K-anonymity : The devil's advocate

Lilit Yenokyan
Patricio Massaro

Introduction

With increasing amount of *public data*, increases the need for *data privacy*!

Anonymization

K-Anonymization: general idea

- **Direct identifier** attributes
- **Quasi–identifier** attributes

- **Generalization**
  - Generalization hierarchies
- **Suppression**

Trade-off

DATA PRIVACY

DATA SCIENCE

Anonymization is primarily a **privacy** concept → it **always** generates information loss.

- How much anonymization is too much?
- How does the anonymization affect the performance of the ML algorithms?
- Can we optimize datasets for model performance while keeping privacy?

Objective

Analyze the impact of anonymization in performance of ML tasks.

- ## Problem

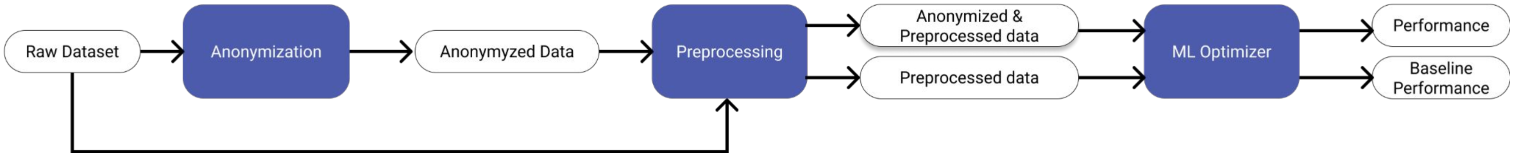
  Previous studies are **limited** in data science aspects:

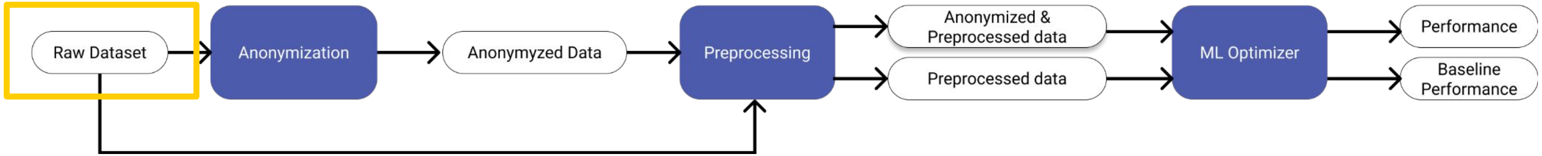
  - Evaluate one dataset, model trained with default hyperparameters!
  - Evaluation conditions are not fair!

Objective

Analyze the impact of anonymization in performance of ML tasks.

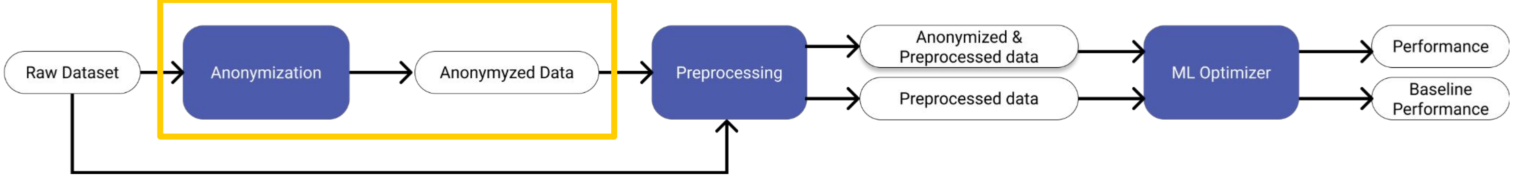
- ## Solution
  - **Three datasets** with diversity of domain and tasks
  - **Two ML** algorithms
  - Search for the **best hyperparams** for each ***K***, using Bayesian optimization
  - **Lots** of models!





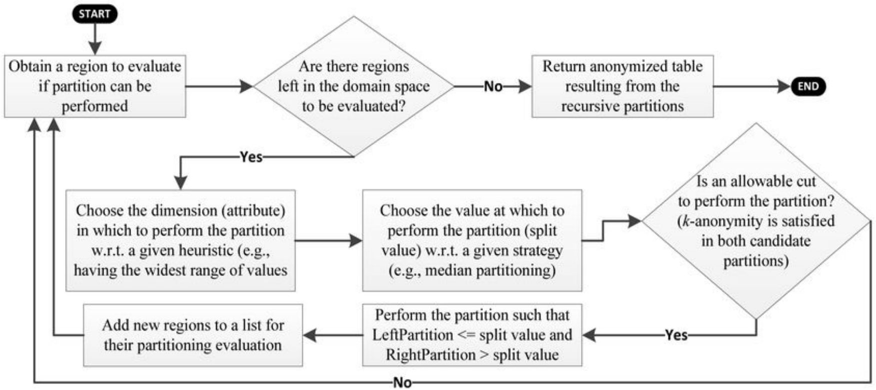
Datasets

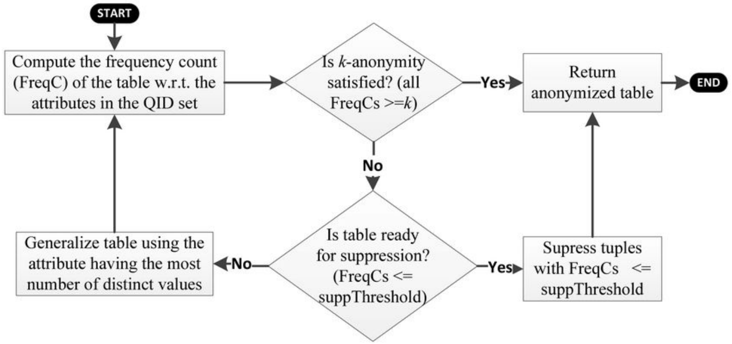
- **ADULT**
  - target variable: income (1= “<= 50K”, 2=” 50K”)
  - QIDs: “age”, “education”, “marital status”, “native country”, “occupation”, “race”, “sex”, and “work class”
  - Size: 45K
- **California Housing**
  - target variable: median house value (continues values)
  - QIDs: “latitude”, “longitude”, “housing median age”, “median income”
  - Size: 20K
- **Contraceptive Methods Choice**
  - target variable: choice of contraceptive method (1=None, 2=Short-term, 3=Long-term)
  - QIDs: “education”, “age” and the “number of children ever born”
  - Size: 1473

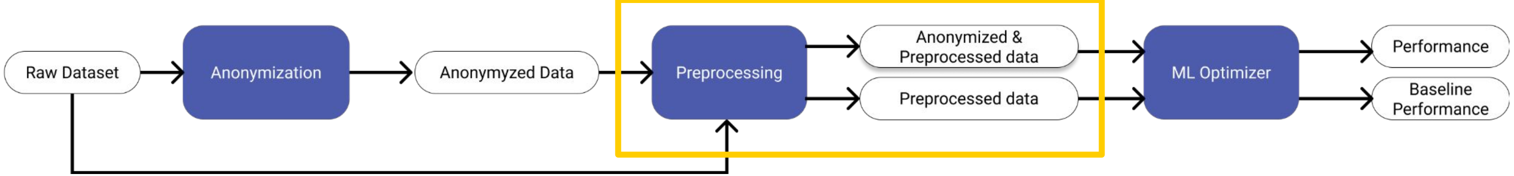


K-Anonymization: methods

- Classic & Basic Mondrian
- Datafly







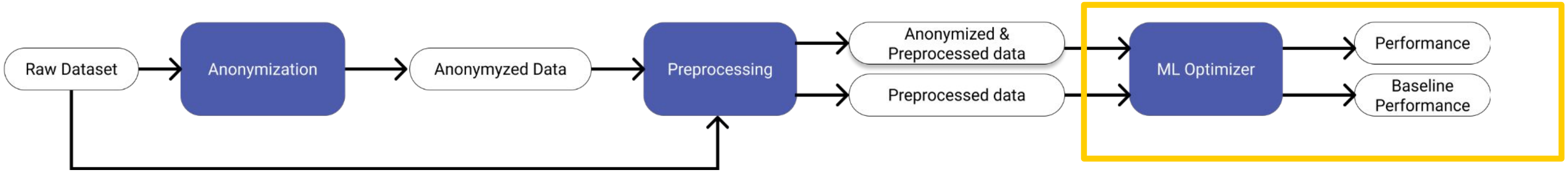
Preprocessing

- Initial:
  - *Removal of the irrelevant variables*
  - *Encoding of the categorical variables*

- After anonymization:
  - *Complete suppression*
  - *Overlapping categories*
    - Numerical values – solved with mean imputation
    - Categorical values – tricky problem

ID	Age	Imputed
1	20~36	28
2	31	31
3	20~30	25

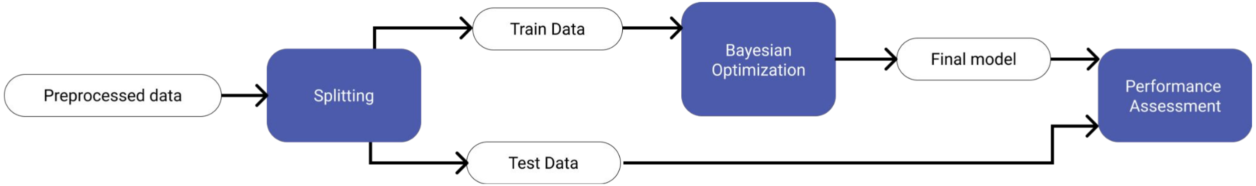
ID	Occupation	Imputed
1	Programmer	?
2	Tech-sector	?
3	Private-sector	?



ML tasks

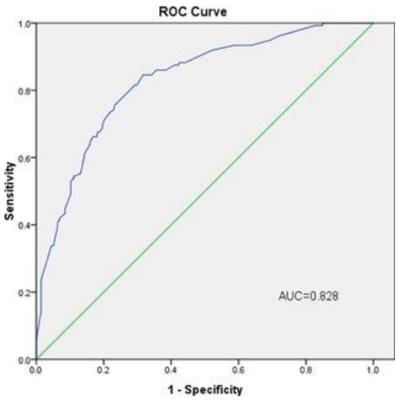
- Algorithms
  - Random Forest
  - XGBoost

- Bayesian Optimization implemented with *Hyperopt*
  - 20 search rounds
  - 4-fold CV in the training set for each round



Metrics

- Classification
- Regression
- Multiclass

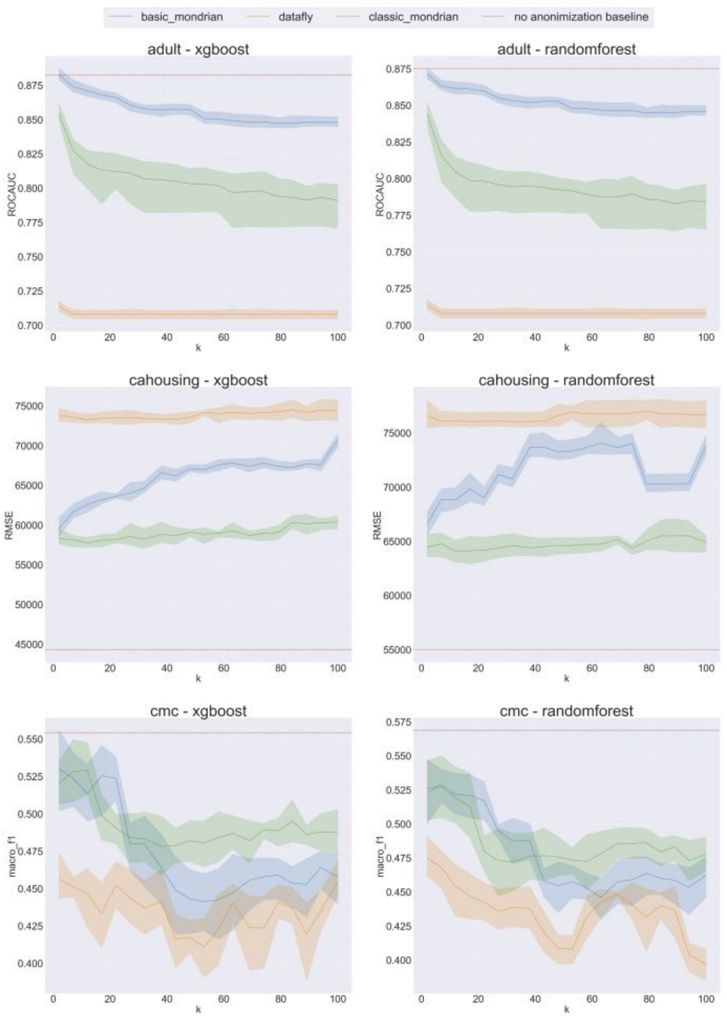


$$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{N}(y-\hat{y})^2}$$

$$F1_{macro} = \frac{1}{M}\sum_{i=1}^{M} F1_i$$

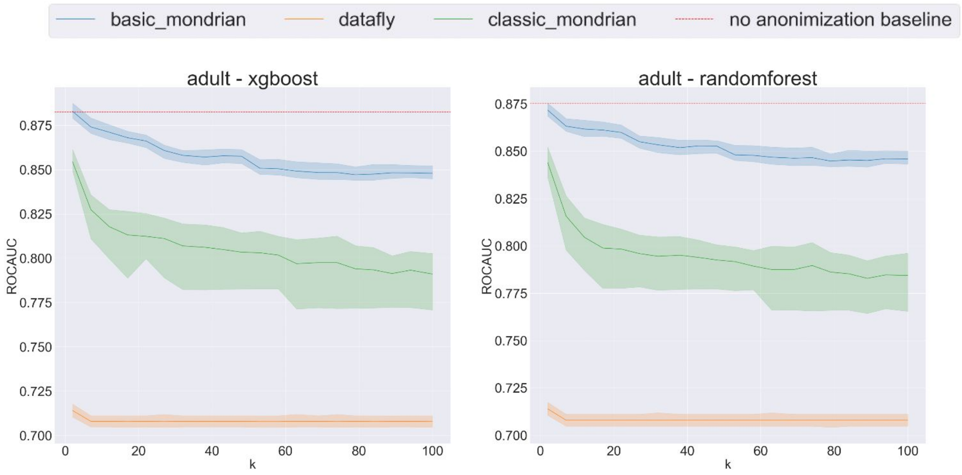
Results

Components	Count
ML Algorithms	2
Datasets	3
Ks	21
Bayesian opt. iterations	20
Cross-validation	4
Pipeline runs	4
Total models	40320



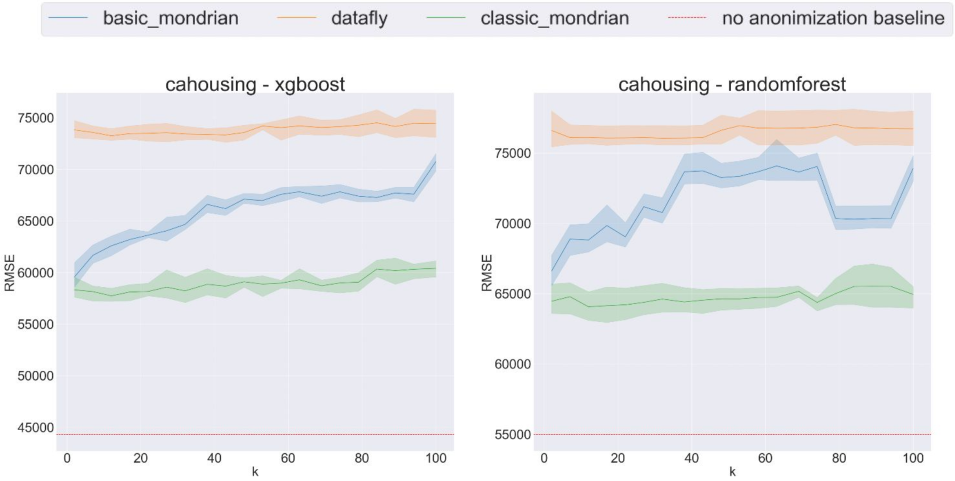
Discussion - ADULT

- **Datafly**
  - Performance drop
  - Overgeneralization
- **Basic Mondrian**
  - Best performance
  - Well-suited for categorical QIDs
- **Classic Mondrian**
  - Worse than Basic Mondrian
  - Ill-suited for categorical QIDs



Discussion - CA Housing

- **Performance gap w.r.t. Baseline**
  - High cardinality of attribute values
- **Datafly**
  - Overgeneralization
- **Classic Mondrian**
  - Best performance
  - Well-suited for numerical QIDs
- **Basic Mondrian**
  - Worse than Classic Mondrian
  - Generalization hierarchies not granular enough

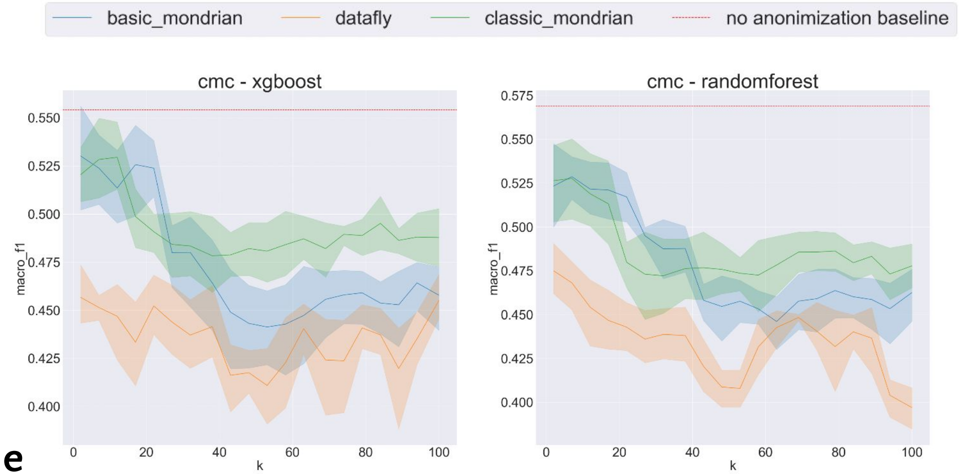


Discussion - CMC

- **Dataset size**
  - Not satisfactory baseline performance
  - High variation
- **High impact of *K***

Dataset	Maximum K	% of the dataset
ADULT	100	0.307
CAH	100	0.484
CMC	100	6.789

- **Smaller performance gap w.r.t. baseline**
  - Low cardinality numerical attributes
  - Granular enough generalization hierarchy



Conclusions

- Important factors **we knew** :
    - Increase K → Information loss → Performance decrease
    - Datafly problem: overgeneralization tendency
- Important factors **we found**:
    - Hierarchy granularity
        - High cardinality attributes, i.e. numerical
    - Type of QID
        - Numerical → Ordering based
        - Categorical → Generalization hierarchy based

Conclusions

DATA PRIVACY

DATA SCIENCE

References

1. Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. 1998.
2. Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In 22nd International conference on data engineering (ICDE'06), pages 25–25. IEEE, 2006.
3. Jon Louis Bentley. Multidimensional binary search trees used for associative searching. Communications of the ACM, 18(9):509–517, 1975.
4. Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Workloadaware anonymization. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 277–286, 2006.
5. Vanessa Ayala-Rivera, Patrick Mcdonagh, Thomas Cerqueus, and Liam Murphy. A systematic comparison and evaluation of k-anonymization algorithms for practitioners. Transactions on Data Privacy, 7:337–370, 12 2014.
6. Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.
7. Cam Nugent. California housing prices, Nov 2017.
8. Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.
9. Tin Kam Ho. Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition, volume 1, pages 278–282. IEEE, 1995.
10. Djordje Slijepˇcevi´c, Maximilian Henzl, Lukas Daniel Klausner, Tobias Dam, Peter Kieseberg, and Matthias Zeppelzauer. k-anonymity in practice: How generalisation and suppression affect machine learning classifiers, 2021.
11. [11] James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins, and David D Cox. Hyperopt: a python library for model selection and hyperparameter optimization. Computational Science Discovery, 8(1):014008, 2015.