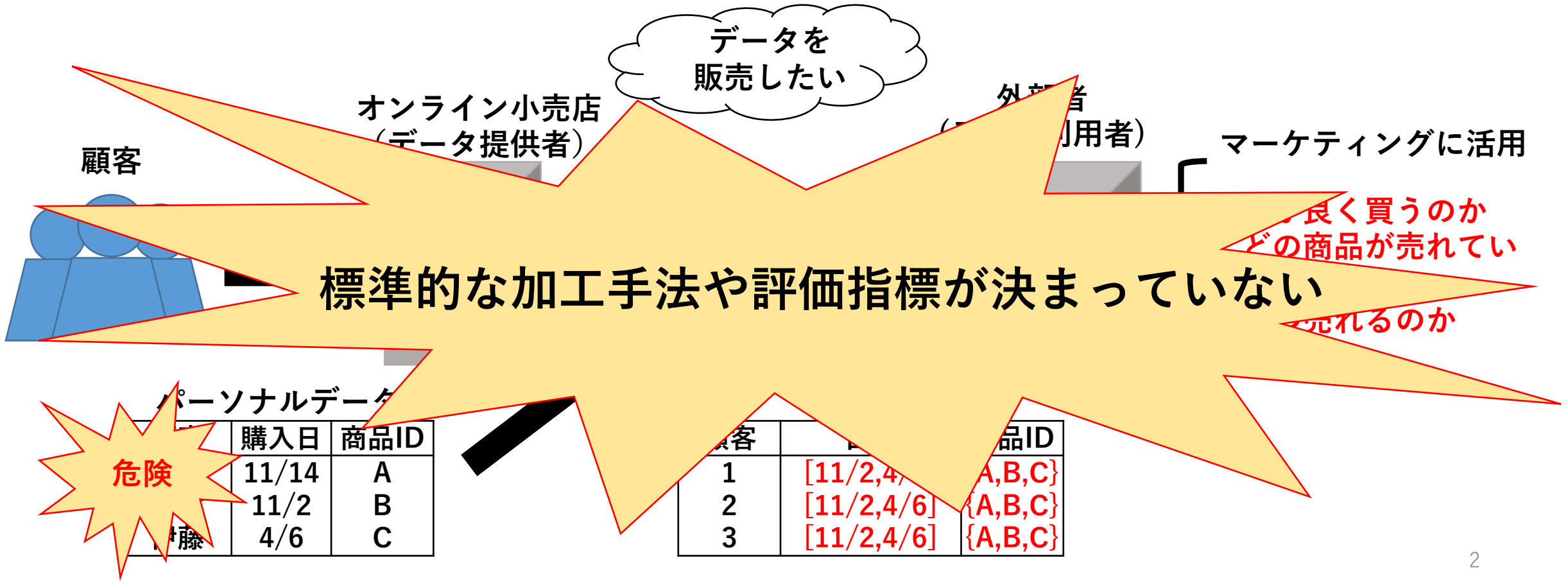
一般化匿名加工された購買履歴データのRFM分析有用性評価

明治大学総合数理学部
小林祐貴

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



研究背景（PWSCUPとは）

- **匿名加工・再識別コンテストPWSCUP**
  - 匿名加工データの活用のために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

顧客	購買日	商品ID	単価	購買数量
小林	11/14	A	1	1
中村	11/2	B	2	2
伊藤	4/6	C	3	10

一般化

一般化匿名加工データ

顧客ID	購買日	商品ID	単価	購買数量
1	[11/2,4/6]	{A,B,C}	[1,3]	[1,10]
2	[11/2,4/6]	{A,B,C}	[1,3]	[1,10]
3	[11/2,4/6]	{A,B,C}	[1,3]	[1,10]

- [11/2,4/6]→11/2から4/6のいずれかの日付が購買日

問題点・解決策

- 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. **一般化匿名加工データの分析が困難である**

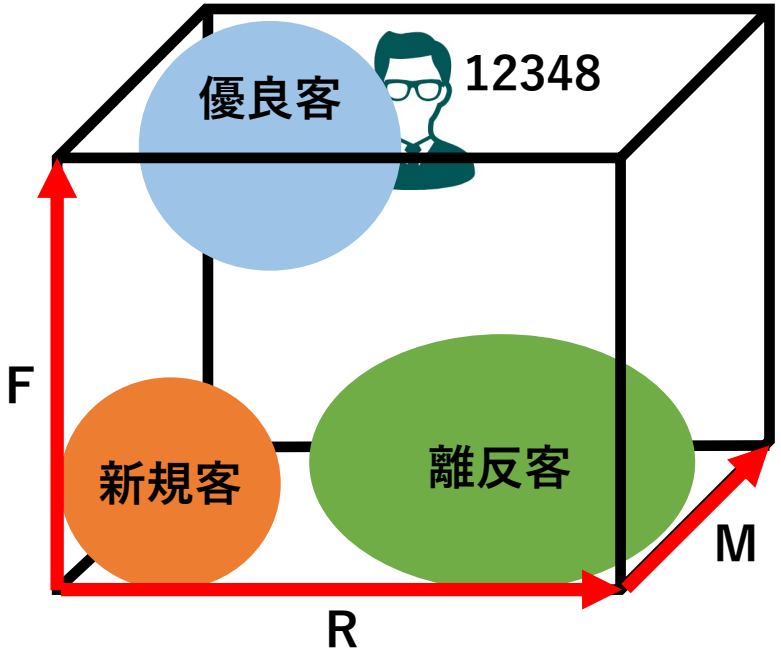
顧客ID	購買日	商品ID	単価	購買数量
1	[11/2,4/6]	{A,B,C}	[1;3]	[1;10]
2	[11/2,4/6]	{A,B,C}	[1;3]	[1;10]
3	[11/2,4/6]	{A,B,C}	[1;3]	[1;10]

ユースケース
- **誰が良く買うのか**
- **どの商品が売れているのか**
- **いつ売れるのか**

- 解決策
  1. RFM分析の観点からユースケースを検討
  2. 一般化匿名加工データからRFMを計算するシステムの開発
  3. ユースケースに基づいた有用性評価を行う

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

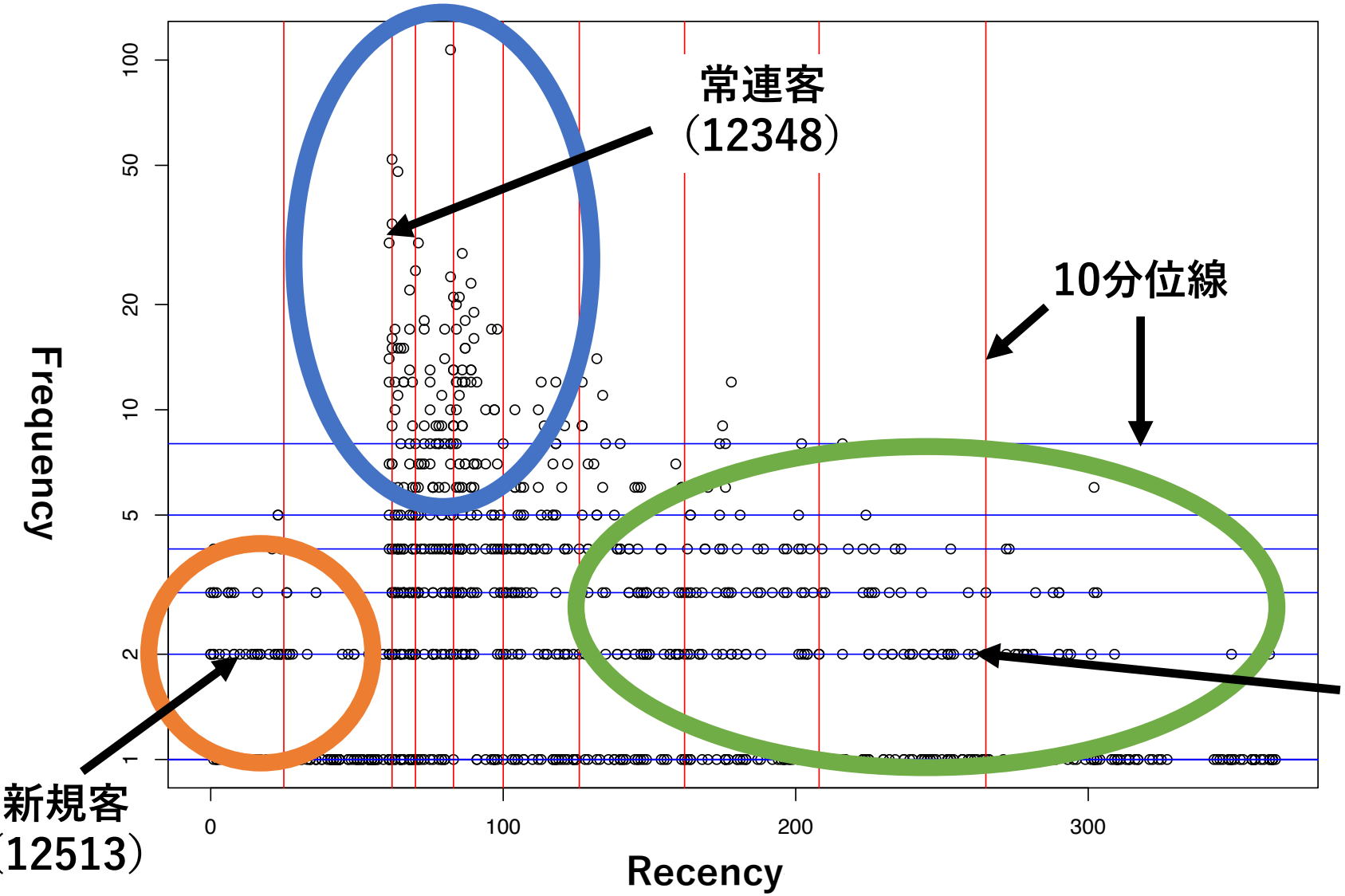


購買履歴データ

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	1797.24
12513	9	2	1757.55
12678	260	2	1079.4

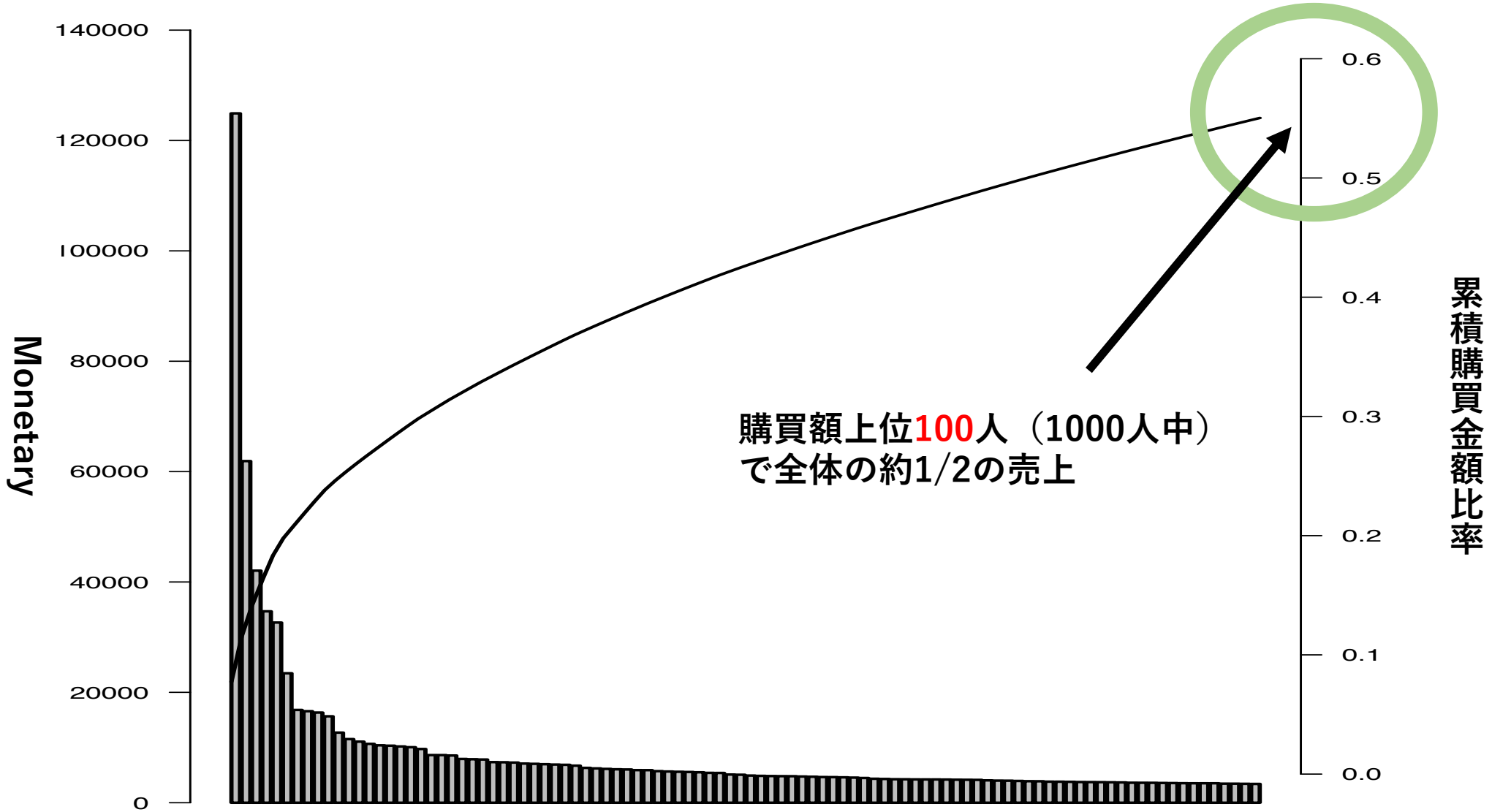
最新購買日(R)と購買頻度(F)の散布図



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

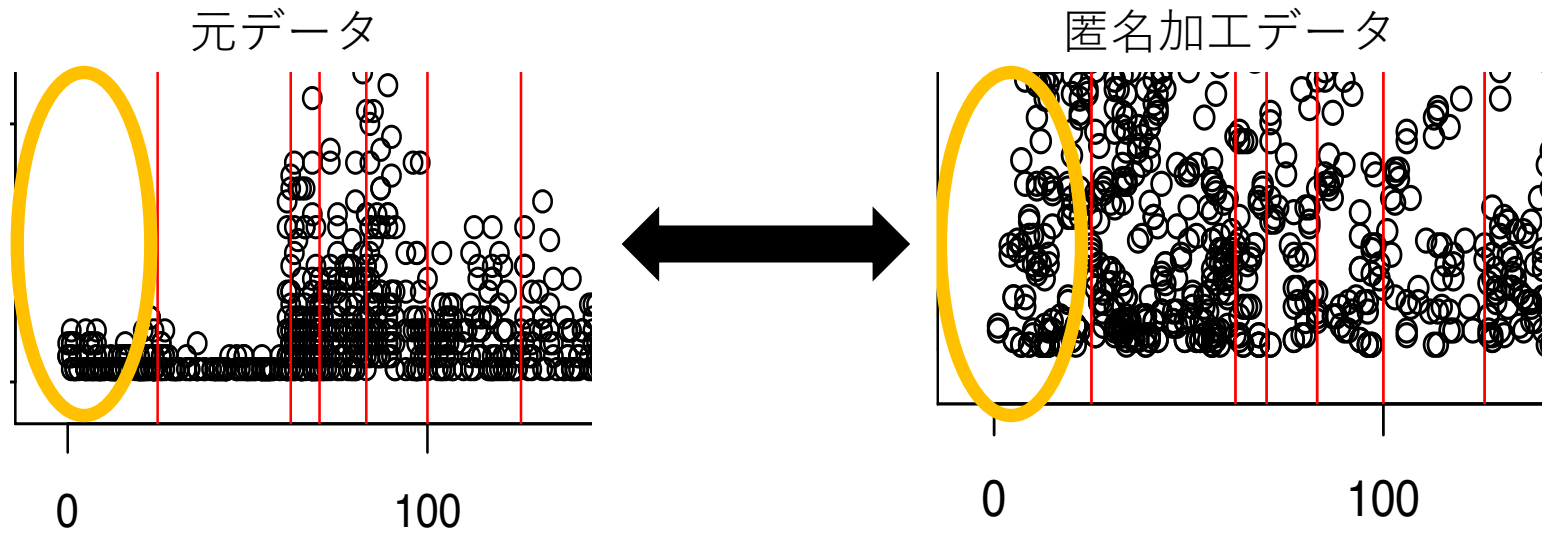
- R,F,Mをそれぞれ10分位値でクラスタリング

年間購買総額(M)と累積購買金額比率



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でクラスタリングすることでそれぞれのクラスタの顧客に対して最適なマーケティングを行うことができる
- **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

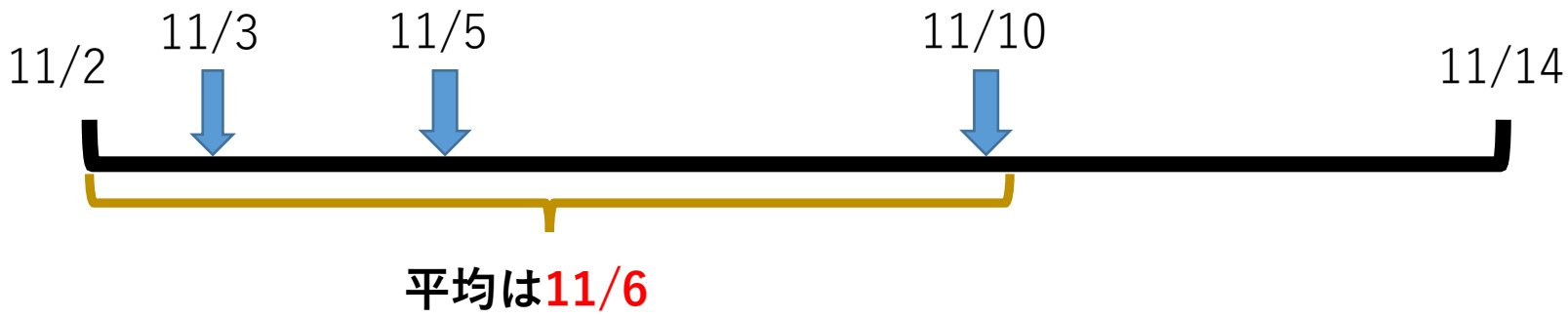
- 匿名加工データのRFMを計算するために特定の値が必要

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

→

顧客ID	R	F	M
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

顧客ID	購買日	商品ID	単価	購買数量
12348	11/14	21	3.7	1
12513	11/2	23	2	2

匿名加工データ

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]

- *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

結果・考察

k	有用性(R)	有用性(F)	有用性(M)	有用性(RFM)	安全性
2	0.270	0.463	0.352	0.097	0.5
3	0.214	0.343	0.301	0.040	0.33
4	0.155	0.287	0.288	0.026	0.25

- **kの値が大きくなるほど有用性が下降、安全性が上昇**
- **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
- **R,F,M は独立ではない**
  - RFMの有用性がR,F,Mの積よりも高い
  - k=2の時：**0.27*0.463*0.352=0.044<0.097**

購買日
[8/2,11/14] [6/10,11/14]

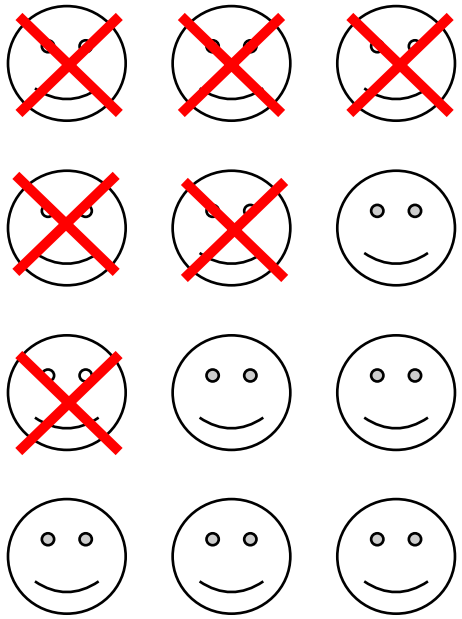
まとめ

- 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

安全性評価

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]

- k=nの時の平均再識別率で評価
- k=3の時



3人中3人が識別される確率→$^1/_6$

3人中2人が識別される確率→0

3人中1人が識別される確率→$^1/_2$

3人中0人が識別される確率→$^1/_3$

平均再識別率は$^1/_3$

- k=2,3,4の時は何も$^1/_k$となった

R,F,Mは独立ではないの説明

k	有用性(R)	有用性(F)	有用性(M)	有用性(RFM)
2	0.270	0.463	0.352	0.097