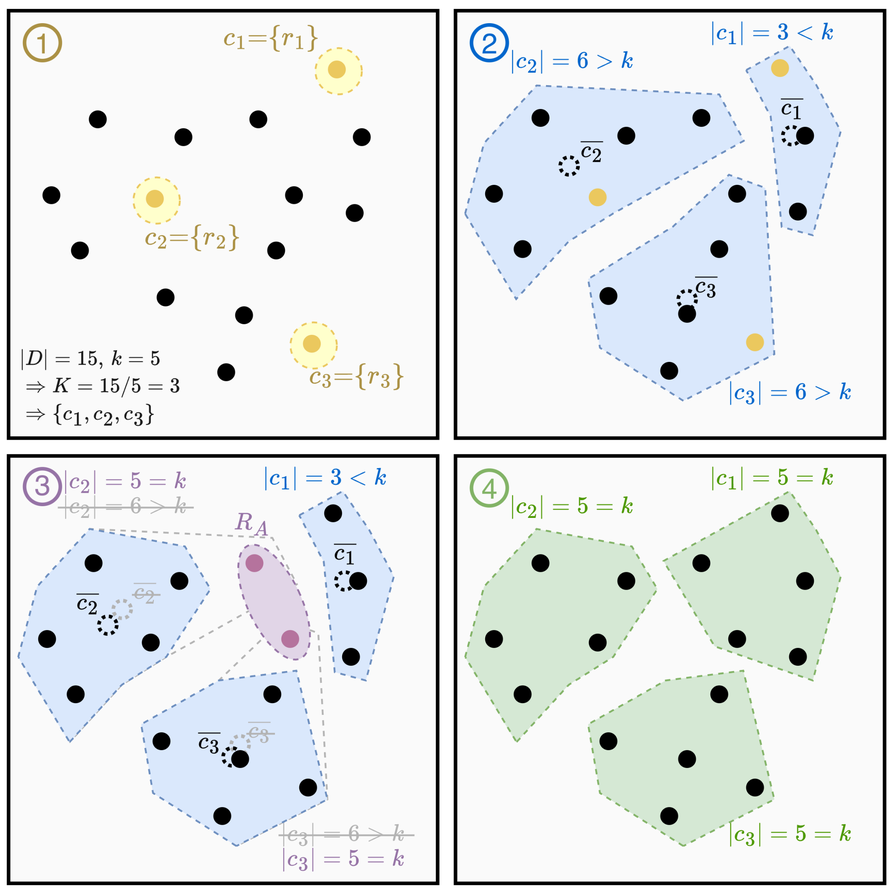
PWS Cup 2025
Team22: HAPPY

○菊池陽[1,2,3], Chanh Tran[1,2], 早川拓実[2,3], 杉山拓海[1,2,3], Wu Liujie[2,3], 南和宏[1,2]
1: データサイエンス共同利用基盤施設, 2:統計数理研究所, 3: 中央大学

1 Anonymization



⑤

	Gender	Age
c_i	M	10
	M	12
	M	11
	F	12
	M	13
$\overline{c_i}$	**M**	**11.6**
	Mode	Mean

$\Rightarrow$

	Gender	Age
c_i'	M	11.6
	M	11.6
	M	11.6
	M	11.6
	M	11.6

$D'=\{c_1',c_2',c_3'\}$

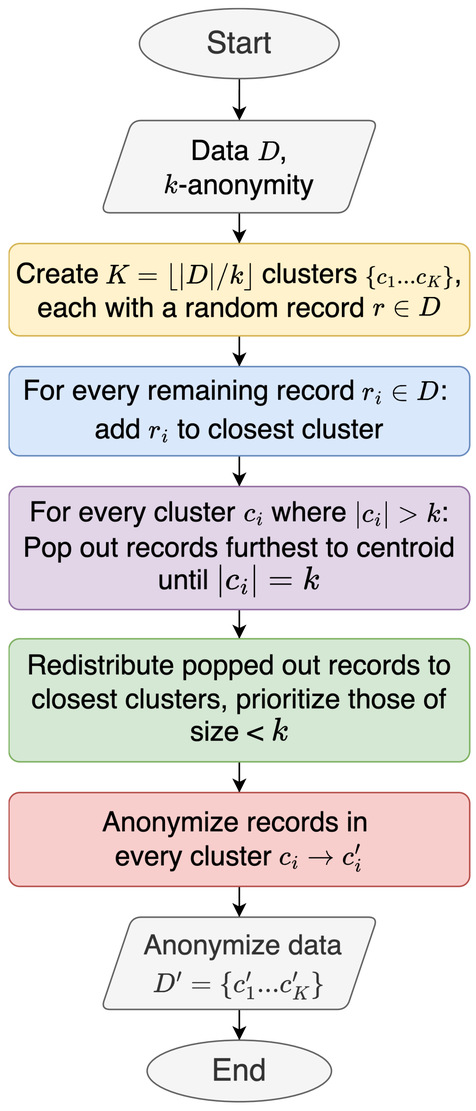
Distance between 2 records r_i and r_j

$$\Delta(r_i,r_j)=\sum_{x=1}^{n}\Delta_N^x(r_i,r_j)+\sum_{y=1}^{c}\Delta_C^y(r_i,r_j)$$

Distance of the x^{th} numerical attribute: $\Delta_N^x(r_i,r_j)=\frac{|N_x(r_i)-N_x(r_j)|}{\max(N_x)-\min(N_x)}$

Distance of the y^{th} categorical attribute: $\Delta_C^y(r_i,r_j)=\begin{cases}1, \text{if } C_y^i\neq C_y^j\\ 0, \text{if } C_y^i=C_y^j\end{cases}$

One-pass K-Means Clustering



2 Machine Learning

Our strategy:
- With enough data, empirical estimates approximate the true underlying distribution.

ML Model trained on all **A**(all data used in prep.), and **BB**(original data) achieved high balance of utility and anonymity.

A data from all 21 teams

GENDER	AGE	RACE	…	mean_weight
F	58	white	…	76.21
F	59	white	…	77.9
…	…	…	…	…
M	50	other	…	79.1

✚

BB data of our team

GENDER	AGE	RACE	…	mean_weight
F	58	white	…	76.21
F	59	white	…	77.9
…	…	…	…	…
M	50	other	…	79.1

ML model accuracy

Train/Valid	BB	A/BB mixed(50% each)
B	0.92	0.87
C	0.79	0.78
All_A	0.88	0.90
All_A+B	0.90	0.92

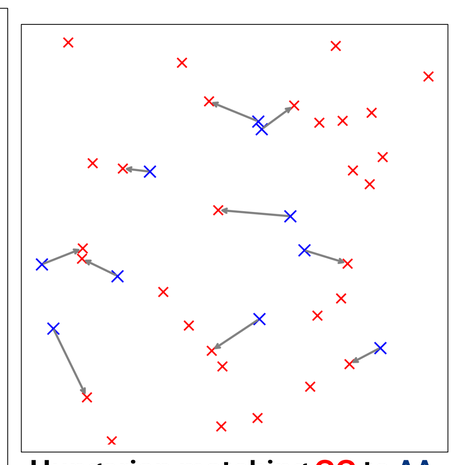
Method:
- **Training ML model(DD) with all A(all data in prep.) and BB(original data) mixed data.**

3 Attack Method

・Target of Attack Phase
→Find 10k records(original data: **BB**) from 100k records(all data: **AA**) with anonymized 10k records(anonymized data: **CC**).

Assumption 1:
- The **closest record** in **AA** from each record in **CC** is membership of **BB**.

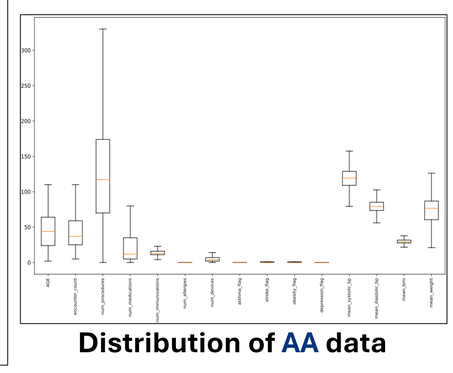
Following this assumption, we can infer all of **BB** records by 1 on 1, **CC** to **AA**, "Hungarian matching".



Hungarian matching CC to AA

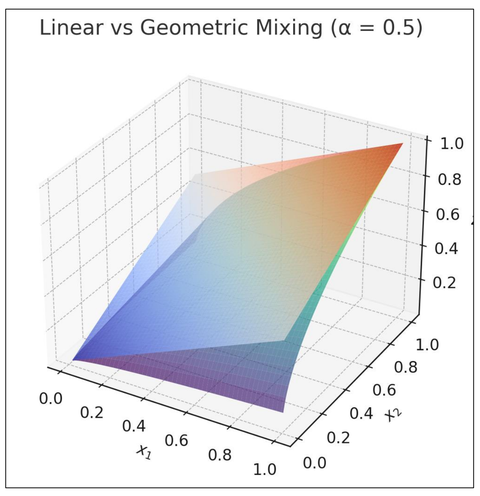
Assumption 2:
- Hard to anonymize variate attribute without utility loss.

E.g. Age, immunizations…



Distribution of AA data

Method:
- **Hungarian matching(CC to AA) with weight for each attributes by each entropy.**



Notes:
- Weights are applied by **"geometric"** and **"linear"** calculation to adapt each team's anonymization.

4 Analysis for Society

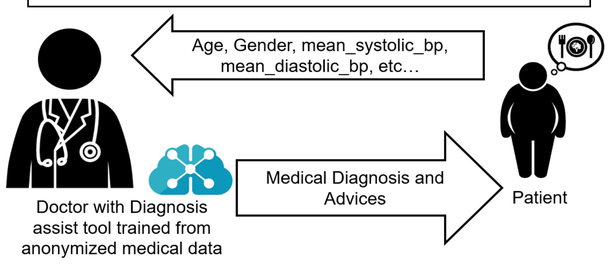
How can we use our anonymized data for society?

Any applications?

↓

Diagnosis Assist tool for Town (Local) Doctors. (Or May be any Doctors)

Patients may state incorrect or unreliable information to the local doctor (rather than their actual current health conditions).

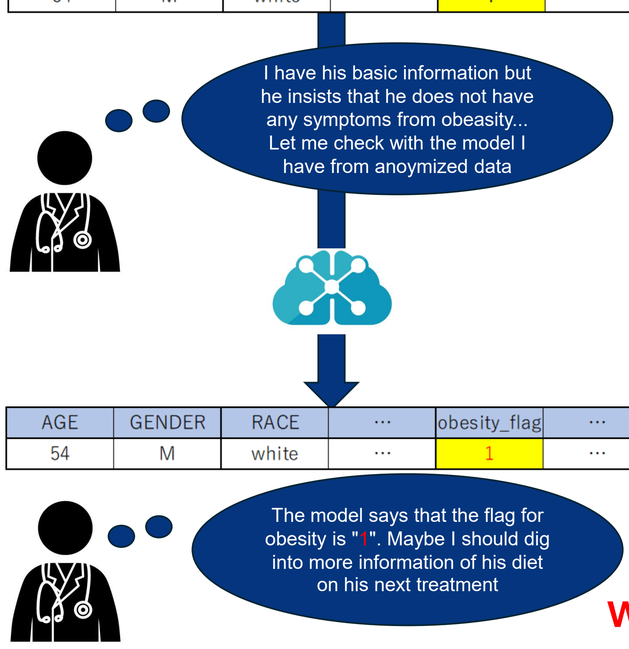


With the Assist tool ...

Town (Local) Doctors can make medical diagnosis using an trained model from anonymized medical data as a support.

Patient A (Example)

AGE	GENDER	RACE	…	obesity_flag	…
54	M	white	…	?	…



AGE	GENDER	RACE	…	obesity_flag	…
54	M	white	…	1	…

TEST

Target: NHANES
(Real U.S. /Obesity(BMI>30))

Method:
- Logistic Regression
- Random Forest Classifier
- XGBoost model

Model	Best Threshold	Accuracy	Precision	Recall	F1	ROC-AUC
Random Forest	0.38	0.856	0.649	**0.954 **	**0.773 **	0.9219
Logistic Regression	0.415 (best F1)	**0.903 **	0.8	0.827	**0.813 **	**0.9599 **
XGBoost	0.315 (best F1)	0.84	0.622	**0.960 **	0.755	0.915

Who Said Anonymized Data Can't Perform?