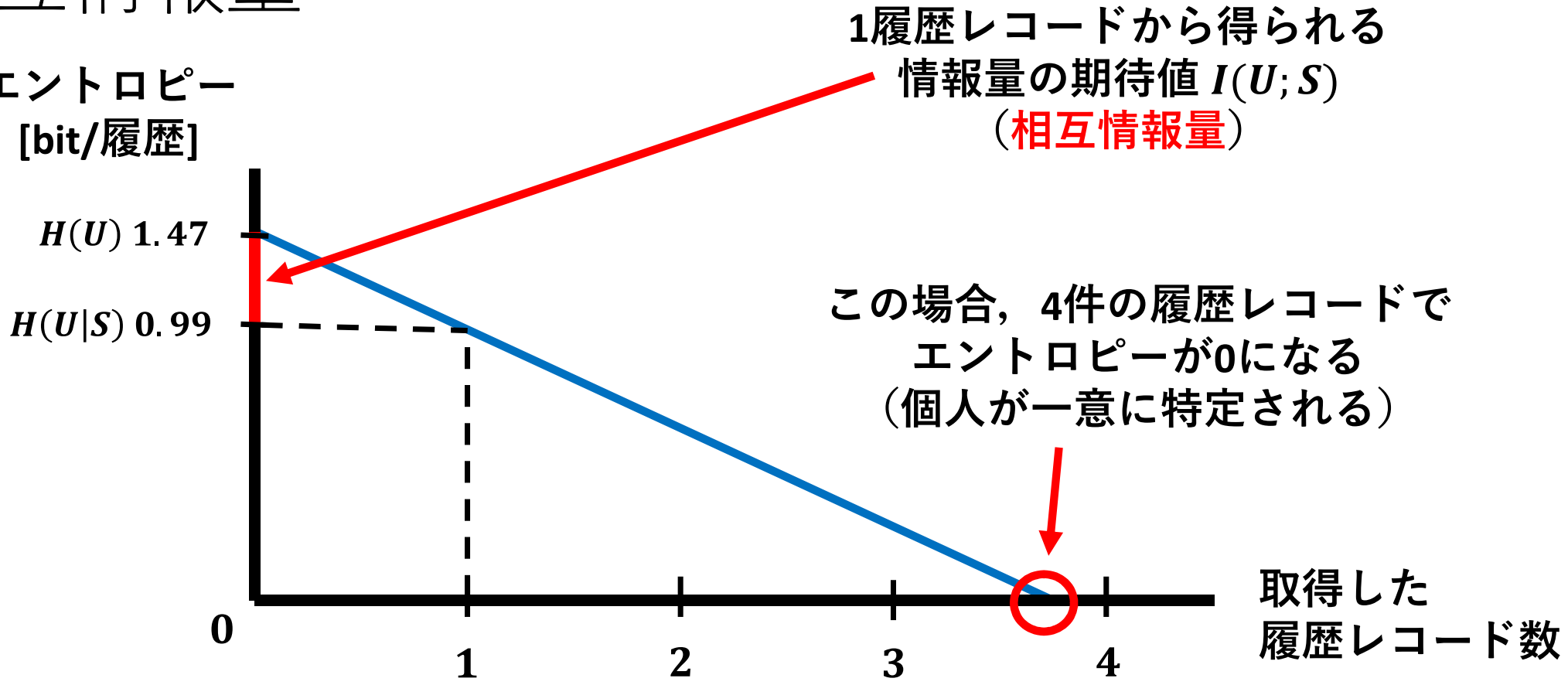
1.52　　0.72　　0

期待値
$H(U|S) = 0.99$

識別リスク: 低

識別リスク: 高

相互情報量

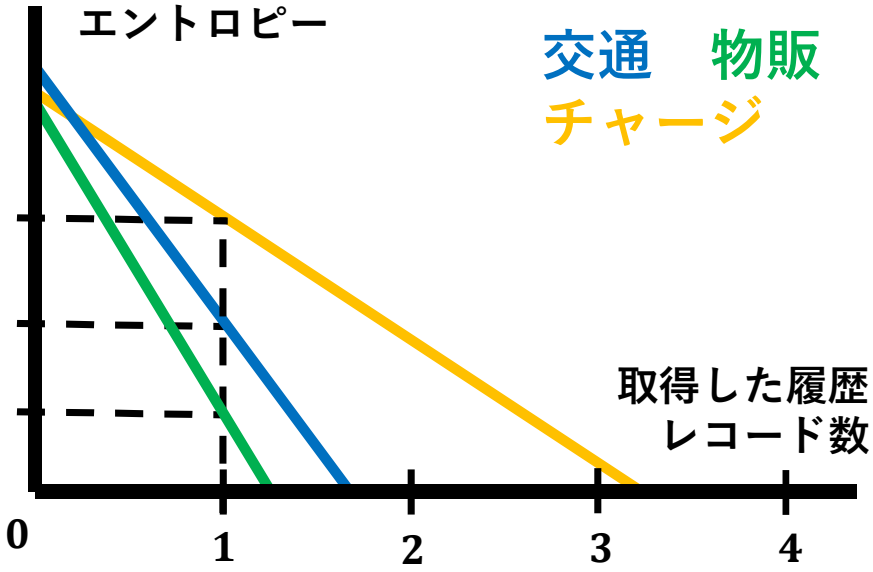


個人が識別される平均確率は$(1/2)$エントロピーの値と等しい

交通ICカードデータのリスク分析

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件判明すると個人を一意に特定できてしまう
- 物販履歴の情報量が3.389で最も大きく，リスクが高い

複数用途の組み合わせ

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・物販履歴を組み合わせた集計表例

	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

複数用途の組み合わせによるリスク

2つの用途を組み合わせた時のエントロピー

	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

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%まで増加する

リサーチクエスチョンの答え

1. 交通ICカードデータはどのような特性を持っているか？
   →5種類の用途からなっており，そのうち交通用途の割合は62％であった

2. 個人を一意に識別するために，どれだけのレコードが必要か？
   →交通・物販履歴は2件，チャージ履歴は4件で個人を一意に識別できる

3. 交通データと購買データ，どちらの方がリスクが高いか？
   →物販履歴の情報量が最も大きく(3.389)，リスクが高い

4. 複数用途のデータが組み合わされたとき，識別リスクはどう変化するか？
   →交通・物販履歴が組み合わされた時，識別リスクは88.1％まで増加した

まとめ

- **匿名加工は国内外で注目されている技術であるが，データを加工する際には個人が識別されるリスクの評価を行う必要がある**

- **乗降履歴や購買履歴のような単一用途のデータについてのリスクは評価されていたが，複数用途のデータを組み合わせた際のリスクは不明であった**

- **複数用途からなる交通ICカードデータのリスクをエントロピーを用いて評価した結果，乗降履歴と購買履歴を組み合わせると88.1%で個人が一意に識別されることが分かった**