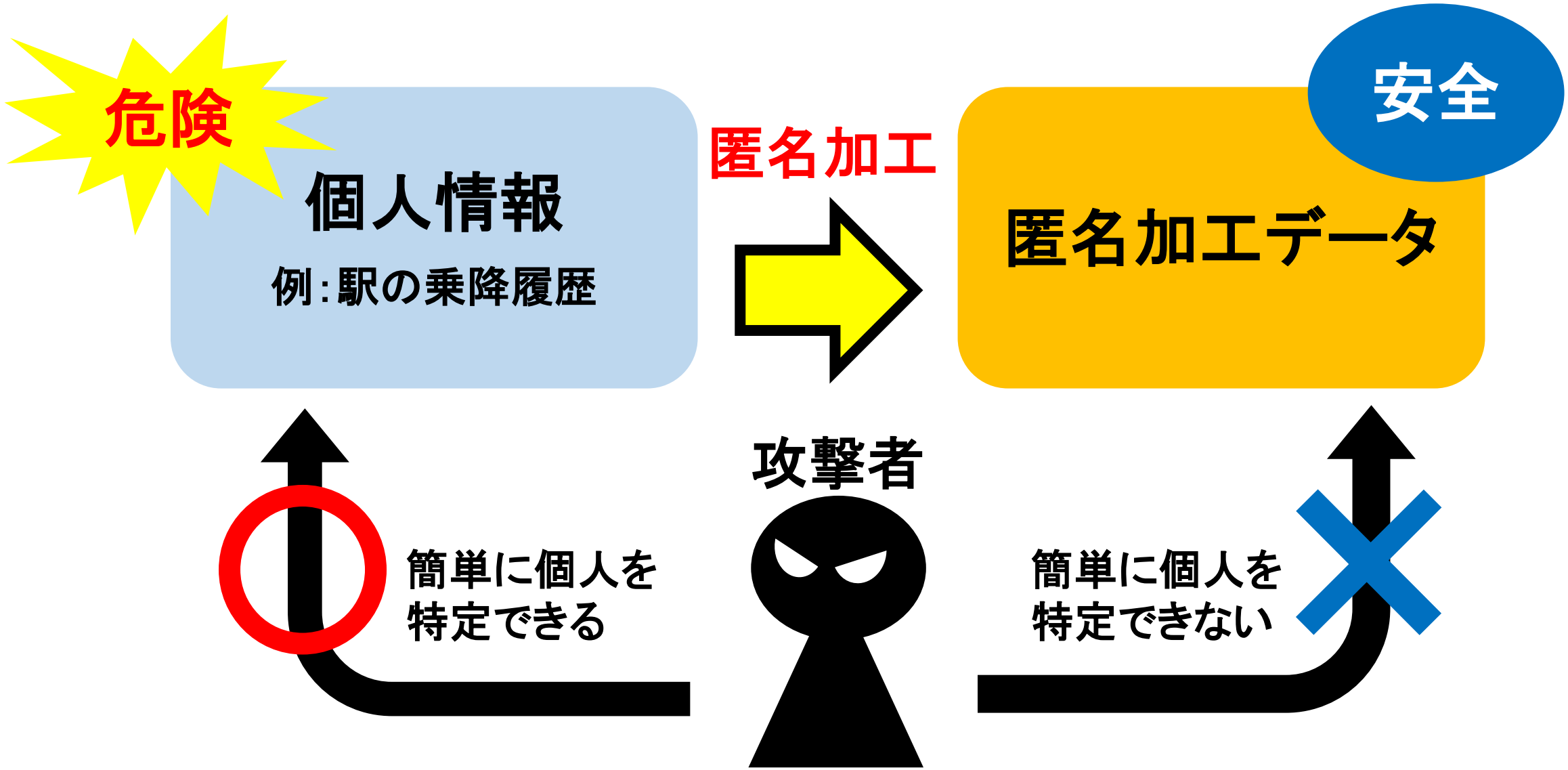
K410　菊池研･斉藤研合同発表会　2017年2月4日

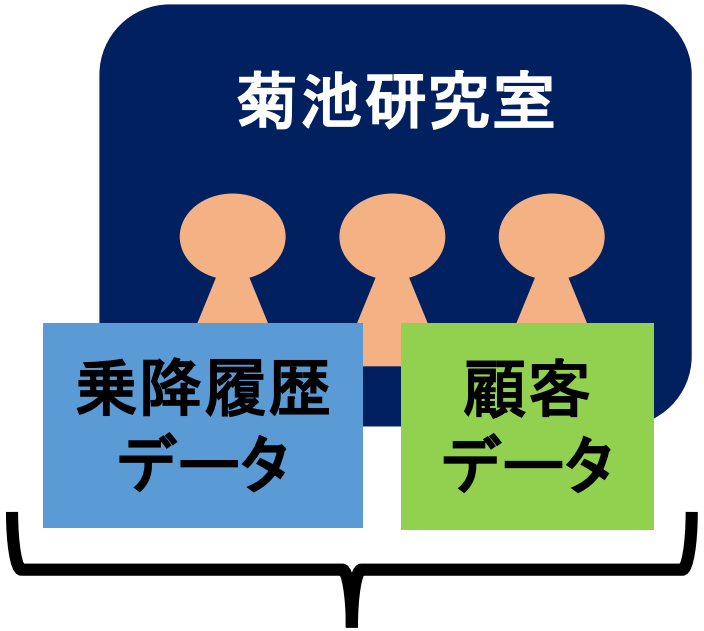
乗降履歴データの
有用性評価指標と匿名加工

明治大学　菊池研究室　4年
伊藤聡志

研究背景


危険
個人情報
例：駅の乗降履歴
匿名加工
安全
匿名加工データ
攻撃者
簡単に個人を
特定できる
簡単に個人を
特定できない

想定されるケース



菊池研究室に所属する
学生のICカード利用履歴
そのまま提供するのは危険

匿名加工した
データの提供

外部組織

菊池研究室のデータから

1. どんな学生がどのくらい駅を利用するのか？
2. どの駅がよく利用されているのか？
3. 菊池研究室にはどのような
学生がいるのか？

という情報を得たい

問題点

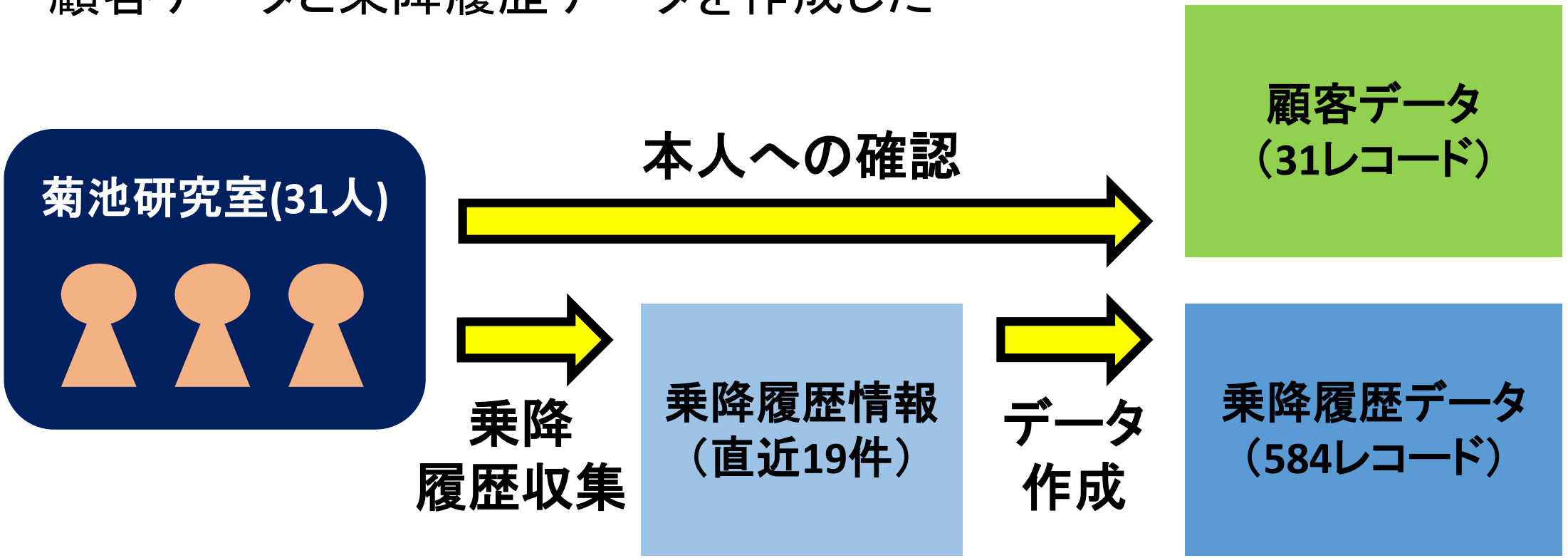
1. 企業内の個人情報はアクセスが許されない
2. データに対してどういった匿名加工をしたらよいのか不明

研究方法

1. 実際に個人情報データを取得し，分析する
2. 取得したデータを用いて匿名加工を行う

菊池研究室の乗降履歴データ

菊池研究室に所属する31人の情報や交通ICカードから
顧客データと乗降履歴データを作成した



データの内容

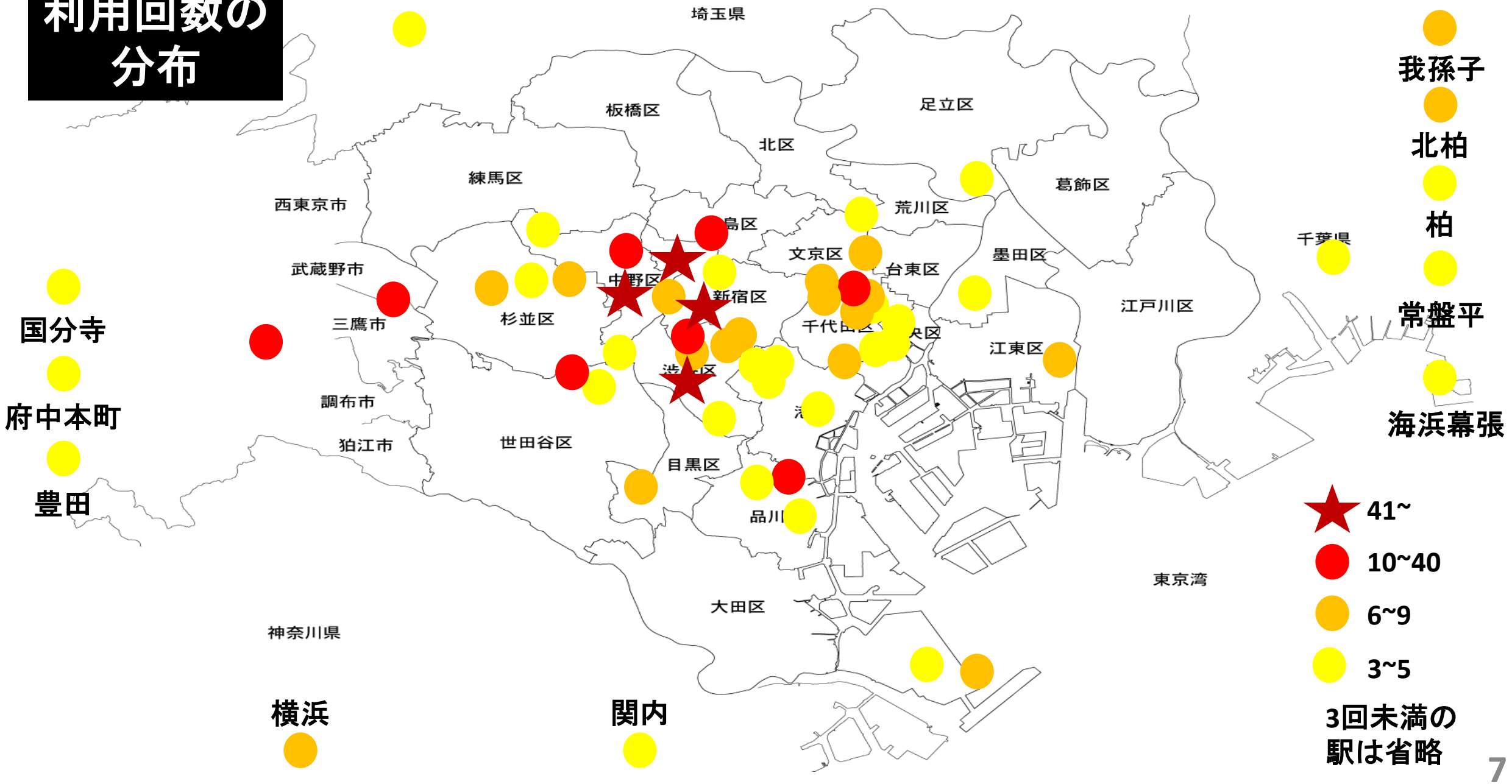
顧客データ例

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

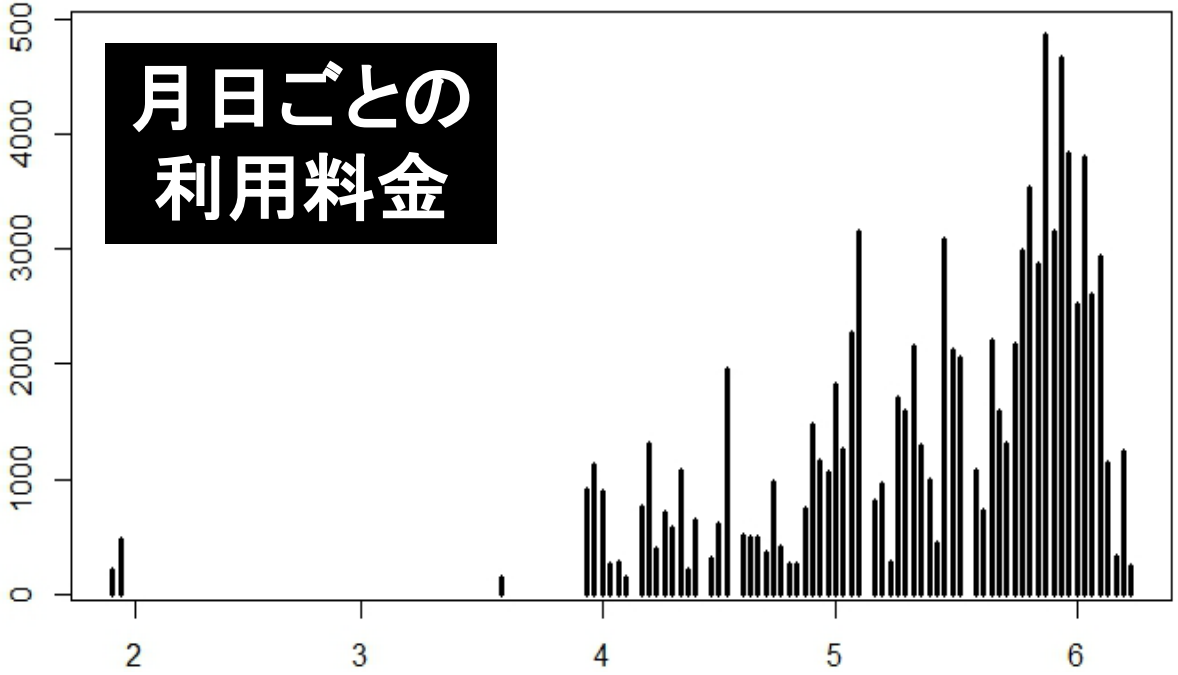
乗降履歴データ例

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
1	2016/10/8	1	NA	NA	NA	NA	チャージ	券売機	2000

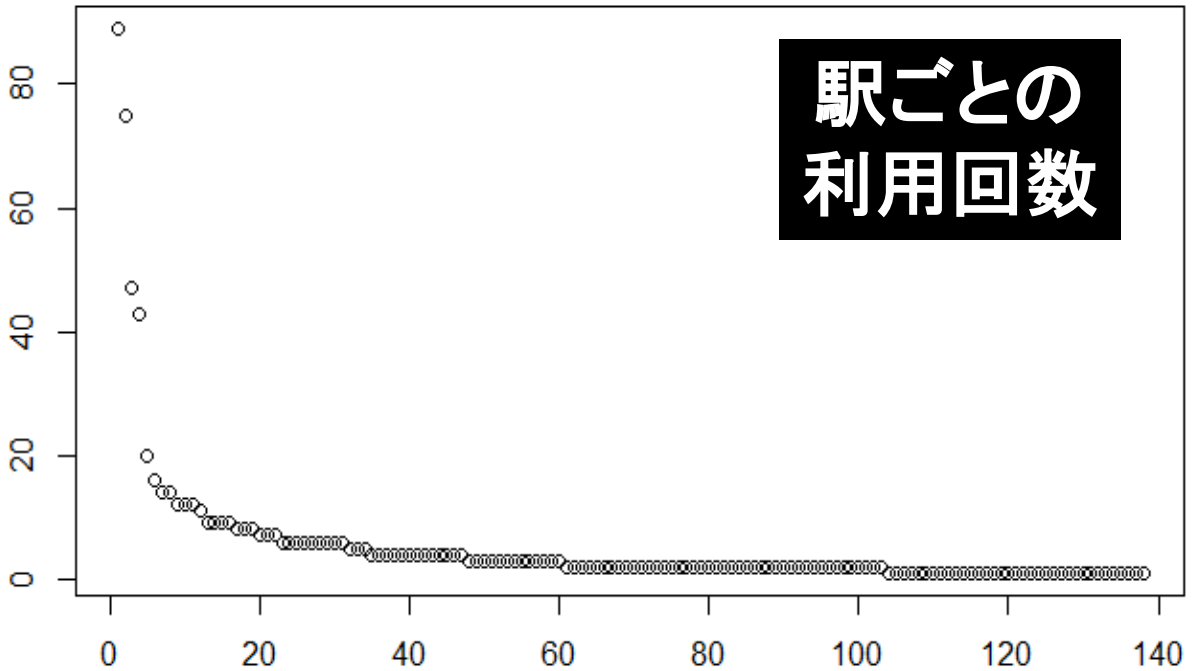
利用回数の分布


埼玉県
板橋区
足立区
北区
練馬区
葛飾区
西東京市
荒川区
武蔵野市
文京区
墨田区
台東区
三鷹市
杉並区
新宿区
江戸川区
江東区
調布市
狛江市
世田谷区
目黒区
品川
東京湾
大田区
神奈川県
我孫子
北柏
柏
常盤平
海浜幕張
国分寺
府中本町
豊田
横浜
関内
41~
10~40
6~9
3~5
3回未満の駅は省略

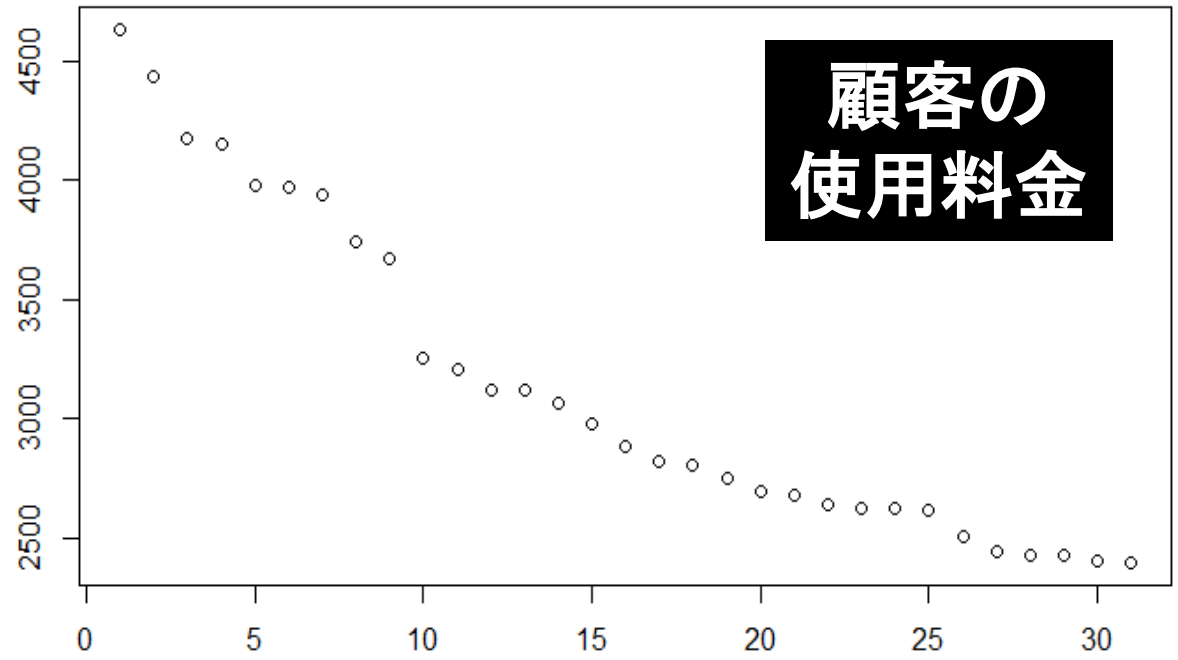
月日ごとの利用料金



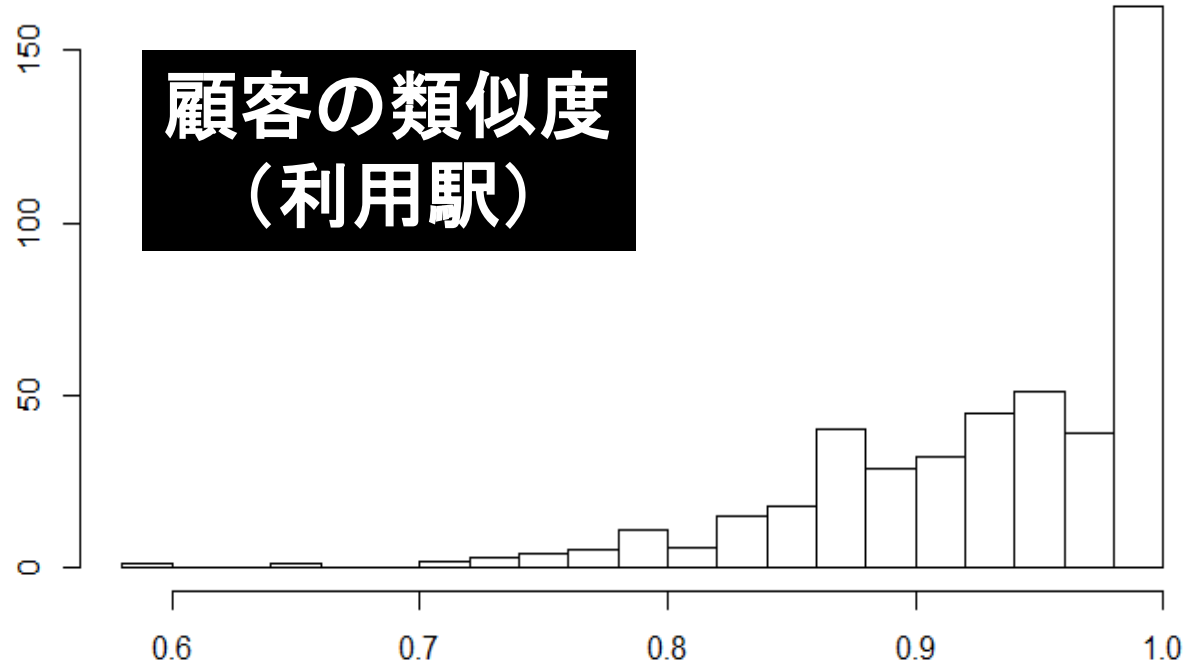
駅ごとの利用回数



顧客の使用料金



顧客の類似度（利用駅）



顧客/乗降履歴データに対する匿名加工

今回の場合，提供先の外部組織が知りたいのは

1. どんな学生がどのくらい駅を利用するのか？
2. どの駅がよく利用されているのか？
3. 菊池研究室にはどのような学生がいるのか？

であるため，これらの情報を保持した匿名加工をする

情報を保持できているかどうかは**有用性指標**を用いて評価し，
データが安全かどうかは**安全性指標**を用いて評価する

顧客/乗降履歴データの有用性/安全性指標

データの有用性を評価する指標を3つ用意した（U1, U2, U3）

有用性指標（U1） 顧客属性ごとの駅利用回数	有用性指標2（U2） 駅利用回数の順位上位	有用性指標3（U3） 顧客属性のクロス集計の人数

データの安全性を評価する指標を2つ用意した（S1, S2）

安全性指標（S1） 特殊な利用駅がない	安全性指標（S2） 特殊な顧客がいない

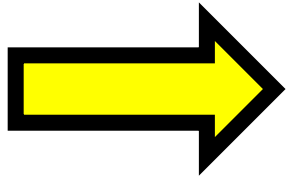
これらの指標の評価をもとに,
有用性を保ちつつ安全性を上げる匿名加工をする

顧客/乗降履歴データの匿名加工　前

簡易顧客データ M

顧客ID	性別	学年	group
1	男	1	A
2	男	1	A
3	男	2	B
4	男	2	B
5	女	4	C

匿名加工



加工された簡易顧客データ M*

顧客ID	性別	学年	group
1	男	1	A
2	男	1	A
3	男	2	B
4	男	2	B
5	男	2	B

特殊な顧客の
属性を書き換える

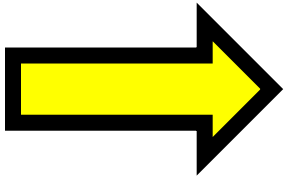
顧客/乗降履歴データの匿名加工　後

利用駅をグループ内でシャッフルする

簡易乗降履歴デー

顧客ID	乗車駅	降車駅	
1	新宿	品川	A
1	品川	新宿	A
2	高田馬場	新宿	A
2	新宿	中野	A
3	中野	新宿	B
3	新宿	中野	B
4	高田馬場	品川	B
4	品川	熱海	
5	中野	東京	
5	東京	中野	

匿名加工



れた簡易乗降履歴データT*

	乗車駅	降車駅	group
1	新宿	新宿	A
1	高田馬場	品川	A
2	品川	中野	A
2	新宿	新宿	A
3	品川	新宿	B
3	中野	品川	B
4	高田馬場	中野	B
	新宿	新宿	B
	中野	東京	B
	東京	中野	B

特殊な駅を最頻値に書き換える

加工手法の評価

	加工前	加工後
U1	0	2.67
U2	0	0
U3	0	0.2
S1	×	○
S2	×	○

データを加工したことにより
有用性が少し下がった
（数値が低いほうが有用性が高い）

データを加工したことにより
安全性が上がった

まとめ

- 31人の交通ICカードから顧客データと乗降履歴データを取得し，それらのデータの分析を行った

- 取得したデータに対し，想定したケースに対応する評価指標と加工手法を検討した

質疑応答用スライド

乗降履歴データの匿名加工　1

簡易顧客データ M

顧客ID	性別	学年	group
1	男	1	A
2	男	1	A
3	男	2	B
4	男	2	B
5	女	4	C

危険な場合1
利用駅をそのまま
データを提供する

簡易乗降履歴データ T

顧客ID	乗車駅	降車駅	group
1	新宿	品川	A
1	品川	新宿	A
2	高田馬場	新宿	A
2	新宿	中野	A
3	中野	新宿	B
3	新宿	中野	B
4	高田馬場	品川	B
4	品川	熱海	B
5	中野	東京	C
5	東京	中野	C

乗降履歴データの匿名加工　２

簡易顧客データ M

顧客ID	性別	学年	group
1	男	1	A
2	男	1	A
3	男	2	B
4	男	2	B
5	女	4	C

簡易乗降履歴データ T*

顧客ID	乗車駅	降車駅	group
1	新宿*	新宿*	A
1	高田馬場*	品川*	A
2	品川*	中野*	A
2	新宿*	新宿*	A
3	品川*	新宿*	B
3			
4			
4			
5			
5			

危険な場合２
顧客属性の
組み合わせが
独特のものがある

顧客属性が同じグループ内で
利用駅をシャッフルする
（U1~U3は損なわない）

乗降履歴データの匿名加工　３

簡易顧客データ M*

顧客ID	性別		
1	男		
2	男		
3	男		
4	男	2	B
5	男*	2*	B*

危険な場合３
特殊な利用駅
（利用回数が少ない・場所が遠い）がある

簡易乗降履歴データ T*

	乗車駅	降車駅	group
1	新宿	新宿	A
2	高田馬場	品川	A
3	品川	中野	A
2	新宿	新宿	A
3	品川	新宿	B
3	中野	品川	B
4	高田馬場	中野	B
4	新宿	熱海	B
5	中野	東京	B*
5	東京	中野	B*

顧客属性の組み合わせが独特なユーザーを別グループに変える
（U1, U3を少し損なう）

乗降履歴データの匿名加工　４

簡易顧客データ M*

顧客ID	性別	学年	group
1	男	1	A
2	男	1	A
3	男	2	B
4	男	2	B
5	男	2	B

特殊な駅（利用回数が少ない・場所が遠い）を利用回数1位の駅に書き換える
（U1~U3は損なわない）

簡易乗降履歴データ T**

顧客ID	乗車駅	降車駅	group
1	新宿	新宿	A
1	高田馬場	品川	A
2	品川	中野	A
2	新宿	新宿	A
3	品川	新宿	B
3	中野	品川	B
4	高田馬場	中野	B
4	新宿	新宿*	B
5	中野	東京	B
5	東京	中野	B

有用性指標の詳細

$$U_1(M, T, M^*, T^*) = \frac{\sum_{g}^{i=1} |T_{station}(X_i) - T^*_{station}(X_i)|}{g}$$

$$U_2(M, T, M^*, T^*) = 5 - |\{s \in S_5(rank(T, s) = rank(T^*, s))\}|$$

$$U_3(M, T, M^*, T^*) = \frac{\sum_{num(sex)}^{i=1} \sum_{num(grade)}^{j=1} |Cross_{sex,grade}(i, j) - Cross^*_{sex,grade}(i, j)|}{num(sex) * num(grade)}$$

M, T : 元データ
M*, T* : 加工されたデータ
$T_{station}(X_i)$: TについてのグループX_iの駅利用総回数
g : Tのグループ数
S_5 : 上位5駅の集合
rank(T, s) : 駅sのTにおける利用回数順位
$Cross_{sex,grade}$: Mの(性別，学年)属性についてのクロス集計値
num(属性名) : 属性の種類数
これらの有用性指標の値が0に近いほど，データ(T*,M*)の有用性は高い

有用性指標の計算例(U1)

$$\mathbf{U_1}(\mathbf{M}, \mathbf{T}, \mathbf{M}^*, \mathbf{T}^*) = \frac{\sum_{\mathbf{g}}^{\mathbf{i=1}} |\mathbf{T_{station}}(\mathbf{X_i}) - \mathbf{T^*_{station}}(\mathbf{X_i})|}{\mathbf{g}}$$

$$= \frac{|\mathbf{T_{station}}(\boldsymbol{A}) - \mathbf{T^*_{station}}(\boldsymbol{A})| + |\mathbf{T_{station}}(\boldsymbol{B}) - \mathbf{T^*_{station}}(\boldsymbol{B})| + |\mathbf{T_{station}}(\boldsymbol{C}) - \mathbf{T^*_{station}}(\boldsymbol{C})|}{\mathbf{3}}$$

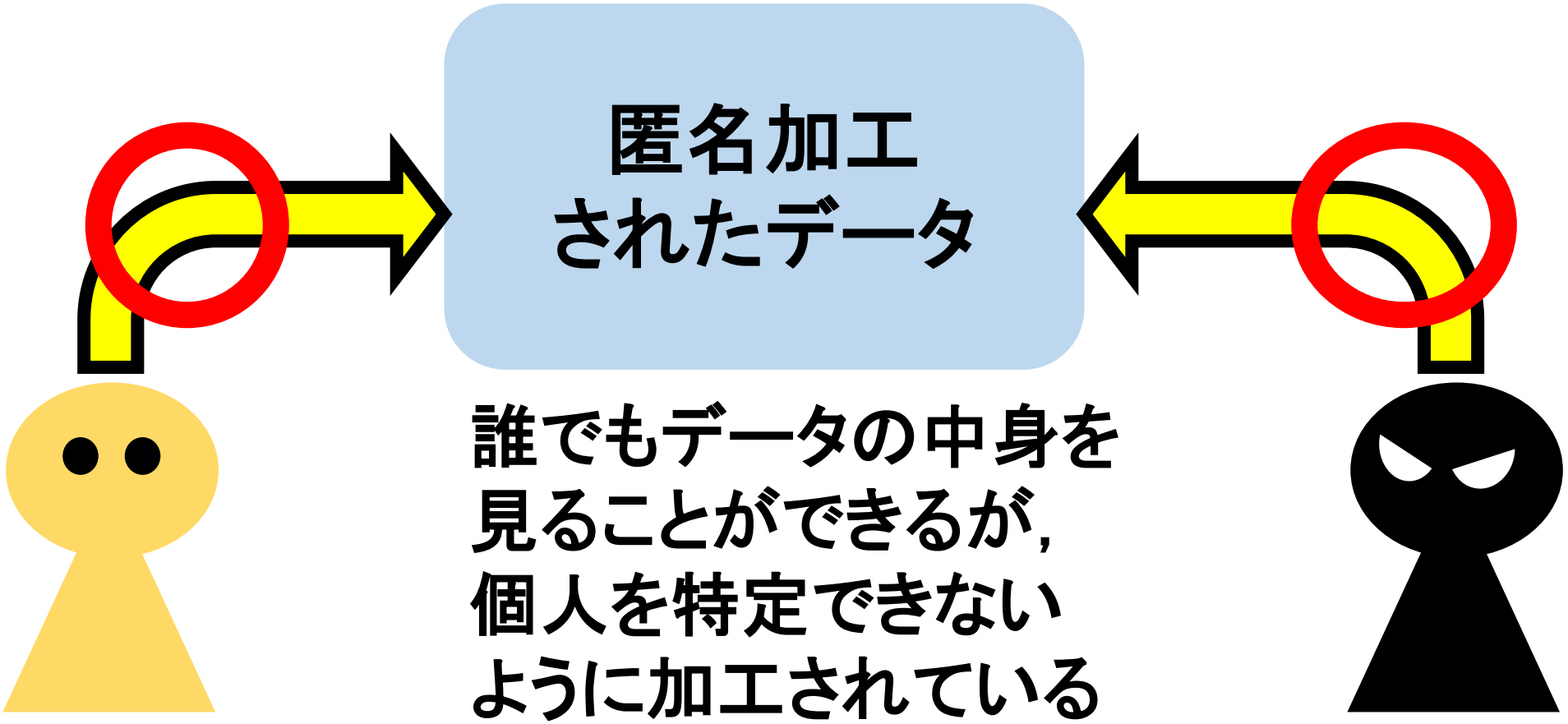
$$= \frac{|\mathbf{8} - \mathbf{8}| + |\mathbf{8} - \mathbf{12}| + |\mathbf{4} - \mathbf{0}|}{\mathbf{3}}$$

$$= \frac{\mathbf{0} + \mathbf{4} + \mathbf{4}}{\mathbf{3}}$$

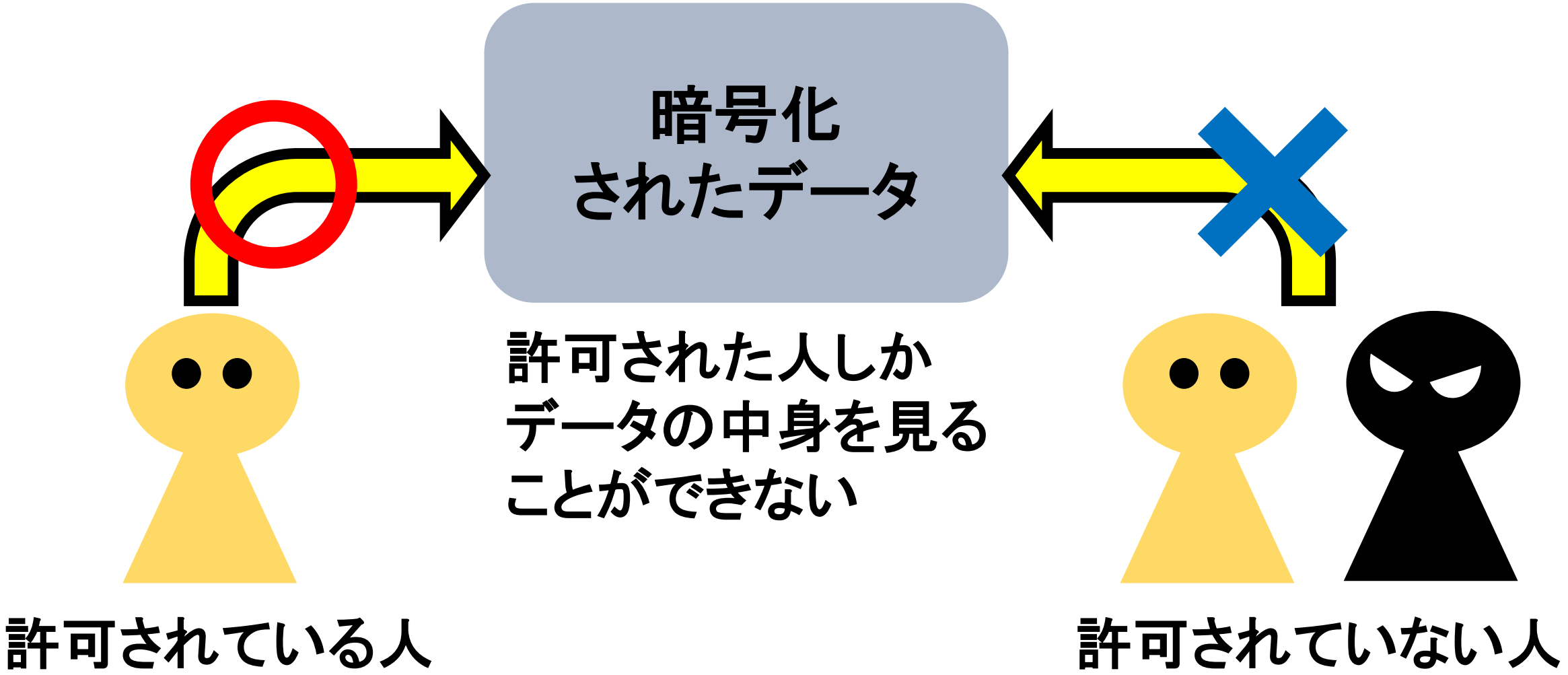
$$= \frac{\mathbf{8}}{\mathbf{3}}$$

$$= \mathbf{2.66\ldots} \approx \mathbf{2.67}$$

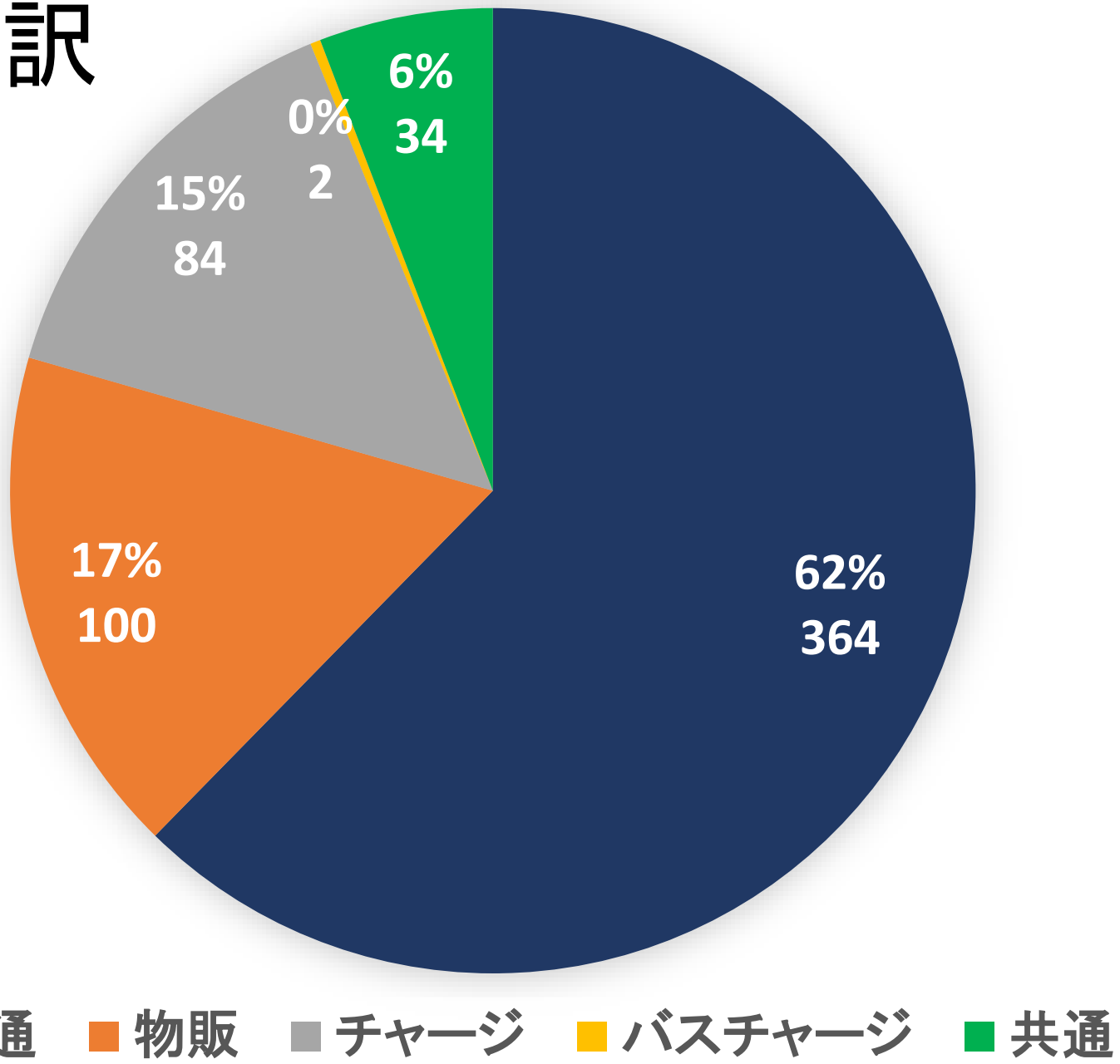
匿名加工と暗号化の違い　1



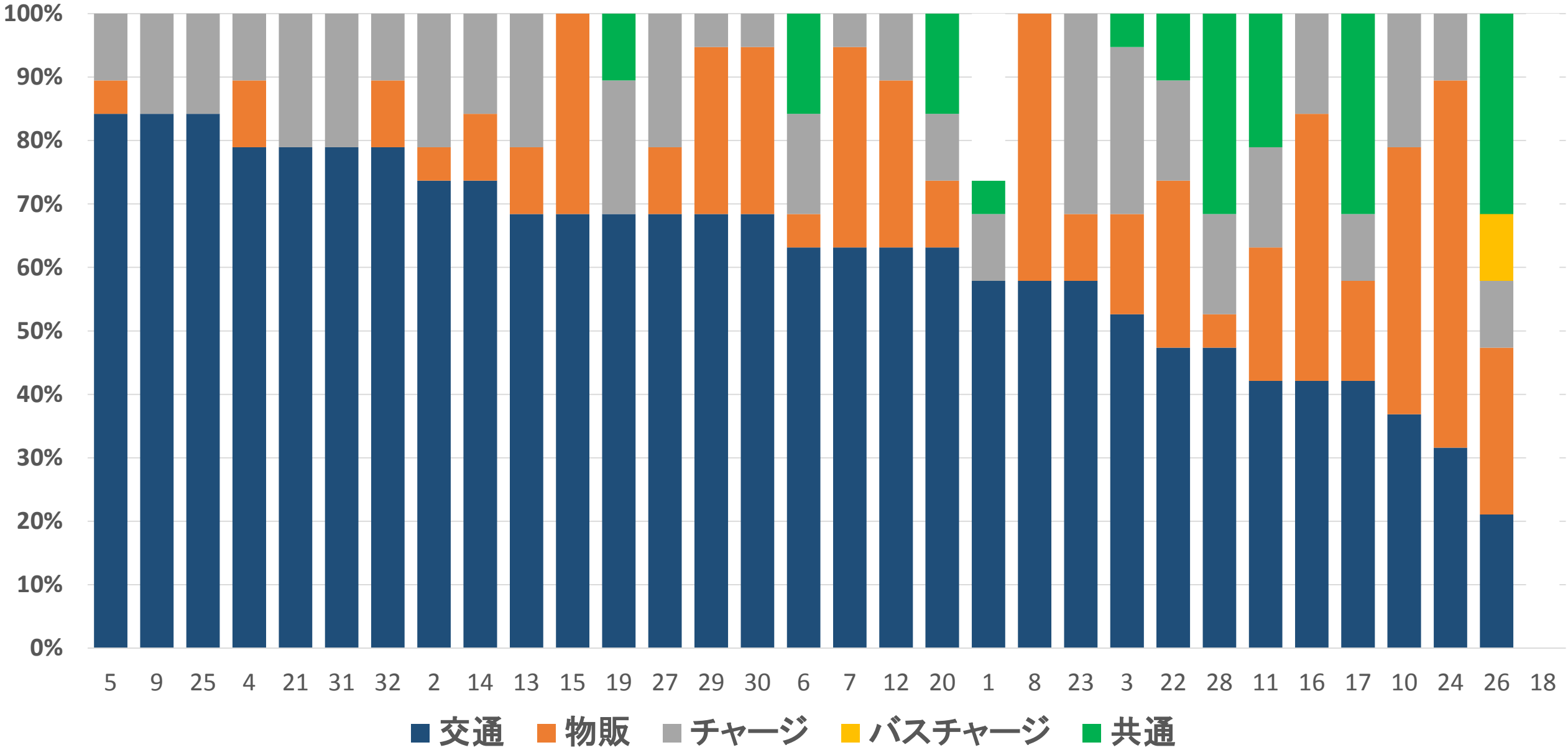
匿名加工と暗号化の違い　２



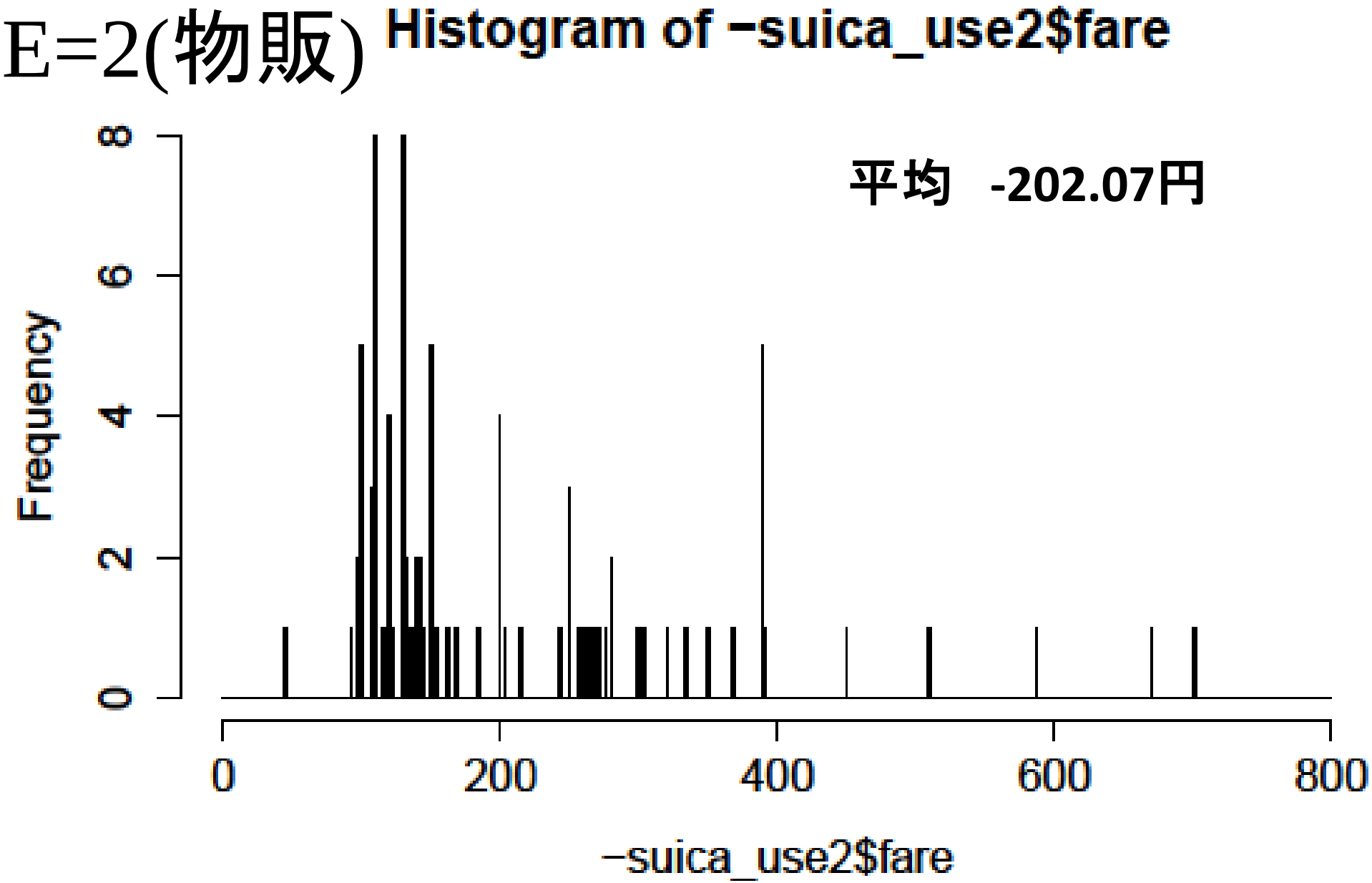
suicaデータの内訳



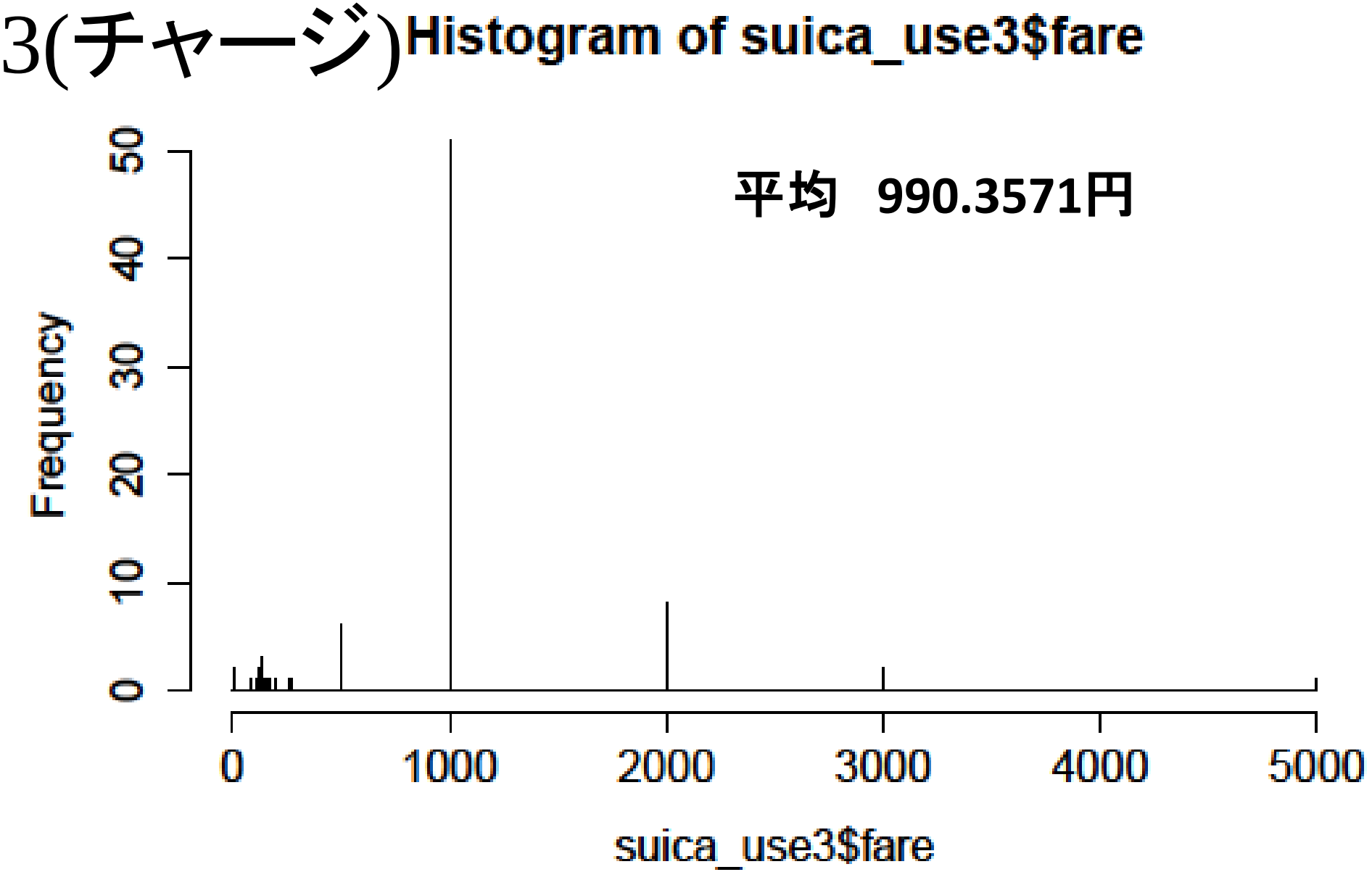
ユーザごとの用途の内訳



USE=2(物販)


Histogram of −suica_use2$fare
平均　-202.07円
Frequency
0
2
4
6
8
0
200
400
600
800
−suica_use2$fare

USE=3(チャージ)


Histogram of suica_use3$fare
平均　990.3571円
Frequency
0
10
20
30
40
50
0
1000
2000
3000
4000
5000
suica_use3$fare