Anonymization and Re-identification for Personal Transaction Data

Hiroshi Nakagawa

（The University of Tokyo／Riken AIP）

Privacy Concern

- In EU, GDPR focuses on this privacy protection issue legally, technically aiming at IT businesses.

- In Japan, 2016
- The private data protection acts (revised)
  - The new concept of "anonymized private data."

- Anonymized private data can be treated as if they are not personal data any more,
- they are even transferred to the third party without data subject's consent.

- The way to transform personal data into anonymized private data
  - clearly defined at least in technical sense.
- We have to estimate how easily an anonymized personal data is re-identified, in order to give the technical evaluation to legal authorities who make the definition of anonymized private data.

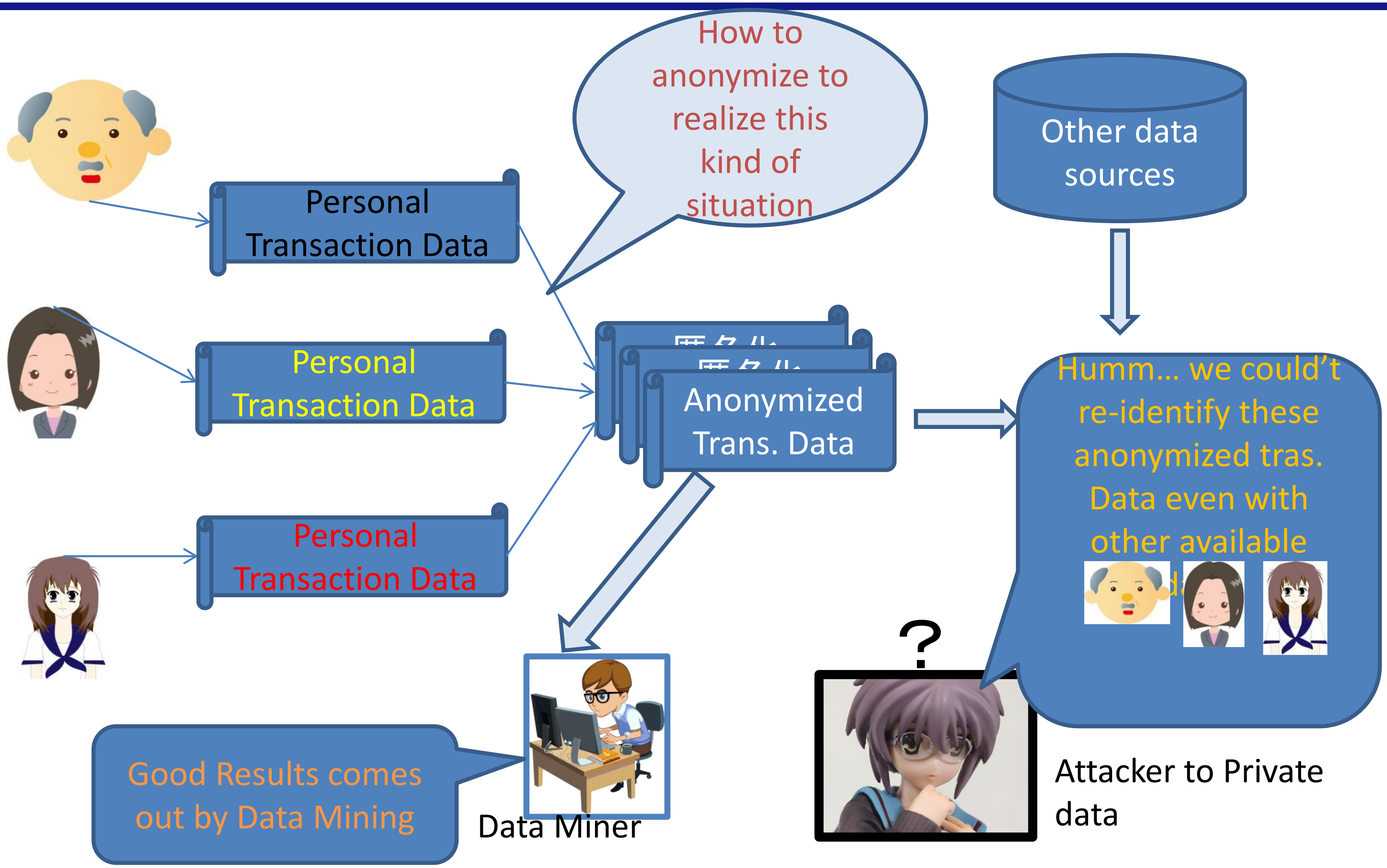
PWSCUP 2015,2016

- For this purpose, we organized PWSCUP last October.
- The competition of PWSCUP was: for given transaction data (400 people transaction of purchasing for one year period),
- 1) 15 teams submitted anonymized transaction data by their own methods.
- 2) Each team tried to re-identify other teams’ anonymized transaction date.

Winner

The highest score of utility
+ # of non-re-identified person.

The situation we want to work out by anonymization



How to anonymize to realize this kind of situation

Other data sources

Personal Transaction Data

Personal Transaction Data

Personal Transaction Data

匿名化

匿名化

Anonymized Trans. Data

Humm… we could’t re-identify these anonymized tras. Data even with other available da

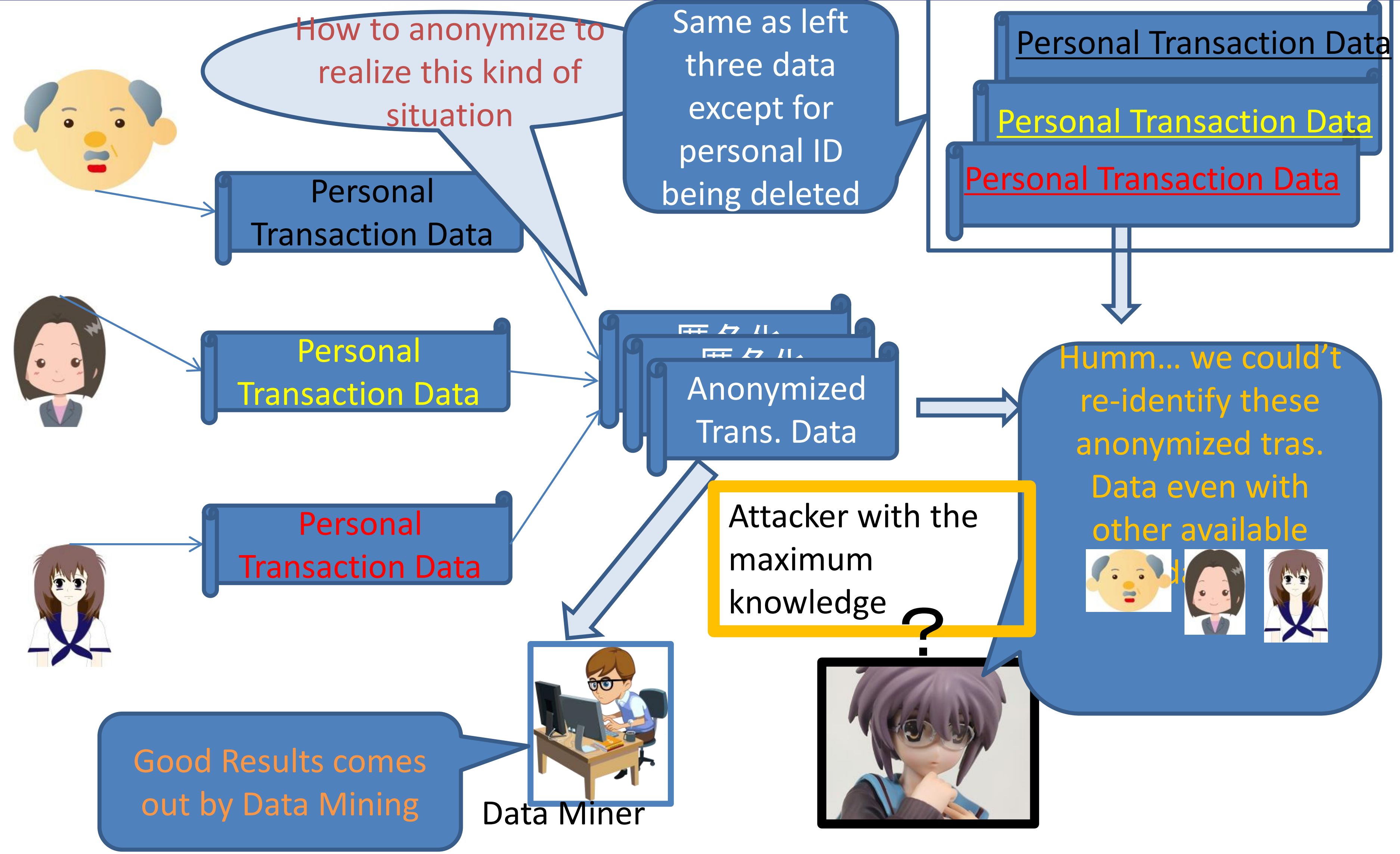
?

Good Results comes out by Data Mining

Data Miner

Attacker to Private data

The situation we want to work out by anonymization



How to anonymize to realize this kind of situation

Same as left three data except for personal ID being deleted

Personal Transaction Data

Personal Transaction Data

Personal Transaction Data

Personal Transaction Data

Personal Transaction Data

Personal Transaction Data

Anonymized Trans. Data

Humm… we could’t re-identify these anonymized tras. Data even with other available d

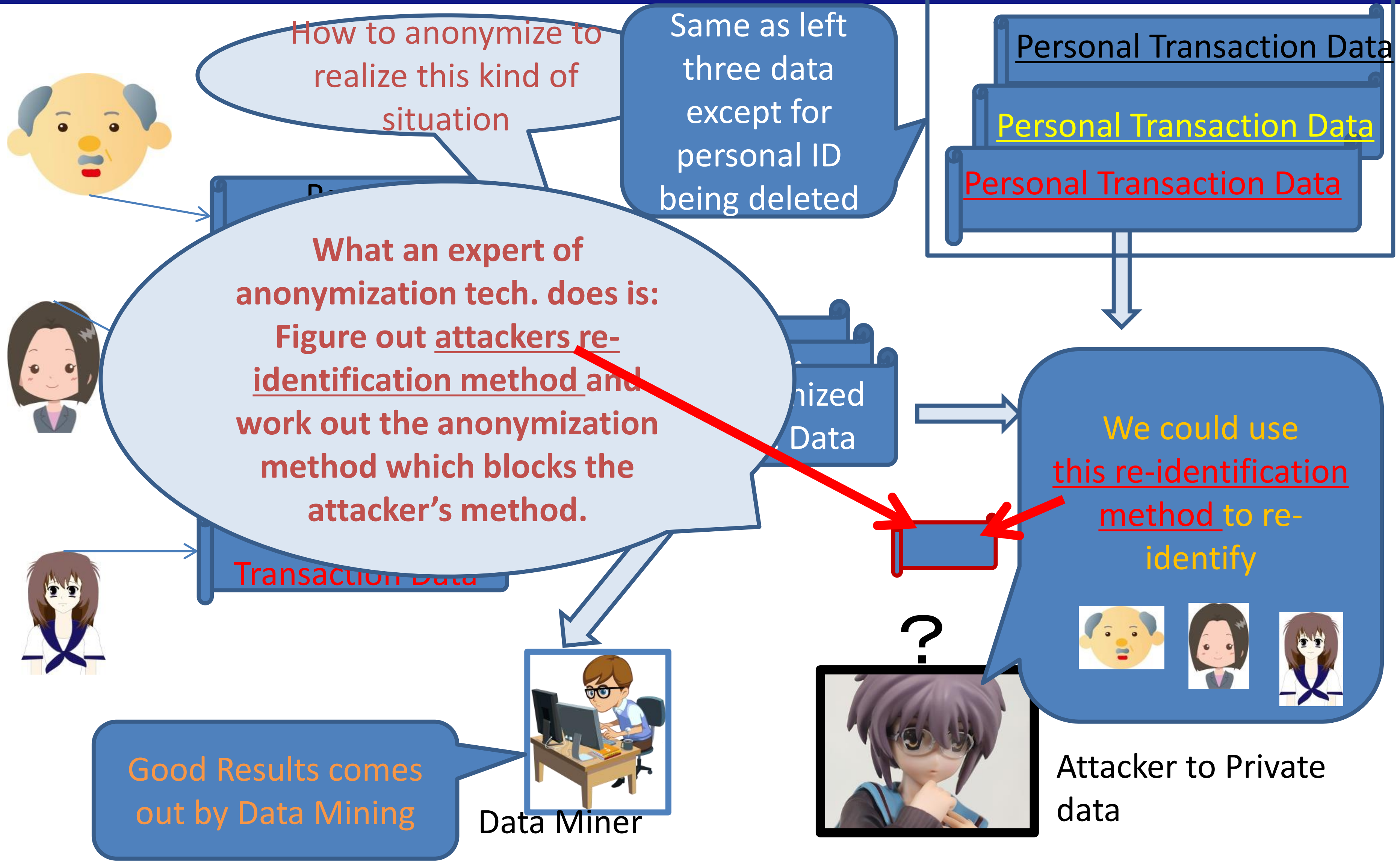
Attacker with the maximum knowledge

?

Good Results comes out by Data Mining

Data Miner

PWSCUP：Expert of anonym. tech. does this way！



How to anonymize to realize this kind of situation

Same as left three data except for personal ID being deleted

Personal Transaction Data

Personal Transaction Data

Personal Transaction Data

What an expert of anonymization tech. does is: Figure out attackers re-identification method and work out the anonymization method which blocks the attacker's method.

nized Data

Transaction Data

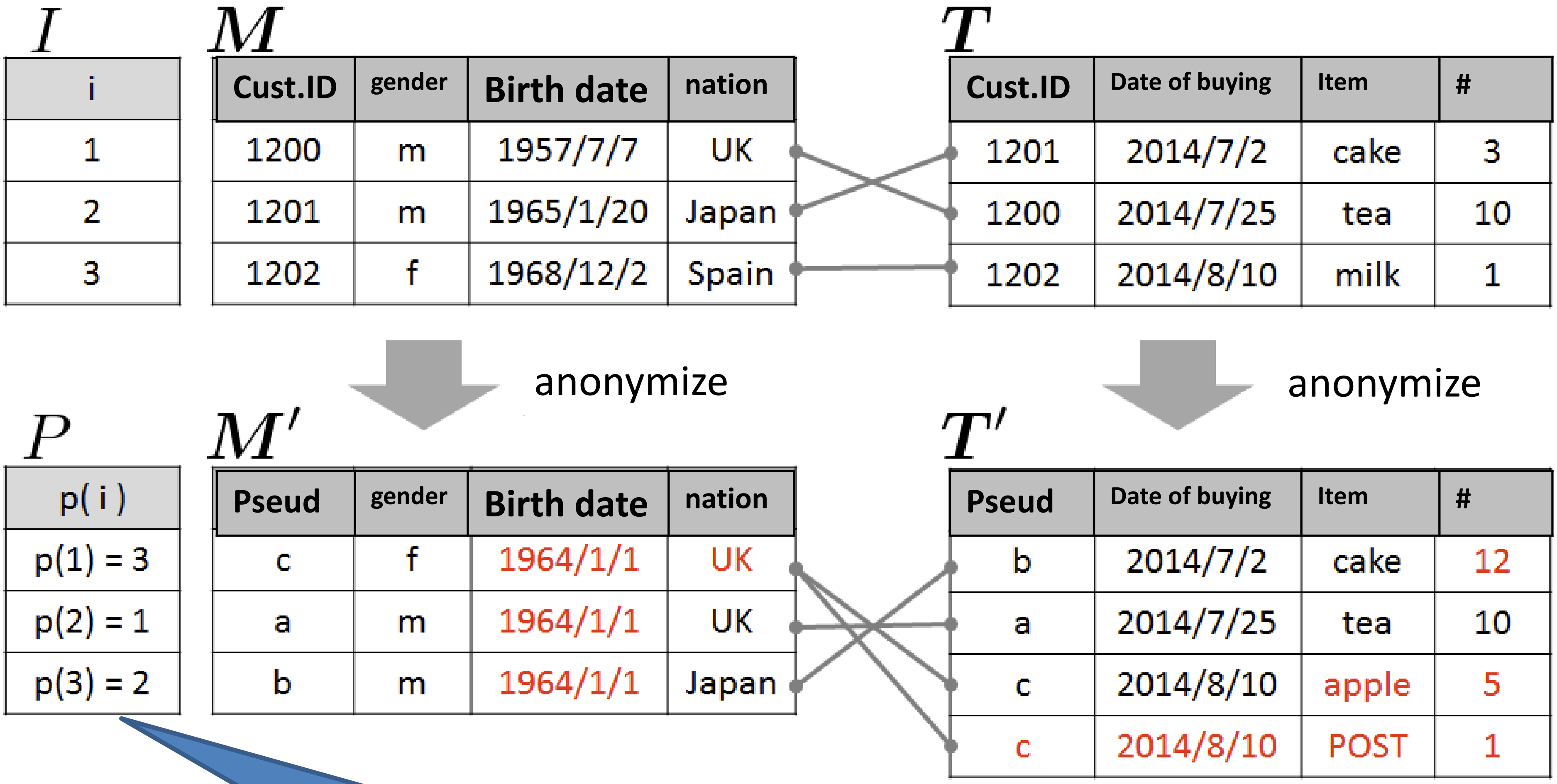
We could use this re-identification method to re-identify

Good Results comes out by Data Mining

Data Miner

Attacker to Private data

Record of Purchase DB used at PWSCUP



I

i
1
2
3

M

Cust.ID	gender	Birth date	nation
1200	m	1957/7/7	UK
1201	m	1965/1/20	Japan
1202	f	1968/12/2	Spain

T

Cust.ID	Date of buying	Item	#
1201	2014/7/2	cake	3
1200	2014/7/25	tea	10
1202	2014/8/10	milk	1

P

p(i)
p(1) = 3
p(2) = 1
p(3) = 2

M'

Pseud	gender	Birth date	nation
c	f	1964/1/1	UK
a	m	1964/1/1	UK
b	m	1964/1/1	Japan

T'

Pseud	Date of buying	Item	#
b	2014/7/2	cake	12
a	2014/7/25	tea	10
c	2014/8/10	apple	5
c	2014/8/10	POST	1

p(i) : order of records
＝permutation of row # of table data

Attackers with Maximum Knowledge Model and PWSCUP task

- Attacker, who does re-identification, knows M and T.
- Then, try to figure out the permutation {p(i), i=1,n} from anonymized M'and T',

  which is re-identification

  - Re-identification rate is the ratio of being properly re-identified.

Utility Measures (in Kikuchi)

- How similar M’,T’(anonymized data) with M,T(original data)
- *cmae*: Cluster based similarity
  - Customers are clustered by nationality and gender.
- *subset*: The maximum value of difference between average of total purchase of X and that of X’, for consecutive 30 days

Utility measure : RFM(M, M', T, T’)

- Customers M / M‘ are clustered by

  Recency (last purchasing date),

  Frequency(frequency of purchasing) and

  Monetary (amount of money paid) of T / T’.

- Then RFM(M, M', T, T’) is the normalized RMS between these two clusters is .

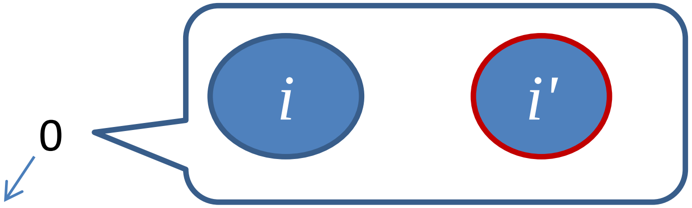
Utility measure : ut-jaccard → important!

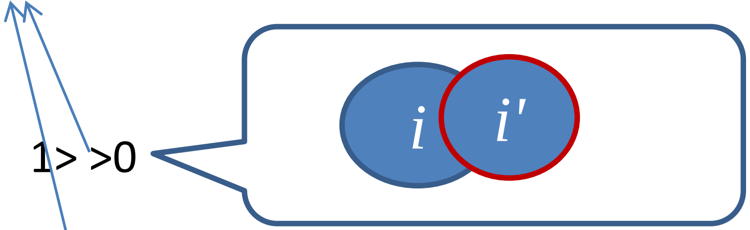
- $S(T,i)$: a set of items purchased by customer c_i , described in T.
- $S(T',i)$: a set of items purchased by customer c_i , described in T’.
- Jaccard coeffcient :
- $d\left(S(T,i),S(T',i)\right) = 1 - \frac{|S(T,i) \cap S(T',i)|}{|S(T,i) \cup S(T',i)|}$
- Sum of d within M :

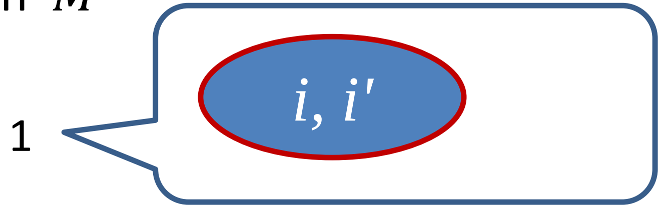
$$ut - jaccard(M,M',T,T',p)$$

$$= \frac{1}{n'}\sum_{i=1}^{n'} d\left(S(T,i),S(T',i)\right)$$

where n' is a number of records in M'







Imposed condition on utility measures and anonymization schema

- $subset \leq 50000$

  and ut-jaccard $\leq 0.7 \cdot (\# \; of \; records \; in \; T)$

- The condition on ut-jaccard is severe,

because we could not do big change of data value or shuffling records order.

Imposed condition on utility measures and anonymization schema

1. Anonymizers try to work out anonymization method which satisfies the condition on ut-jaccard as tightly as possible.
2. Attackers try to work out re-identification method considering the above mentioned anonymization method.
3. The anonymizers try to develop anonymization methods that overcome the above mentioned re-identification methods.

First of all, how to design effective re-identification method?

- Each team submits anonymized data which preserve purchased item set of each customer to high extent.

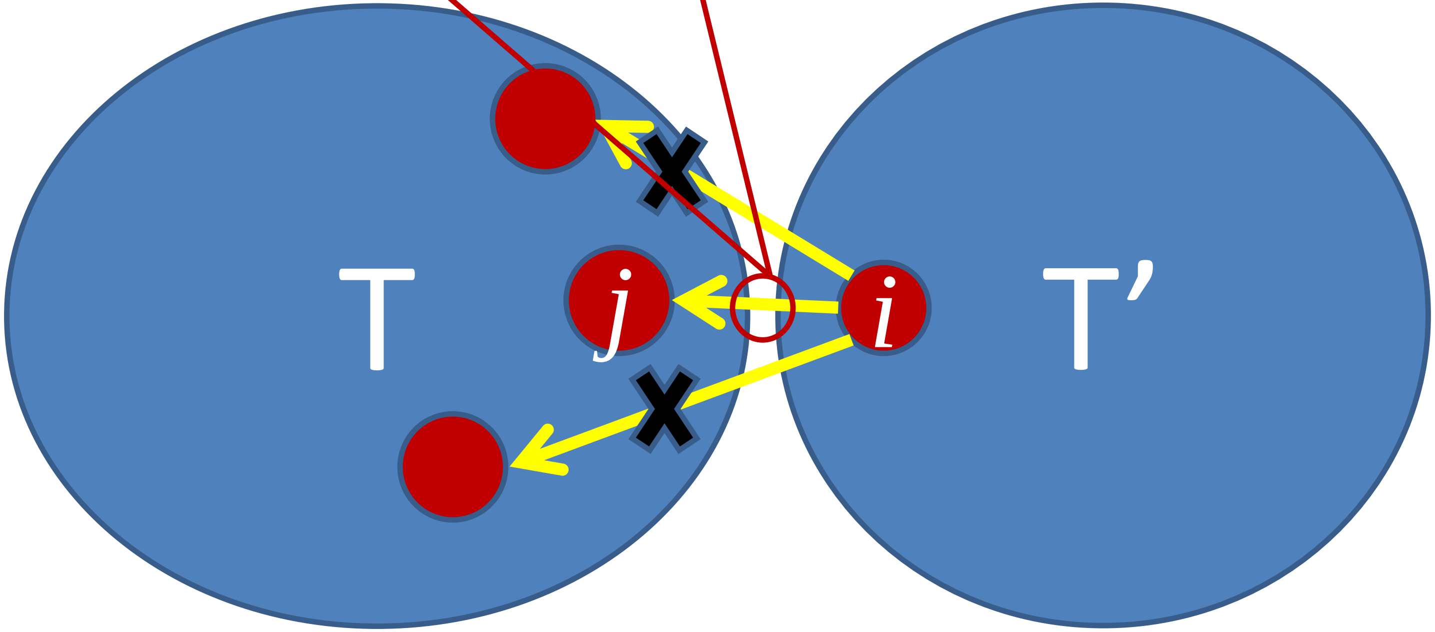
- Customers’ purchased item sets are very diverse.
- Then it is hard to make re-identification difficult while maintaining the condition of ut-jaccard.

- Considering this, we proposed the re-identification method: ***re-itemset*** shown in the next slide.

Effective re-identification method: *re-itemset*

The most similar S(T,j) to S(T’,i) in terms of ut-jaccard

=S(T,i)’s counterpart



Outline of anti “*re-itemset*”

1. Make a *ci* centered cluster which consists customers *cj(j≠i)* whose *S*(*T* ; *j*) is similar to *S*(*T* ; *i*). → Precisely described later

2. Modify *cj*’s items in order to make all customers within *ci* centered cluster have the same item set ,
   - all customers in *ci* centered are regarded as *ci*.
   - → At most one customer is re-identified within one cluster, say *ci*.
   - Then, we want to minimize the number of clusters under the condition of utility measures such as “ut-jaccard≤0.7”

Expected re-identification rate and the results of PWSCUP competition

- Our anonymization algorithm satisfies

  “ut-jaccard≤0.7･(# of records in T) as well as other utility conditions.

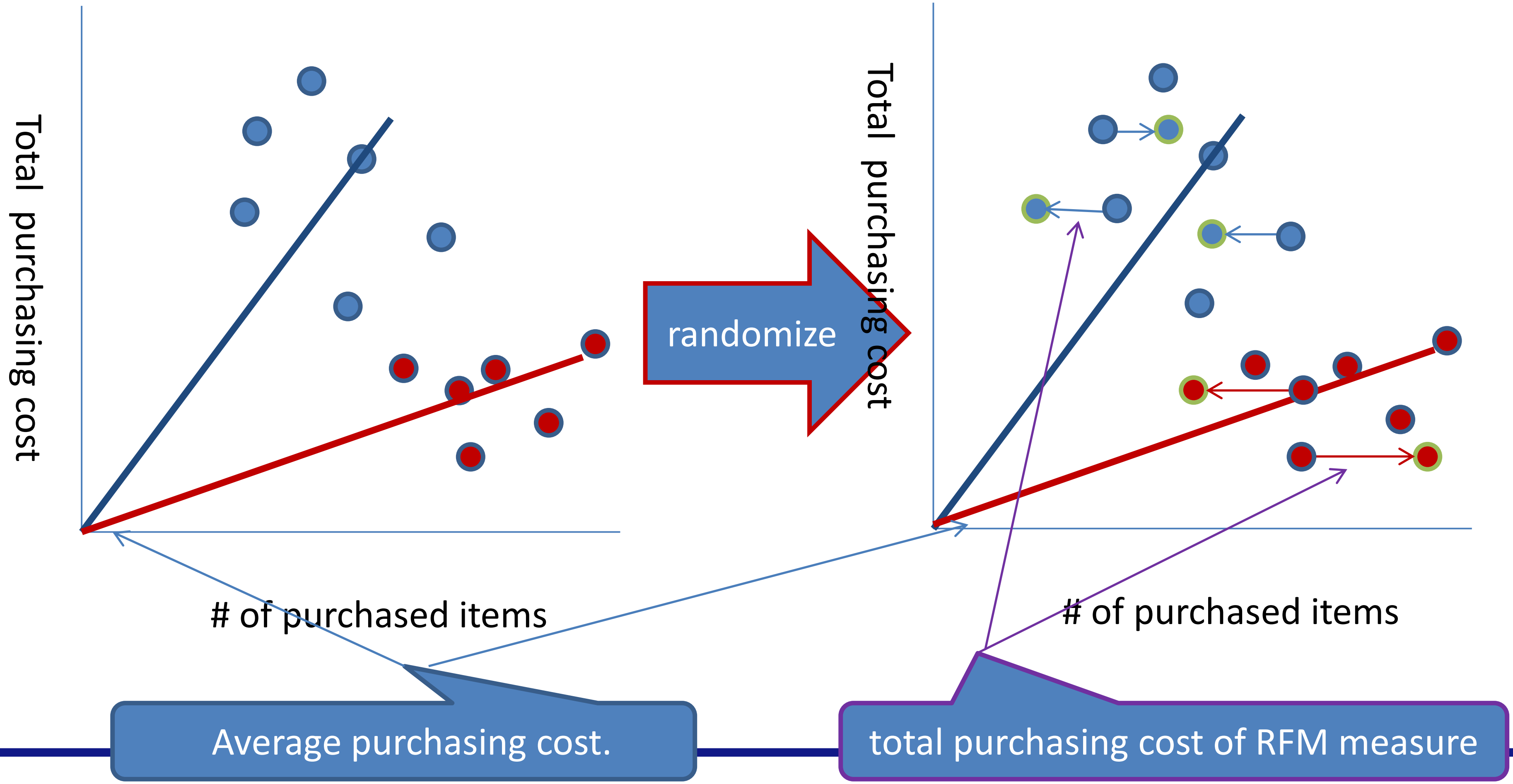
- In PWSCUP, 400 customers are divided into 89 clusters with ut-jaccard =0.699

- We expect that re-itemset algorithm does not re-identify more than 90 customer if more than one customers within one cluster are re-identified as we planned.

- Great!! At most 89 customers are re-identified on PWSCUP re-identification phase.

Sketch of randomization

- Randomize not to be re-identified within the cluster
- Keep utility measures as invariant as possible



Summary of PWSCUP

- Many teams seem to employ ***re-itemset*** tuned to ut-jaccard as re-identification method.

- At PWSCUP re-identification phase, at most 89 customer (22.5% of 400 customers) of our team's anonymized data got re-identified as we expected.

- As explained, 89 is the upper bound of ***re-itemset*** tuned to ut-jaccard.

- **Note that the value of this 22.5% depends on**
  - **employed utility measures**
  - **nature of target data base.**

- Thus, 22,5% is to be regarded as a reference value of this PWSCUP contest. → We do not have a one fits all approach!

Prospects

- We have to design anonymization method considering the following three conditions:

- ➢ Maintenance and management of ID of data subjects and pseudonym (psuedo ID)
- ➢ Anonymization which prevents re-identification such as proposed at PWSCUP
- ➢ Quality and quantity an attacker has.
  - ➢ A long transaction data is dangerous because some of action described in it might be observed and used by the attacker.

Appendix

The details of

1. Re-identification algorithm
2. Randomization sketch

How to develop anonymization method given the lower bound of re-identification rate

```
while{re-identification rate > Theshold}
    create a new anonymization method:A
    Apply A to personal DB:D and get the result:A(D)
    if {A(D) satisfys the predetermined utility condition:C }
        work out a new re-identification method R against A(D)
        calculate re-identification rate by applying R to A(D)
    end if
end while
return anonymization method:A
```

Utility measure : *cmae*

- Clustering customer by gender and nationality
  - Notation
  - {C}: The whole cluster . *s*: Subset of C. *p: permutation*
  - T|s : customer data of T which is in *s* of T
  - *tj* :j-th record of T

Average cost of item in cluster *s*:$\mu_{up}(T|s) = \frac{\sum_{tj \in T|s} unit\ cost\ of\ tj\ \ \cdot \#\ of\ t_j}{\sum_{tj \in T|s} \#\ of\ t_j}$

Average absolute error for the whole cluster

C: $cmae(M, M', T, T') = \sum_{s \in C} \frac{|\mu_{up}(T|s)| - |\mu_{up}(T\prime|s)|}{|C|}$

Utility measure : subset

- X’ is a set of 10 selected customers from M’.
- X is a counter part of X’ in M.
- The following subset means the maximum value of difference between average of total purchase of X and that of X’, for consecutive 30 days: $subset((M,T),(M'T'),p) = max_{X',D}(\left|\mu_{tp}(X',D,T')\right| - \left|\mu_{tp}(X,D,T)\right|)$

Randomizing customer's item set in clustering of anonymization

- In order that less than 90 customers within one cluster are re-identified, we may highly randomize customer's item set in one cluster or clustering itself.

- But, too much randomization degrades utilities.

➢ We need the method including both of randomization of clustering and item set and maintaining utilities.

Effective re-identification method: ***re-itemset***

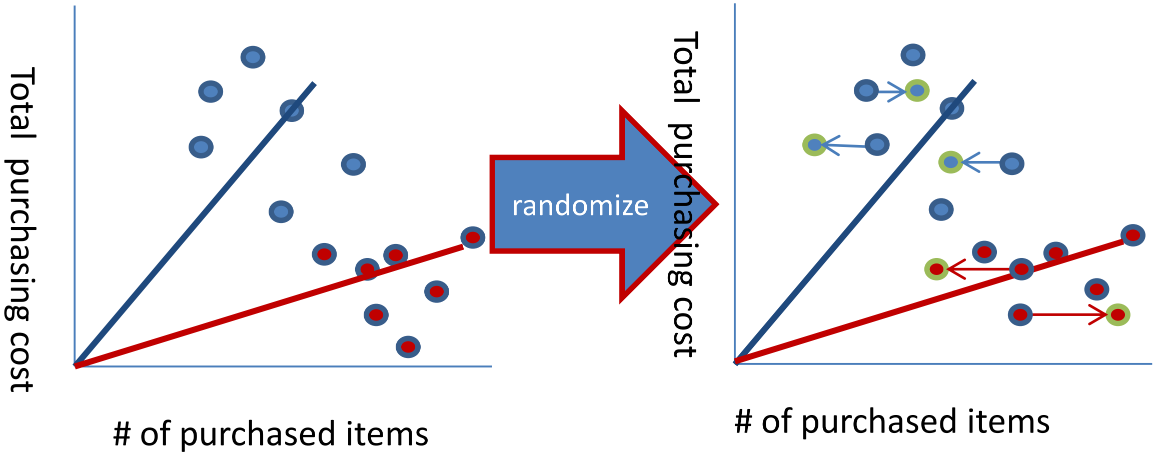
```
$n' \leftarrow |M'|$
for$\{i = 1\ to\ n'\}$
    $q(i) \leftarrow argmin_j d\left(S(T,j), S(T',i)\right)$
end for
return $Q = \left(q(1), \cdots . q(n')\right)$
```

- The majority of teams employs this ***re-itemset*** , which is actually the most powerful re-identification method meaning it re-identifies the highest number of re-identified customers.

Clustering method of anonymization

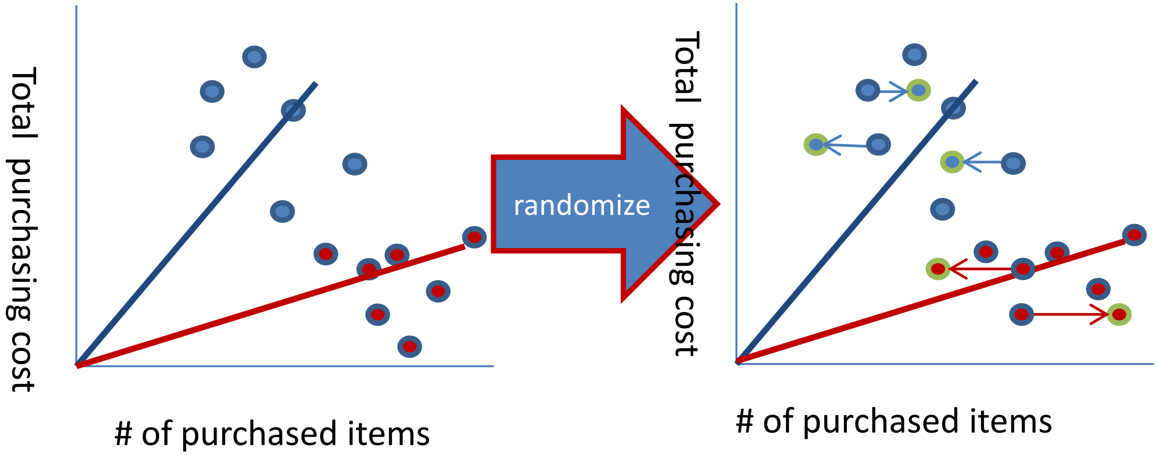
- Step 1 Randomize some customer’s purchasing data in a cluster.
- Step 2 Adjust other customer’s purchasing data to maintain utilities.
- Step 3 Re-build T’ based on adjusted purchasing data.

Step 1



- Some customer ● and ● are randomly and horizontally shifted to ● and ● .
  - The purpose of “Randomly” means making hard to identify the corresponding original data.
  - The Purpose of “horizontally” is the following:
  - If one ● (●)or is shifted in the right direction, another ● (●) is shifted in the opposite direction in order to total purchasing cost of RFM measure be invariant within a cluster.

Step 2



- To prevent big degradation of the utility measure of average absolute error : ***cmae***,

  a center of gravity of each cluster should be kept as possible .
  - a center of gravity of each cluster means average purchasing cost.

- Under this condition, customers are randomly moved.
- However, each customers can move only once at Step2.

- Suppose *U* is a cluster whose non moved customer is smallest.
- Non moved customers in *U* are moved to adjust (= keep) the average value of utilities of the cluster.