北京大学前沿计算研究中心
Center on Frontiers of Computing Studies, Peking University

静园5号院
青 年 讲 座

# Truthful Dataset Valuation by Pointwise Mutual Information


## Dr. Shuran Zheng

Institute for Interdisciplinary Information Sciences
Tsinghua University


Host：孔雨晴 助理教授

2024年3月5日 星期二 14:00

静园五院204室


## Abstract

In the age of artificial intelligence (AI), data serves as the lifeblood that fuels innovation and development. A common way to evaluate a dataset in ML involves training a model on this dataset and assessing the model's performance on a test set. However, this approach has two issues: (1) it may incentivize undesirable data manipulation in data marketplaces, as the self-interested data providers seek to modify the dataset to maximize their evaluation scores; (2) it may select datasets that overfit to potentially small test sets. We propose a new data valuation method that provably guarantees the following: data providers always maximize their expected score by truthfully reporting their observed data. Any manipulation of the data, including but not limited to data duplication, adding random data, data removal, or re-weighting data from different groups, cannot increase their expected score. Our valuation score measures the *pointwise mutual information* of the test dataset and the evaluated dataset. We show that this score has useful information theoretic properties and show how to efficiently estimate it for certain Bayesian settings. Finally, we test by simulations the effectiveness of our data valuation method in identifying the top datasets among multiple data providers. Our method consistently outperforms the standard approach of selecting datasets based on trained model's test performance, suggesting that our evaluation score, in addition to disincentivizing data manipulation, is also more robust to overfitting.


## Biography

Shuran Zheng is a tenure-track Assistant Professor in the Institute for Interdisciplinary Information Sciences at Tsinghua University. Before coming to Tsinghua, she obtained her Ph.D. in Computer Science at Harvard University in 2022. After that, she spent one year as a postdoctoral researcher at Carnegie Mellon University. During the fall of 2022, she was a Student Researcher at Google Research NYC. Broadly speaking, her research is situated at the intersection of Economics and Computer Science. In particular, she is interested in understanding the value of data and information. Her research uses concepts and tools from Economics (especially Mechanism Design), Machine Learning, and Algorithm Design.