计算机科学与软件工程学院
School of Computer Science and Software Engineering

華東師範大學
EAST CHINA NORMAL UNIVERSITY

Using Anonymized Data for Regression with Hyper-Rectangle Pruning

Kun Liu
workkun@outlook.com
East China Normal University

https://github.com/build2last/UHRP

Outline

- Background
- Problem definition
- Solution
- Experiments

Anonymization

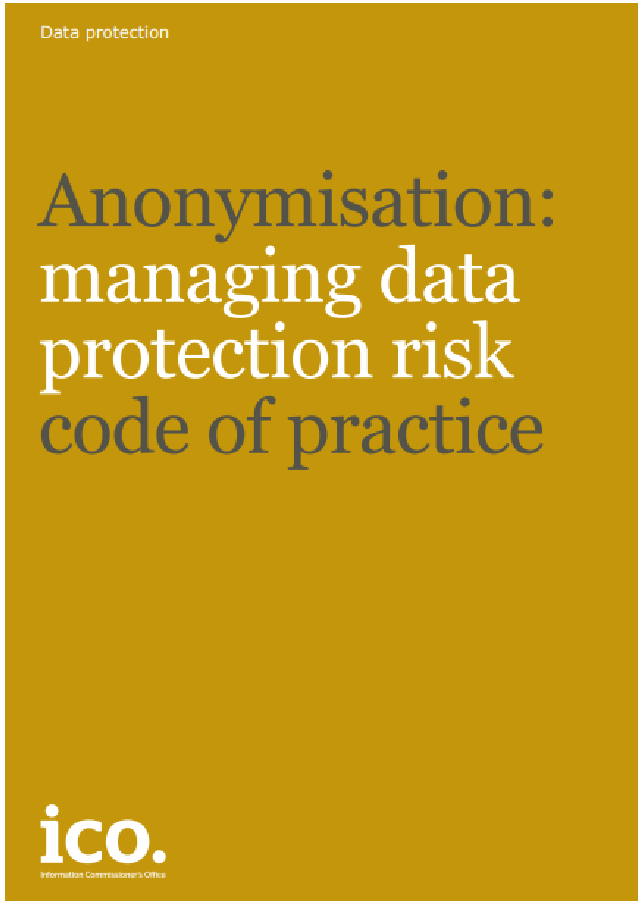
Widely adopted by Google, ICO and …

HOW GOOGLE ANONYMIZES DATA

Anonymization is a data processing technique that removes or modifies personally identifiable information; it results in anonymized data that cannot be associated with any one individual. It's also a critical component of Google's commitment to privacy.

By analyzing anonymized data, we are able to build safe and valuable products and features, like autocompletion of an entered search query, and better detect security threats, like phishing and malware sites, all while protecting user identities. We can also safely share anonymized data externally, making it useful for others without putting the privacy of our users at risk.

Google Privacy & Terms



Information Commissioner's Office, UK

Anonymized Data

Anonymize data with generalization.

Original Data

Engine	Horsepower	Acceleration	Weight	Miles Per Gallon
L	130	12	3504	18
V	165	11.5	3693	15
W	150	11	3436	18

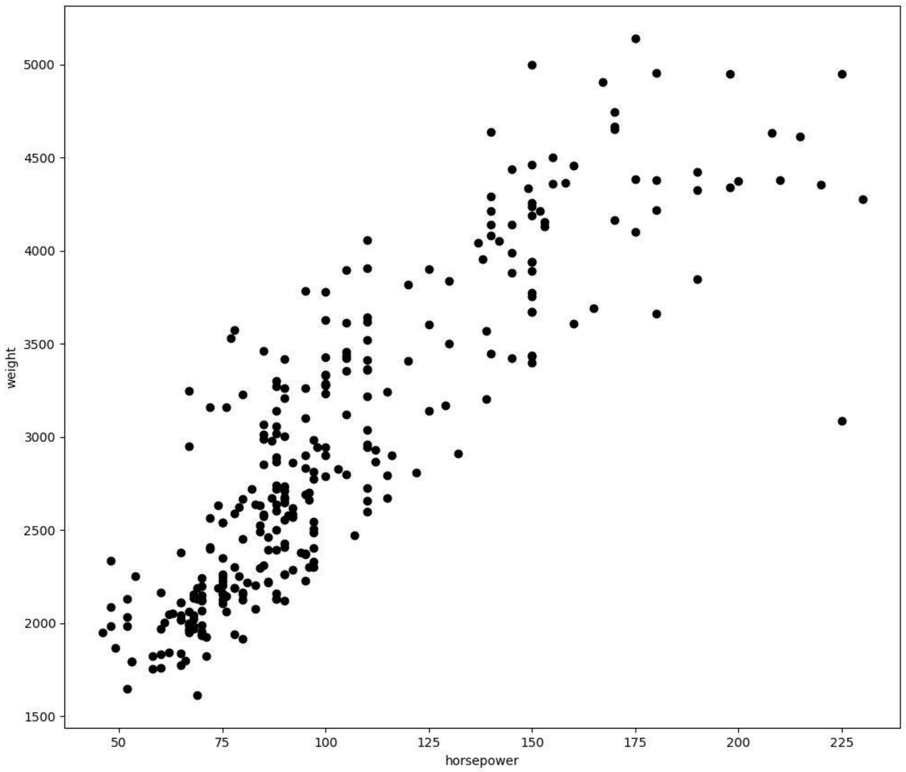
Anonymized Data

Engine	Horsepower	Acceleration	Weight	Miles Per Gallon
L	[130, 150)	[11.5, 12)	[3436, 3504)	[15, 18)
{W, V}	[150, 165)	[11.5, 12)	[3504, 3693)	[15, 18)
{W, V}	[150, 165)	[11, 11.5)	[3436, 3504)	[15, 18)

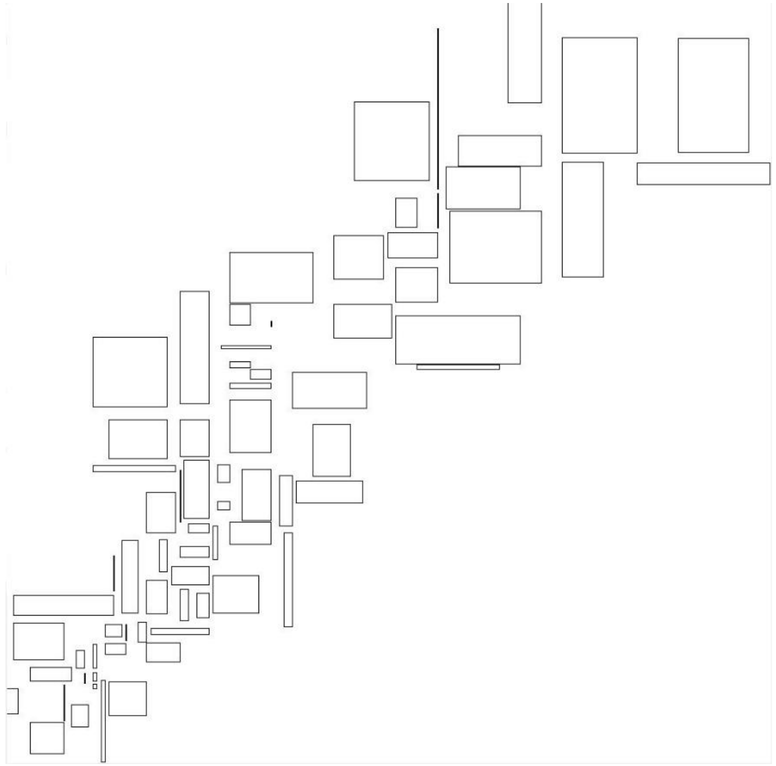
Transform specific data to Interval-valued or set-valued data

School of Computer Science and Software Engineering
EAST CHINA NORMAL UNIVERSITY

Changes after generalization of a 2-dimension dataset



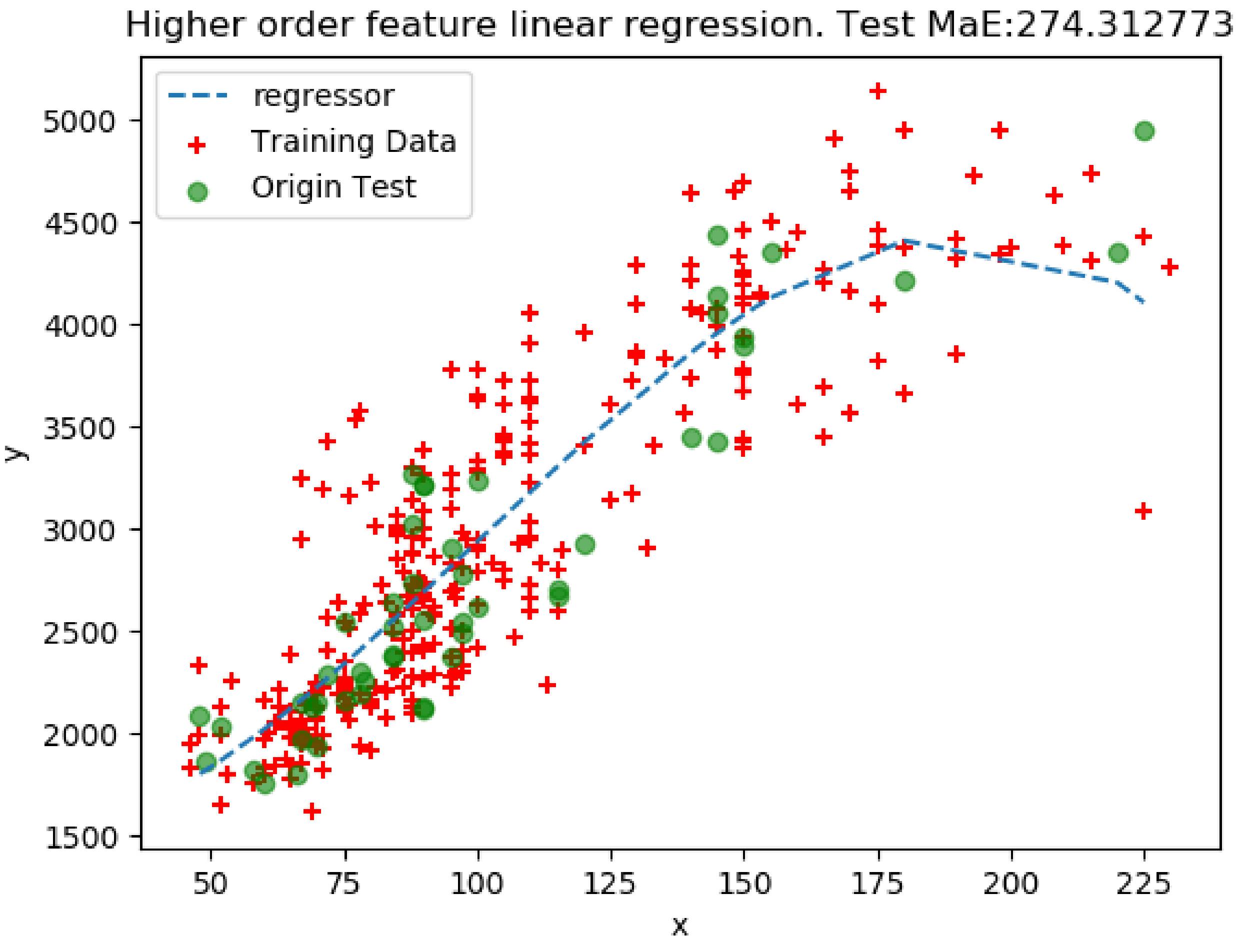
Original Data Samples

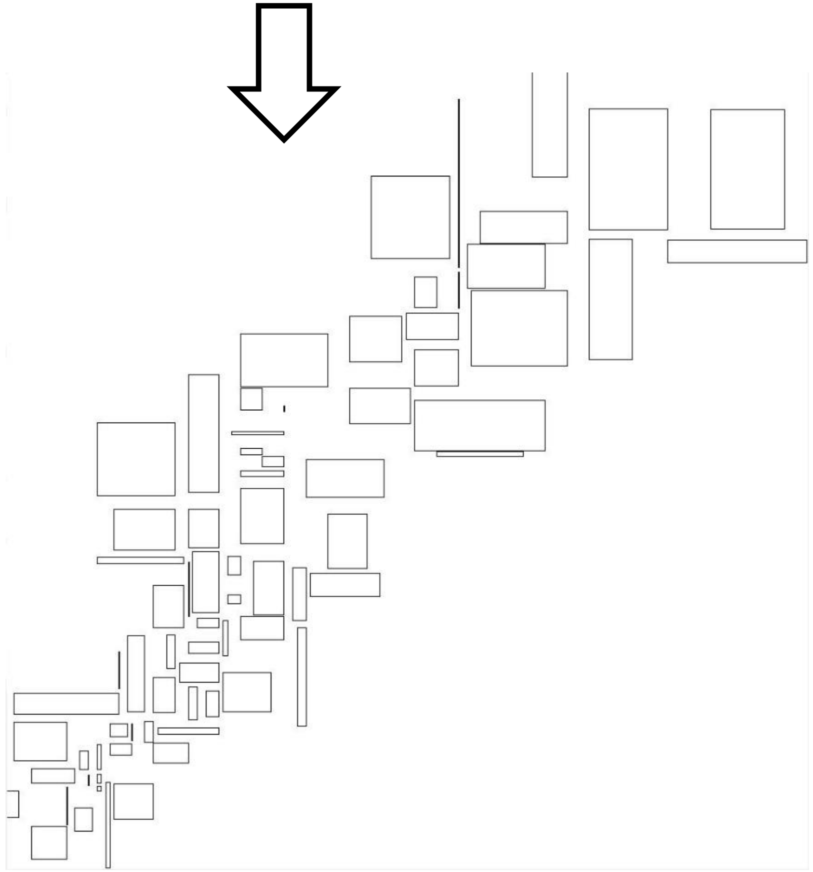


Anonymized Data Samples

Regression task

But, how to train on Anonymized data ??





Regression task trained on original dataset

Motivation and Challenges

- Motivation:
  - Using anonymized dataset training regression model which can be used on both original and anonymized sample prediction.

- Challenges:
  - How to **represent unspecified data** ?
  - How to **reduce the impact of uncertainty (or noise)** ?

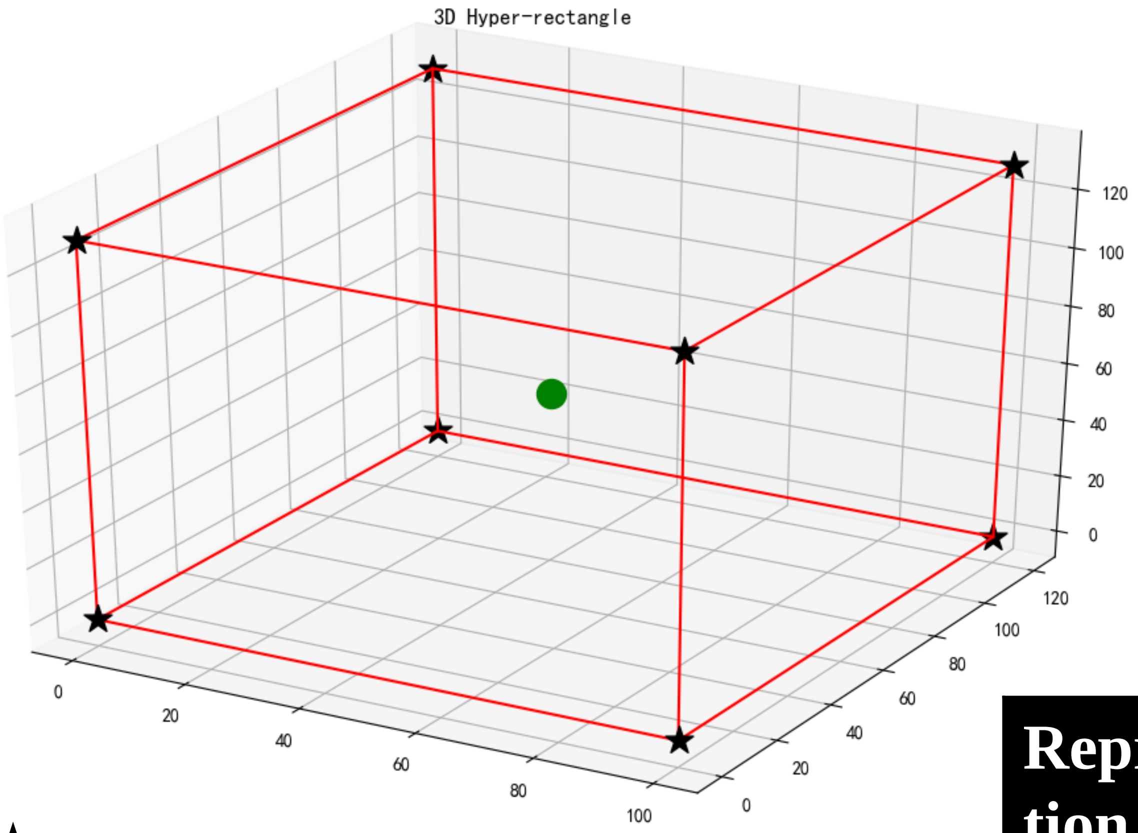
Anonymized data representation methods

- Related work：

ICDE 2009 Using anonymized data for classification

1. FR1: Each data interval is represented by its average.
2. FR2: A hyper-rectangle is represented by its center point
3. FR3: Represented by the upper and lower bounds of the interval.

Hyper-rectangle representation



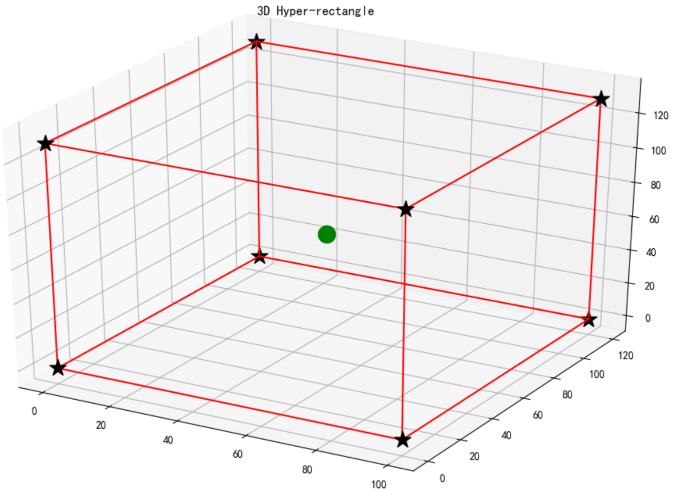
★ Corner point

● Central point

- With point
  1. Central point
  2. Corner points and central point
- With range

Representation	Feature 1 (axis-x)	Feature 2 (axis-y)	Label (axis-z)
With point	50	60	60
With range	[0, 100]	[0, 120]	[0, 120]

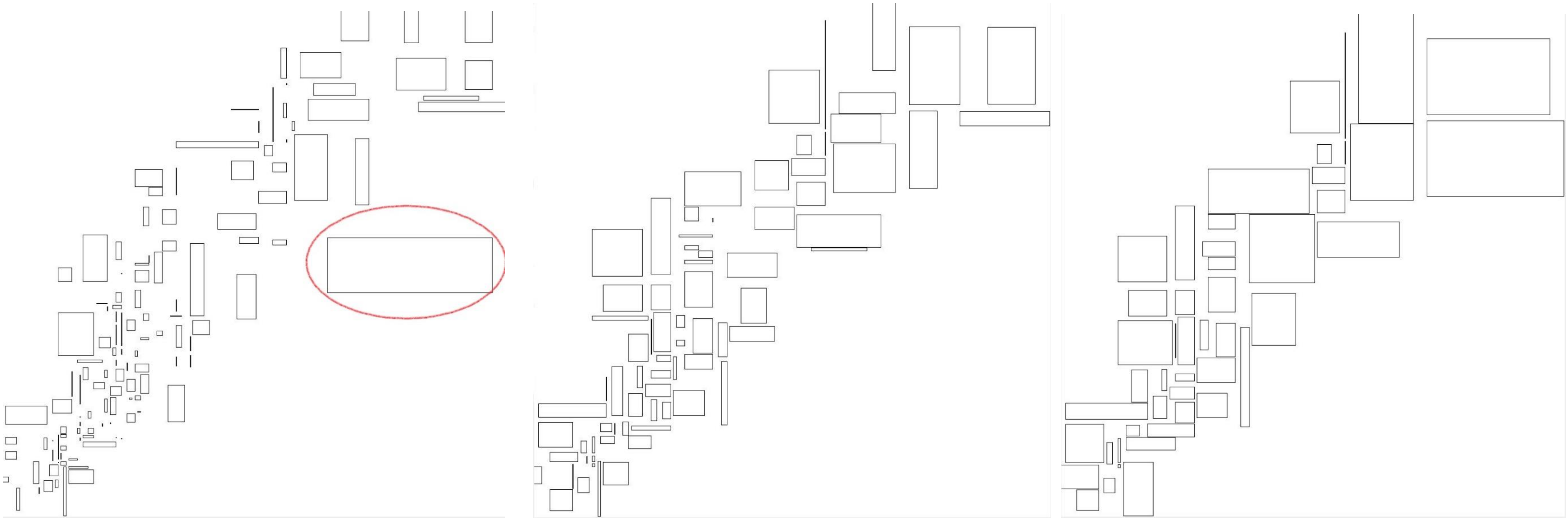
Feature Representation methods



EAST CHINA NORMAL UNIVERSITY

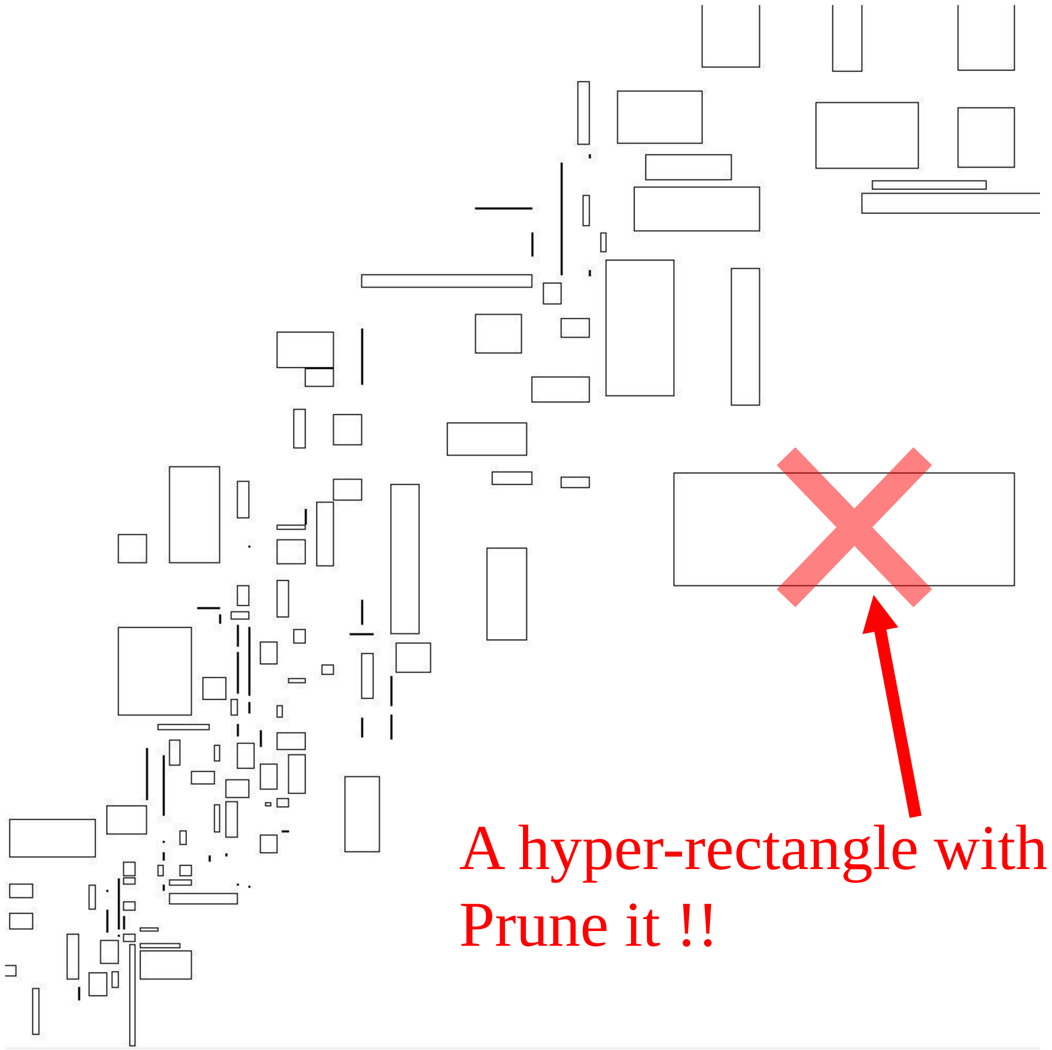
	Feature 1	Feature 2	Label	Feature Vector and label
Original data	70	60	90	[70 60] 90
Anonymized data	50~80	50~70	80~100	
FR1	65	60	90	[65 60] 90
FR2 (Eight point)	50 65 80	50 60 70	80 90 100	[50 50] 80 [65 60] 90 [80 70] 100
FR3	[50 80]	[50 70]	90	[50 80 50 70] 90

The `average size` of the hyper-rectangles increases with K



K = 2 K = 3 K = 5

Hyper-rectangle Pruning



Model uncertainty factors for regression：
Model + Data

Theory and motivation:
Reduce the average data uncertainty

Before pruning,
How to calculate the
uncertainty of a hyper-rectangle ?

1. Multiply all attributes‘ uncertainty
2. Add up all attributes‘ uncertainty

A hyper-rectangle with too much uncertainty !!
Prune it !!

Calculating the uncertainty of a hyper-rectangle

Uncertainty-base Hyper-Rectangle Pruning

$$U(x_i) = \prod_{j=1}^{q} u_{ij} \qquad (1)$$

$$U(x_i) = \sum_{j=1}^{q} u_{ij} \qquad (2)$$

Formula (1) meets trouble when $u_{ij} = 0$. So we choose (2).

We pruning the Hyper-Rectangle with biggest ‘Uncertainty’ to a scale before training.

Experiments

- **Datasets: UCI Machine Learning**

Dataset	Data dimention
Air Quality	9000 * 6
AUTO-MPG	392 * 7

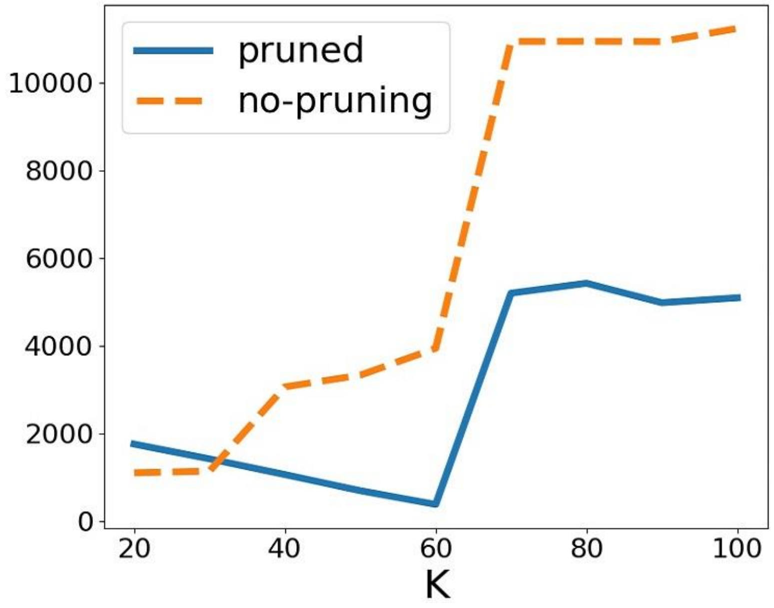
- **Anonymity Algorithm: Mondrian**
- **Model : Linear Regression model in scikit-learn**

Model performance on anonymized data with various K and feature representation methods.

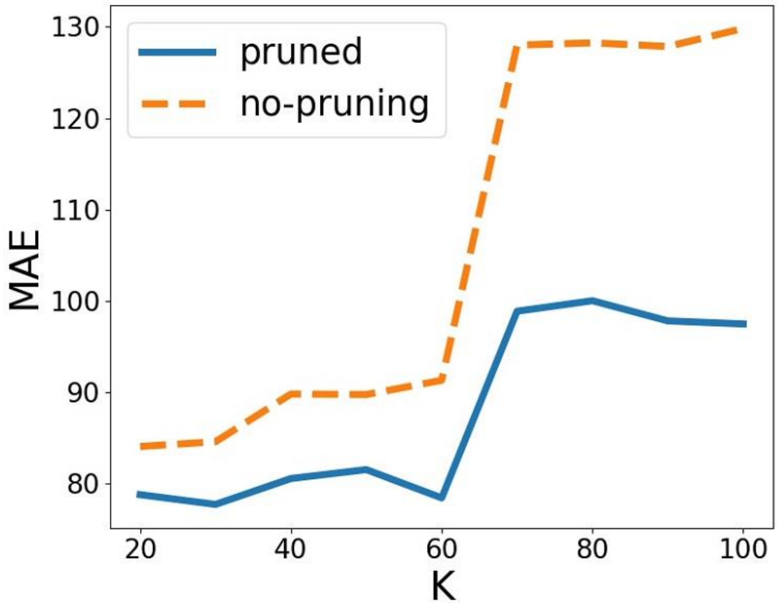
Utility of anonymized data

Data Set	k	*FR1*	*FR1-H*	*FR2*	*FR2-H*	*FR3*	*FR3-H*
Auto MPG (size 300+)	1(origin)	6.54	103.31	4.22	3.57	3.82	3.25
	2	4.82	98.77	4.13	3.63	4.06	3.55
	3	7.31	121.32	4.14	3.71	3.99	3.71
	4	8.4	130.36	4.12	3.81	4.07	3.42
	5	4.13	85.65	4.08	3.91	4.4	5.48
AirQuality (size 9000+)	1(origin)	461.05	297.22	97.22	99.95	86.93	86.9
	8	362.81	383.17	94	95.31	79.09	85.13
	32	191.17	1044.96	92.73	94.72	79.64	117.03
	64	221.54	5220.48	101.04	121.03	118.2	184.03
	128	194.62	2740.58	130.17	141.9	146.41	150.76

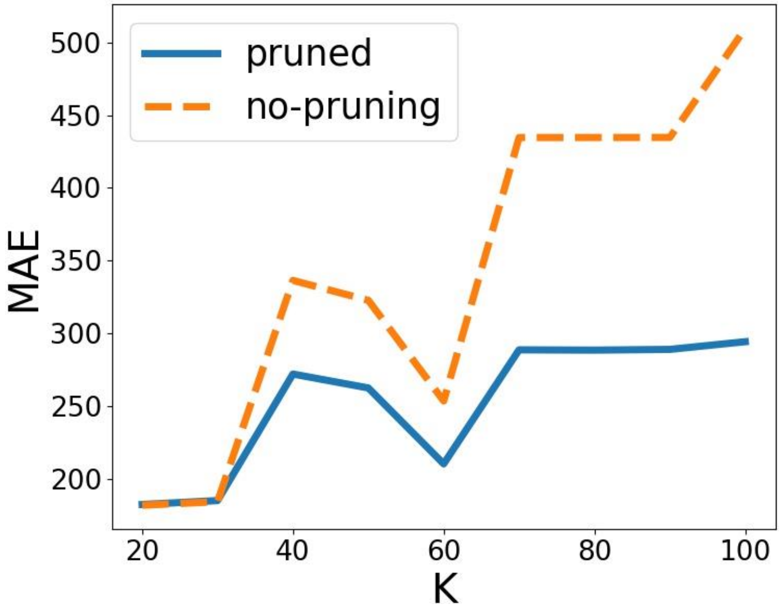
Pruning efficiency experiment



FR1-H
Remainder-Ratio = 0.8



FR2-H
Remainder-Ratio = 0.8



FR3-H
Remainder-Ratio = 0.99

Conclusion

- Regression model trained with anonymized data can be expected to do as well as the model trained on original dataset under certain conditions.

- Our UHRP method can improve regression model performance

- Future work
  - How to better evaluate uncertainty of anonymized data.
  - How to find a good Remainder-Ratio need more research.

Q & A