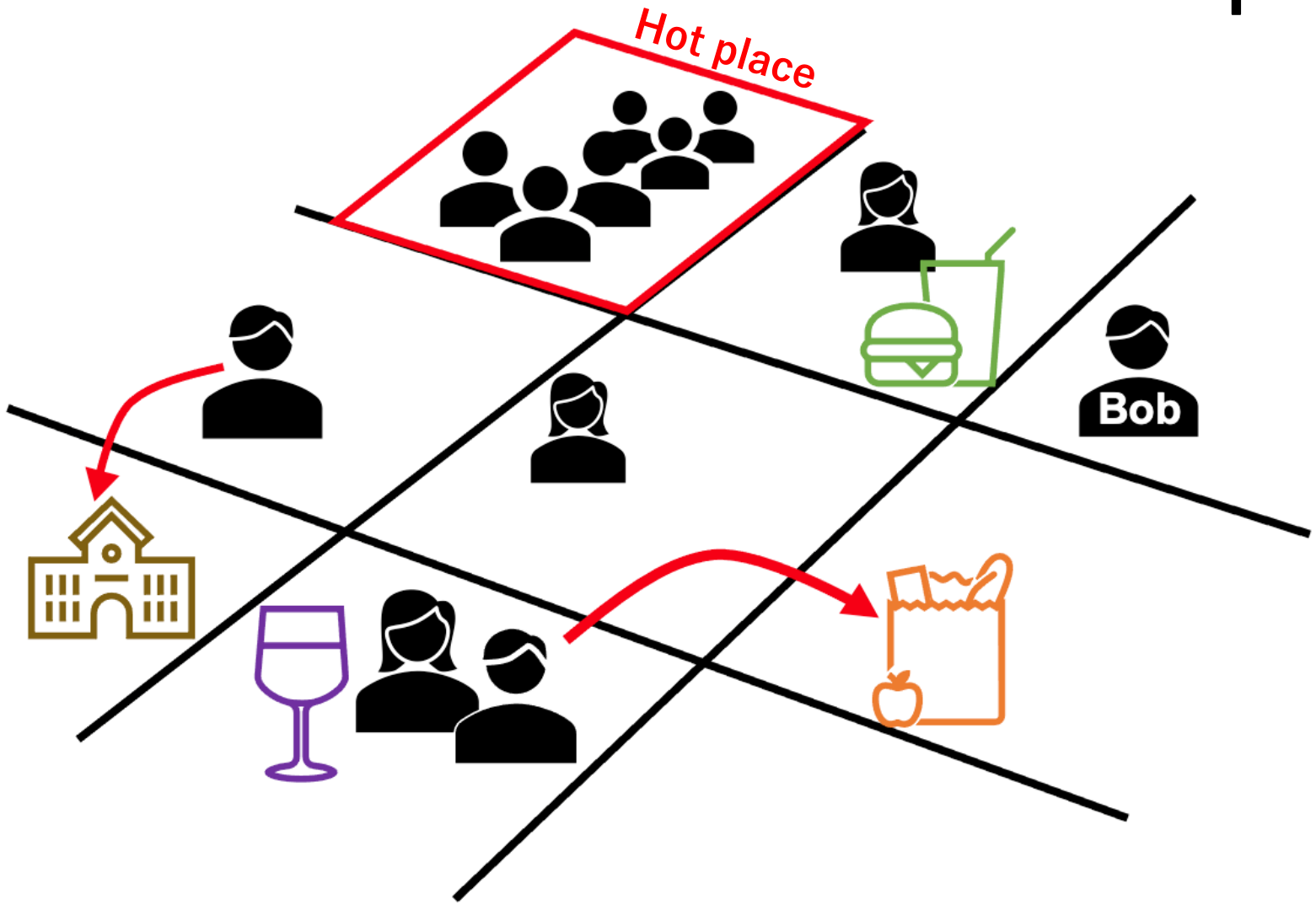
Improvement of Estimate Distribution with Local Differential Privacy

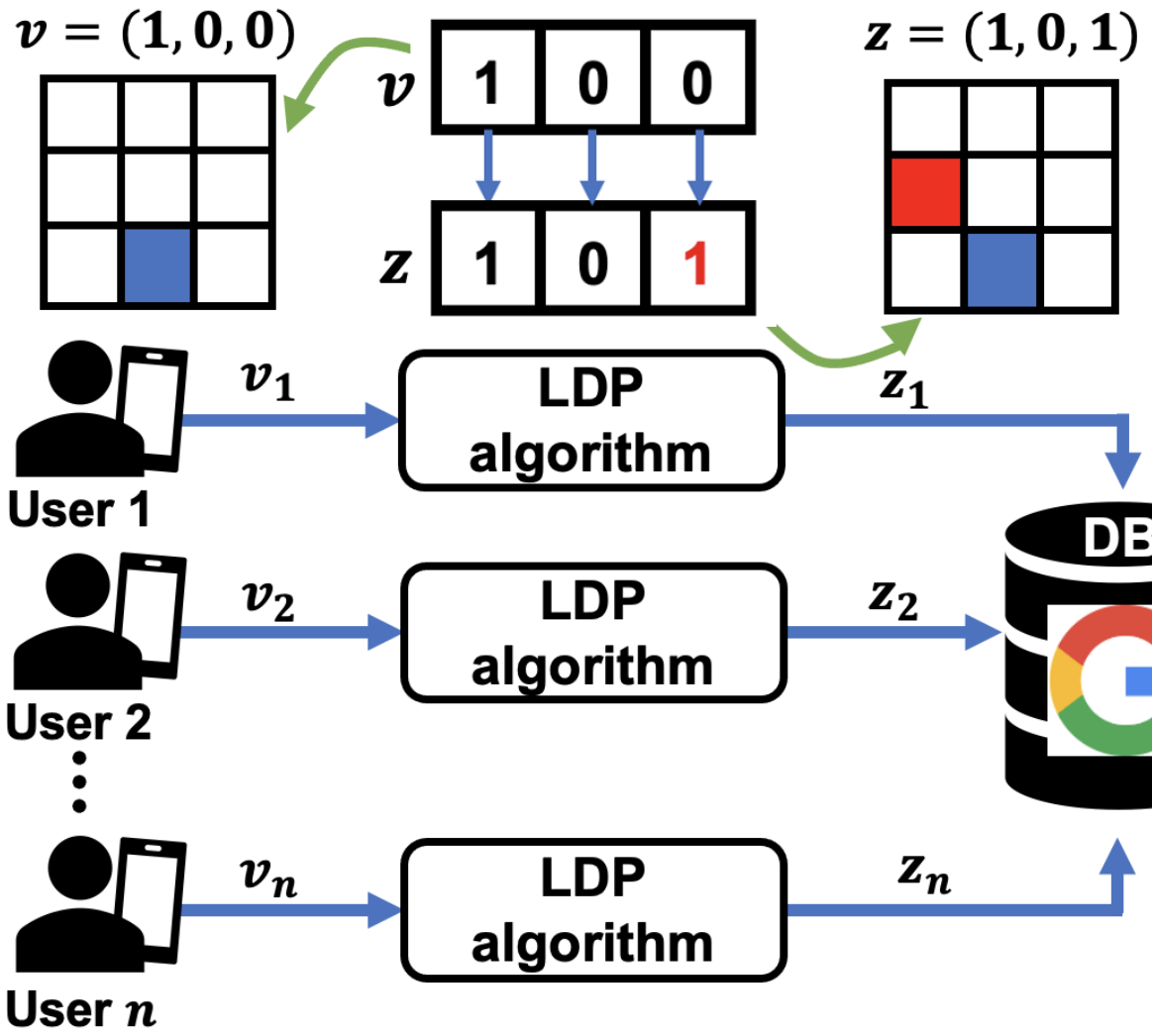
Hikaru Horigome and Hiroaki Kikuchi

Meiji University

Background :
Location based service and location privacy

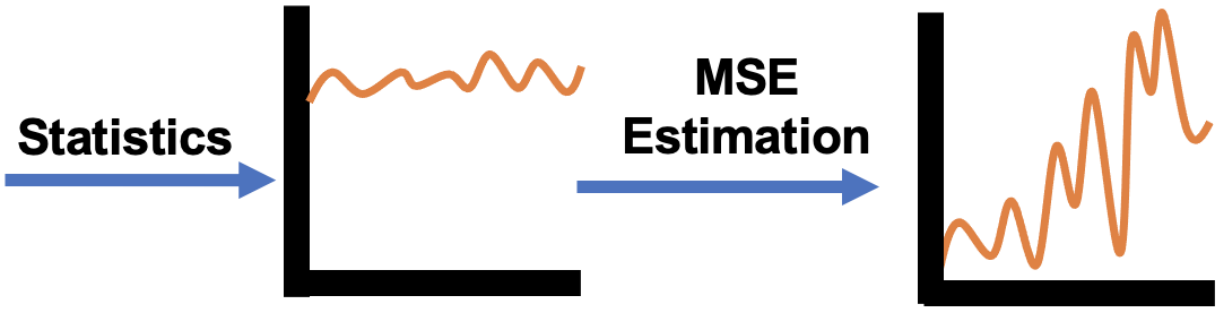


Local Differential Privacy (LDP)



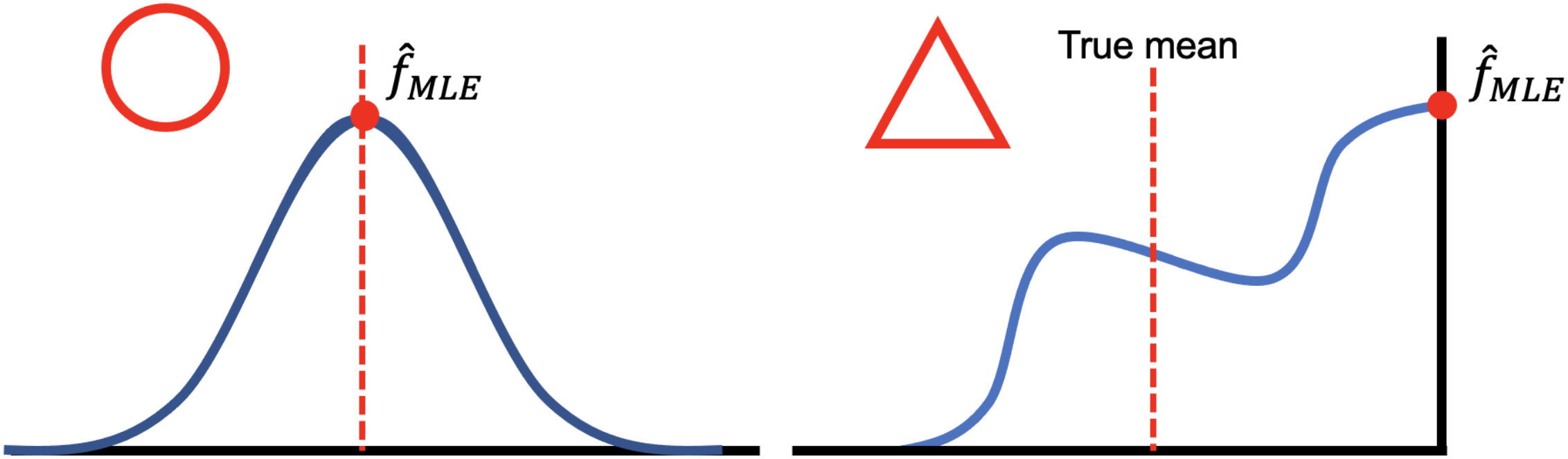
- **A randomized algorithm Q satisfies ε-LDP if any pair of inputs v, v' and any output z**

$$\frac{Pr[Q(v) = z]}{Pr[Q(v') = z]} \leq e^{\varepsilon}$$



Google : $RAPPOR$ [Erlingsson, etc., 2014]

Problems of Most Likelihood Estimation (MLE)



Our idea

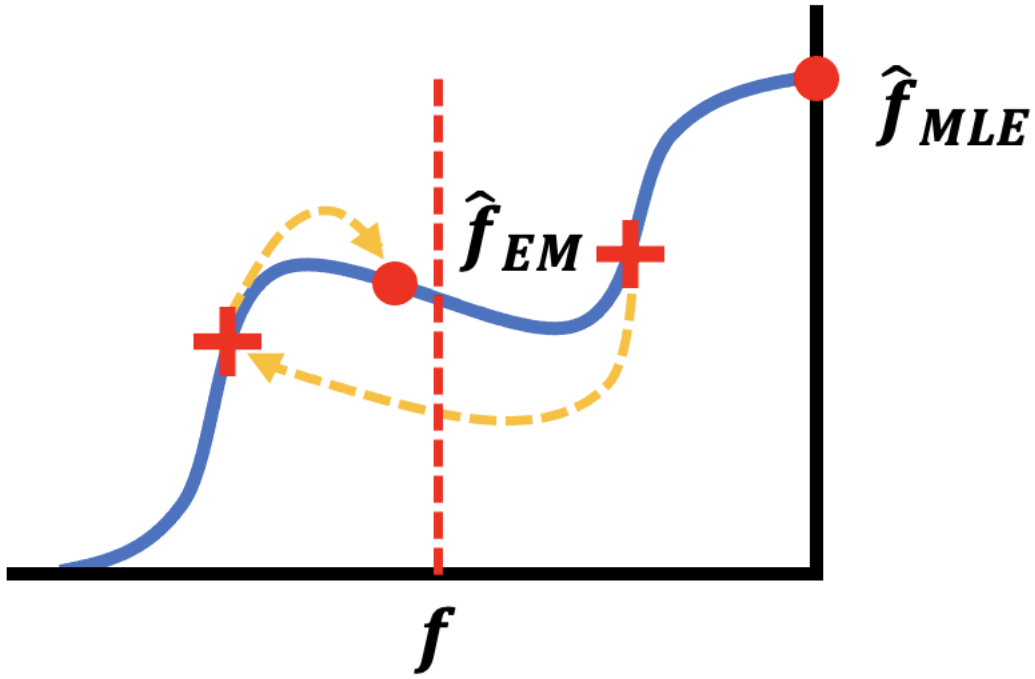
- **Our Approach**
  - **Expectaion Mazimization (EM) algorithm estimation**
- **Proposal**
  - **We apply the EM algorithm to the randomization used in RAPPOR**

	Randomization	Estimation
Google RAPPOR[1]	Bloom Filter + Randomized Response (RR)	MLE
Ours		EM

Why EM algorithm works better?

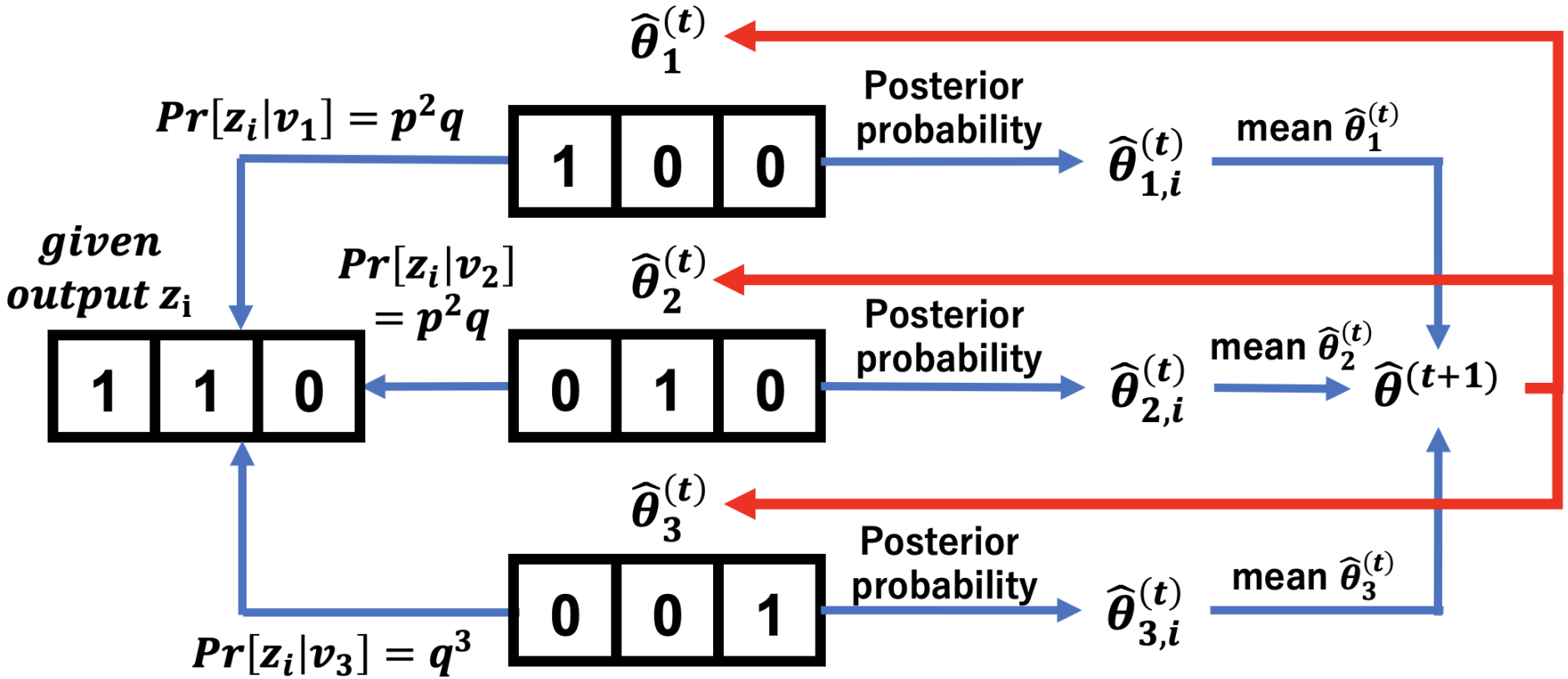
- **EM algorithm [Dempster, et al., 1997]**

  - Iterative process for which posterior probabilities are updated based on Bayes' theorem.



How to apply the EM algorithm to RAPPOR

marginal probability $\widehat{\boldsymbol{\theta}}^{(t)} = (\mathbf{Pr}[\mathbf{1},\mathbf{0},\mathbf{0}], \mathbf{Pr}[\mathbf{0},\mathbf{0},\mathbf{1}], \mathbf{Pr}[\mathbf{0},\mathbf{0},\mathbf{1}])$



Research Questions

- **RQ1.**
  - **Dose the proposed method estimate the frequency more accurately than MLE used in RAPPOR?**

- **RQ2.**
  - **Is there any correlation between privacy badget ε and the estimation accuracy?**

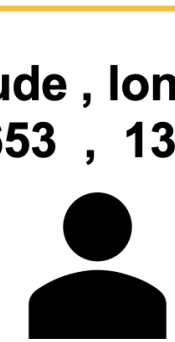
Experiment

- **Data : Nightley Dataset**

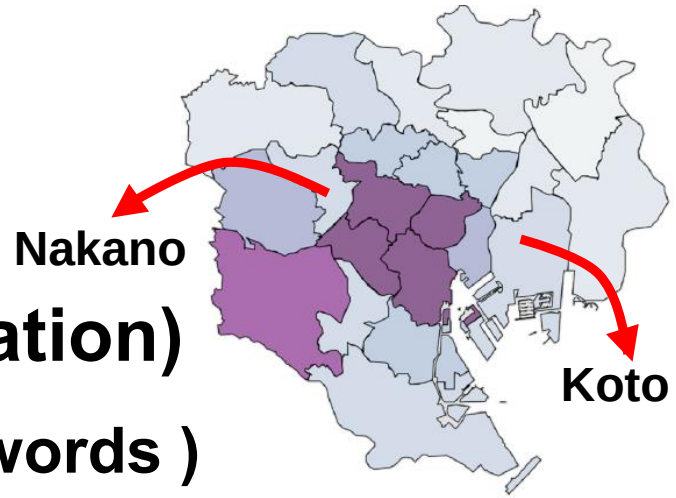
(6,258 users' location information)

input $v = \{0,1\}^{|23|}$ **(Tokyo has 23 words)**

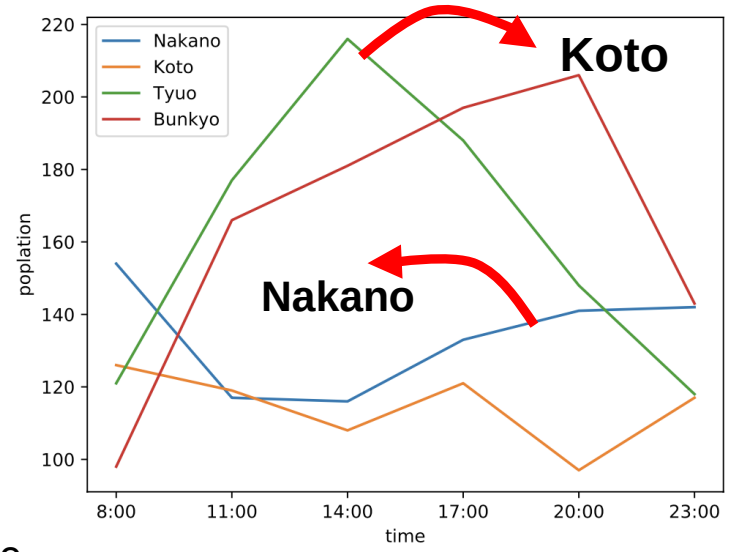
[latitude , longitude
35.653 , 139.711] → **" Nakano "**



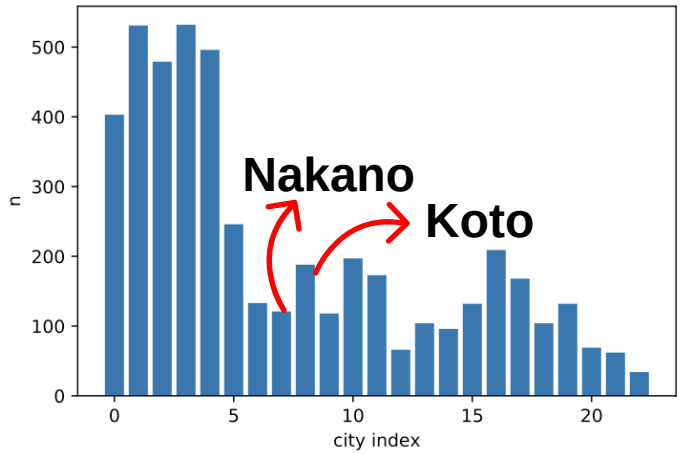
$$v = (v_{Setagaya}, v_{Nakano}, \dots, v_{Edogawa}) = (0, 1, 0, \dots, 0)$$



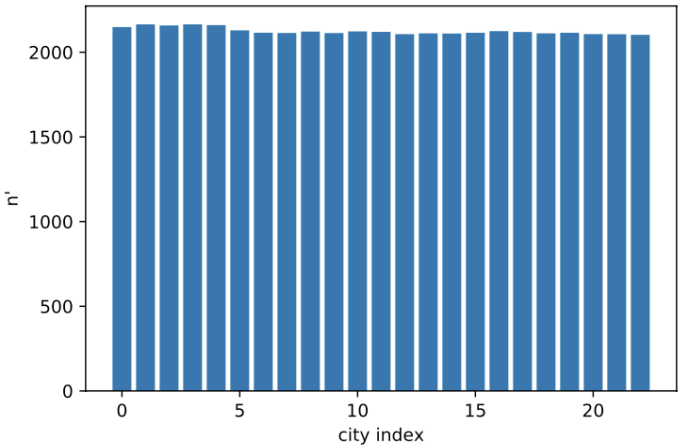
Heat map of Tokyo at 14:00



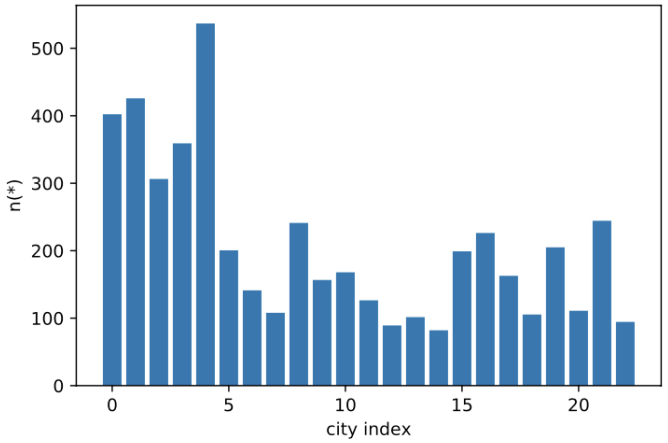
true



randomize

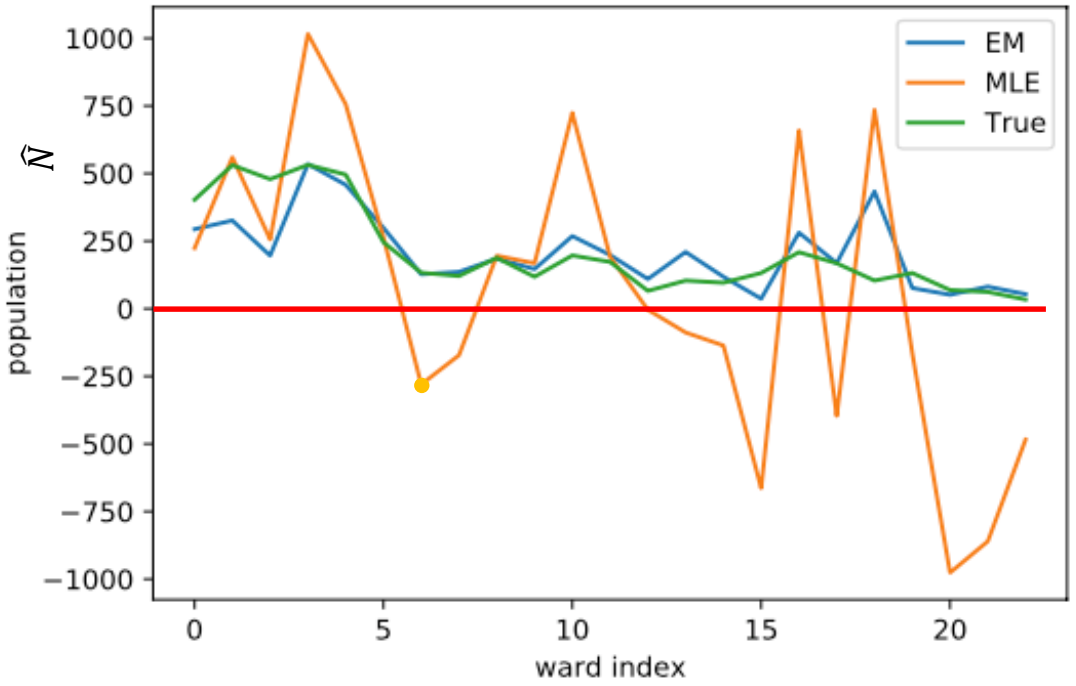


estimate

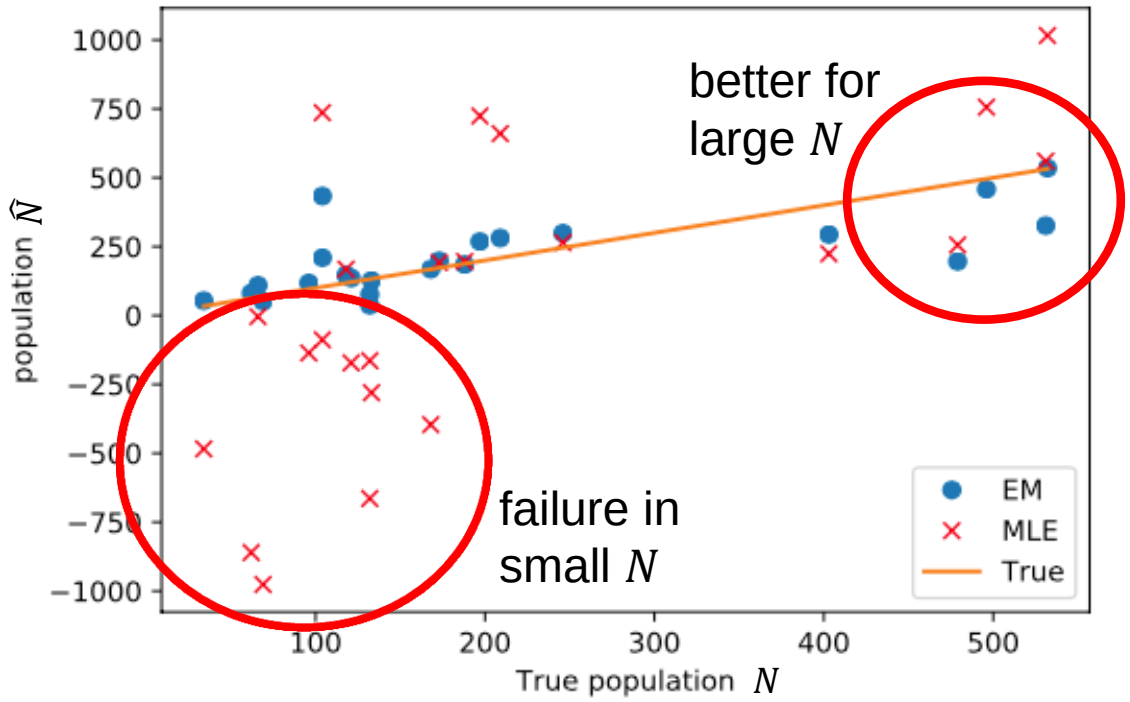


Result(1) : Estimation Distribution

Privacy budget $\varepsilon = 0.5$



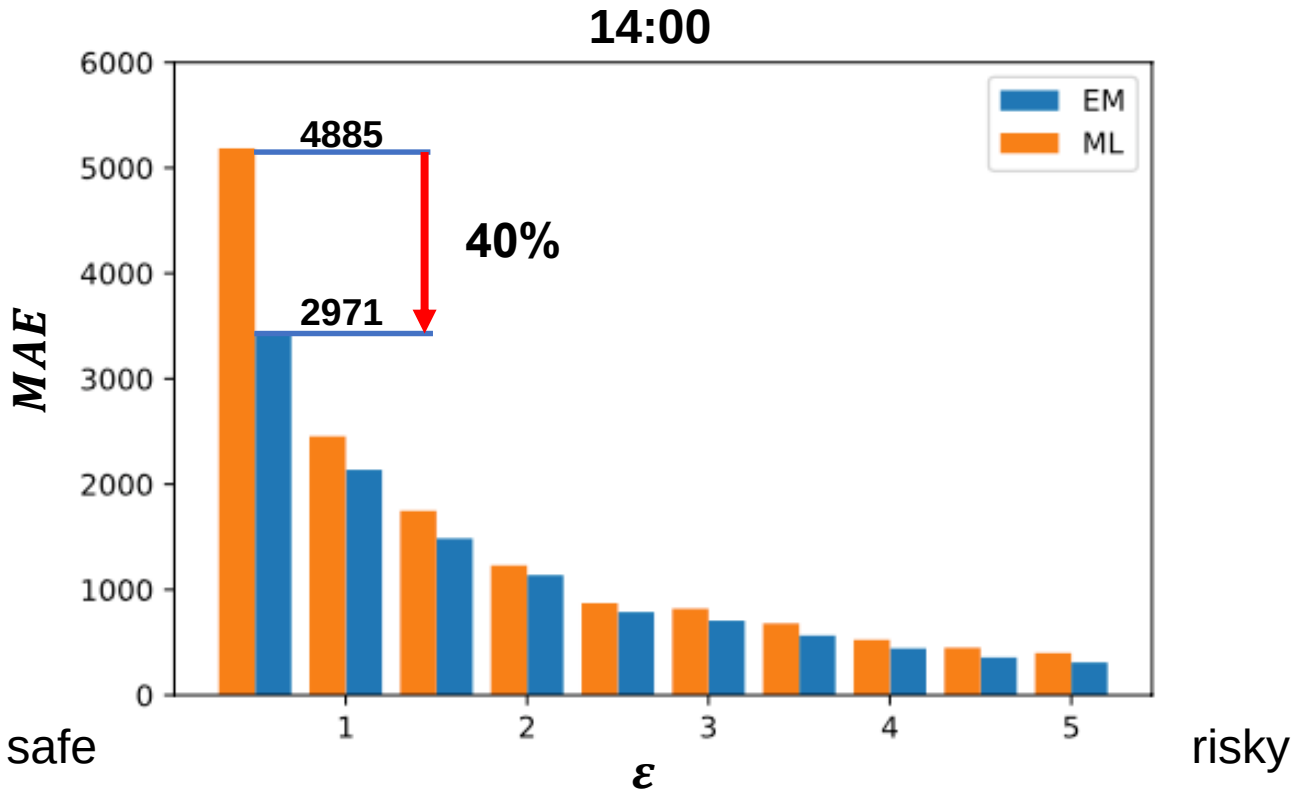
Estimated populations for 23 wards in Tokyo at 14:00



Scatter-plot between true and estimated populations (MLE, EM)

Result (2) : Mean Absolute Error (MAE)

RQ2. Is there any effect of safety parameter ε on the estimation accuracy?



Regardless of the value of ε, the proposed method estimated more accurately.

Conclusions

- **We studied the privacy preservation of time-series location traces using LDP algorithm RAPPOR and proposed the EM for estimating distributions.**

- **Our experiment using 6,528 individuals' location traces in Tokyo demonstrates that the proposed algorithm performs better than the MLE used in RAPPOR for any privacy budgets ε.**