MEIJI UNIVERSITY

Design for a data Anonymization Competition 2018

Hiroaki Kikuchi (Meiji Univ.)

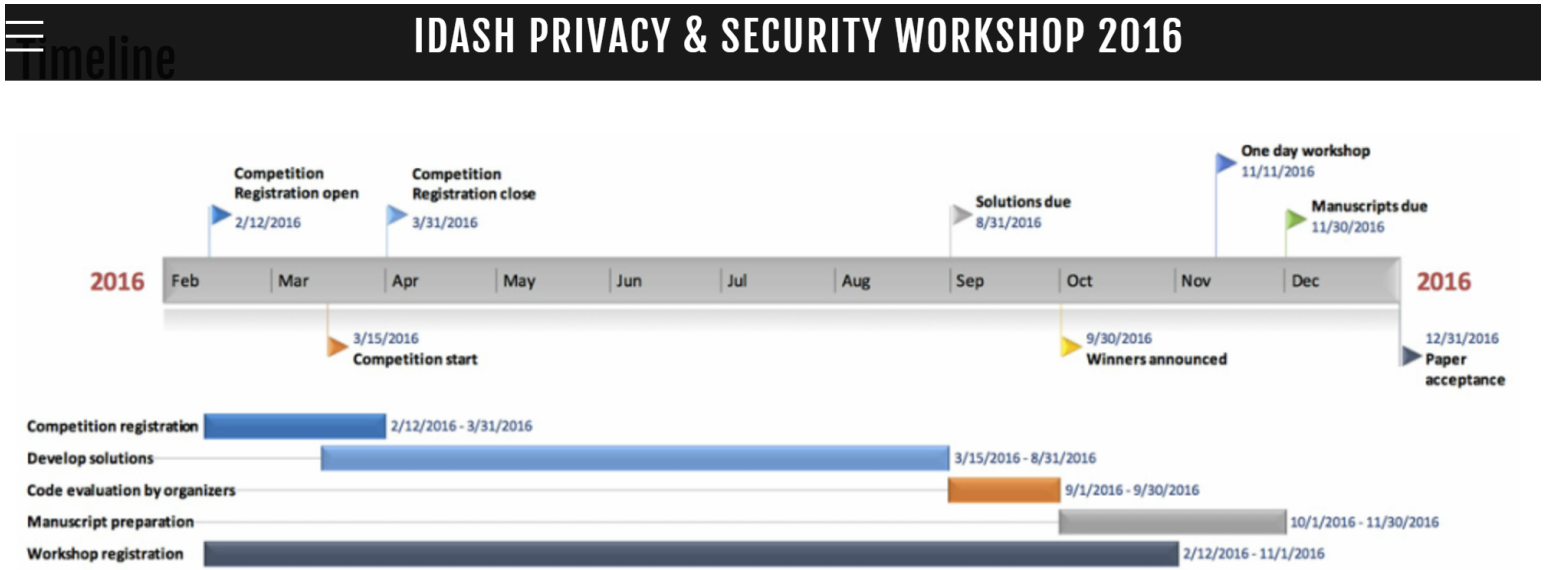
PETS 2017, Minneapolis, US

Criticize to past PWSCUP

- 1. Hidden algorithm
  - Players submit the anonymized data without showing source or algorithm. Not able to analyze the process for details.
- 2. Max-knowledge assumption is too strong.
  - It is far from reality.
- 3. Record-linkage challenge is problematic.
  - Instead, why don’t us to attribute estimation?
- 4. Synchronized fashion of games
  - Arbitrarily attack and defense is more exciting, like the CTF style.

Open-Source style

- iDash Privacy and Security WS



Important time points

02/12/2016 Competition registration open

~~**03/15/2016**~~ (delayed to 3/18/2016) Competition start (data release to registered teams)

03/31/2016 Early competition registration

05/31/2016 Workshop registration deadline (onsite attendance)

09/07/2016 Solutions submission registration due (No more registration allowed after 09/07/2016). Here

1. Pros and Cons for Open-Source style

- Pros
  - Allows deep analysis
  - Can be re-used for anonymizing other dataset.
  - Fair and reliable. Allows to trace the steps one by one.
  - “cheating” can be denied.
  - No need high-performance
- Cons
  - Revealing method is prohibited by Japanese low
  - Most companies does not allow to submit their source since it has IP.
  - Not processed in a single source. Often used internal library.

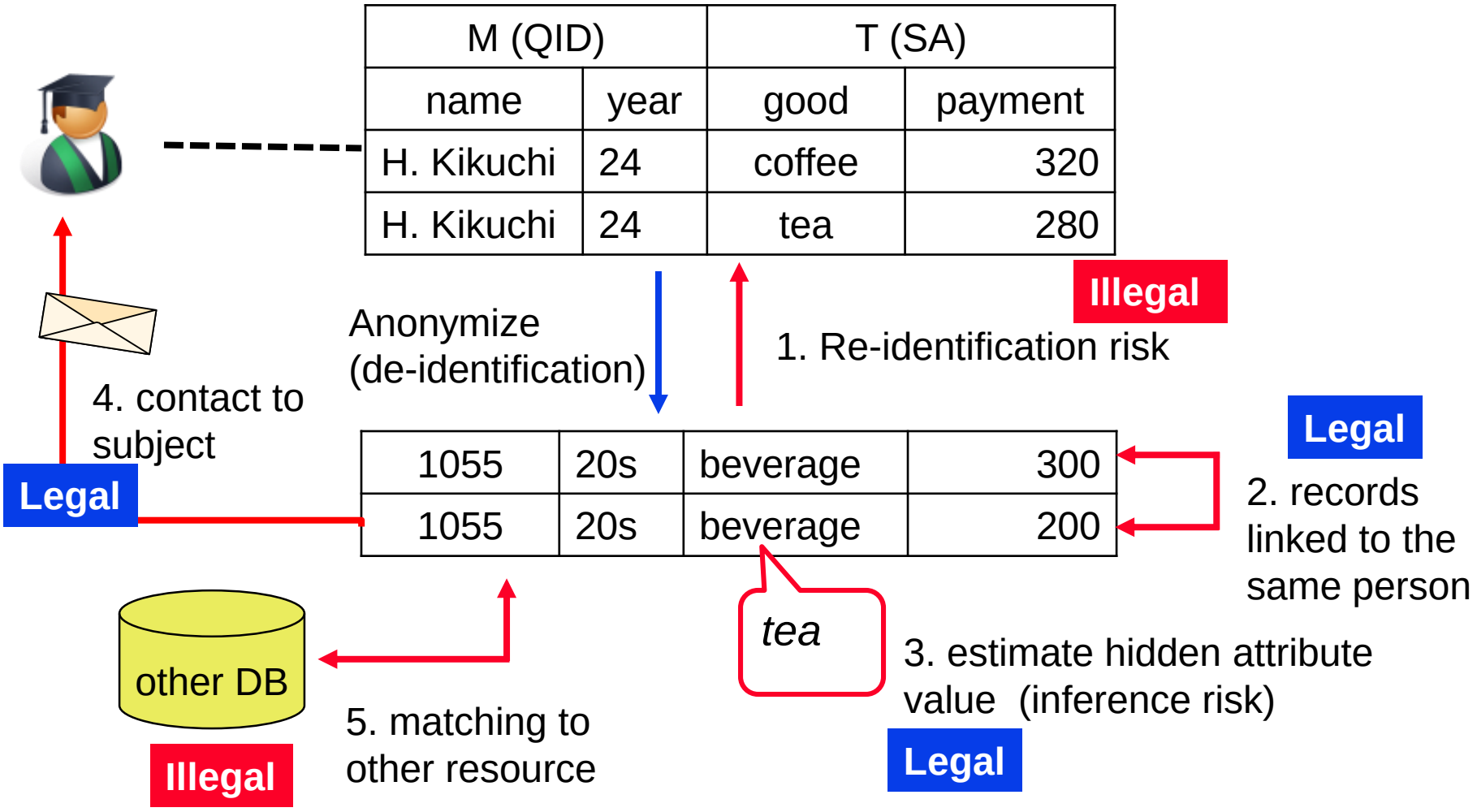
Our Suggestion to 1.

- We should have a closed-source (PWSCUP) style so that industry teams can participate.
- Alternatively, we may have an additional open-source style completion as well as the closed-style.

2. Why we assume the Max-knowledge adversary

- Reasons
  - It is simple. If some algorithm was better than others in the Max-knowledge adversary, it could be safe against a moderate adversary.
  - Many requests to join both anonymizing and re-identifying. (including committee members)
  - It is hard to provide exactly equal knowledge to all parties. The risk may quite depend on the (partial) knowledge.

3. Why we did not study attribute estimation in the past PWSCUP



Our new competition

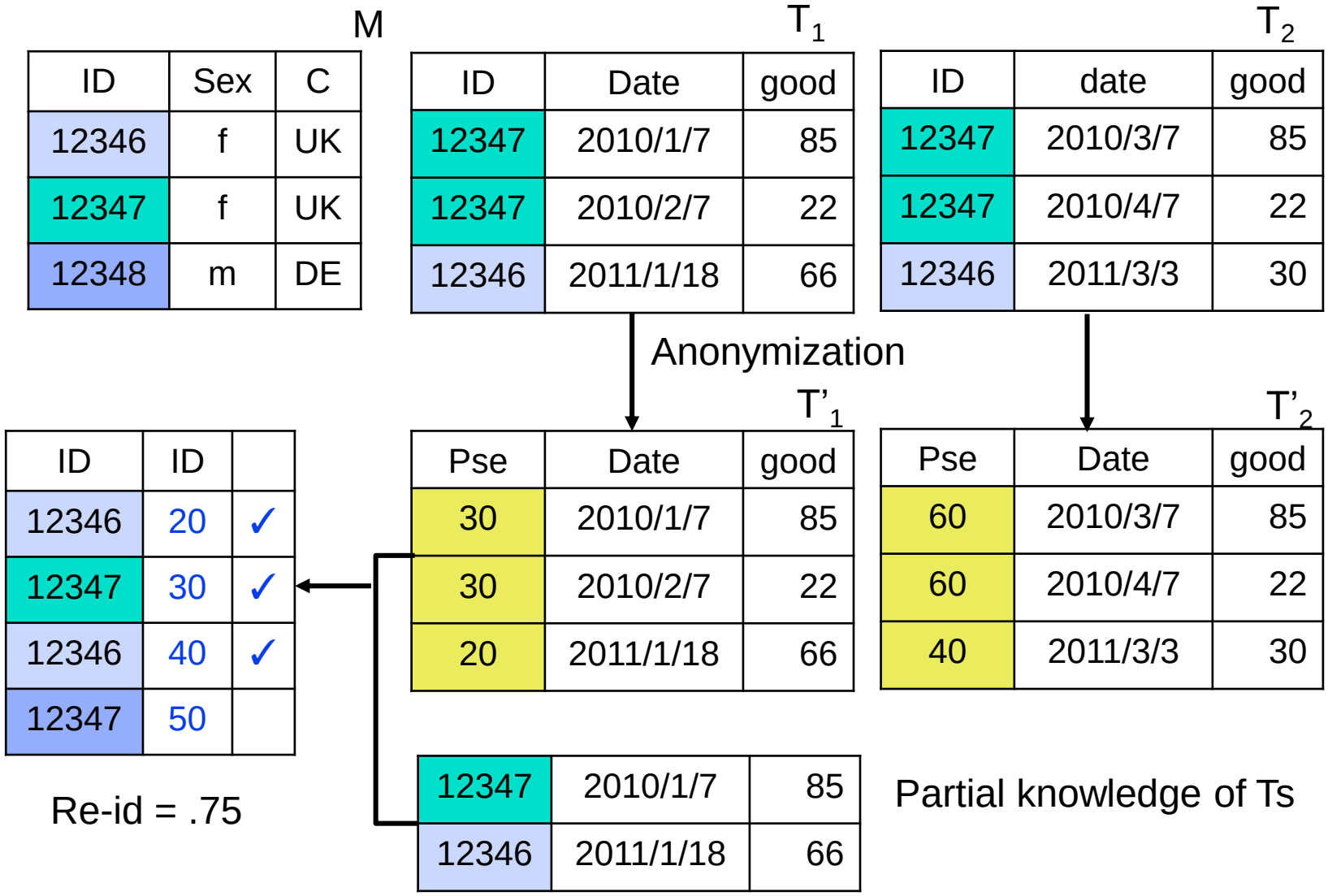
Update PWSCUP 2017

PWS CUP 2017

- Oct. 23-25
- Yamagata Int. Hotel
- Call
  (July 24-Aug. 21)
- Privacy Workshop 2017 (IPSJ, Sig. CSEC)

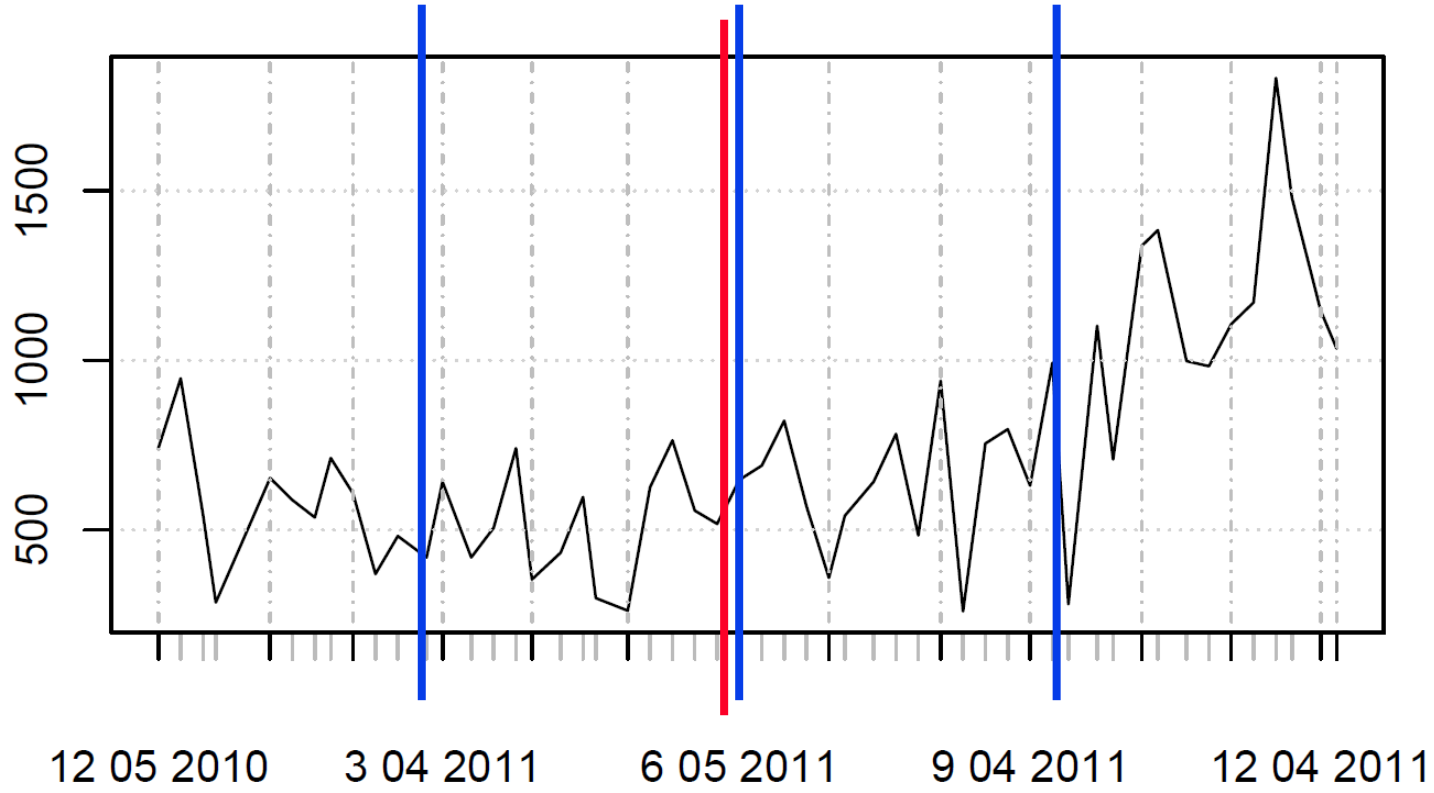
2017 Outline

Anonymize: submit T'1, T'2, T'3, …
Identification: given T'1, T'2, guess IDs



1-year History divided

- cnt <- zoo(t400$V7, d400)
  cnt.weekly <- apply.weekly(as.xts(cnt), length)



Changes in 2017

- 1. anonymization of long history
  - Allows multiple pseudonyms per one person so that re-identification becomes harder
  - The more pseudonym, the more secure. But, it accordingly loses the utility.
- 2. weaken the adversary's knowledge
  - Given (some) partial transaction records, try to estimate model and guess the assignment

Some plans for Competition

Proposal of completion 2018

- Plan A. NSTAC synthesized data
- Plan B. Online Retail
- Plan C. Online Retail with pseudonyms
- Plan D. Open Algorithms completion
- Plan E. Trajectory Data

Plan A "Pseudo Micro Data"

- NSTAC (National Statistics Center)
  - Real statistics about income and expenditure for Japanese household in 2004.

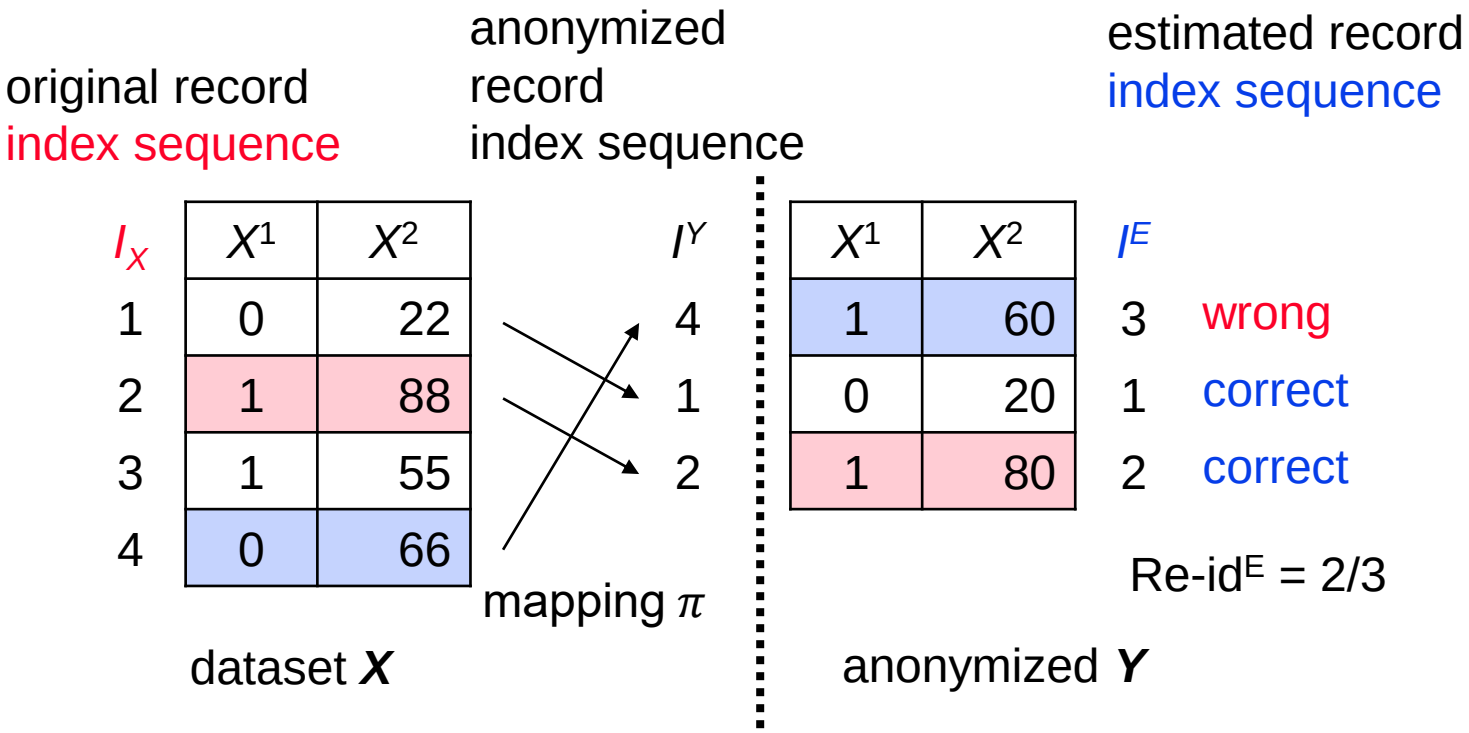
Dataset	# of records	QI	SA	
	n	m	(exp)	(inc)
Full	32,027	14	149	34
Simple	8,333	14	11	N/A

http://www.nstac.go.jp/services/giji-microdata.html#P

Pseudo Micro Data (Tbl. VII)

No	Attribute	# of value	Average	Example	Type
1	Type	1	1	1 (empied)	QID
2	# of people	1	4	4	QID
3	# of employed	1	1.504	1	QID
4	Accom. Type	5	1	1 (wooden)	QID
5	Bldg. type	7	1	1 (detached)	QID
6	Owner	8	1	1 (owned)	QID
7	Sex	1	1	1 (male)	QID
8	Age	11	5	1 (1-18 Y/O)	QID
…					QID
14	Weight	8333	15.741	13.2	SA
15	Total Expenditures	8333	324,525	155,006	SA
16	Foods	8333	74,639	25,227	SA
17	Accom.	8333	14,686	2000	SA
14	Lightning	8333	19,733	18,333	SA
…					SA
25	Others	8333	62,227	20,455	SA

Record Re-identification



Re-identification Ratio:

Re-id $^{IE}(I^Y, I^E) = |\{j \text{ in } \{1,\ldots,n'\} \mid i_j^Y = i_j^E\}/n'$

Plan B: Online Retail

- Dataset
  - UCI Machine Learning, “Online Retail”
- Task
  - Identify secret permutation P(M) from anonymized data M’ and T’
- Limitation
  - Assign one pseudonym to one customer

Plan C: Online Retail with Many Pseudonyms

- Dataset
  - UCI Machine Learning, “Online Retail”
- Task
  - Identify owners of records from anonymized history T’ using partial knowledge
- Limitation
  - Assign one pseudonym to one customer

Plan D: Open-source style competition

- Data:

Plan E: Trajectory Data Competition