INTEGRAL PRIVACY COMPLIANT STATISTICS COMPUTATION

NAVODA SENAVIRATHNE – UNIVERSITY OF SKÖVDE, SWEDEN

VICENÇ TORRA – UNIVERSITY OF MAYNOOTH, IRELAND

CONTENT

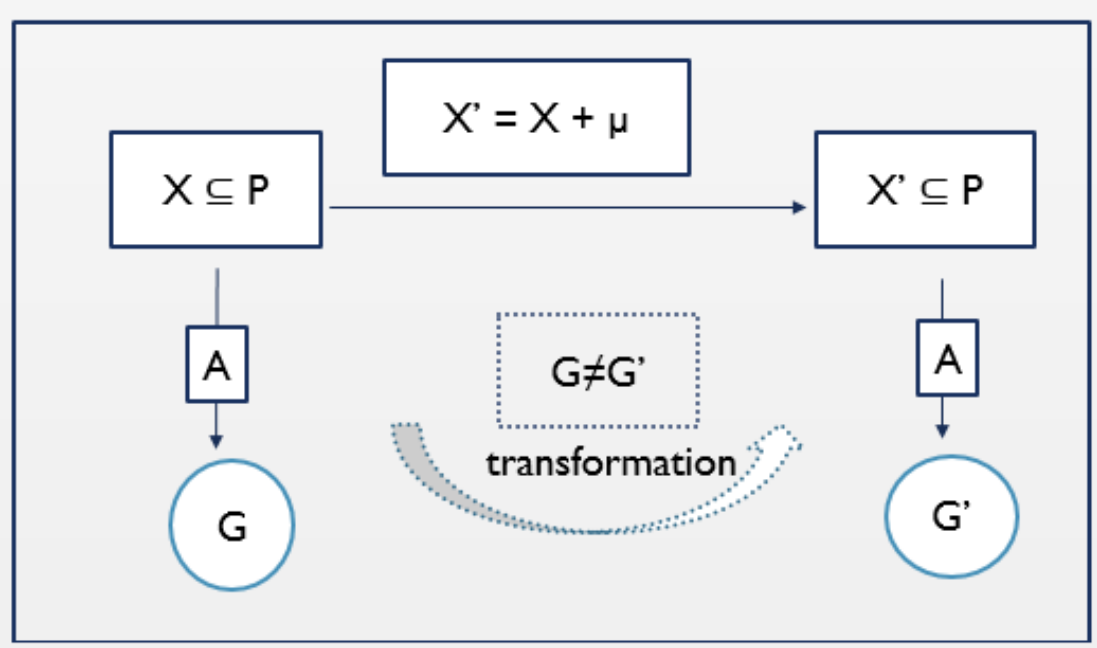
- Privacy Preserving Data Analysis
- Integral Privacy
- Differential privacy
- Methodology
- Results
- Discussion
- Conclusion

PRIVACY PRESERVING DATA ANALYSIS

- Requirement for privacy in data analytics arises when **sensitive data** are used in the process.
- Main objective of Privacy Preserving Data Analysis is to ensure a **degree of privacy is provided** while maintaining the **analytical utility of the results**.

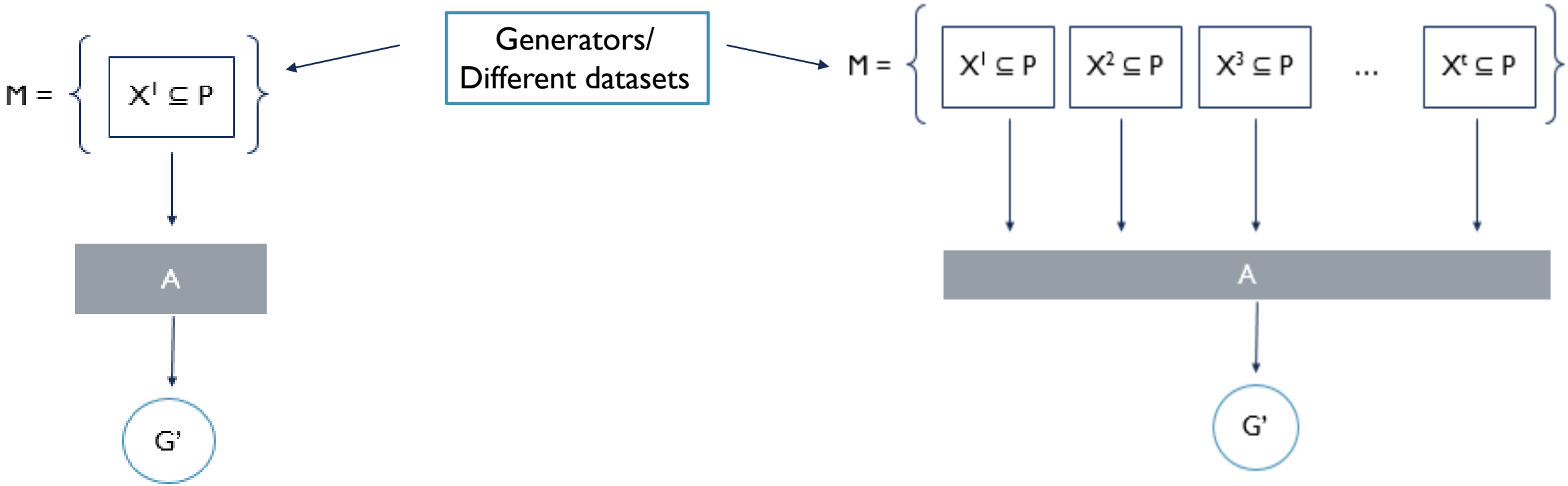
INTUITION OF INTEGRAL PRIVACY

- When the data are modified we may be required to **re-compute** the inferences/ answers to a given function.



- In case, if the intruder has access to G, G' with some background knowledge on P or X, can we ensure the **privacy of the set of modifications (μ) is guaranteed?**

PRIVACY PROBLEM



1:1 relationship – Less uncertainty for the intruder

M :1 relationship – High uncertainty for the intruder

INTEGRAL PRIVACY

- Integral privacy is defined when the **set of modifications (M) is large** ($|M| \geq k$)

integral privacy

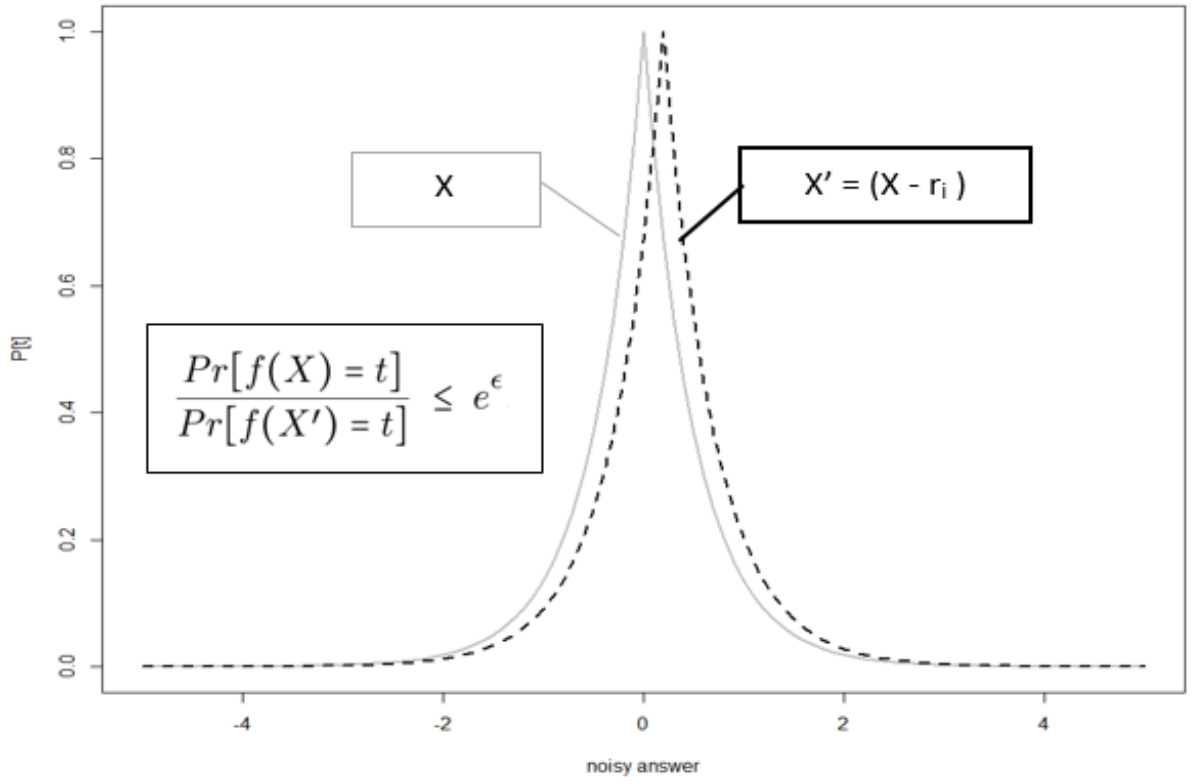
$$M = \{\mu \mid G = A(X) \text{ and } G' = A(X + \mu)\}$$

- And the **intersection is empty**.

$$\cap_{\mu \epsilon M} \mu = \emptyset$$

DIFFERENTIAL PRIVACY

- $DP\ answer = A(X) + Lap(\frac{\Delta A}{\varepsilon}), \text{for } \varepsilon > 0$



NOTION OF STABILITY

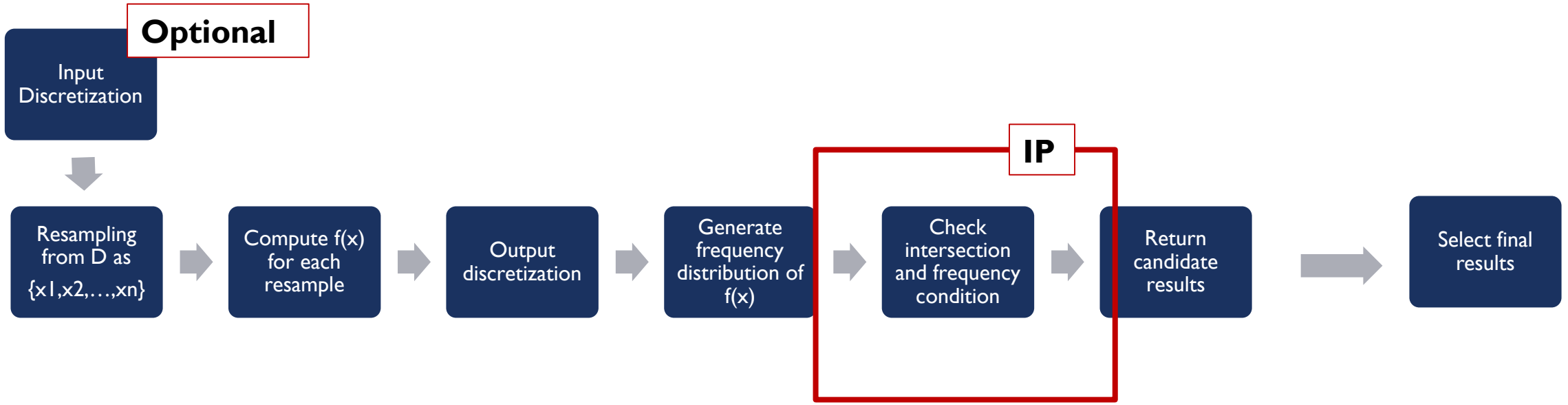
- Stable results : Less susceptible towards the perturbation done on the input data.
- **Integral Privacy :**
  - Stability is explained in terms of **recurring results** that can be generated by **different generators**.
  - Stability = **Relative frequency of different results** that complies with IP conditions.

- **Differential Privacy :**
  - With respect to neighboring datasets the result of given function is **not largely affected by the presence or absence of a particular data record**.

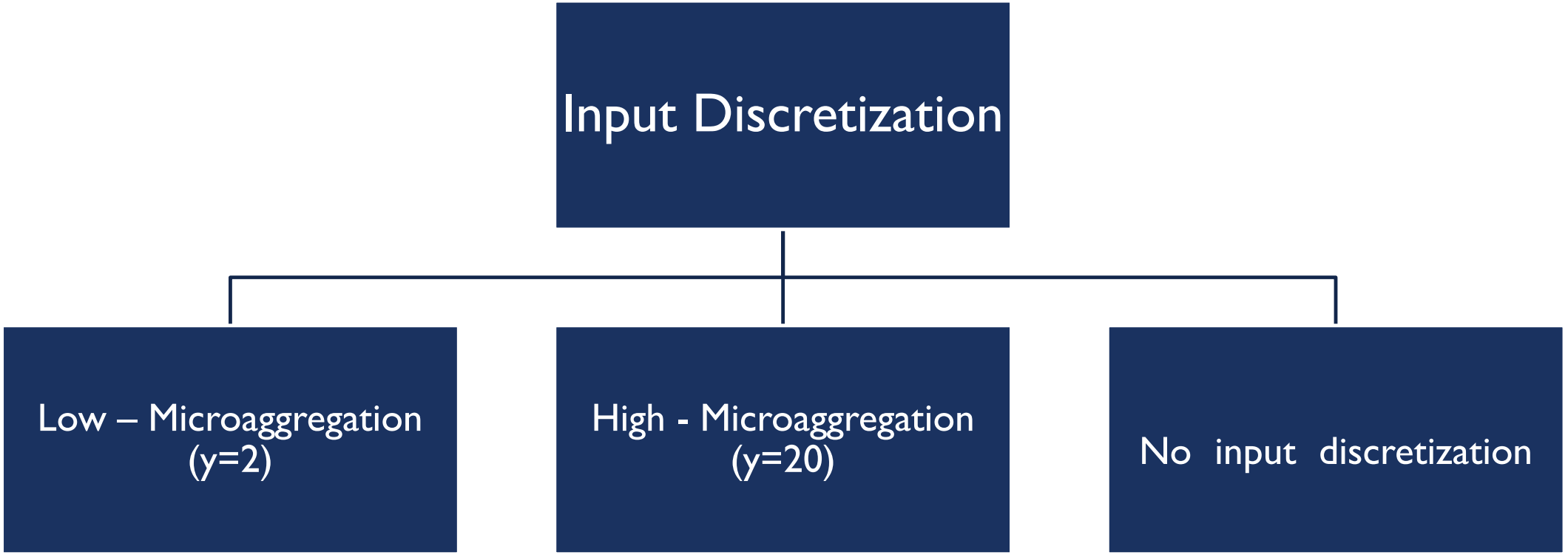
MOTIVATION

- To adopt the **notion of stability presented in IP** in the context of descriptive statistics computation?
  - Mean, median, IQR, standard deviation, variance, count, sum, min and max
  - Achieved through resampling and discretization based method.
- Can it be used to address the limitations of Differential Privacy?

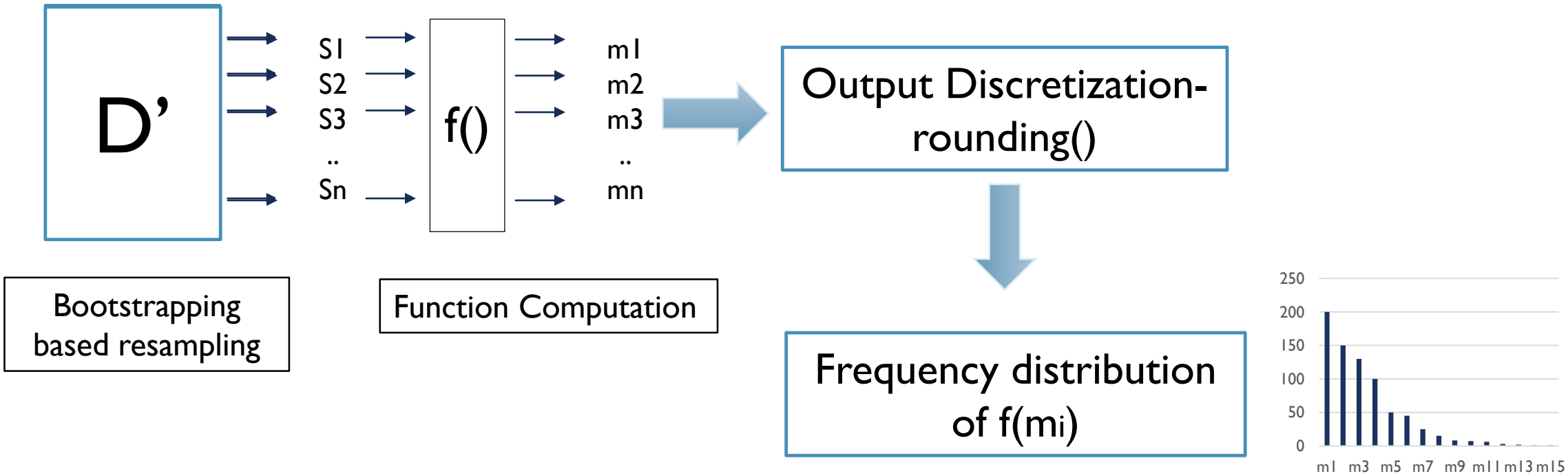
METHODOLOGY



INPUT DISCRETIZATION



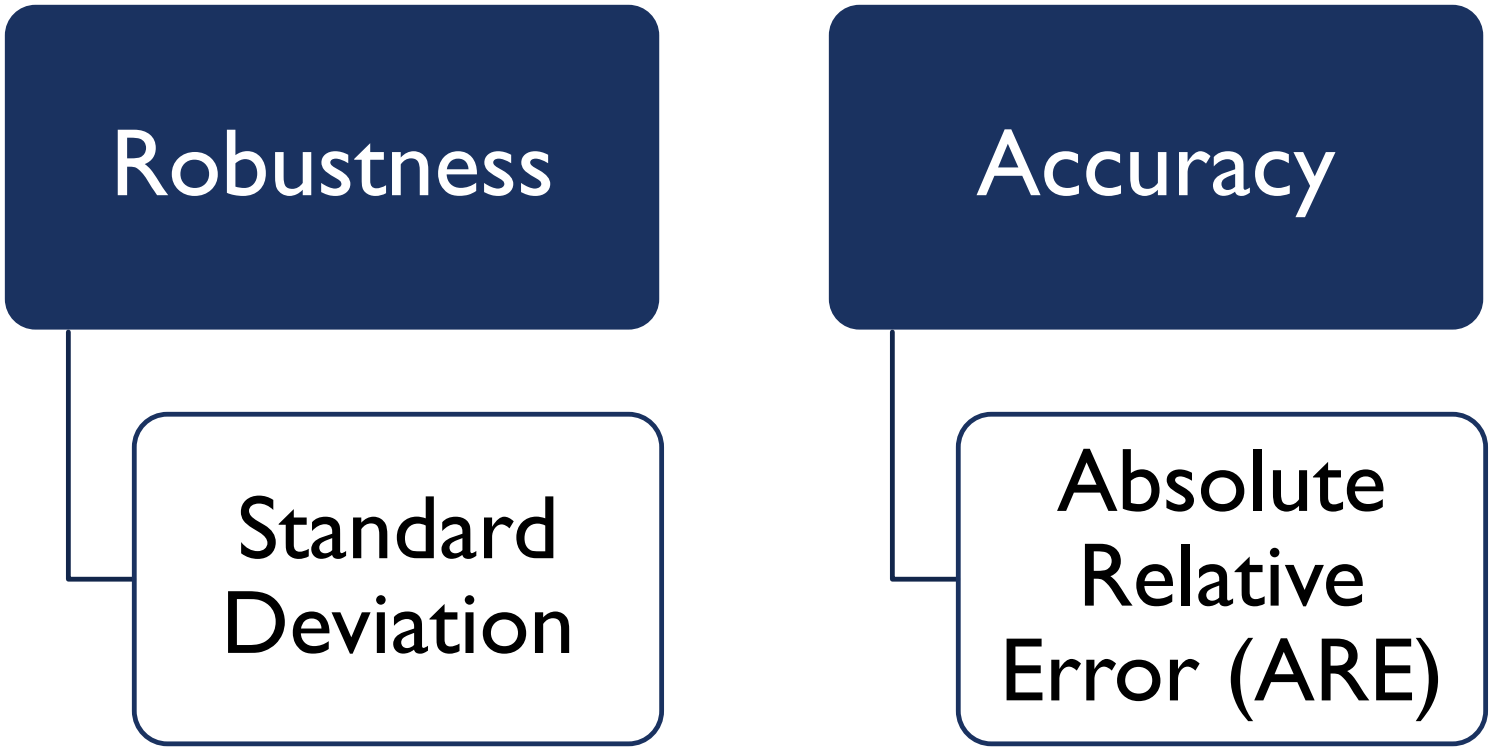
RESAMPLING, OUTPUT DISCRETIZATION AND FREQUENCY DISTRIBUTION



INTEGRAL PRIVACY CONDITIONS

- From the **"Distribution of Results"** select the results with a **frequency of occurrence >= k**
- From the selected results filter the ones with **no intersection among their generators; = "Candidate Results"**
- If **multiple "Candidate Results"** are available select the final result which has,
    - Highest Accuracy → high utility
    - Highest Frequency → high privacy

EVALUATION CRITERIA

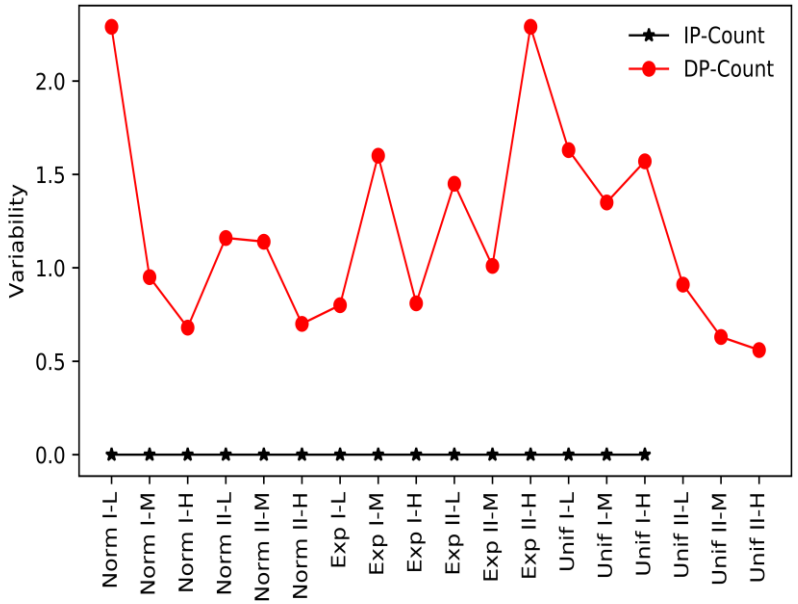


DATA

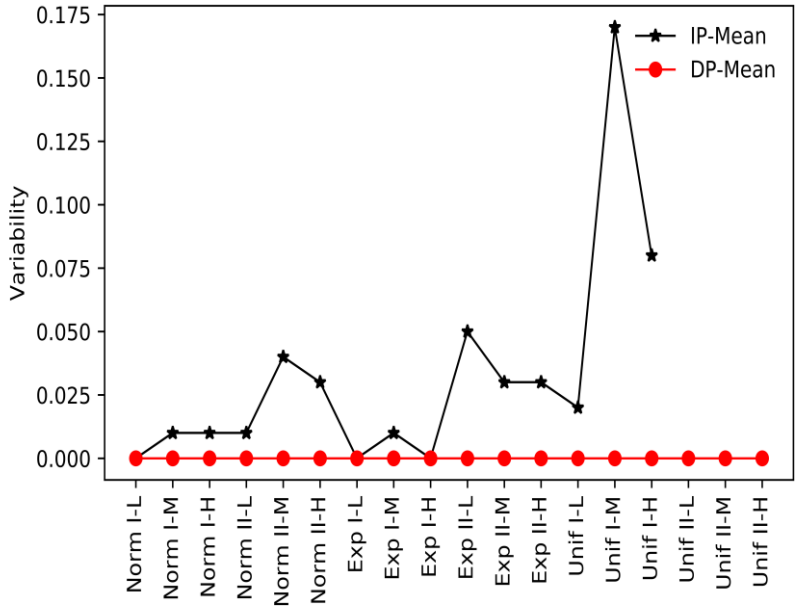
Dataset	Instances × Columns	Description
Norm I	1000×1	Normally distributed with $\mu = 1, \sigma = 1$
Norm II	1000×1	Normally distributed with $\mu = 1, \sigma = 5$
Exp I	1000×1	Exponentially distributed with $\lambda = 1$
Exp II	1000×1	Exponentially distributed with $\lambda = 0.2$
Unif I	1000×1	Uniformly distributed in range (min=0,max=100)
Unif II	1000×1	Uniformly distributed in range (min=0,max=1000)
Abalone Dataset	4177×8	UCI data repository
Breast Cancer	683×7	UCI data repository

THEORETICAL DISTRIBUTIONS

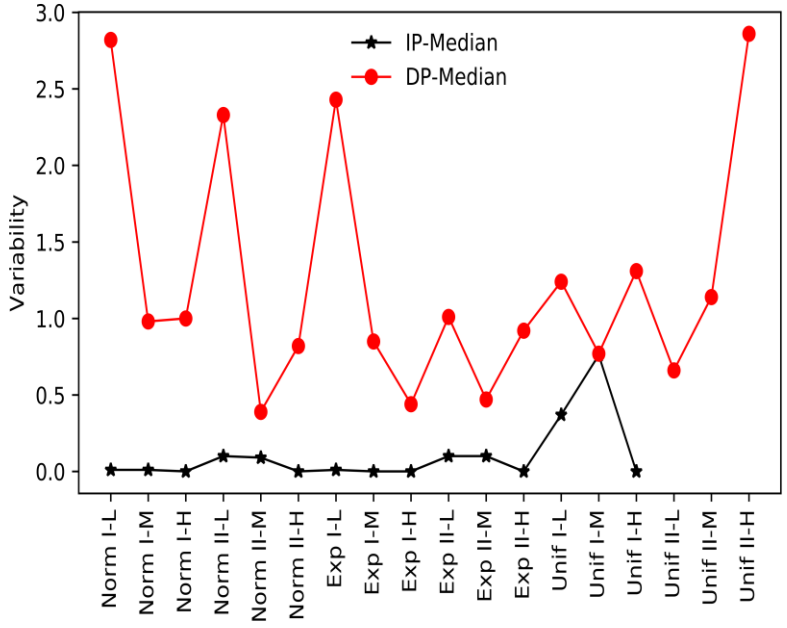
ROBUSTNESS OF THE RESULTS



1. Count

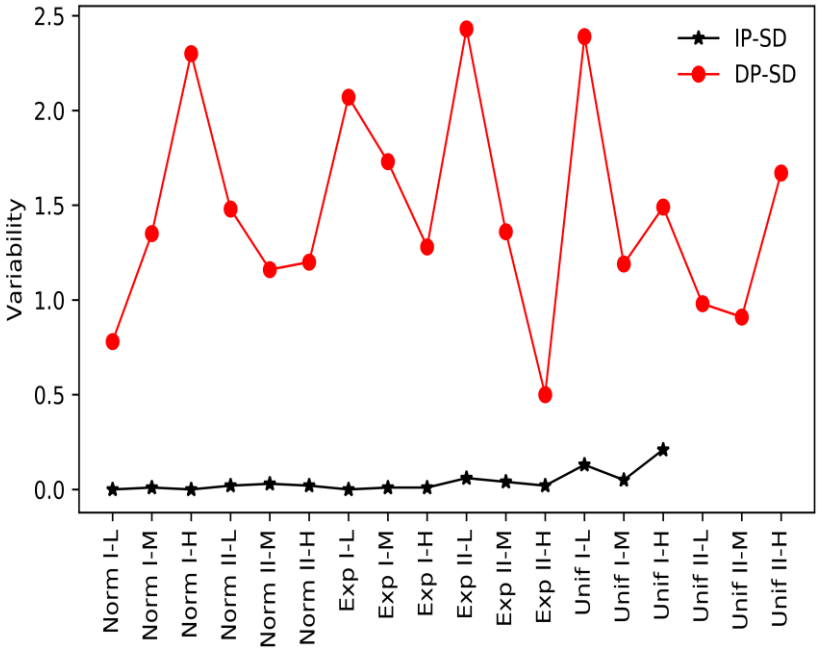


2. Mean

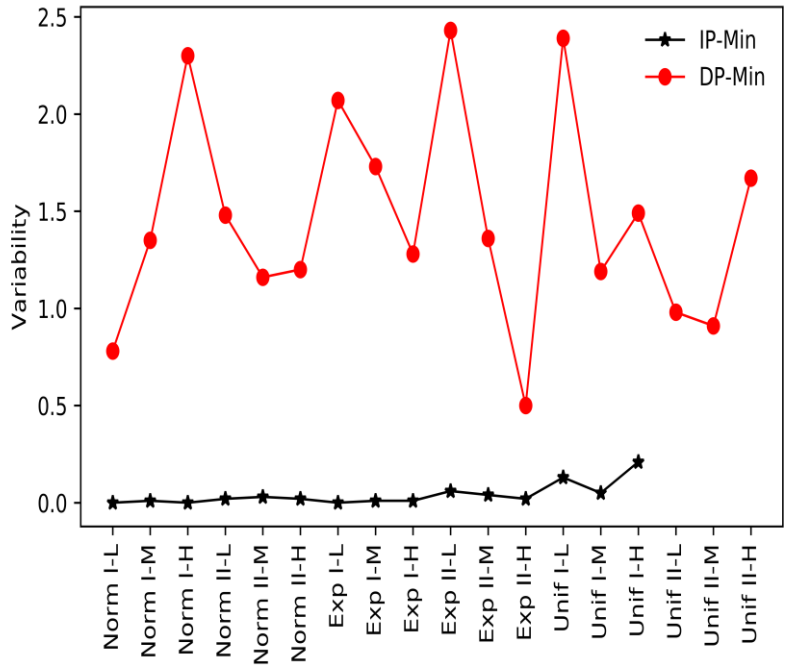


3. Median

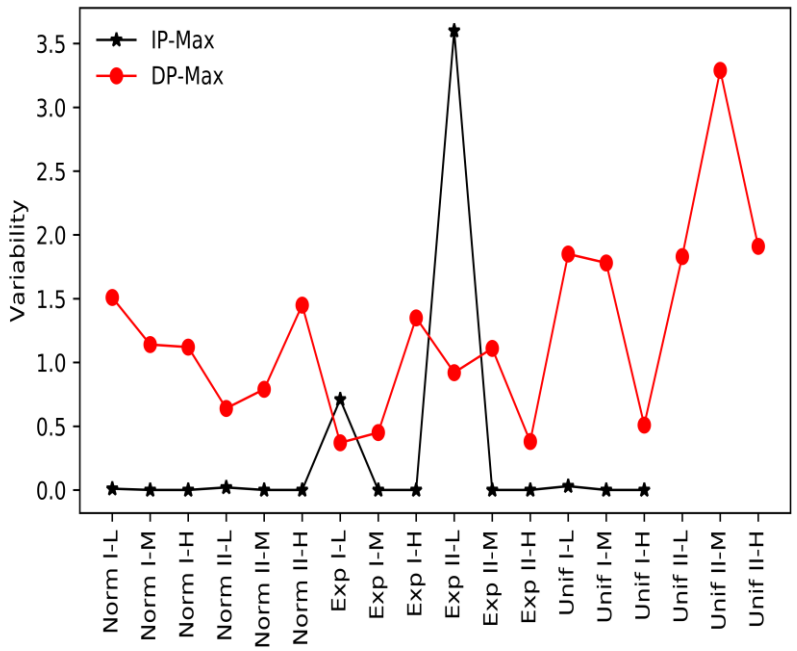
ROBUSTNESS OF THE RESULTS CONT.



4. SD

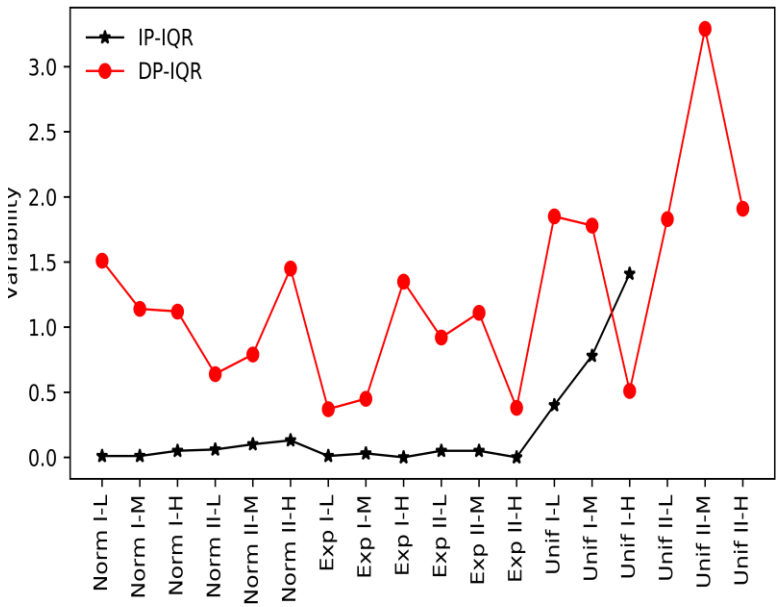


5. Min

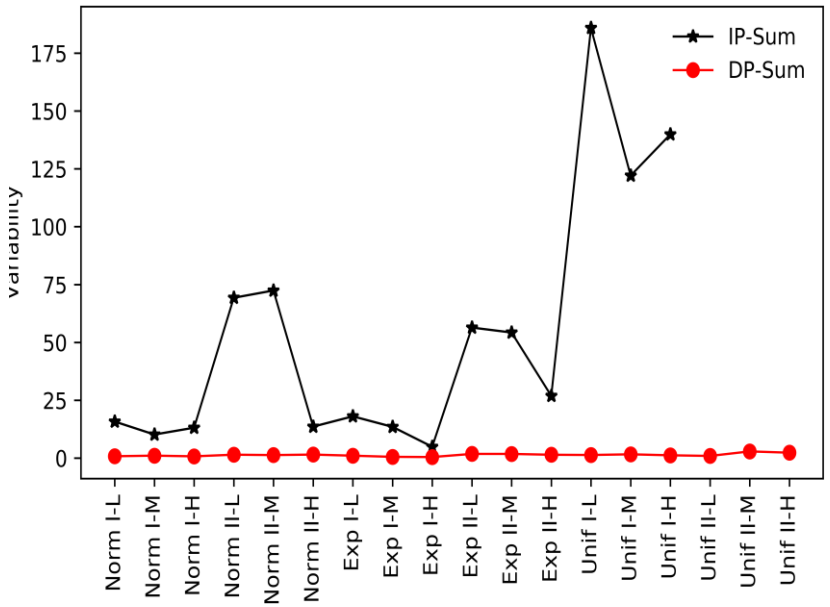


6. Max

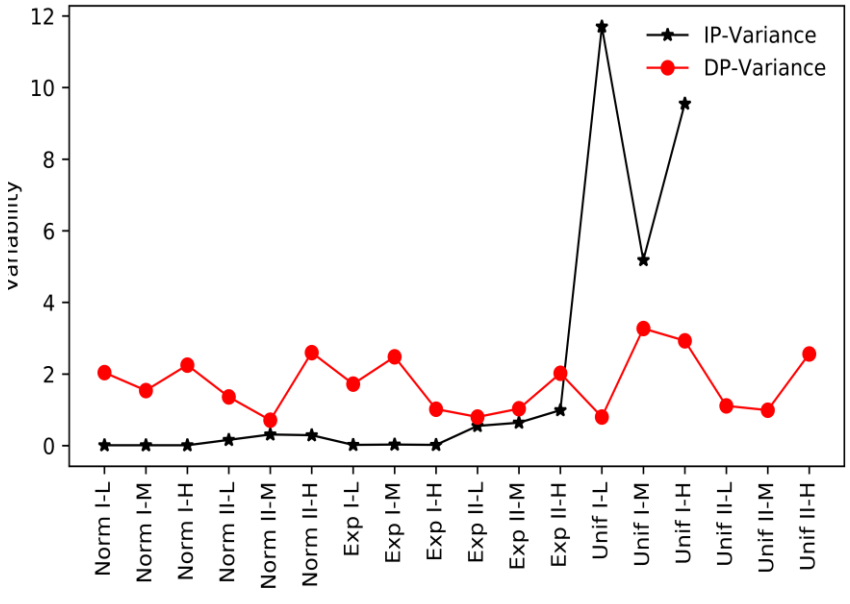
ROBUSTNESS OF THE RESULTS CONT.



7. IQR



8. Sum



9. Variance

ACCURACY - ABSOLUTE RELATIVE ERROR (ARE)

Dataset	Count-IP	Count-DP
Norm I Out Dis:	0	0.1
Norm I in/out Dis:(L)	0	0
Norm I in/out Dis:(H)	0	0
Norm II Out Dis:	0	0.1
Norm II in/out Dis:(L)	0	0
Norm II in/out Dis:(H)	0	0
Exp I Out Dis:	0	0.1
Exp I in/out Dis:(L)	0	0
Exp I in/out Dis:(H)	0	0
Exp II Out Dis:	0	0.1
Exp II in/out Dis:(L)	0	0
Exp II in/out Dis:(H)	0	0
Unif I Out Dis:	0	0.1
Unif I in/out Dis:(L)	0	0
Unif I in/out Dis:(H)	0	0
Unif II Out Dis:	0	0.1
Unif II in/out Dis:(L)	0	0
Unif II in/out Dis:(H)	0	0

1. Count

Dataset	Mean-IP	Mean-DP
Norm I Out Dis:	0	0.43
Norm I in/out Dis:(L)	0	1
Norm I in/out Dis:(H)	0	1
Norm II Out Dis:	0.01	2.42
Norm II in/out Dis:(L)	0	0.94
Norm II in/out Dis:(H)	0	0.95
Exp I Out Dis:	0	0.16
Exp I in/out Dis:(L)	0	1.02
Exp I in/out Dis:(H)	0	1.03
Exp II Out Dis:	0.01	1.34
Exp II in/out Dis:(L)	0.01	5.11
Exp II in/out Dis:(H)	0.02	5.12
Unif I Out Dis:	0.06	39.13
Unif I in/out Dis:(L)	0.06	48.73
Unif I in/out Dis:(H)	0.12	48.75
Unif II Out Dis:	0.94	373.32
Unif II in/out Dis:(L)	2.63	469.02
Unif II in/out Dis:(H)	0.89	469.26

2. Mean

Dataset	Median-IP	Median-DP
Norm I Out Dis:	0.01	0.44
Norm I in/out Dis:(L)	0.01	0.38
Norm I in/out Dis:(H)	0	0.63
Norm II Out Dis:	0.03	0.58
Norm II in/out Dis:(L)	0.04	0.53
Norm II in/out Dis:(H)	0.09	0.33
Exp I Out Dis:	0.01	0.1
Exp I in/out Dis:(L)	0.01	0.57
Exp I in/out Dis:(H)	0	0.66
Exp II Out Dis:	0.07	0.18
Exp II in/out Dis:(L)	0.04	0.01
Exp II in/out Dis:(H)	0.09	0.78
Unif I Out Dis:	0.05	6.38
Unif I in/out Dis:(L)	0.17	0.57
Unif I in/out Dis:(H)	0.41	0.04
Unif II Out Dis:	3.22	111.39
Unif II in/out Dis:(L)	3.11	0.05
Unif II in/out Dis:(H)	10.22	2.77

3. Median

ACCURACY CONT.

Dataset	SD-IP	SD-DP
Norm I Out Dis:	0	19.03
Norm I in/out Dis:(L)	0.01	0.2
Norm I in/out Dis:(H)	0	0.9
Norm II Out Dis:	0.01	112.32
Norm II in/out Dis:(L)	0.01	0.54
Norm II in/out Dis:(H)	0.03	0.59
Exp I Out Dis:	0.01	29.73
Exp I in/out Dis:(L)	0.01	0.24
Exp I in/out Dis:(H)	0.01	0.14
Exp II Out Dis:	0.01	123.7
Exp II in/out Dis:(L)	0.01	0.19
Exp II in/out Dis:(H)	0	0.99
Unif I Out Dis:	0.03	320.44
Unif I in/out Dis:(L)	0.09	2.11
Unif I in/out Dis:(H)	0.07	1.2
Unif II Out Dis:	0.18	3193.41
Unif II in/out Dis:(L)	0.4	16.21
Unif II in/out Dis:(H)	0.5	9.35

4. SD

Dataset	Min-IP	Min-DP
Norm I Out Dis:	0.05	5.56
Norm I in/out Dis:(L)	0.05	0.33
Norm I in/out Dis:(H)	0.28	0.01
Norm II Out Dis:	1.09	272.23
Norm II in/out Dis:(L)	1.71	0.82
Norm II in/out Dis:(H)	3.82	0.85
Exp I Out Dis:	0	0.04
Exp I in/out Dis:(L)	0	0.32
Exp I in/out Dis:(H)	0.01	0.1
Exp II Out Dis:	0	0.53
Exp II in/out Dis:(L)	0	0.52
Exp II in/out Dis:(H)	0.03	0.35
Unif I Out Dis:	0.03	4.15
Unif I in/out Dis:(L)	0.03	0.01
Unif I in/out Dis:(H)	0.86	0.21
Unif II Out Dis:	0.03	6.62
Unif II in/out Dis:(L)	0.01	0.31
Unif II in/out Dis:(H)	2.58	0.11

5. Min

Dataset	Max-IP	Max-DP
Norm I Out Dis:	0	0.97
Norm I in/out Dis:(L)	0.01	0.02
Norm I in/out Dis:(H)	0.15	0.23
Norm II Out Dis:	0.01	48.09
Norm II in/out Dis:(L)	0.01	0.23
Norm II in/out Dis:(H)	0.97	0.1
Exp I Out Dis:	0.08	39.72
Exp I in/out Dis:(L)	0.2	0.09
Exp I in/out Dis:(H)	0.92	0.02
Exp II Out Dis:	1.61	201.52
Exp II in/out Dis:(L)	1.01	1
Exp II in/out Dis:(H)	3.48	0.68
Unif I Out Dis:	0.01	1.85
Unif I in/out Dis:(L)	0.01	0.03
Unif I in/out Dis:(H)	0.22	0.11
Unif II Out Dis:	0.03	27.34
Unif II in/out Dis:(L)	0.12	0.46
Unif II in/out Dis:(H)	2.39	0.05

6. Max

ACCURACY CONT.

Dataset	IQR-IP	IQR-DP
Norm I Out Dis:	0	4.89
Norm I in/out Dis:(L)	0.01	1.94
Norm I in/out Dis:(H)	0.02	2.68
Norm II Out Dis:	0	151.27
Norm II in/out Dis:(L)	0.02	8.26
Norm II in/out Dis:(H)	0.04	7.88
Exp I Out Dis:	0	124.53
Exp I in/out Dis:(L)	0	6.12
Exp I in/out Dis:(H)	0.02	5.79
Exp II Out Dis:	0.01	627.16
Exp II in/out Dis:(L)	0.01	27.97
Exp II in/out Dis:(H)	0.12	27
Unif I Out Dis:	0.09	41.6
Unif I in/out Dis:(L)	0.15	36.15
Unif I in/out Dis:(H)	0.38	35.74
Unif II Out Dis:	0.23	420.31
Unif II in/out Dis:(L)	4.54	339.99
Unif II in/out Dis:(H)	5.14	338.46

7. IQR

Dataset	Sum-IP	Sum-DP
Norm I Out Dis:	0.35	0.39
Norm I in/out Dis:(L)	0.35	0
Norm I in/out Dis:(H)	0.35	0
Norm II Out Dis:	0.3	1.6
Norm II in/out Dis:(L)	0.32	0.01
Norm II in/out Dis:(H)	0.27	0
Exp I Out Dis:	0.36	0.87
Exp I in/out Dis:(L)	0.37	0
Exp I in/out Dis:(H)	0.37	0
Exp II Out Dis:	1.79	3.79
Exp II in/out Dis:(L)	1.79	0.02
Exp II in/out Dis:(H)	1.82	0.01
Unif I Out Dis:	NA	9.64
Unif I in/out Dis:(L)	NA	0.05
Unif I in/out Dis:(H)	NA	0.02
Unif II Out Dis:	NA	96.24
Unif II in/out Dis:(L)	NA	0.48
Unif II in/out Dis:(H)	NA	0.24

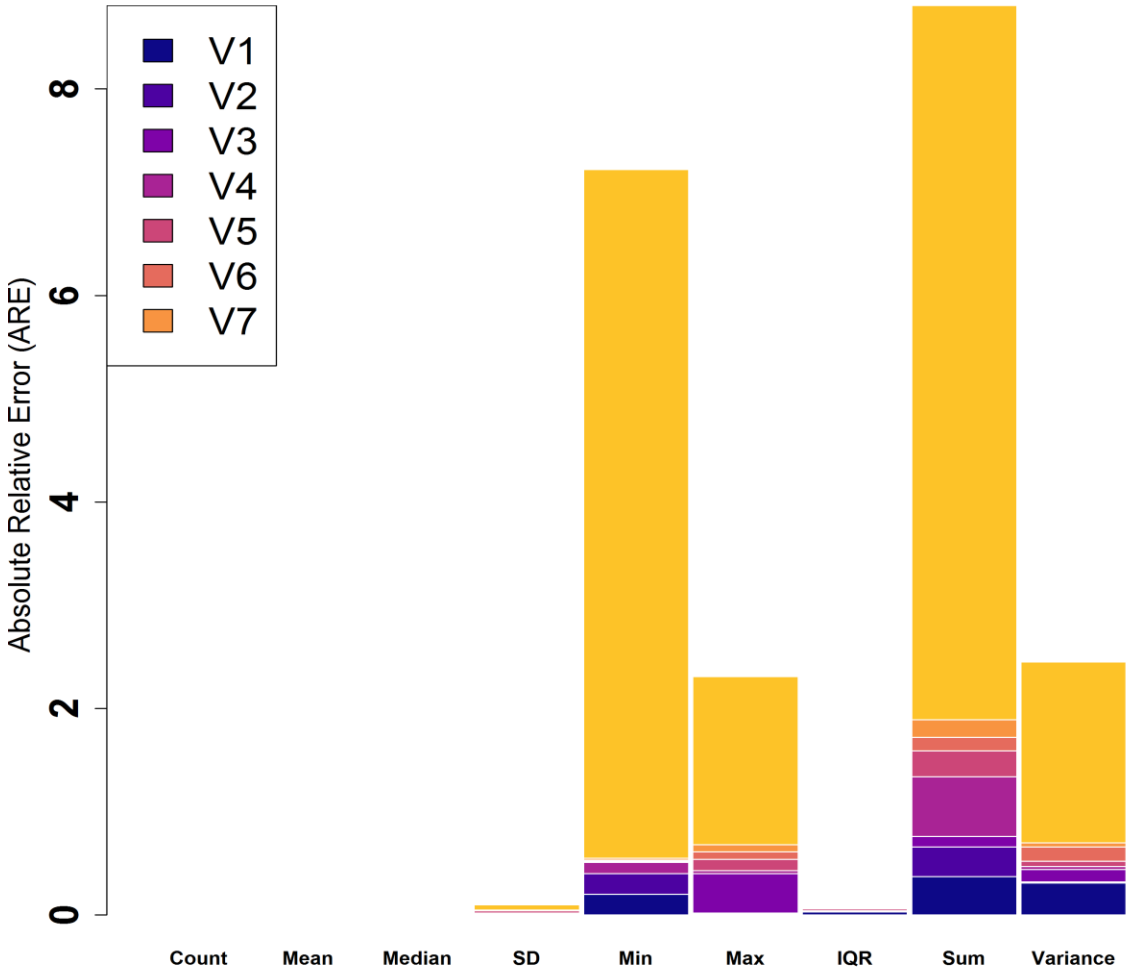
8. Sum

Dataset	Variance-IP	Variance-DP
Norm I Out Dis:	0.01	4.76
Norm I in/out Dis:(L)	0.01	1
Norm I in/out Dis:(H)	0.01	1.48
Norm II Out Dis:	0.04	126.32
Norm II in/out Dis:(L)	0.1	0.99
Norm II in/out Dis:(H)	0.21	0.89
Exp I Out Dis:	0.02	7.28
Exp I in/out Dis:(L)	0.02	1.47
Exp I in/out Dis:(H)	0.04	0.68
Exp II Out Dis:	0.14	158.97
Exp II in/out Dis:(L)	0.32	0.54
Exp II in/out Dis:(H)	0	1.65
Unif I Out Dis:	5.54	1024.67
Unif I in/out Dis:(L)	4.5	4.32
Unif I in/out Dis:(H)	1.79	3.41
Unif II Out Dis:	NA	101990.65
Unif II in/out Dis:(L)	NA	510.8
Unif II in/out Dis:(H)	NA	256.13

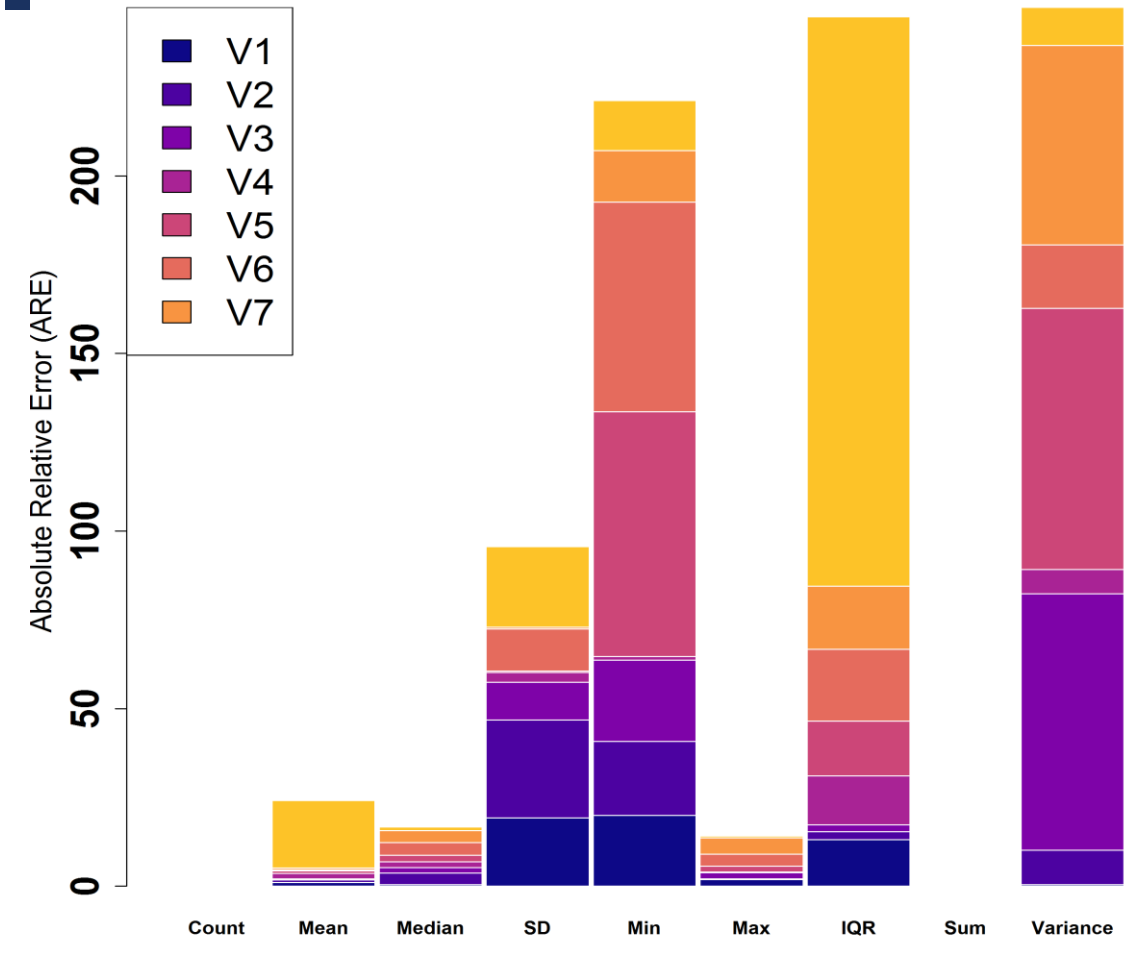
9. Variance

REAL WORLD DATASETS

ABALONE DATA

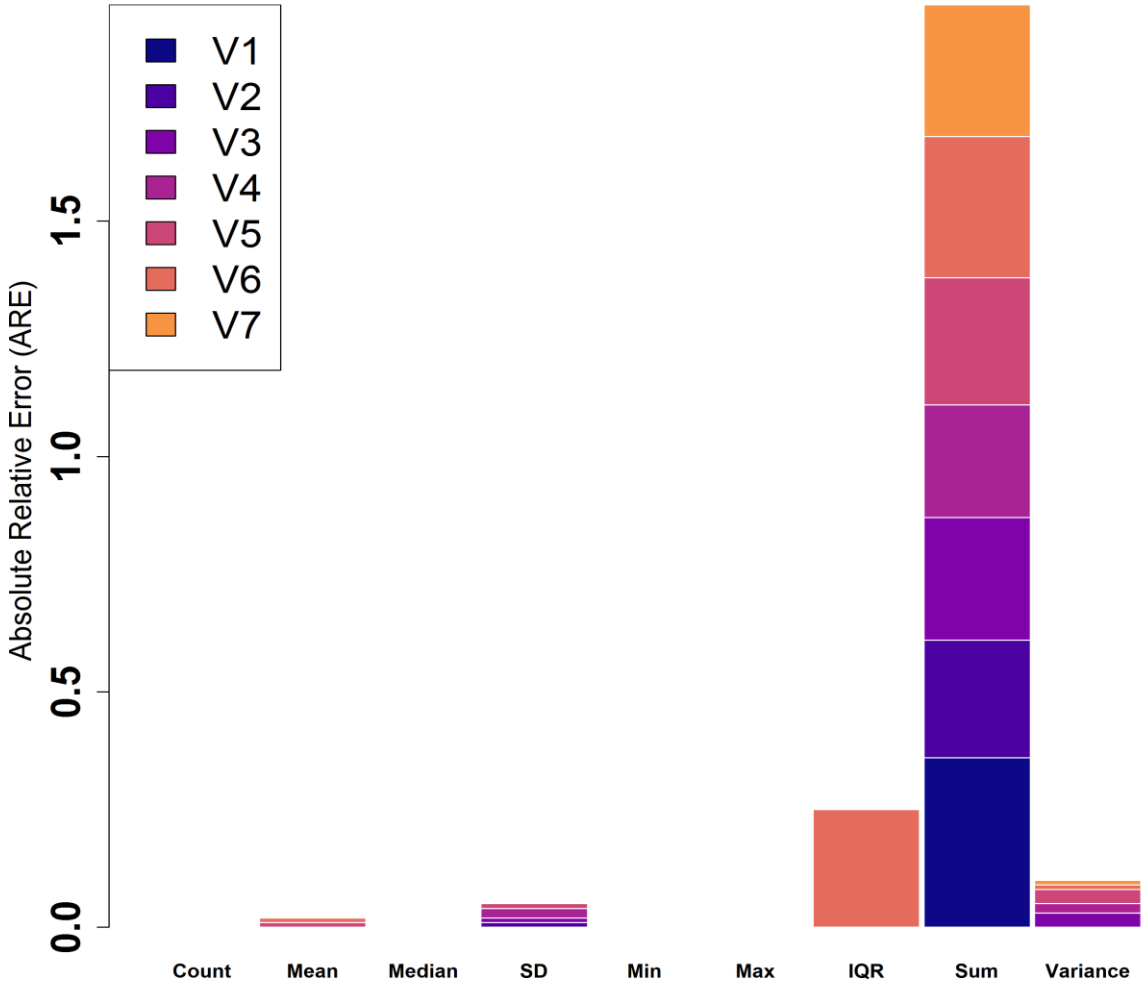


Integral Privacy (k=highest)

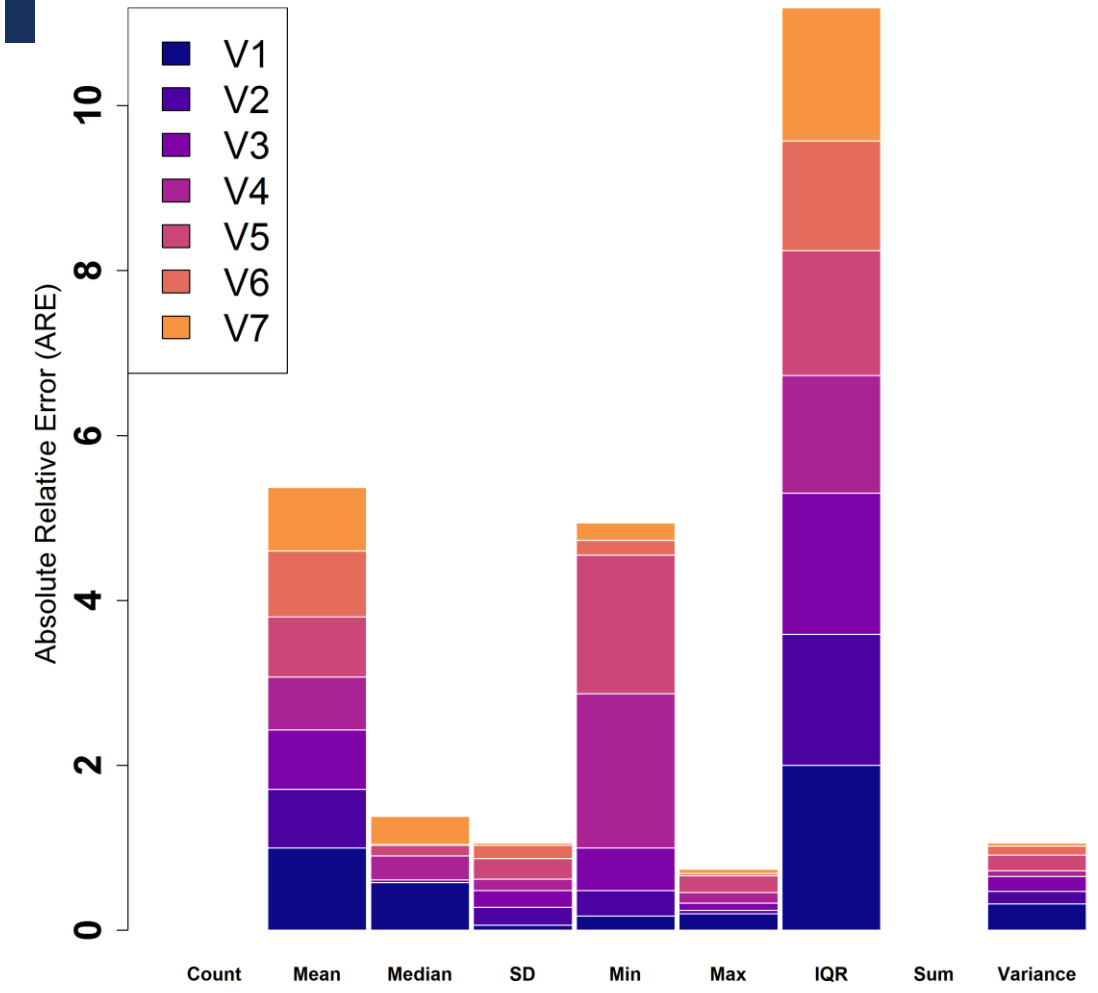


Differential Privacy (ε=4)

BREAST CANCER DATA



Integral Privacy (k=highest)



Differential Privacy (ε=4)

LIMITATIONS

- Not suitable for all the descriptive statistics i.e., sum
- IP answers might not be available with every set of parameters selected (number of resamples, input discretization params, output discretization params, frequency threshold).
- Computational complexity might limit its applicability w.r.t large datasets.

CONCLUSION AND FUTURE WORK

- High accuracy and robustness of the IP results compared to DP in most of the instances.
- Appropriate in the context of small datasets where DP does not perform well.
- Only output discretization is *sufficient* to generate the IP compliant results.
- Efficient resampling methods are required.

Thank You!

Q/A

REFERENCES

1. Vicenc Torra and Guillermo Navarro-Arribas. Integral privacy. In Sara Foresti and Giuseppe Persiano, editors, Cryptology and Network Security, pages 661{669,Cham, 2016. Springer International Publishing.

2. Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography Conference, pages 265{284. Springer, 2006.

3. Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the sulq framework. In Proceedings of the twenty-fourth ACM SIGMOD- SIGACT-SIGART symposium on Principles of database systems, pages 128{138.ACM, 2005.

4. Cynthia Dwork and Adam Smith. Dierential privacy for statistics:What we know and what we want to learn. Journal of Privacy and Condentiality, 1(2), 2010.

5. Cynthia Dwork and Jing Lei. Dierential privacy and robust statistics. In STOC, volume 9, pages 371{380, 2009.

6. Chris Clifton and Tamir Tassa. On syntactic anonymity and dierential privacy. In 2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW), pages 88{93. IEEE, 2013.