# Validation Free and Replication Robust Volume-based Data Valuation

Xinyi Xu*[1], Zhaoxuan Wu*[2], Chuan Sheng Foo[3], Bryan Kian Hsiang Low[1]

Department of Computer Science, National University of Singapore[1]
Institute of Data Science, National University of Singapore[2]
Institute of Infocomm Research, Agency for Science, Technology and Research (A*STAR)[3]

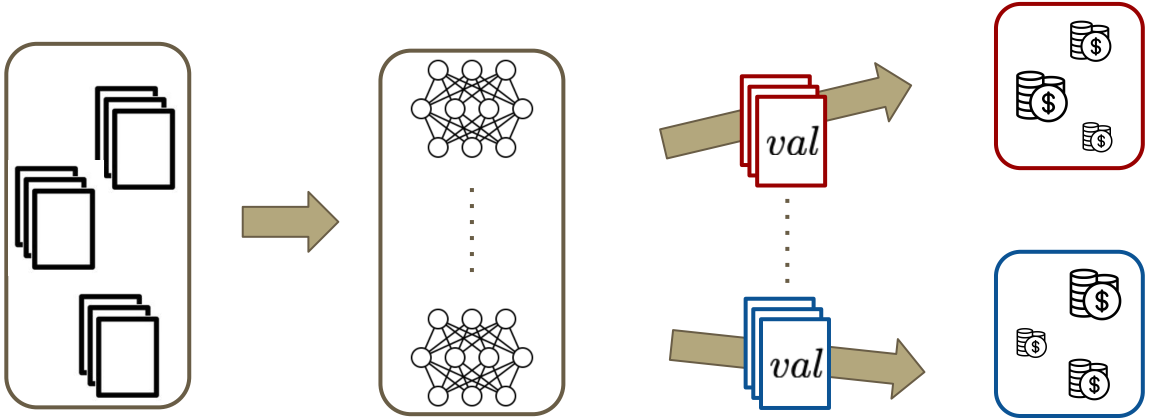
NEURAL INFORMATION PROCESSING SYSTEMS

NUS National University of Singapore

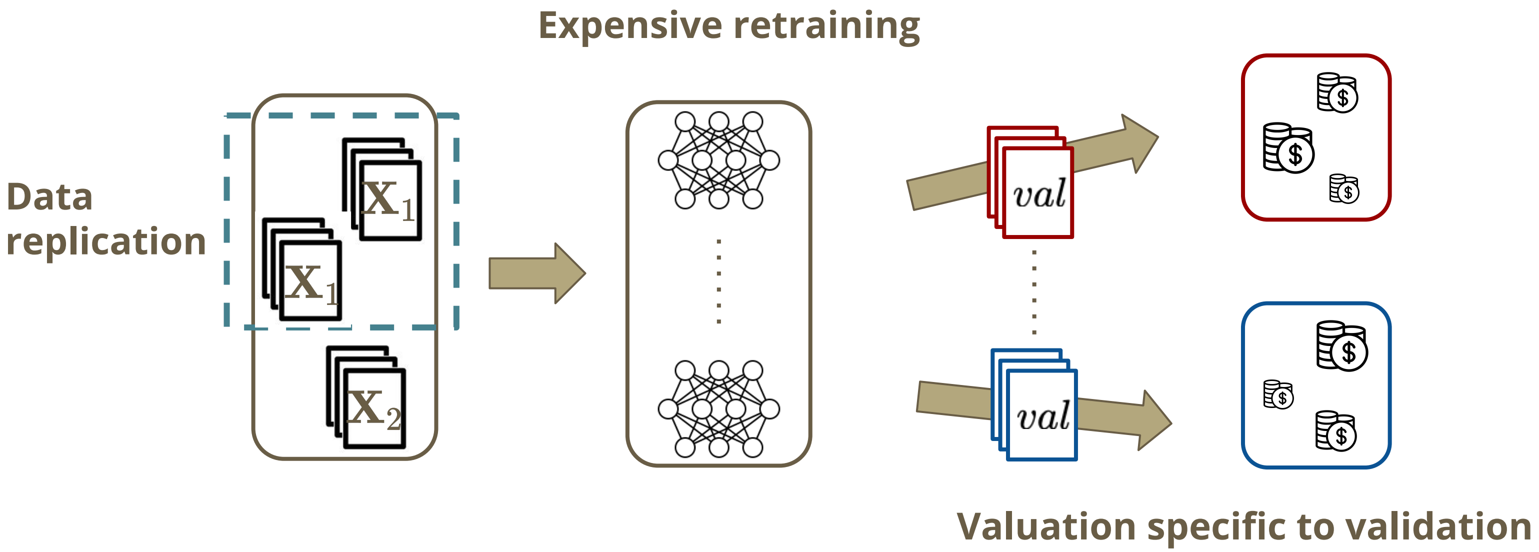
NUS National University of Singapore Institute of Data Science

SINGAPORE Agency for Science, Technology and Research

# Data Valuation in Machine Learning

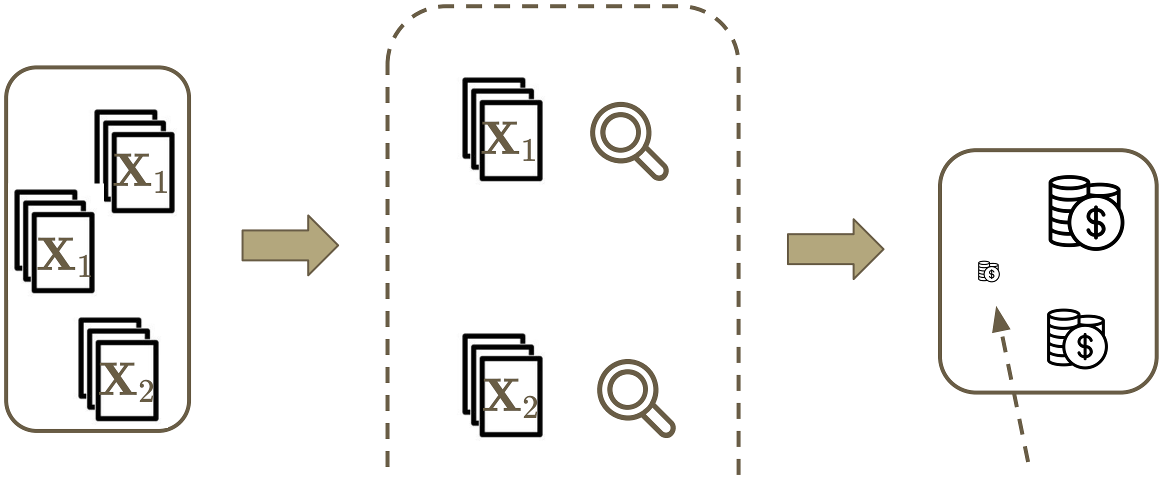


# Motivation & Goal



# Data Valuation via Diversity of Data

***Better diversity in data can result in better learning performance.***




**Examine diversity in the data**

Insignificant value due to
a lack of increase in diversity

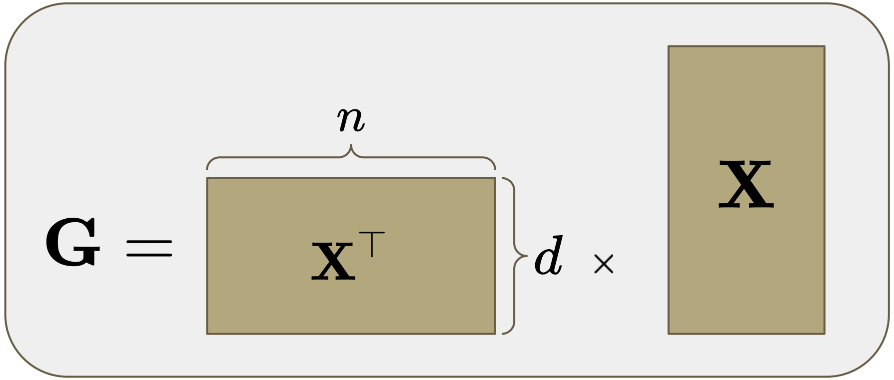
# Data Valuation via Diversity of Data

***Better diversity in data can result in better learning performance.***

- Intuition
  - More ***inherent diversity*** in data → better generalizability of learner → higher value.
- Connection between the determinant of data matrix and diversity.
  - Determinantal Point Processes (DPPs) [1]
  - Geometric interpretation
- Interestingly, we also eliminate the need for a validation when using diversity.

[1] Alex Kulesza. “Determinantal Point Processes for Machine Learning”. In: Foundations and Trends® in Machine Learning 5.2-3 (2012), pp. 123–286.

# Data diversity via *Volume*

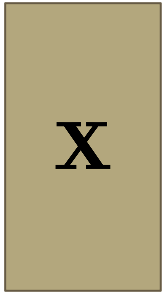


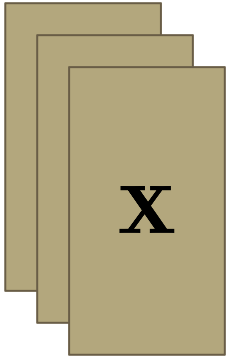

Definition 1 (Volume)

$\mathbf{X} \in \mathbb{R}^{n\times d}, \mathrm{Vol}(\mathbf{X}) := \sqrt{|(\mathbf{X}^\top \mathbf{X})|} = \sqrt{|\mathbf{G}|}$

- Higher volume (diversity) ⇔ better learning performance ⇔ higher value.
  - Larger volume ⇔ more accurate pseudo-inverse (Propositions. 1,2).
  - Larger volume ⇔ lower mean squared error (MSE) for $d = 1$ (Proposition. 3).
- Additional properties: (Proposition. 4)
  - Non-negativity
  - Monotonicity (Lemma. 1)

# Data Replication

- Suppose the value of 

is $\nu(\mathbf{X})$, what should be the value of 

?

- If data replication via direct copying ***<u>strictly increases</u>*** the total value, then a dishonest data provider may exploit the valuation method by replication.

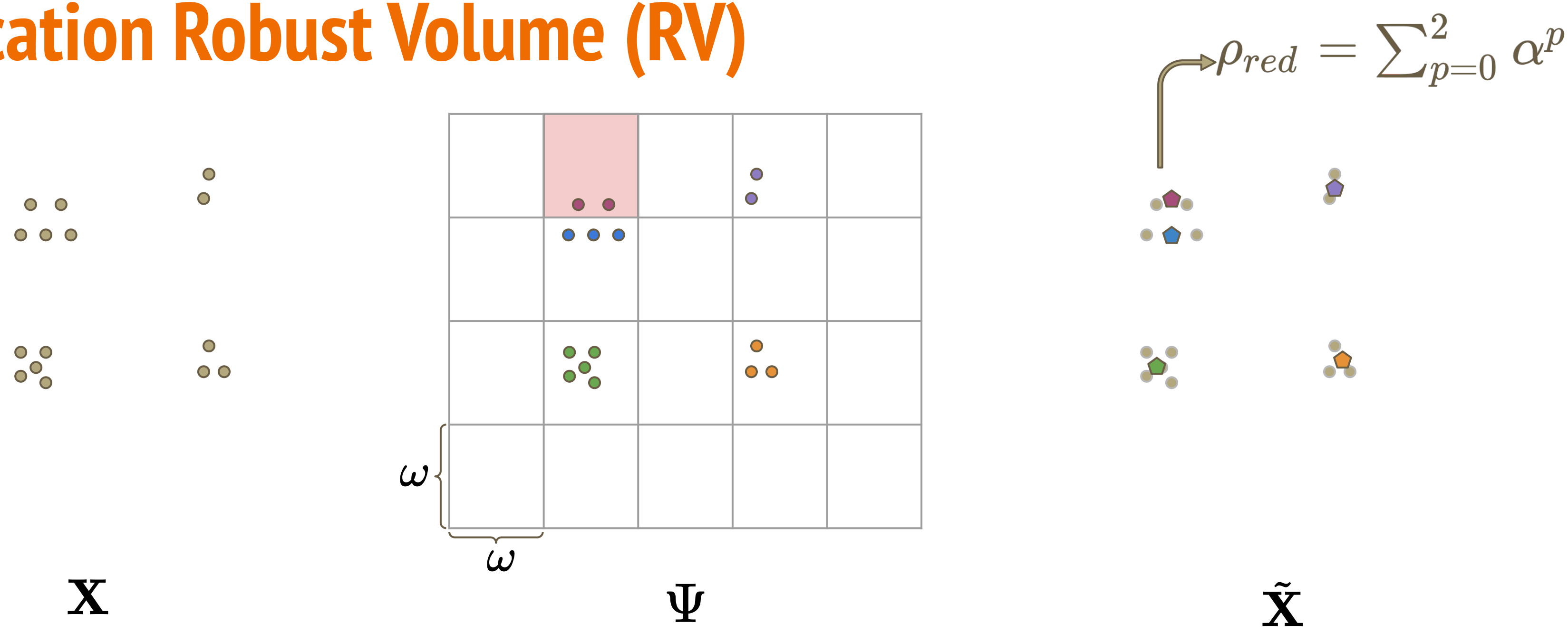
# Replication Robust Volume (RV)

- Propose a robust definition to balance the value of diversity and repetition.
  - Construct a 'compressed' version of the original data matrix $\mathbf{X}$ by grouping and representing data points via discretized cubes of the input space.

$$\mathrm{RV}(\mathbf{X};\omega) := \mathrm{Vol}(\tilde{\mathbf{X}}) \times \prod_{i\in\Psi} \rho_i, \text{ where } \rho_i := \sum_{p=0}^{\phi_i} \alpha^p, \alpha \in (0,1).$$

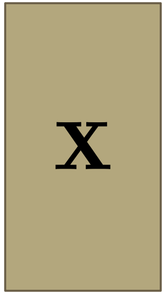
  - Discretize the input domain with a coefficient $\omega$.
  - For each discretized cell $i \in \Psi$,
    - Compute a statistic (e.g. mean) for all data points in it and use it to construct $\tilde{\mathbf{X}}$.
    - Count the number of data points in it, $\phi_i$, and use it to compute the multiplicative coefficient $\rho_i$.

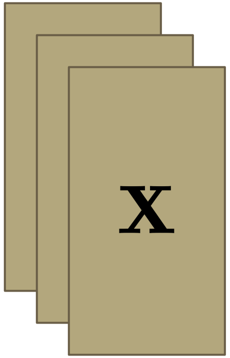
# Replication Robust Volume (RV)




$$\mathrm{RV}(\mathbf{X};\omega) := \mathrm{Vol}(\tilde{\mathbf{X}}) \times \prod_{i\in\Psi} \rho_i, \text{ where } \rho_i := \sum_{p=0}^{\phi_i} \alpha^p, \alpha \in (0,1).$$

# Replication Robustness Defined via Inflation

- Suppose the value of 

is $\nu(\mathbf{X})$, and be value of 

is $\nu(\mathbf{X}, 3)$.

- Define inflation caused by replication of $c$ times as: $\text{inflation}(\mathbf{X}, c) = \frac{\nu(\mathbf{X}, c)}{\nu(\mathbf{X})}$.

- Define replication robustness as: $\gamma_\nu = \frac{\nu(\mathbf{X})}{\sup_{c \geq 1} \nu(\mathbf{X}, c)}$ .

***High robustness should curb inflation from replication.***

# Replication Robust Volume (RV)

- RV is robust (Proposition. 6).
  - $\gamma_{\text{RV}} \geq (1-\alpha)^{|\Psi|}$
- RV is flexible between $\gamma = 0$ and the optimal $\gamma = 1$(Proposition. 7)
- RV is similar to the original volume formulation in terms of relative values (Proposition. 5).
  - High RV indicates high diversity and thus better learning performance.

# Experiments

1. Validating volume/robust volume is a good measure for learning performance via diversity

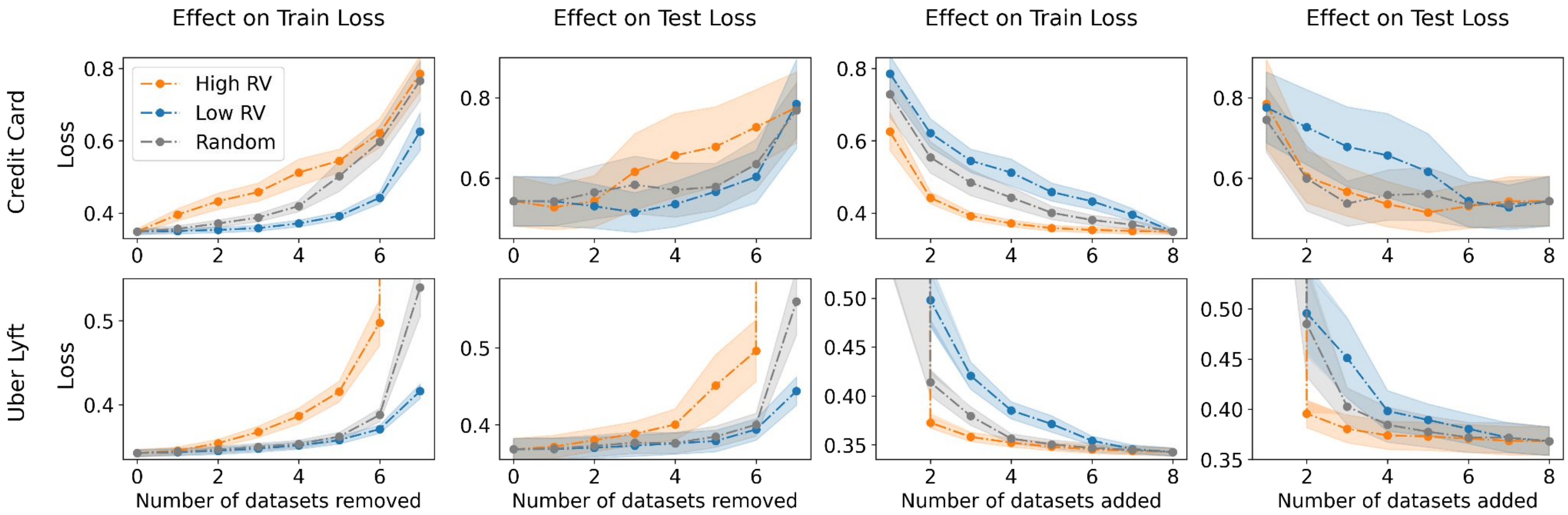
2. Demonstrating RV produces consistent valuation with existing baselines, *without requiring validation*

3. Replication robustness

# Experiments - High RV means High performance

- Datasets:
  - Credit card transaction prediction (8);
  - Uber & Lyft carpool ride price prediction (12);
  - UK Used Car price prediction (5);
  - TripAdvisor Hotel review rating prediction (8).
  - The numbers represent the dimension of the standardized features.
- 8 data providers, so 8 data submatrices.
- Setting: we gradually add/remove the submatrices one at a time and monitor the performance of the current learner.
- Ordering: highest RV first, lowest RV first and random.

# Experiments - High RV means High performance




- *Removing* high RV data *increases* both train/test losses quickly.
- *Adding* high RV data *reduces* both train/test losses quickly.

# Experiments - RV Shapley Value v.s. Baselines

- For a fair comparison, we extend RV to Shapley formulation

$$\mathrm{RVSV}_m = \frac{1}{M!}\sum_{\mathcal{C}\subseteq\mathcal{M}\setminus\{S_m\}}[|\mathcal{C}|!\times(M-|\mathcal{C}|-1)!]\times[\mathrm{RV}(\mathbf{X}_{\mathcal{C}\cup\{S_m\}};\omega)-\mathrm{RV}(\mathbf{X}_{\mathcal{C}};\omega)]$$

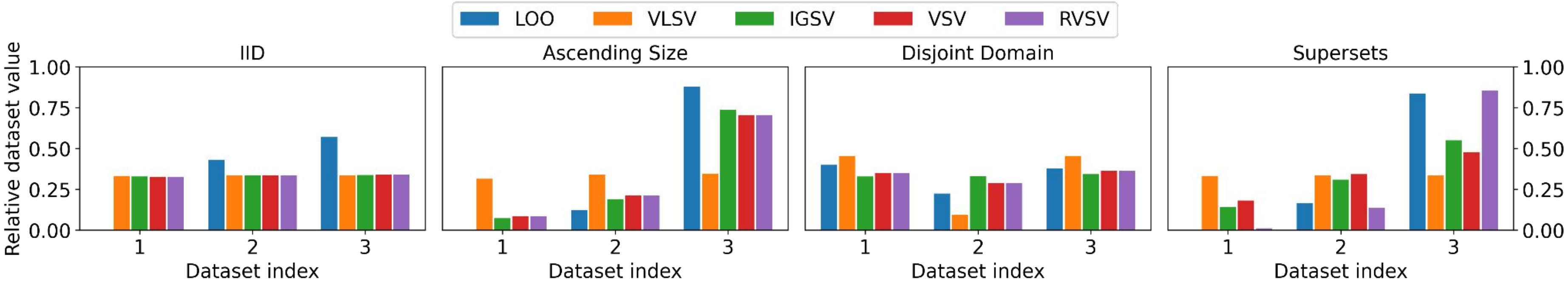
where $\mathcal{C}\subseteq\mathcal{M}:=\{S_1,\ldots,S_M\}$

- We compare with
  - Leave-One-Out (LOO) value
  - Validation Loss Shapley Value (VLSV)
  - Information Gain Shapley Value (IGSV) [2]

[2] Rachael Hwee Ling Sim, Yehong Zhang, Mun Choon Chan, and Bryan Kian Hsiang Low. Collaborative machine learning with incentive-aware model rewards. In Proc. ICML, pages 8927–8936, 2020.

# Experiments - RV Shapley Value v.s. Baselines

- We consider the 6D Hartmann Function [5] defined over $[0,1]^6$ and four baseline data distributions:
  - [***i.i.d.***] where 3 data submatrices each contains 200 samples.
  - [***ascending size***] where 3 data submatrices contains 20, 50 and 200 i.i.d. samples resp.
  - [***disjoint domains***] where $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$ & $\mathbf{X}_{S_3}$ sample from $[0,1/3]^6$, $[1/3,2/3]^6$, $[2/3,1]^6$ input domains resp.
  - [***supersets***] where $\mathbf{X}_{S_1} \subset \mathbf{X}_{S_2} \subset \mathbf{X}_{S_3}$ with sizes 200, 400 and 600 resp.

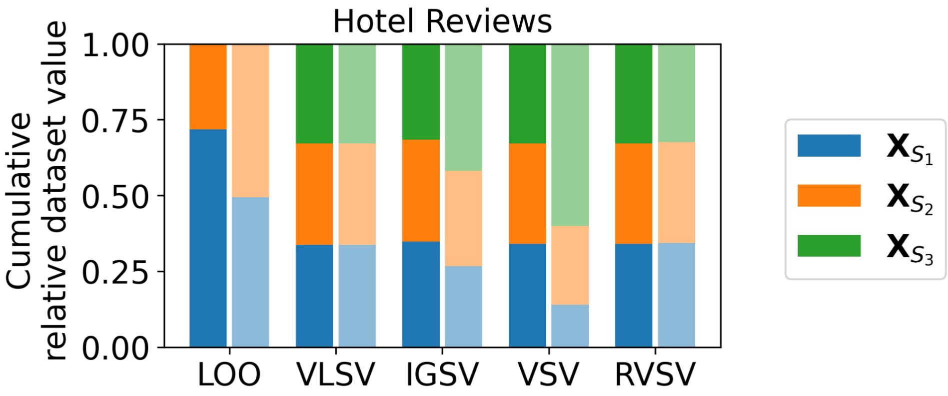


# Experiments - Replication Robustness

- Datasets:
  - TripAdvisor Hotel review rating (8)
  - California housing price prediction (CaliH) (10)
  - Kings county housing sales prediction (KingH) (10)
  - US census income prediction (USCensus) (16)
  - Age estimation from facial images (FaceA) (10)
  - The numbers represent the dimension of the standardized features.
- 3 data providers, so 3 data submatrices.
- Comparison baselines: LOO, VLSV, LOO, VSV and RVSV.

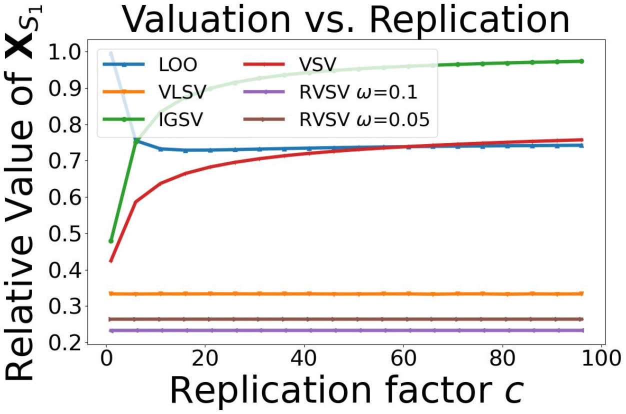
# Experiments - Replication Robustness

- TripAdvisor Hotel Review Text Dataset.
- We utilize GloVe[7] word embedding and a bidirectional LSTM with FC of 8 hidden units.
- i.i.d. Sample $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, $\mathbf{X}_{S_3}$, replicated for 0, 2, 10 times respectively.
- Darker/lighter shades denote the valuations before/after replication.
- Both IGSV & VSV are not robust to replication as the value for $\mathbf{X}_{S_3}$ increased due to replication.

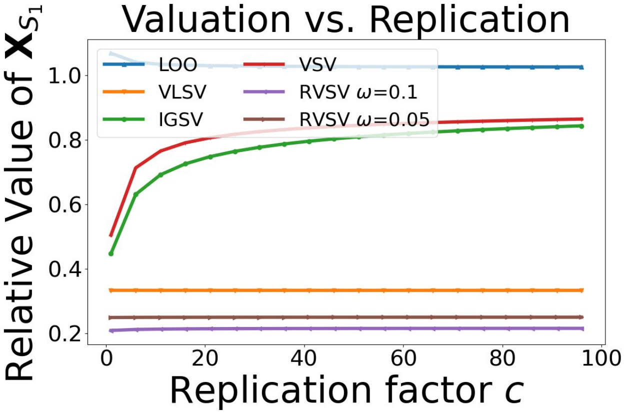


# Experiments - Asymptotic Replication Robustness

- Value of $\mathbf{X}_{S_1}$ vs. the replication factor $c$ up to 100 under i.i.d. distribution.
- A more stable curve means better robustness.
- RVSV is robust as well as VLSV, while IGSV and VSV increase with replication $c$.




CaliH




FaceA

# Conclusion

- We proposed and designed Robust Volume (RV) valuation that is
  - [***validation free***] Decoupled valuation task from validation, which has developed as a norm in current literature.
  - [***replication robust***] Circumvented unbounded scaling of replication in naive volume.
  - [***theoretically sound***] Theoretically show that larger volume leads to better learning performance.
  - [***efficient***] No model retraining is required.
  - [***versatile***] Can be combined with Shapley value to enhance fairness.
  - [***interpretable***] Assigns higher value to data that lead to high performance.
  - [***useful in practice***] Empirically works well even in complex models including DNNs.

# Thank you!

- See you at the conference!