

# 一般化匿名加工された購買履歴 データのRFM分析有用性評価

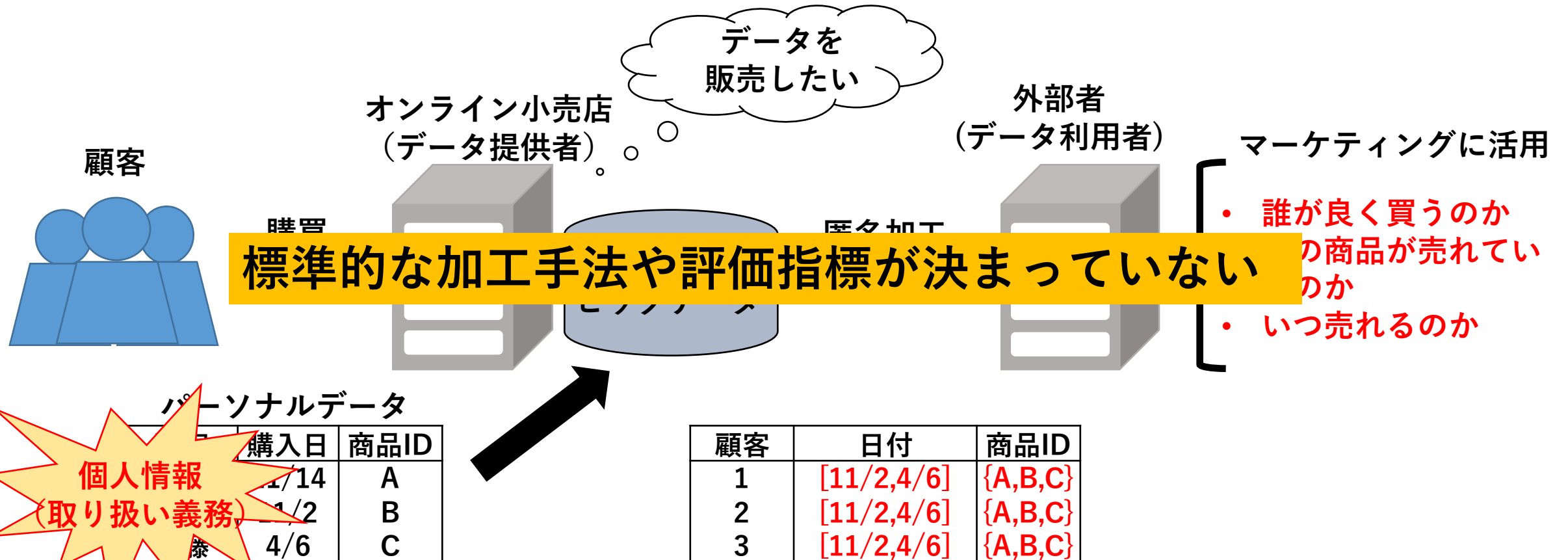
明治大学

小林祐貴 中村幸輝 伊藤聡志 菊池浩明

# 研究背景（匿名加工とは）

- 近年パーソナルデータ利活用による匿名加工の必要性

匿名加工：データから個人を特定されないようにデータを加工すること



# PWSCUPとその問題点

## 匿名加工・再識別コンテストPWSCUP

- 匿名加工データの優れた加工手法や評価指標を明らかにするコンテスト
- 2018年は「**一般化**」手法がテーマ。
- 匿名加工データの有用性と安全性を評価するコンテスト

元データ					一般化	一般化匿名加工データ				
顧客	購買日	商品ID	単価	購買数量		顧客ID	購買日	商品ID	単価	購買数量
小林	11/14	A	1	1	1	[11/2,4/6]	{A,B,C}	[1,3]	[1,10]	
中村	11/2	B	2	2	2	[11/2,4/6]	{A,B,C}	[1,3]	[1,10]	

## 問題点

### 1. PWSCUP2018では特定のユースケースに対する有用性は不確か

PWSCUPは元データと加工データの平均誤差で有用性評価

### 2. $k=2$ の匿名加工データが主流で $k \geq 3$ の匿名加工データに対する評価が不十分

$k$ -匿名化：同一のレコードが $k$ 件以上になるように匿名加工を行うことで再識別される確率を $1/k$ にする

# 解決策

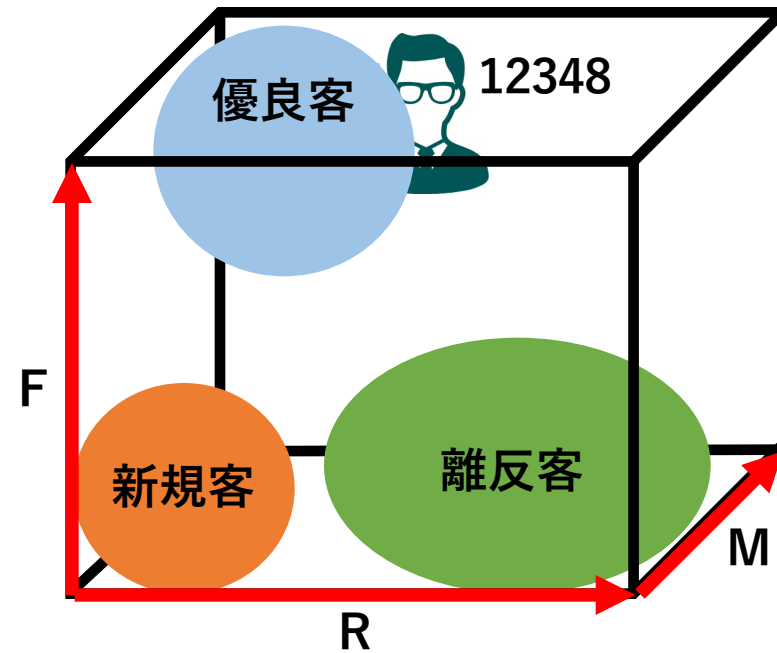
1. RFM分析の観点からユースケースを検討
2.  $k=3,4$ のプログラムを開発し、その安全性と有用性を評価する

# 購買履歴データのRFM分析

- 2010年から1年間の英国のオンライン小売店における購買履歴データ 1000人分を使用

## RFM分析

- R(Recency)：最新購買日**  
2011/12/31(最新日)から何日前か
- F(Frequency)：購買頻度**  
1日に何度も購買していても1回とカウント
- M(Monetary)：購買金額 (ポンド)**  
年間の購買総額



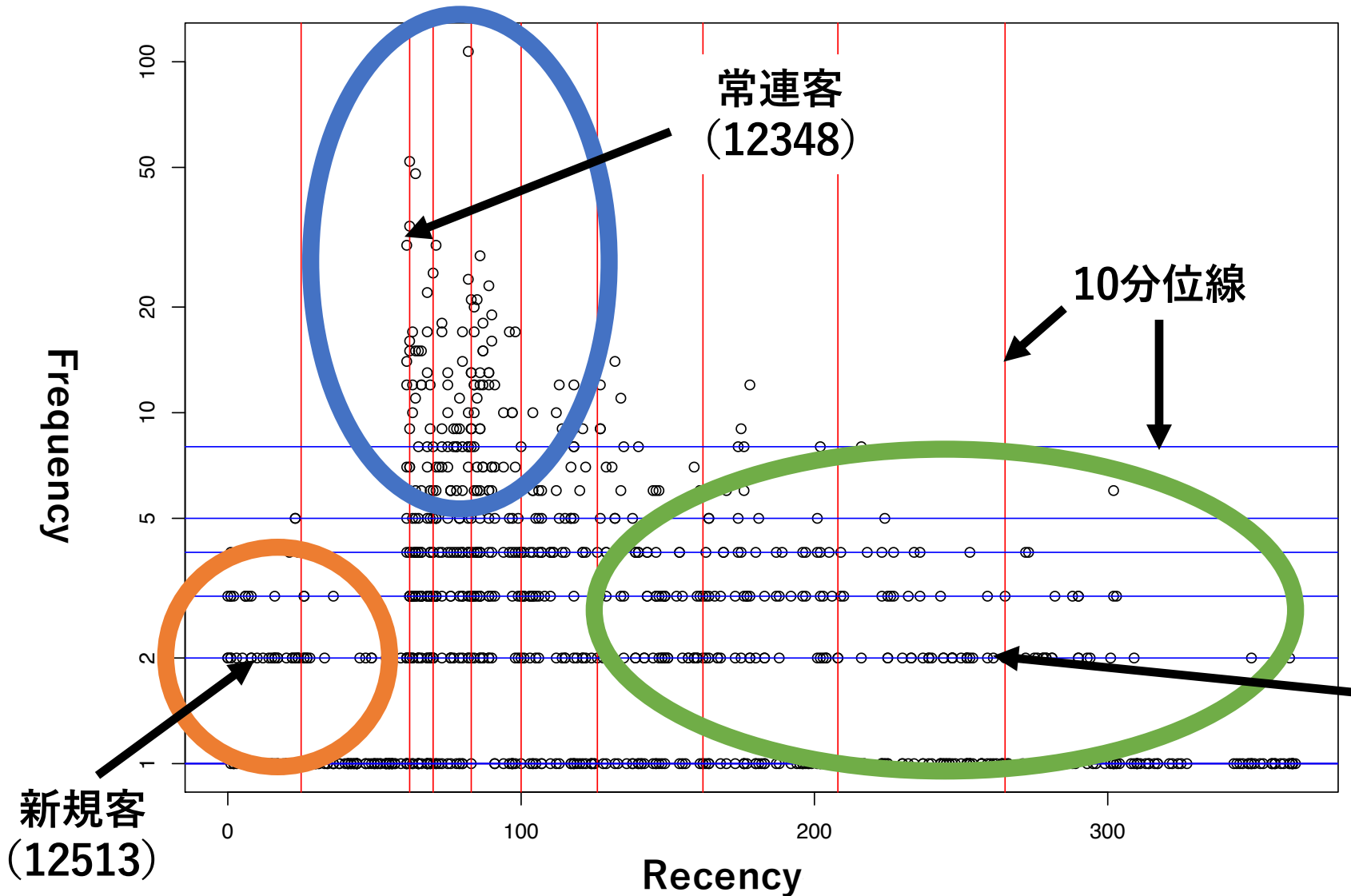
購買履歴データ

顧客ID	購買日	商品	単価	購買数量
12348	11/14	A	3.7	1
12513	11/2	B	2.0	2
12678	4/6	C	0.6	10
12678	4/6	D	2.0	5



顧客ID	R	F	M
12348	66	30	23540
12513	9	2	145.2
12678	260	2	5041.56

# 最新購買日(R)と購買頻度(F)の散布図

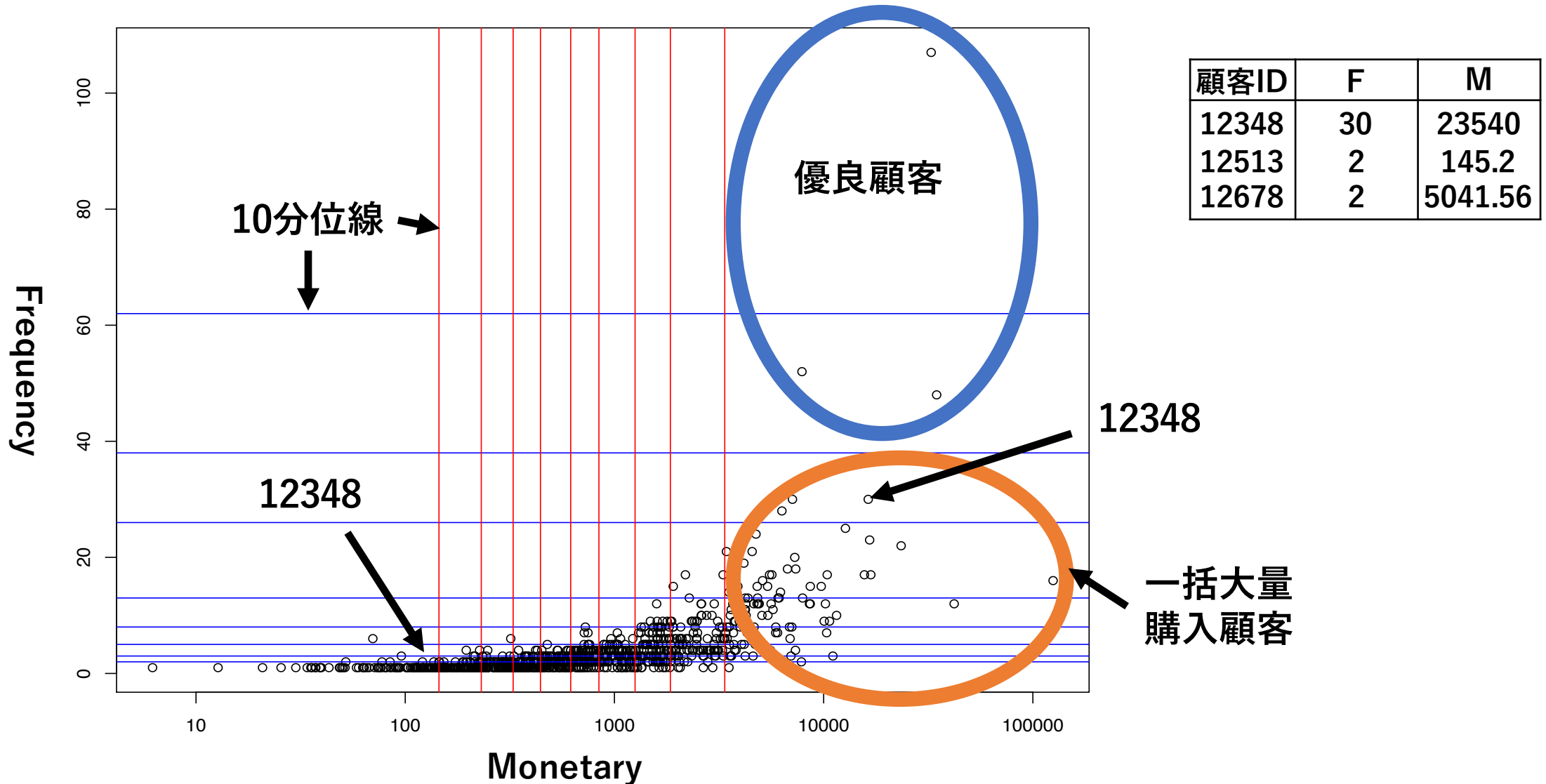


顧客ID	R	F
12348	66	30
12513	9	2
12678	260	2

- R,F,Mをそれぞれ10分位値でクラスタリング

離反客  
(12678)

# 年間購買総額(M)と購買頻度(F)の散布図



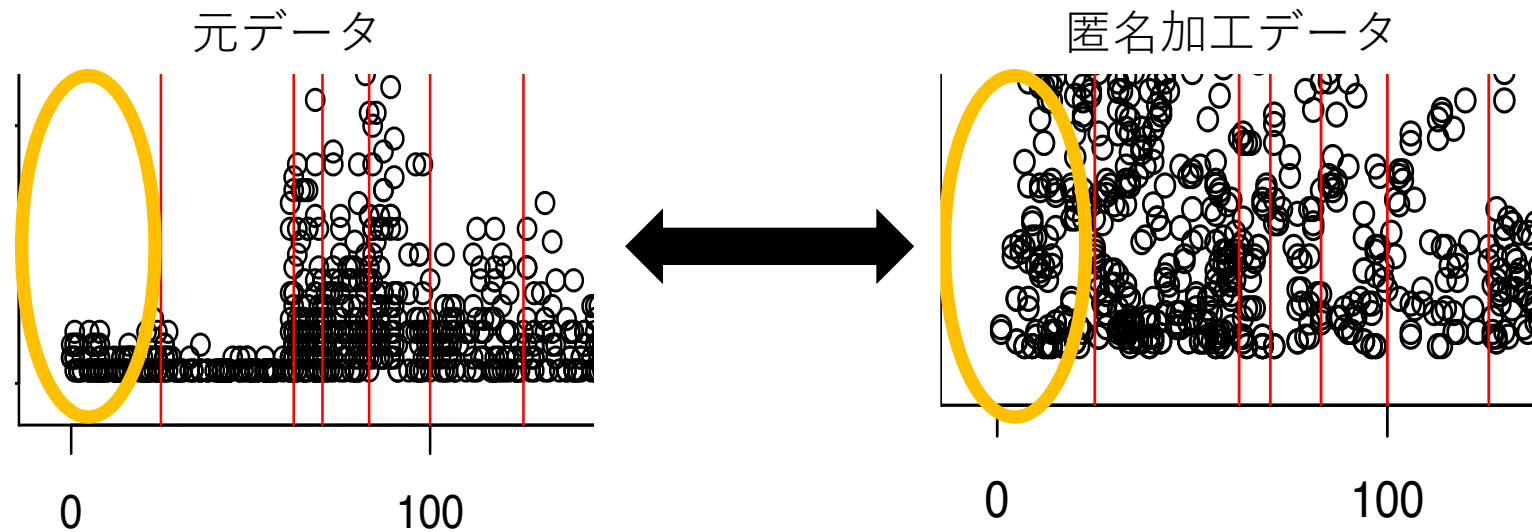
# 購買履歴データのユースケース・有用性評価

- **ユースケース**

- RFMをクラスタリングすることで顧客の性質・分布を知ることができる

- **有用性評価**

- **元データと匿名加工データで顧客のクラスタが一致した数の割合**



- R,F,M一次元の有用性、RFM三次元の有用性を評価する

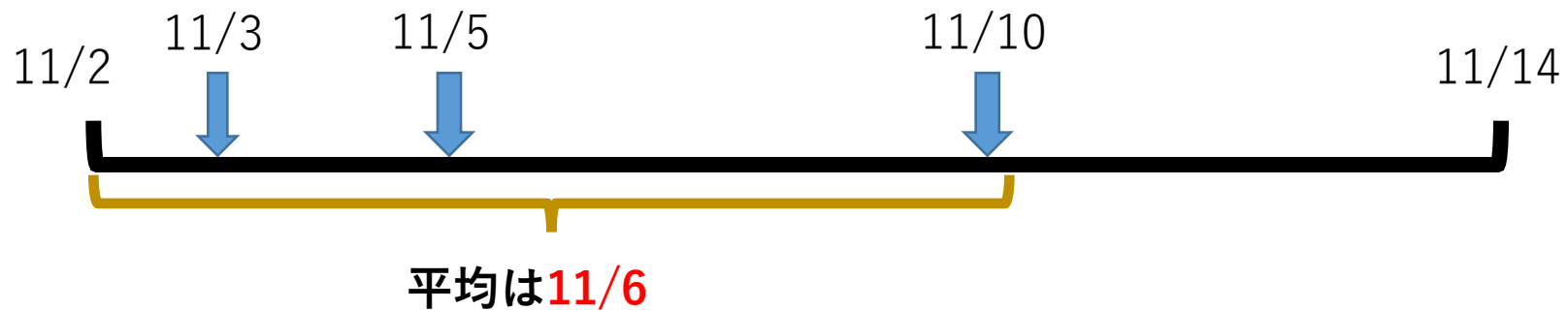


# 匿名加工データのRFM

- 匿名加工データのRFMを計算するために特定の値が必要

顧客ID	購買日	単価	購買数量	顧客ID	R	F	M
23	[11/2,11/14]	[2,3.7]	[1,2]	23	?	?	?

- 区間からランダムにn回選んだ値の平均値でRFMを計算(本研究はn=100)
  - 例：n=3で[11/2,11/14]に一般化されている場合



# 評価実験・使用する匿名加工データ

- PWSCUP2018における加工手法
  - **仮名化**：顧客IDを変更すること
  - **削除**：データの要素を削除し、削除を表す\*に変更する
  - **一般化**：要素を区間や集合へ変更

元データ					匿名加工データ				
顧客ID	購買日	商品ID	単価	購買数量	仮名ID	購買日	商品ID	単価	購買数量
12348	11/14	21	3.7	1	23	[11/2,11/14]	*	[2,3.7]	[1,2]
12513	11/2	23	2	2	40	[11/2,11/14]	*	[2,3.7]	[1,2]
12556	11/10	24	3	1	13	[11/2,11/14]	*	[2,3.7]	[1,2]

- $k$ -匿名化を行う
  - 同一のレコードが $k$ 件以上になるように匿名加工を行うことで再識別される確率を $1/k$ にする
  - 本実験では $k=2,3,4$ の匿名加工データを使用する
- 安全性評価は平均再識別率 $1/k$ とする

# k-匿名化のアルゴリズム

- k=3の例

顧客ID	レコード数	クラス番号	削除レコード数
6	1500	1	500
100	1200	1	200
30	1000	1	0
44	800	2	600
56	500	2	300
62	200	2	0
⋮	⋮	⋮	⋮

1. 1000人の顧客をレコード数順にソートする
2. レコード数上位の顧客から3人ずつマッチング
3. レコード数を合わせるためレコード削除を行う

仮名ID	購買日	商品ID	単価	購買数量
23	[11/2,11/14]	*	[2,3.7]	[1,2]
40	[11/2,11/14]	*	[2,3.7]	[1,2]
13	[11/2,11/14]	*	[2,3.7]	[1,2]

# 結果1：有用性と安全性の関係

$k$	有用性(R)	有用性(F)	有用性(M)	有用性(RFM)	安全性	PWSCUP有用性評価 (誤差の大きさ)
2	0.270	0.463	0.352	0.097	0.5	0.35
3	0.214	0.343	0.301	0.040	0.33	0.47
4	0.155	0.287	0.288	0.026	0.25	0.54

- $k$ の値が大きくなるほど**有用性が下降**、**安全性が上昇**

- R,F,Mの有用性は約3割減少
- RFMの有用性は1割以下に減少

- **Rの有用性がFやMと比べて低い**

平均117日の区間から任意の値を選んでいるため

購買日
[8/2,11/14]
[6/10,11/14]

# 結果2：RFMの独立性について

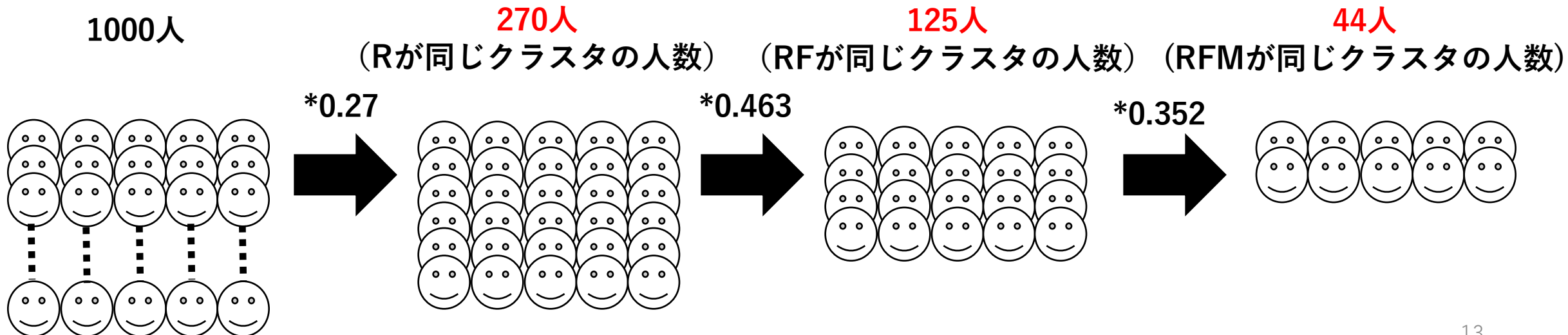
- **R,F,M は独立ではない**

R,F,Mの積よりもRFMの有用性が高いから

$k$	有用性(R)	有用性(F)	有用性(M)	R*F*M	期待値
2	0.270	0.463	0.352	0.044	44人

$k$	有用性(RFM)	人数
2	0.097	97人

- $k=2$ の時

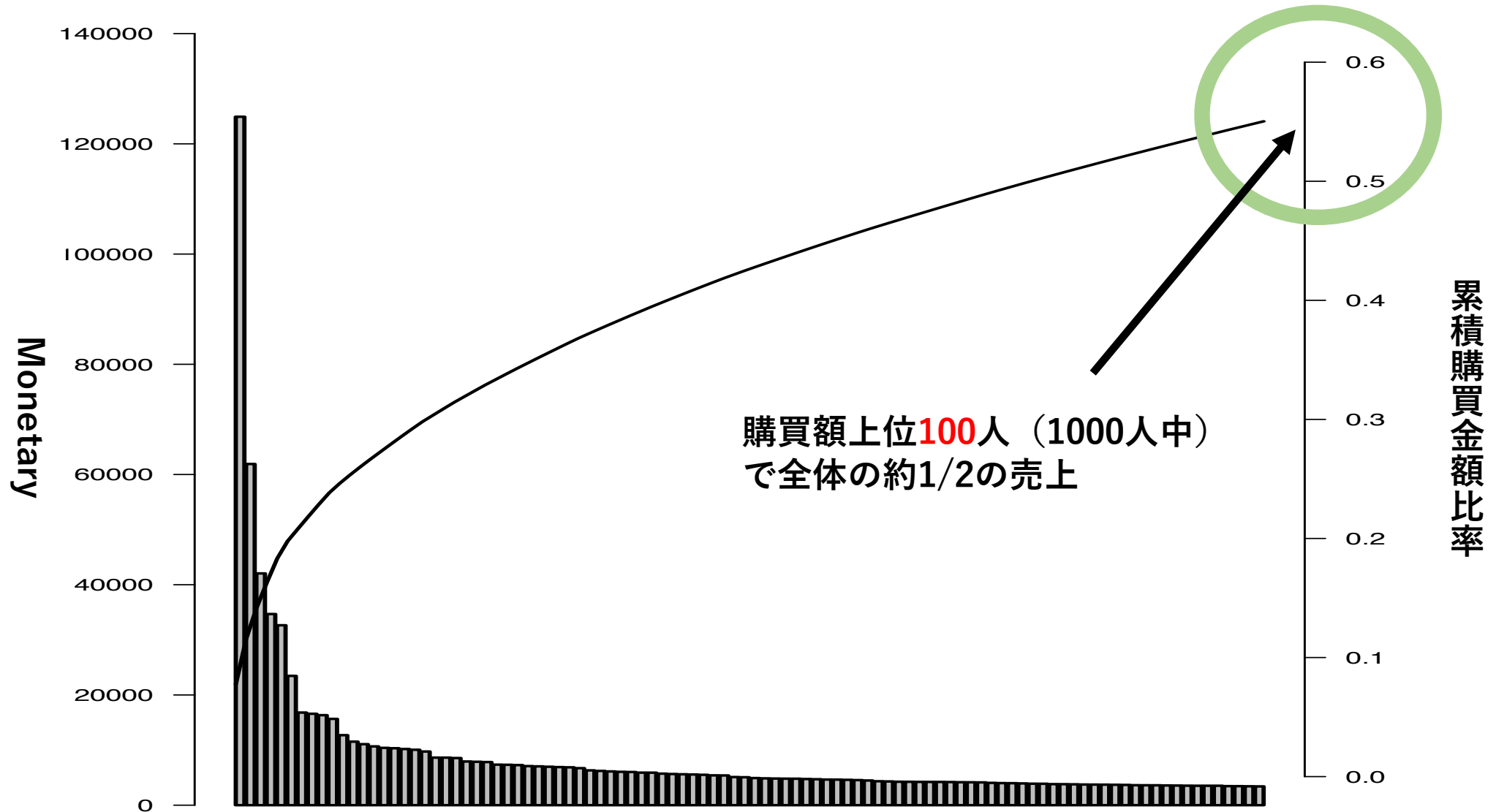


# まとめ

- RFM分析から購買履歴データのユースケースを検討した
- 一般化匿名加工データにおける有用性評価指標を提案した
- 匿名加工することにより有用性は下降し安全性は上昇した
  - R,F,Mの有用性は約3割減少
  - RFMの有用性は1割以下に減少
- R,F,Mは独立ではないことがわかった



# 年間購買総額(M)と累積購買金額比率





# 安全性評価

- $k=n$ の時の平均再識別率で評価
- $k=3$ の時

仮名ID	購買日	商品ID	単価	購買数量
23	[11/2,11/14]	{21,23}	[2,3.7]	[1,2]
40	[11/2,11/14]	{21,23}	[2,3.7]	[1,2]
32	[11/2,11/14]	{21,23}	[2,3.7]	[1,2]



3人中3人が識別される確率 $\rightarrow 1/6$



3人中2人が識別される確率 $\rightarrow 0$



3人中1人が識別される確率 $\rightarrow 1/2$



3人中0人が識別される確率 $\rightarrow 1/3$

平均再識別率は $1/3$

- $k=2,3,4$ の時は何も $1/k$ となった