MEDI-CHAL PRESENTATION

REVIEW OF METRICS FOR SYNTHETIC DATA : PRIVACY AND UTILITY TRADEOFFS

- SALONI DASH

OVERVIEW

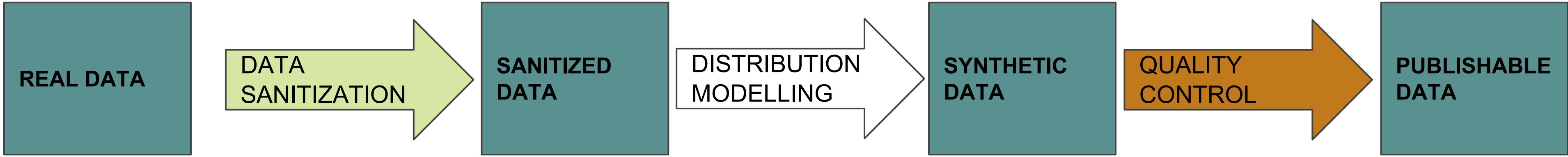
- DATA PUBLISHING CONTEXT
- PRIVACY
  - WHAT IS PRIVACY?
  - PRIVACY PRESERVING TECHNIQUES
  - PRIVACY REQUIREMENTS
- PRIVACY VS DATA UTILITY
  - WHAT IS DATA UTILITY?
  - GENERAL METHODOLOGY
  - DIRECT COMPARISON METHOD
  - EXPERIMENTAL RESULTS
  - CRITICISM OF DIRECT COMPARISON
- CONCLUSION
- REFERENCES

LINK TO PREVIOUS TALK

A STUDY OF PRIVACY METRICS :

https://docs.google.com/presentation/d/1JEBJh3VKOj6XGFcn3_8N584nxO0VDrqMAGRCV4ok6PQ/edit?usp=sharing

DATA PUBLISHING CONTEXT



PRIVACY

WHAT IS PRIVACY?

- In the context of Data Publishing :
    - Organization has confidential records of individuals
- Attack Scenarios :
    - Adversary has QIDs of an individual
    - Adversary builds classifier for sensitive attributes
- Data Anonymization Goals :
    - **Unique Identity Disclosure**
    - **Sensitive Attribute Disclosure**

Key Attribute	Quasi-identifier			Sensitive attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

PRIVACY PRESERVING TECHNIQUES

Name	Gender	Zip code	Age
Reena	Female	444805	45
Shweta	Female	424806	46
Kavita	Female	424806	58
Neha	Female	444806	65

ORIGINAL DATASET

- Generalization

Name	Gender	Zip code	Age
Reena	Person	444805	40-60
Shweta	Person	424806	40-60
Kavita	Person	424806	40-60
Neha	Person	444806	40-60

- Suppression

Gender	Zip code	Age	Disease
Female	444805	**	**
Female	424806	**	**
Female	424806	**	**
Female	444806	**	**

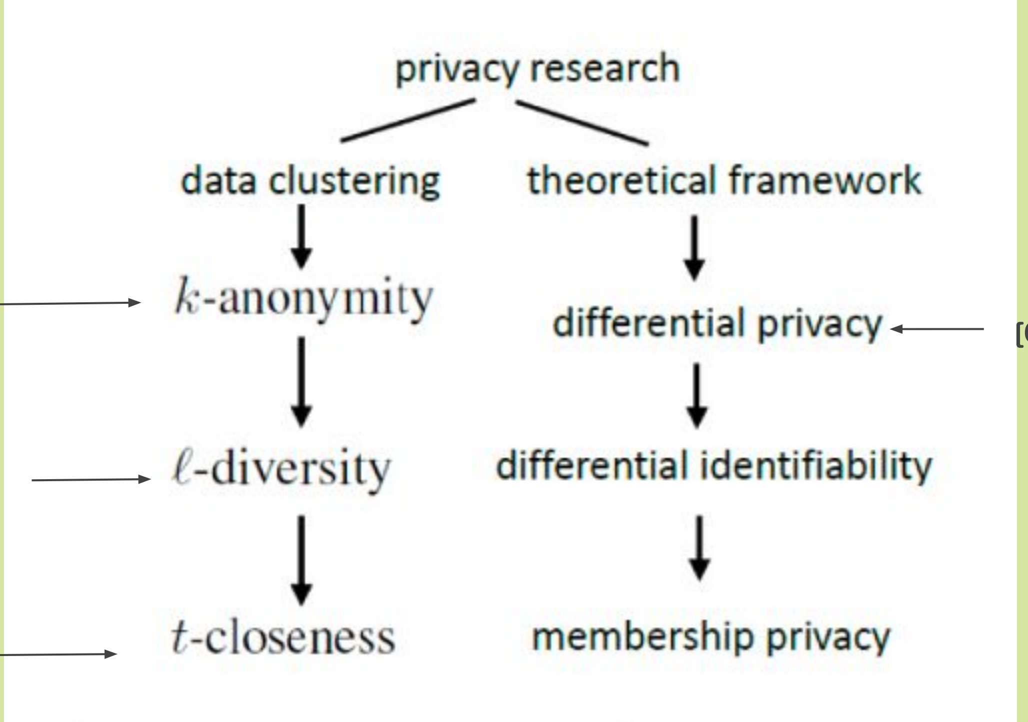
- Bucketization

Gender	Zip code	Age	GID
Female	444805	45	1
Female	424806	46	2
Female	424806	58	3
Female	444806	65	4

a) QId table

Disease	GID
TB	1
Diabetes	2
Fever	3
Cancer	4

CATEGORIES OF PRIVACY RESEARCH



[11] {L. Sweeney et al. , 1998}

[6]{Dwork, C et al. 2006}

[15]{A. Machanavajjhala et al. 2006}

[8]{N. Li et al. 2007}

PRIVACY REQUIREMENTS

- **k-anonymity :**
  - The identifiable attributes of any database record are indistinguishable from at least other k-1 records

	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

Table 1. Original Patients Table

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

Table 2. A 3-Anonymous Version of Table 1

- **l-diversity :**
  - Every equivalence class should abide by the l-diversity principle

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

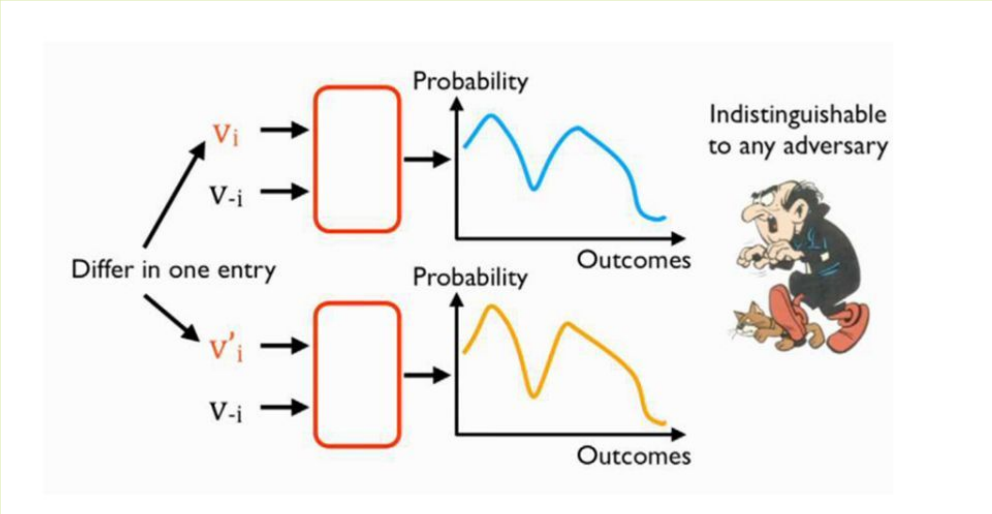
Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

- **t-closeness:**
  - Distance between the distribution of a sensitive attribute in the original table and the distribution of the same attribute in any equivalence class is less or equal to t

	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

Table 5. Table that has 0.167-closeness w.r.t. Salary and 0.278-closeness w.r.t. Disease

- **Differential privacy :**
  - Measure the difference on individual privacy disclosure between the presence and the absence of the individual’s record.
  - Ensures that a single record does not considerably affect the outcome of the analysis over the dataset



PRIVACY VS UTILITY

WHAT IS DATA UTILITY?

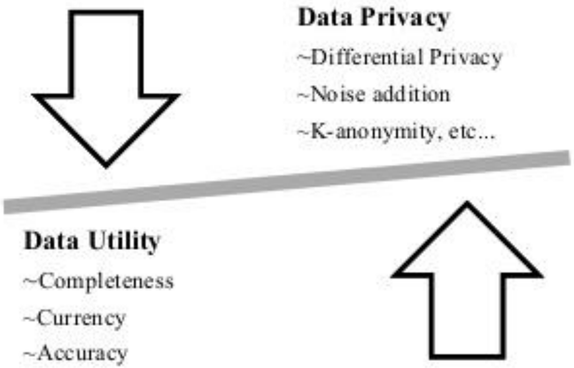
- Gains to society from data publishing:
  - Enables **researchers** and **policy-makers** to analyze the data
  - Important information benefiting the society as a whole can be learned :
    - Diseases
    - Effectiveness of a medicine or treatment
    - Social-economic patterns
- Data Utility Preservation Metrics (Ritik’s talk)
  - For e.g., K.S. tests, Comparison of Covariance Matrices etc.

BALANCING PRIVACY & UTILITY

THE PROBLEM

Finding a user-defined balance between data privacy and utility needs with trade-offs.

- The challenge of ambiguous definitions of privacy and utility.

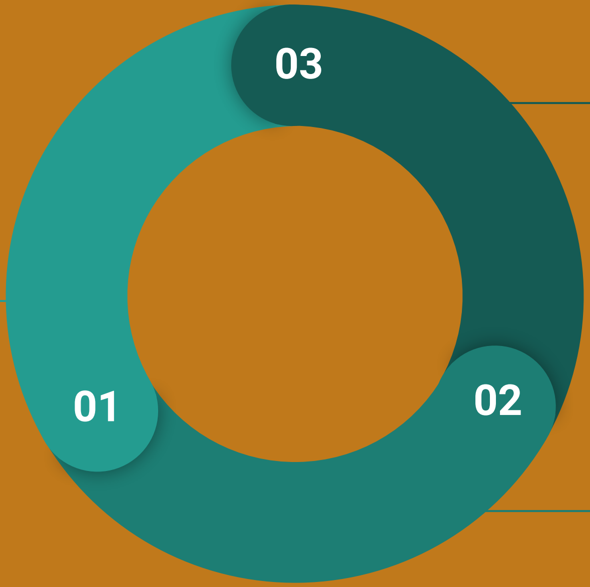


"Perfect privacy can be achieved by publishing nothing at all, but this has no utility; perfect utility can be obtained by publishing the data exactly as received, but this offers no privacy" Cynthia Dwork (2006)

GENERAL METHODOLOGY

PRIVACY REQUIREMENT:

- k-ANONYMITY
- l-DIVERSITY
- t-CLOSENESS
- DIFFERENTIAL PRIVACY



UTILITY METRIC:
- K.S. TESTS
- COMPARISON OF COVARIANCE MATRICES

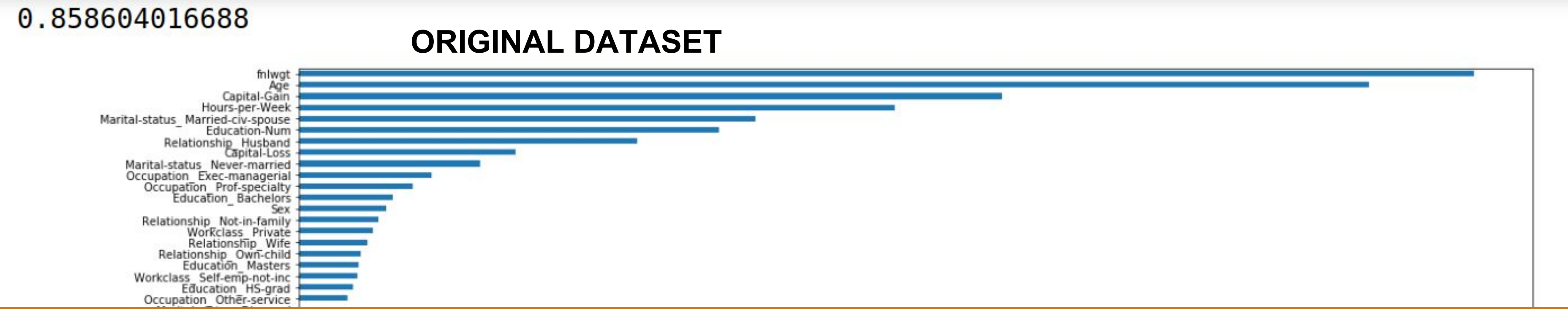
PRIVACY PRESERVING ALGORITHM:

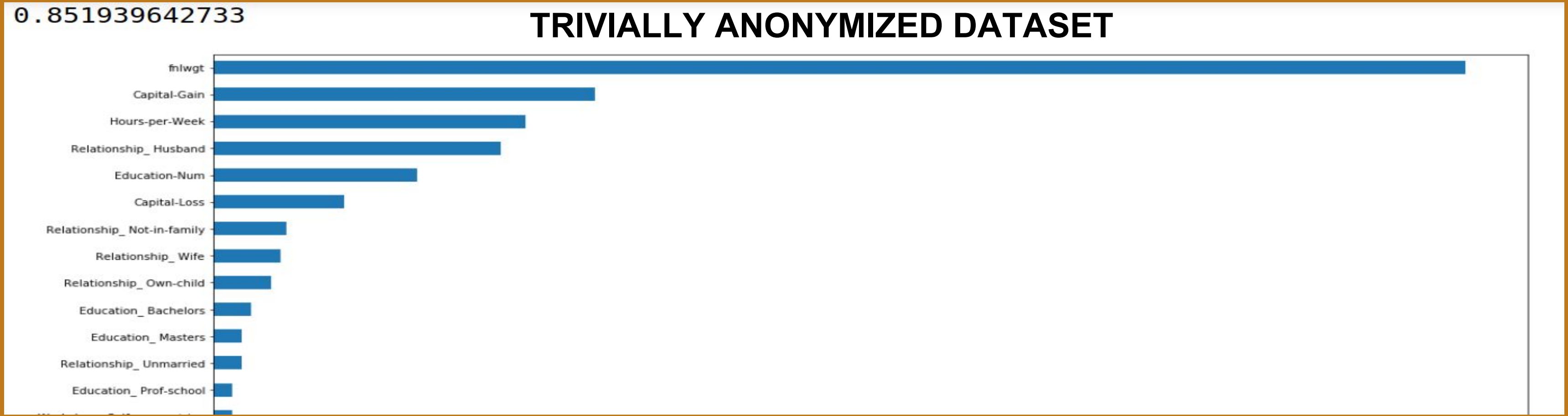
- GENERALIZATION
- SUPPRESSION
- BUCKETIZATION

DIRECT COMPARISON {J. Brickell et al.}

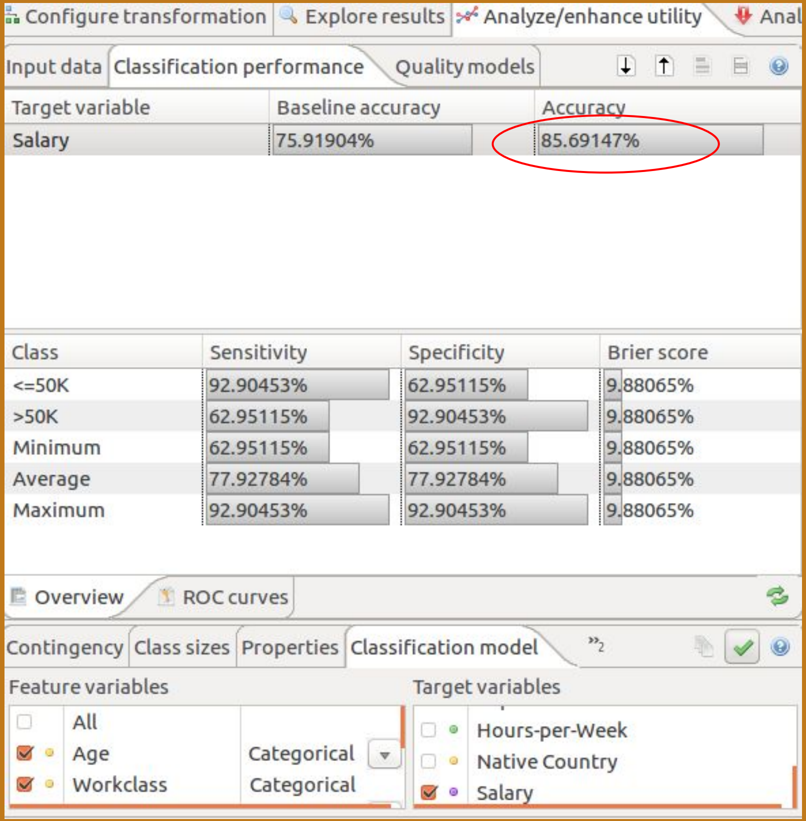
- Privacy loss is measured as the adversary's accuracy improvement in guessing the sensitive attribute value of an individual
- Utility gain is measured as the researcher's accuracy improvement in building a classification model for the sensitive attribute.
- Privacy Loss as well as Data Utility measured against trivialised anonymized dataset.
- Experiments in [18]{J. Brickell and V. Shmatikov.} leads to the intriguing conclusion "even modest privacy gains require almost complete destruction of the data-mining utility

EXPERIMENTAL RESULTS



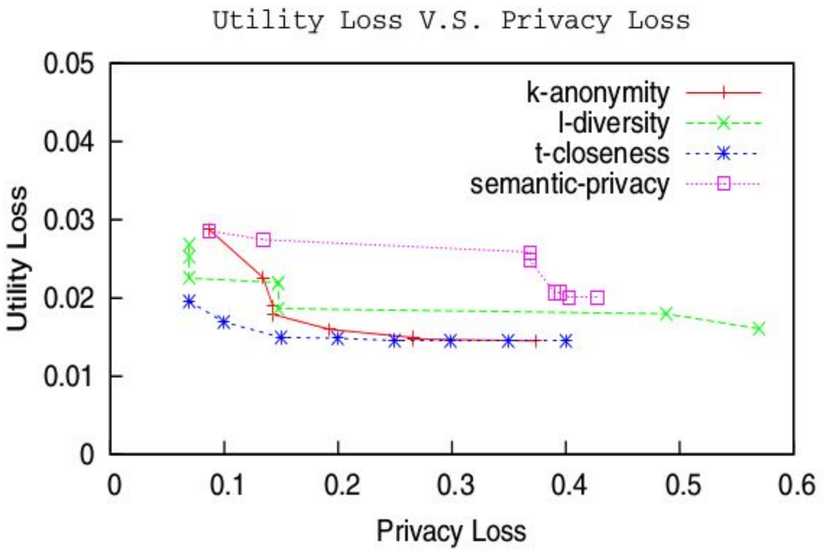


5-ANONYMIZED + 0.001 CLOSE DATASET

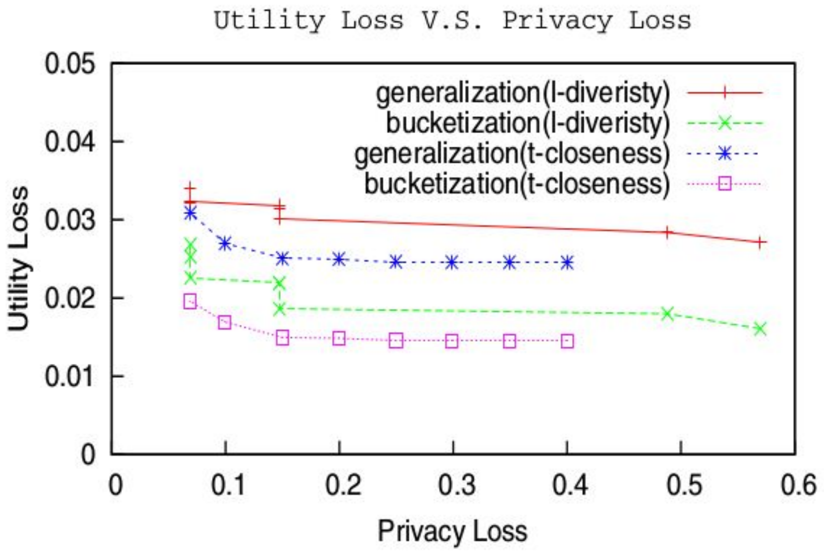


CRITICISM OF DIRECT METHODOLOGY [18]{T. Li et. al, 2009}

- Specific knowledge has a larger impact on privacy, while aggregate information has a larger impact on utility.
- Privacy is an Individual concept, while utility is an aggregate concept
- For privacy, the worst-case privacy loss should be measured. For utility, the aggregated utility should be measured.
- Data utility should be measured against original dataset.



(a) Varied privacy requirements

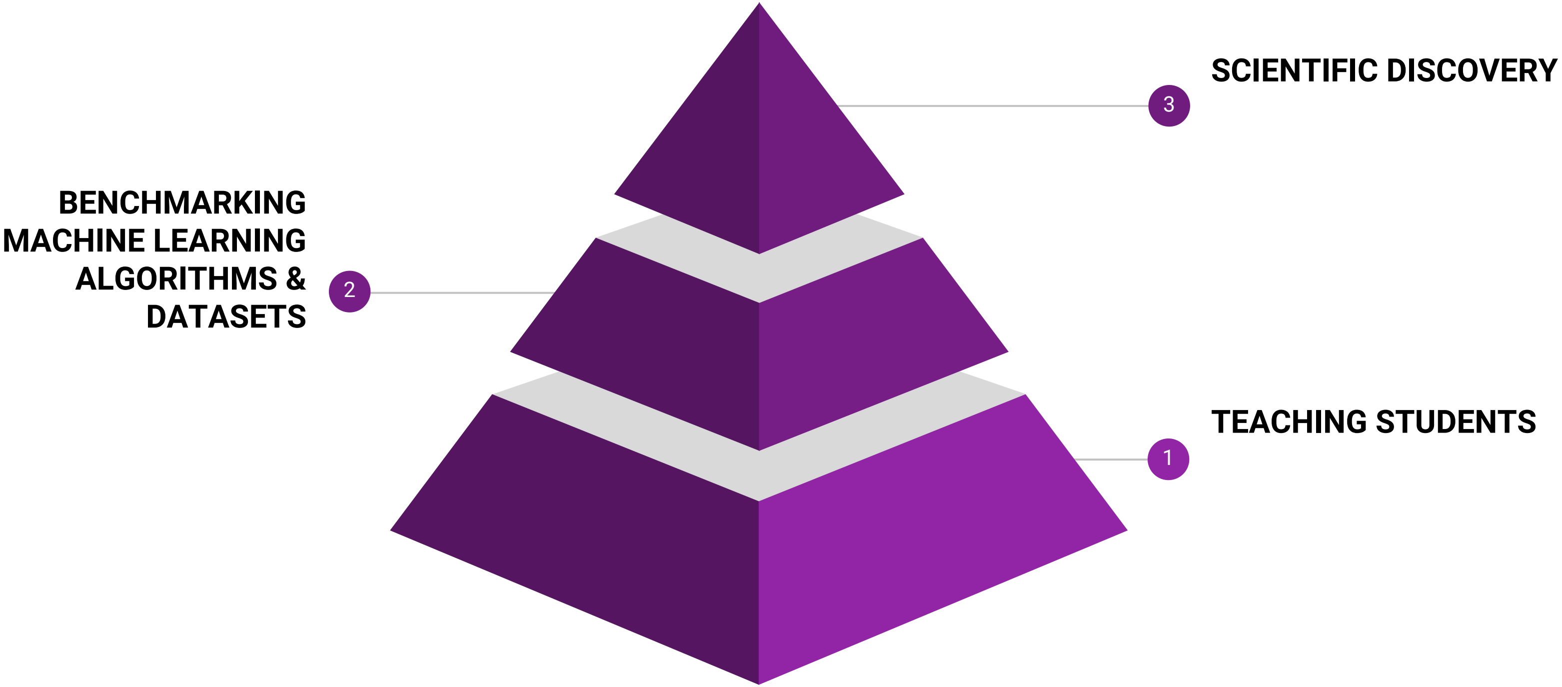


(b) Varied anonymization methods

Distance Metric : J.S. Divergence

CONCLUSION

SYNTHETIC DATA UTILITIES



REFERENCES

REFERENCES

1. Aggarwal, C. C. (2015). Data mining: the textbook (1st ed.). New York, NY, USA: Data Mining: The Textbook.
2. Agrawal, D. & Aggarwal, C. C. (2001). On the design and quantification of privacy preserving data mining algorithms. Proc. 20th ACM SIGMOD- SIGACT-SIGART Symp. Principles Database Sys, 247–255.
3. Agrawal, R. & Srikant, R. (2000). Privacy-preserving data mining. ACM SIGMOD Rec, 29 (2), 439–450.
4. Dankar, F. K. & Emam, K. E. (2013). Practicing differential privacy in health care: a review. TRANSACTIONS ON DATA PRIVACY.
5. J. Brickell and V. Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In KDD, pages 70–78, 2008
6. Dwork, C. (2006). Differential privacy. Automata, Languages and Programming, 4052,1–12.
7. E. Bertino, D. L. & Jiang, W. (2008). A survey of quantification of privacy preserving data mining algorithms. Privacy-Preserving Data Mining, 183–205.
8. MENDES, R. & VILELA, J. P. (2017). Privacy-preserving data mining: methods, metrics, and applications. IEEE Access.

8. N. Li, T. L. & Venkatasubramanian, S. (2007). T-closeness: privacy beyond k-anonymity and l-diversity. Proc. IEEE 23rd Int. Conf. Data Eng. (ICDE), 106–115.
9. Oliveira, S. R. M. & Zaıane, O. R. (2010). Privacy preserving clustering by data transformation. J. Inf. Data Manag, 1 (1), 37.
10. Samarati, P. & Sweeney, L. (1998a). Generalizing data to provide anonymity when disclosing information. Proc. PODS, 188.
11. Samarati, P. & Sweeney, L. (1998b). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. IEEE Symp. Res. Secur. Privacy, 384–393.
12. T. S. Gal, Z. C. & Gangopadhyay, A. (2008). A privacy protection model for patient data with multiple sensitive attributes. Int. J. Inf. Secur. Privacy, 2 (3), 28.
13. Yu, S. (2016). Big privacy: challenges and opportunities of privacy study in the age of big data. IEEE Access, 4, 2751–2763.
14. Ali Borji. “Pros and Cons of GAN Evaluation Measures”. In: CoRR abs/1802.03446 (2018). arXiv: 1802.03446. url: http://arxiv.org/abs/1802.03446.

15. A. Machanavajjhala, M. Venkitasubramaniam, D. Kifer, and J. Gehrke.l-diversity: Privacy beyond k-anonymity. In 22nd International Conference on Data Engineering (ICDE’06)(ICDE), volume 00, page 2404 2006

16. Daniel Jiwoong Im et al. “Quantitatively Evaluating GANs With Divergences Proposed for Training”. In: CoRR abs/1803.01045 (2018). arXiv: 1803.01045. url: http://arxiv.org/abs/1803.01045.
17. L. Theis, A. van den Oord, and M. Bethge. “A note on the evaluation of generative models”. In: ArXiv e-prints (Nov. 2015). arXiv: 1511.01844 [stat.ML].
18. J. Brickell and V. Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In KDD, pages 70–78, 2008.
19. On the Tradeoff Between Privacy and Utility in Data Publishing, Tiancheng Li and Ninghui Li, Department of Computer Science, Purdue University {li83,ninghui}@cs.purdue.edu

THANK YOU!