The 11th Industrial Problem Solving Workshop

2021-08-27

BANQUE NATIONALE

Réalisons vos idées[MC]

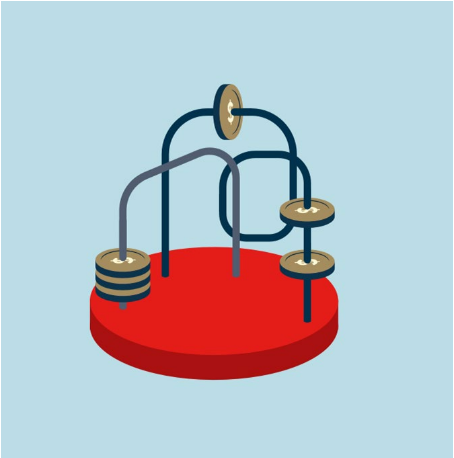
Workshop problematic

Data: Anonymized and raw transactional data *

- ~ 3M anonymized transactions made by 1581 hashed customers.
- ~ 300k raw transactions made by 200 unhashed customers.

Objective: Test the robustness of the anonymization method used

- Re-identify as many individuals as possible using both datasets.
- Rebuild original information from the anonymized dataset.



* the provided data is public data

Data provided

Transaction	Client	Date	Montant	Type	Nom du marchand	Ville du marchand	Ã‰tat du marchand	Code postal	CatÃ©gorie du marchand
0	55fc43d6ab4bb15d7e3e	2018-11-19 11:26:00 - 2018-11-19 11:28:00	18.67	Chip Transaction	-2.32265E+18	Aynor	CA	29511	5912
1	95770a8fb5f3807b5b3c	2019-03-12 09:19:00 - 2019-03-12 09:21:00	90.15285714	Chip Transaction	-5.46792E+18	Brewster	CA	2631	4121
2	8162094451ffe70b8f5e	2018-09-14 12:57:00 - 2018-09-14 12:58:00	24.36625	Chip Transaction	-5.47568E+18	Farmington	CA	4938	5814
3	45880637ee4679a7d7b4	2019-10-01 21:44:00 - 2019-10-01 21:45:00	47.17	Chip Transaction	-4.28247E+18	Birmingham	AL	1527	5812
4	3764c92cf296743fa504	2019-11-17 23:01:00 - 2019-11-17 23:20:00	43.01875	Chip Transaction	-4.76476E+18	Anchorage	CA	8005	5812
5	45eeb24676b688a74e92	2018-01-08 06:02:00 - 2018-01-08 06:03:00	54.51142857	Chip Transaction	1.79919E+18	West Chester	PA	19380	5499
6	2d49d7347b95f6f03f36	2019-03-30 06:44:00 - 2019-03-30 06:47:00	5.893333333	Chip Transaction	-8.42808E+18	Brooklyn	FL	8318	5411
7	3e7d6464ed0452ddf1d8	2019-07-10 16:49:00 - 2019-07-10 16:52:00	74.595	Chip Transaction	-5.16204E+18	Orlando	FL	32825	5541
8	a980b40bf512c6cd8ef5	2019-08-07 07:32:00 - 2019-08-07 07:33:00	36.98571429	Chip Transaction	-8.37441E+18	Burley	ID	13850	5541
9	e970c5f2d190582c823b	2018-05-23 13:17:00 - 2018-05-23 13:17:00	33.08428571	Chip Transaction	-5.7089E+18	Arlington	FL	13126	5411
10	59f5b358c764426ea98c	2019-03-13 10:04:00 - 2019-03-13 10:05:00	68.14625	Chip Transaction	-5.90412E+18	Clifton Springs	IA	14432	4829
11	a1f044[illegible]51ed5e88e3e15	2019-08-20 12:41:00 - 2019-08-20 12:42:00	71.87625	Chip Transaction	2.02755E+18	Atlanta	FL	29203	5541
12	34f4ac[illegible]e668c8368331a	2019-02-01 16:40:00 - 2019-02-01 16:46:00	244.53	Chip Transaction	-5.46792E+18	East Northport	NY	5146	4829
13	f4a217[illegible]0337bbe4a0e94	2018-12-04 12:33:00 - 2018-12-04 12:36:00	6.788	Chip Transaction	-9.19875E+18	Columbus	IL	16801	5411

Find the mapping between the anonymized transactions and the 200 unhashed client IDs

Transacti	Clie	Date	Monta	Type	Nom du marcha	Ville du marchand	Ã‰tat du marcha	Code pos	CatÃ©gorie du marcha
11340358	942	2018-11-29 06:06	99	Chip Transaction	1.79919E+18	Morganton	NC	28655	5499
11340609	942	2019-03-03 20:39	83.1	Chip Transaction	-2.45178E+17	Morganton	NC	28655	5311
11341273	942	2019-11-21 20:39	93.24	Chip Transaction	-5.46792E+18	Morganton	NC	28655	5912
11340416	942	2018-12-26 12:35	5.74	Chip Transaction	6.09178E+18	Morganton	NC	28655	5411
11339635	942	2018-02-21 22:51	41.38	Chip Transaction	6.96863E+18	North Wilkesboro	NC	28659	7832
11339654	942	2018-02-26 13:13	149.81	Chip Transaction	6.21387E+18	Morganton	NC	28655	6300
11339904	942	2018-06-09 12:38	7.58	Chip Transaction	6.09178E+18	Morganton	NC	28655	5411
11340832	942	2019-05-25 20:38	72.63	Chip Transaction	6.09178E+18	Morganton	NC	28655	5411
11339825	942	2018-05-07 22:17	37.84	Chip Transaction	-3.26567E+18	Arden	NC	28704	7832

mission:
impossible?

CHALLENGE
ACCEPTED

The Team

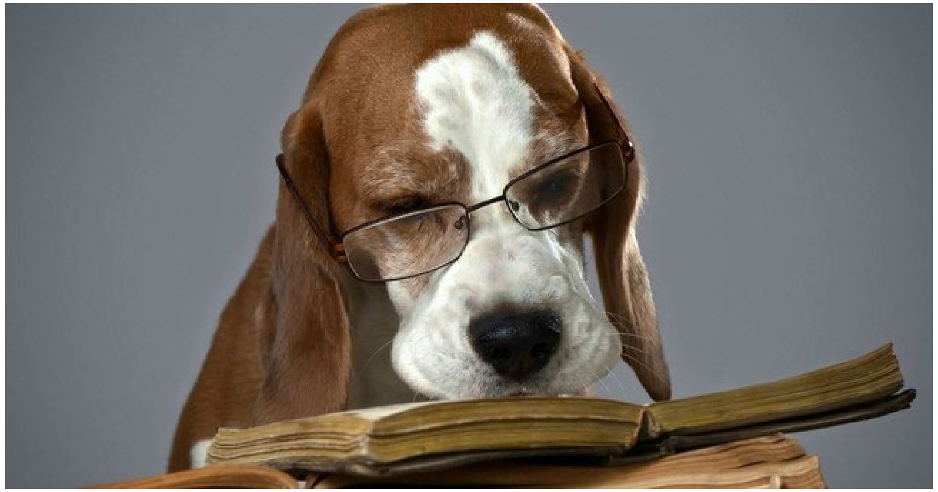
Expertises in:

- Generative Modelling, Computer vision, Quantum Computing
- Data visualization, storytelling.
- Synthetic Data, Data analysis, D.P
- Data generation, Privacy
- Computational Statistics, Data Mining, Applied probability
- Prototyping (programming), seq2seq models, self-supervised learning

Approaches

Supervised Classification using fine-to-coarse feature mapping functions

Stéphane Gazaille & Simon Kassab



Nearest Neighbor

Distance based Method

Elnaz Karimian Sichani

Linear regression model and Euclidean distance method

Leila Vanessa Nombo

Main challenges with the data

Discrepancies between anonymized and deanonymized data spaces

1. Features are sometimes represented differently between the two datasets.

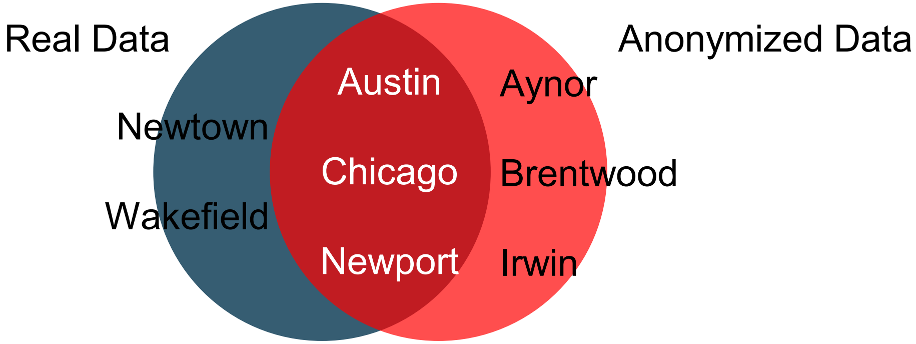
K-Anonymization often requires attributes to be aggregated.

Ex: Dates

- Real Data: 2018-11-29 06:06:00
- Anonymized Data: [2018-11-29 11:26:00, 2018-11-29 11:28:00]

1. Feature domains are different between the two datasets.

Ex: Cities



To classify the anonymized data using a model trained on the deanonymized data, both datasets need to be aligned in format and domain.

Data Preprocessing

Mapping features to a common space (Fine-to-Coarse)

Dates

- Real dates were transformed into Unix time format.
- Anonymized dates were summarized into their average values, then converted to Unix format.

Transaction Amounts

- Floating point amounts were categorized into 8 intervals.
- The final values are ordinal encodings of the associated interval.

Zip Codes

- During the k-anonymization process, Zip Codes are typically altered by sometimes removing the last digit.
- We kept only the first 3 digits and encoded them.

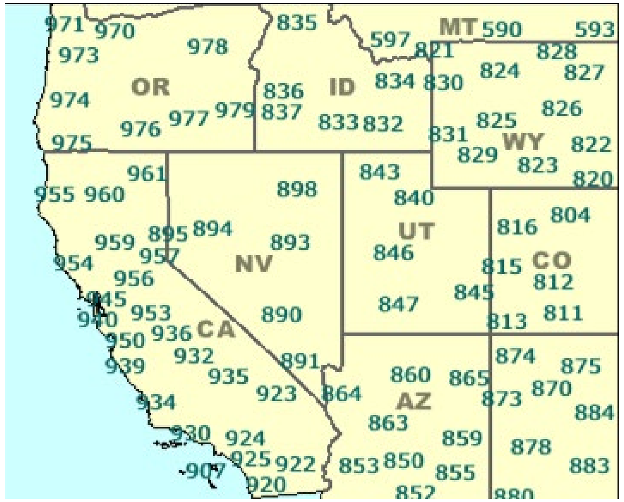
Transaction Type, Merchant Name, Category, State and City

- We were unable to design *fine-to-coarse* mapping functions for these attributes. They were simply encoded.

Transaction Amounts Ex:

Code	Value
0	0$ to 20$
1	20$ to 100$
2	>100$

Information held by first 3 Zip code digits



Classification Model Training

Choosing the model and selecting features

Models

- A validation set was extracted from the auxiliary data.
- Various classification models were evaluated using that validation set.
  - Decision Tree, Extra Tree, K Nearest Neighbors, Nearest Centroid, Logistic Regression, Ridge Classifier

Feature Selection

- We prioritized features for which we developed a *fine-to-coarse* function. (Dates, Amounts, Zip codes).
  - Strong features for which we didn't have a *fine-to-coarse* function were not used.
- SKLearn's feature ranking tool (RFE) was used to identify the strongest features.
- 3 different sets of features were defined.
  1. Dates, Amounts, Zip Codes
  2. Dates, Amounts, Zip Codes, Merchant Category
  3. Dates, Amounts, Zip Codes, Merchant Category, Merchant State

Results

Predicting transaction Clients and building the Re-Identification Hash Map

Building the Re-Identification Hash Map

Anonymized Data

Model

Transaction No.	Client Hash	Features
0	Hash-0	[...]
1	Hash-1	[...]
2	Hash-2	[...]

Model Predictions

Transaction No.	Client ID Predictions
0	ID-0
1	ID-1
2	ID-0

Final Re-Identification Hash Map

Client ID	Most often mapped Client Hash
ID-0	Hash-0
ID-1	Hash-1
ID-2	Hash-4

Results

Feature Sets	Re-Identification Rate (%)	Successfully re-identified clients
Dates, Amounts and Zip Codes	25.50	51
Dates, Amounts, Zip Codes and Merchant Category	22.50	45
Dates, Amounts, Zip Codes, Merchant Category and Merchant State	20.00	40

Conclusion & Future Work

How could we improve from here?

Conclusion

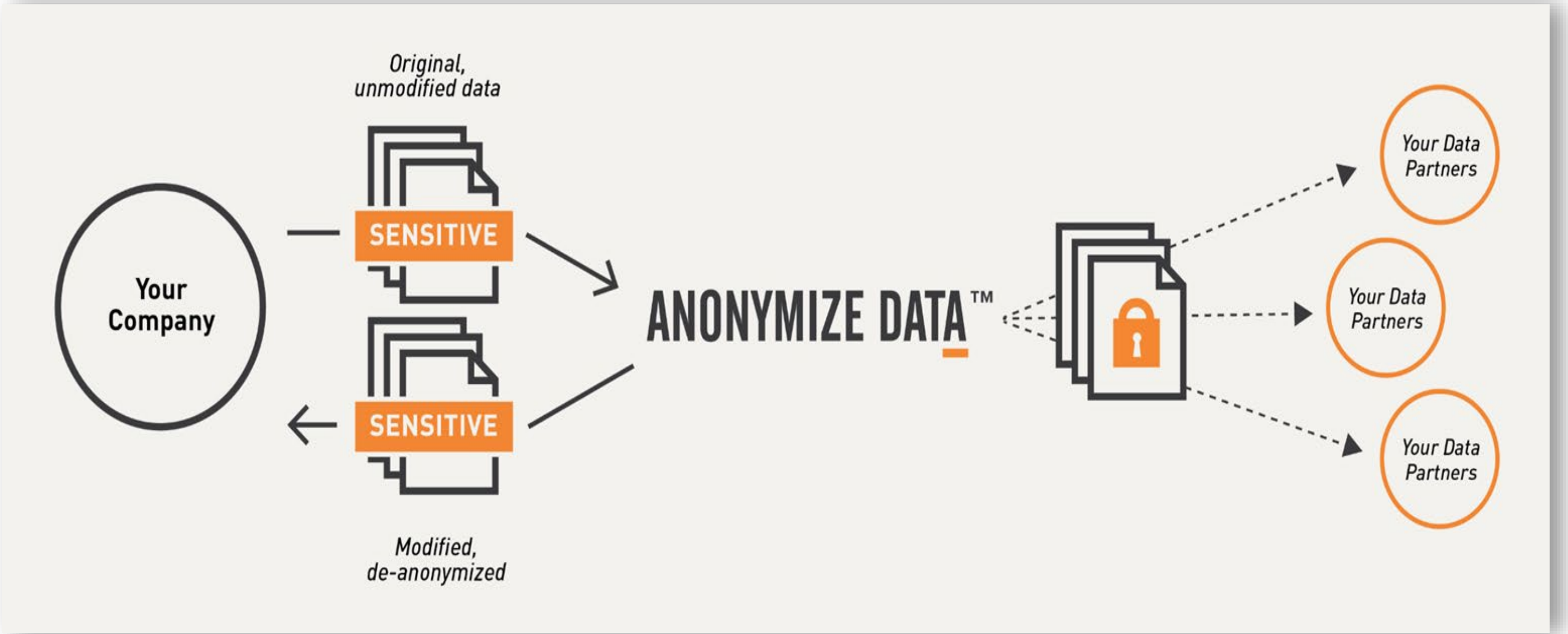
- Results indicate that *fine-to-coarse* data transformation is useful.
- Selecting features that haven't been transformed using a *fine-to-coarse* function significantly decreases the performance.

Future Work

- Improve the re-identification hash map logic of construction.
  - Eliminate duplicate hash values.
  - Make sure all client IDs are mapped to a hash.
- Experiment with clustering algorithms (eg kNN) and their distance functions.
- Improve the preprocessing methods.
  - Test different aggregation methods by trying to reverse engineer k-anonymity data alterations.
- Normalize numerical values.

The goal

Develop and test a methodology for assessing the identity disclosure risks of anonymized data.



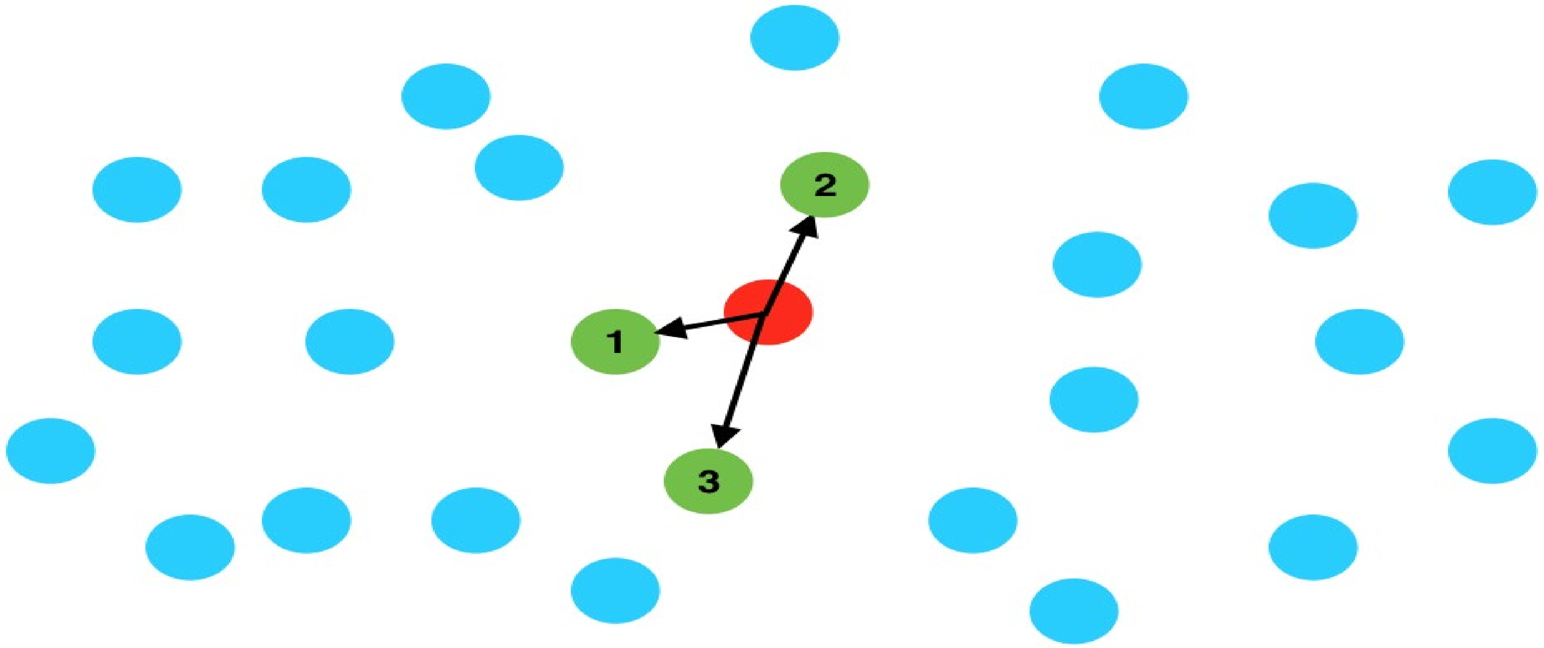
What does it mean for anonymized data to be "identifiable"?

How do we know when they are no longer identifiable?

The Method

Distance based Record Linkage

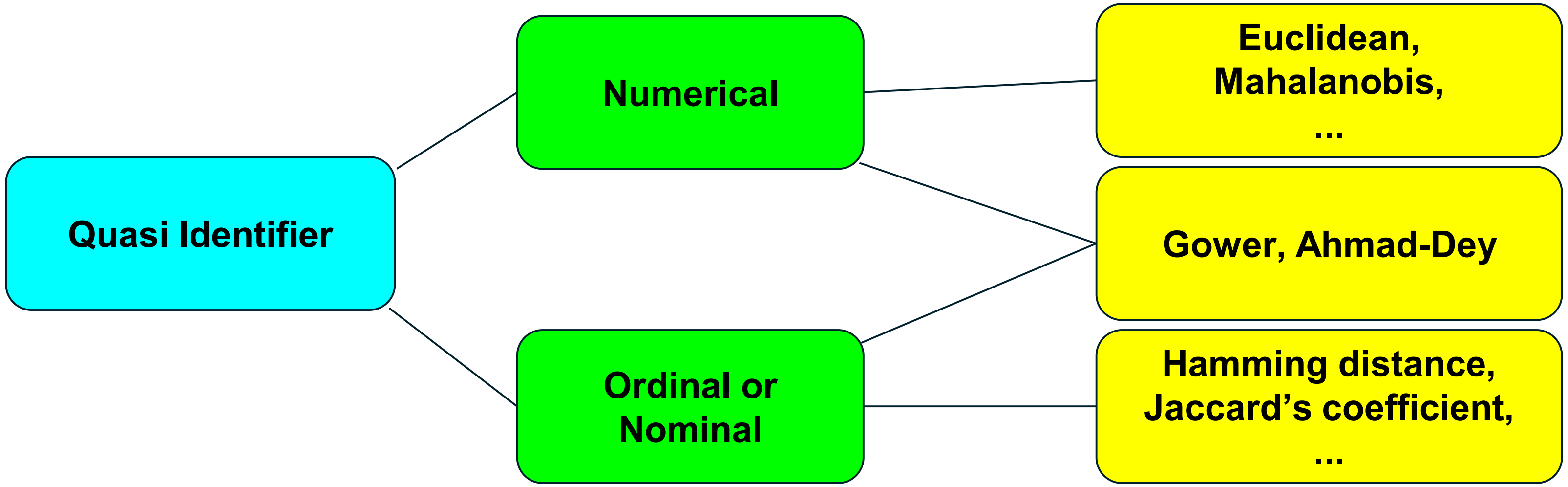
(Nearest Neighbor)



The attacker employs record similarity to discover a probable connection for a particular record in the original data and then infers the sensitive feature from this.

- *For a given record in the original data, find the nearest neighbor from the anonymized data, which is obtained by the distance/dissimilarity of that given record to all the records in the anonymized data.*

- *There are several similarity metrics that can be used to find the distance/dissimilarity between records. The selection of these metrics is determined based on the type of quasi_identifires.*



The Result

Attributes Sets	Re-Identification Rate (%)	Successfully re-identified clients
All	30.5	61

- So far, the distance/dissimilarity based re-identification methods appear to be a highly promising approach for evaluating re-identification and privacy disclosure.

- As a per the above result, adopting a more flexible and appropriate distance metric will almost definitely enhance the accuracy of the algorithm.

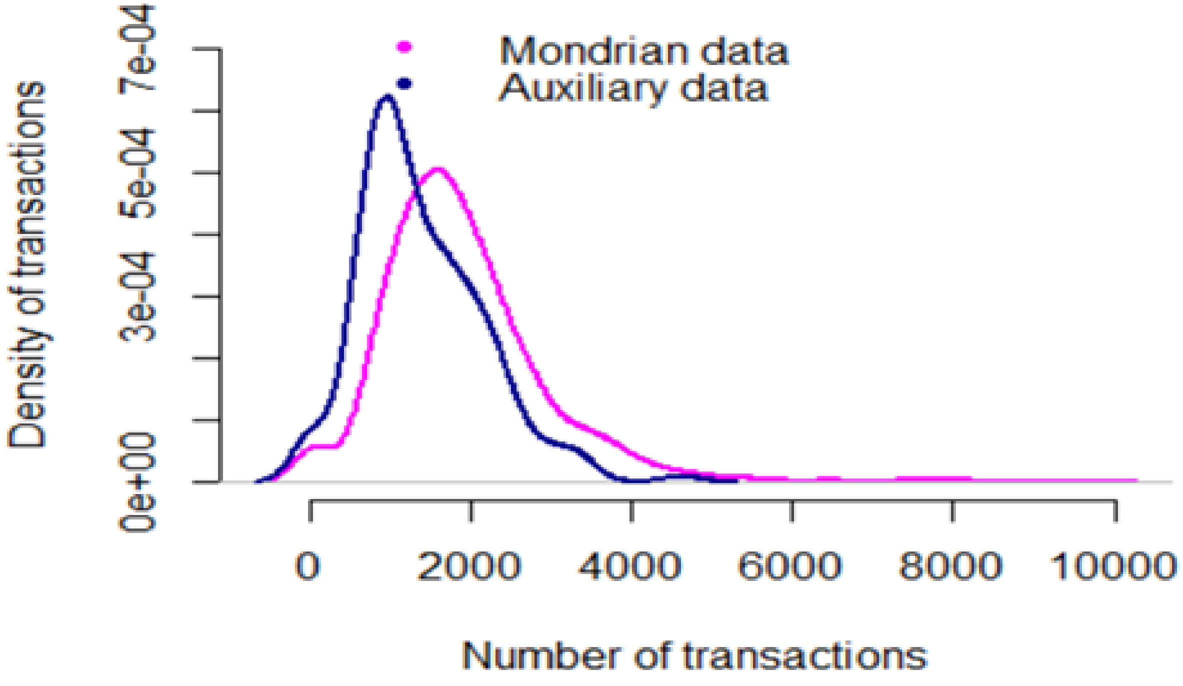
Future Work

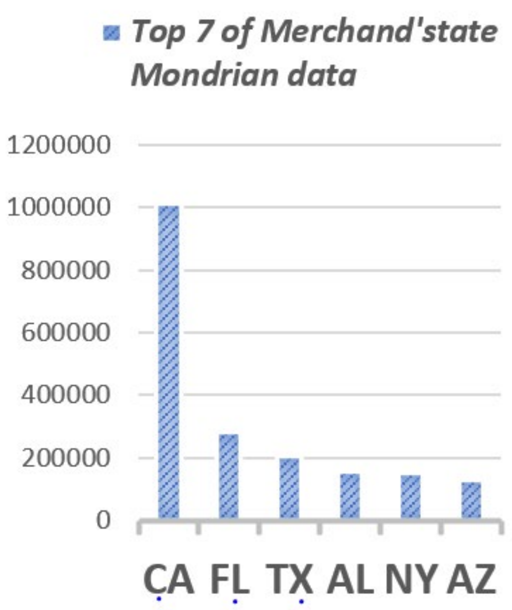
- Improving the method with a distance/dissimilarity metric that can handle a variety of data types and enhance the result.

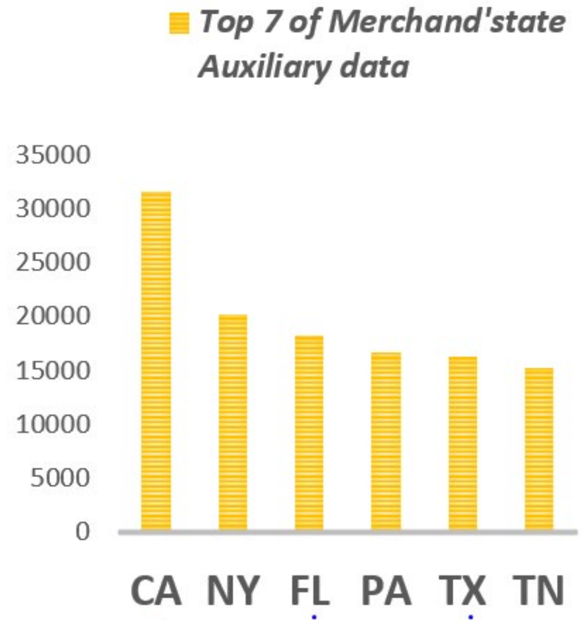
- Developing various post-processing approaches to improve the accuracy.

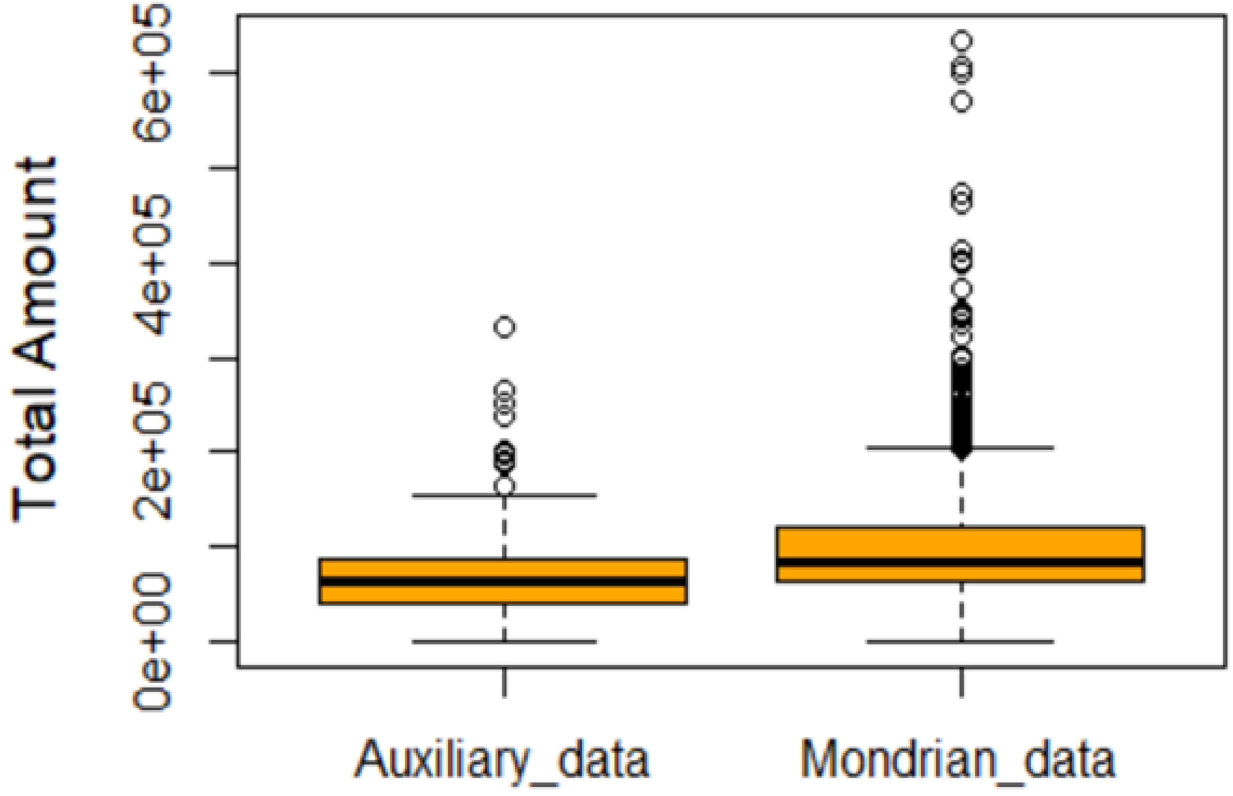
- Furthermore, we may train a ML classifier to anticipate sensitive data based on the anonymised data's quasi-identifier.

First results - Datas’ Exploration

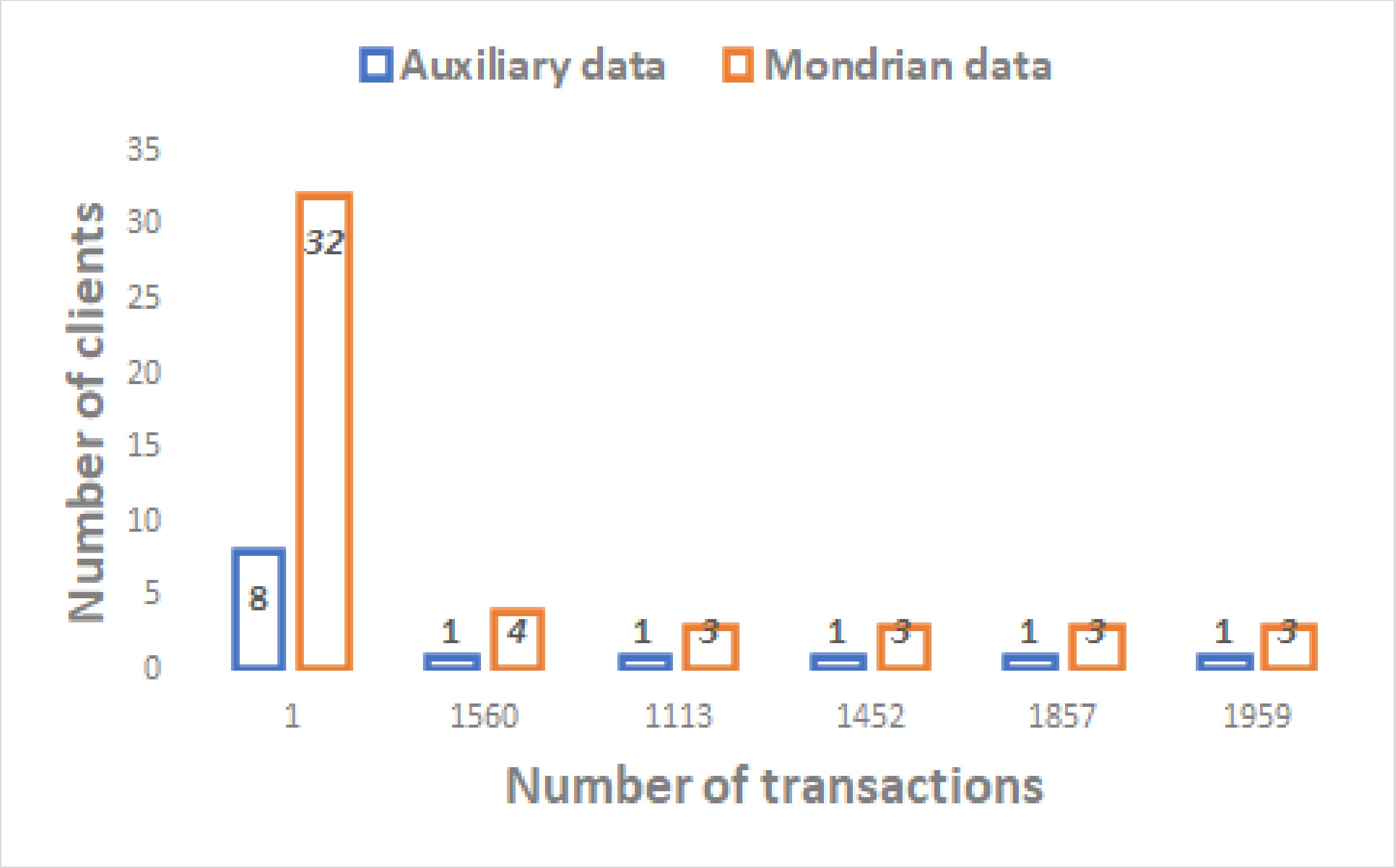








Method1- Results of Datas'Exploration



Number of transactions	Auxiliary data	Mondrian data
752	1	1
785	1	1
790	1	1
807	1	1
811	1	1
826	1	1
850	1	1
899	1	1
916	1	1
947	1	1
968	1	1
978	1	1
981	1	1

- **There is 53 <u>one to one matchings</u> using the number of transactions variable in the two datasets.**

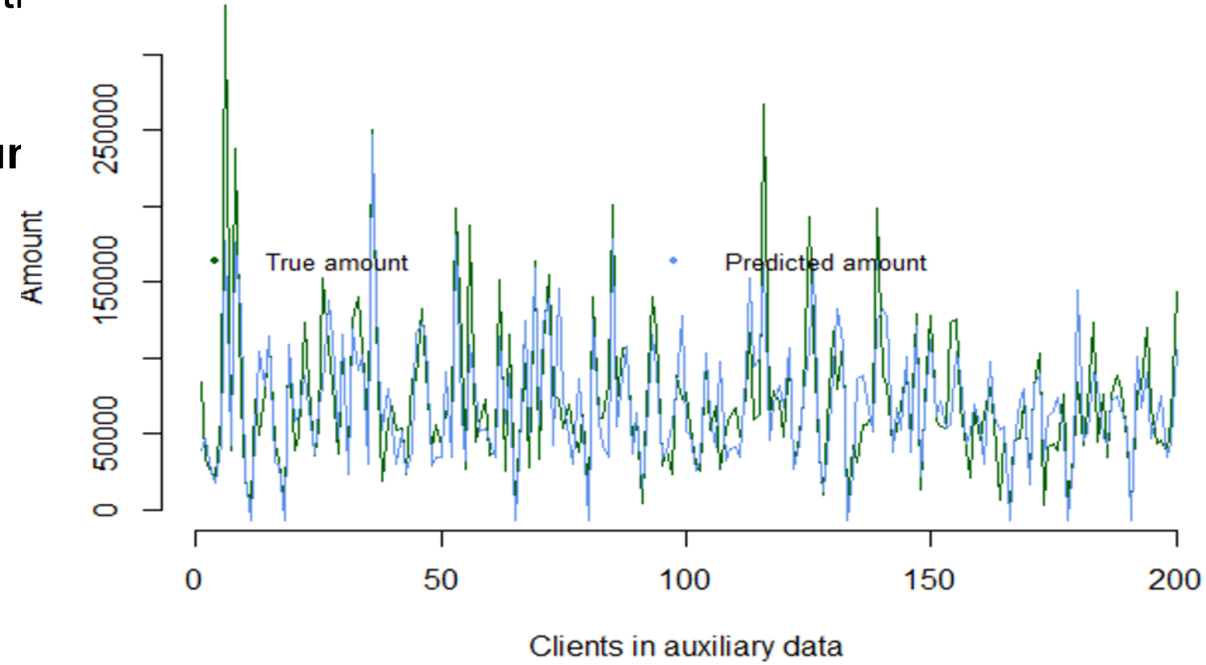
Method2 - Prediction using a linear regression model

1. Transform the two datasets:

- For each client: - the total amount for all transactions
  - The number of transactions
  - The number of type of transactions
  - The number of state where stay their merchants
  - The number of towns where he makes transactions

2-3. Use the Mondrian transform data to model the amour the amount in the auxiliary data.

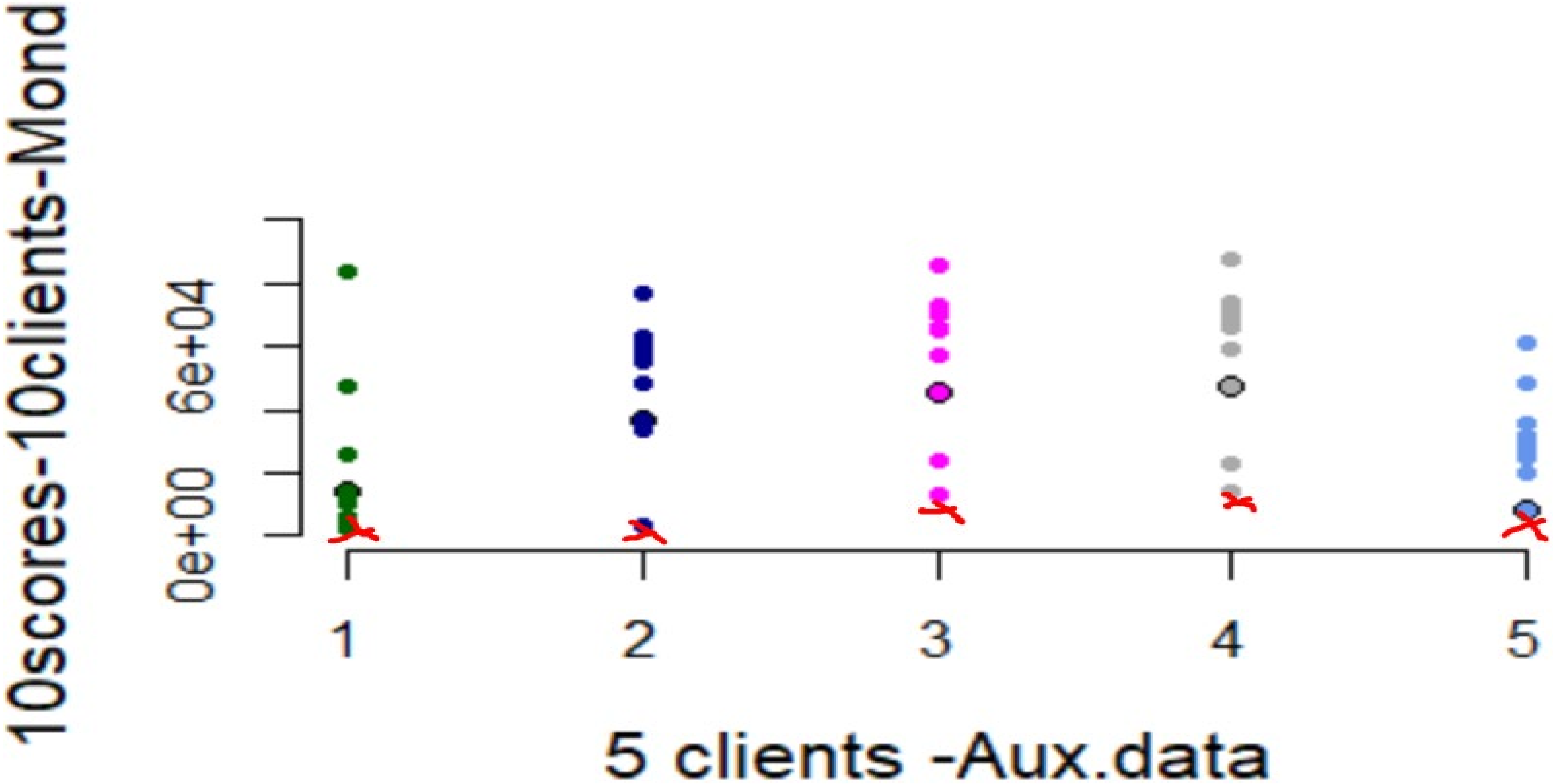
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6255.8857  1367.1451  -4.576 5.11e-06 ***
nbtrans        54.9129     0.6379  86.078  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 25580 on 1579 degrees of freedom
Multiple R-squared:  0.8243,    Adjusted R-squared:  0.8242
F-statistic:  7409 on 1 and 1579 DF,  p-value: < 2.2e-16
```



Method3 - Use of Euclidean distance

1- Calculate the euclidean distance between each record in the auxiliary data and the mondrian data.

2- Choose the ClientID with the minimal distance in the mondrian data



Future work

1- Consider the other variables (date, postal code, …)

2- Test with another distance metric, another model

Conclusion

- ***There is a significant growth in anonymization and data synthesis to enable data sharing for secondary analysis.***

- ***But it might be feasible to re-identify individuals and learn something new about them using the anonymized data.***

- ***Therefore, there is a tremendous need for an assessment privacy disclosure so that we can confidently disclose the data.***

Questions

This message will self destruct in 5 seconds