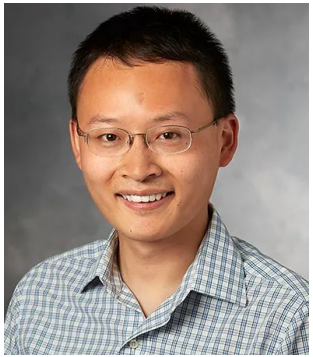
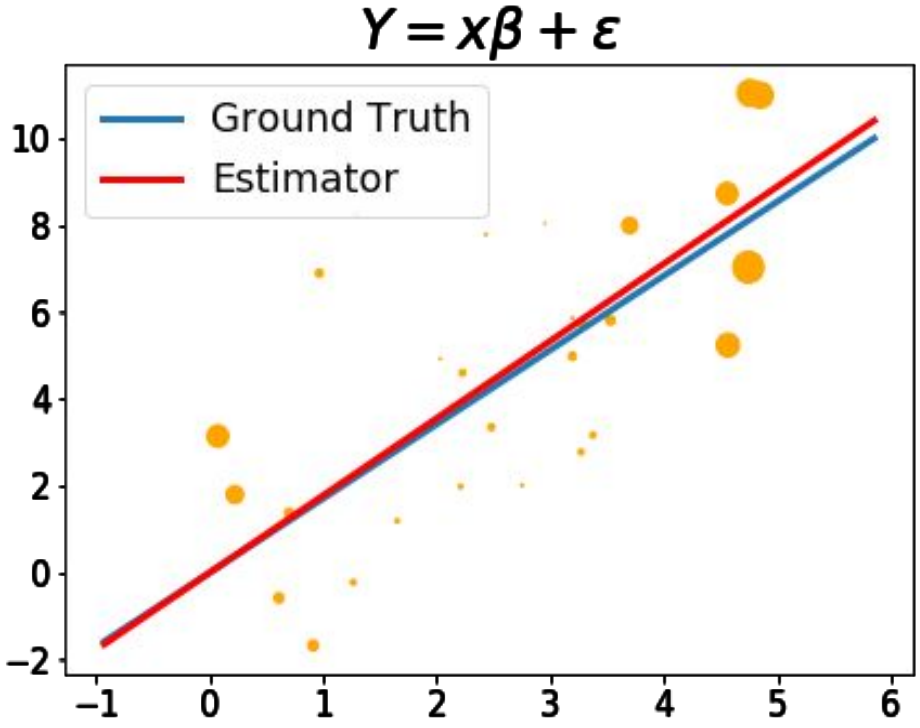
Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning

Yongchan Kwon James Zou
Stanford University



Data valuation: how to quantify the *contribution* of each training data point on my prediction task?



Data value as noisy detector

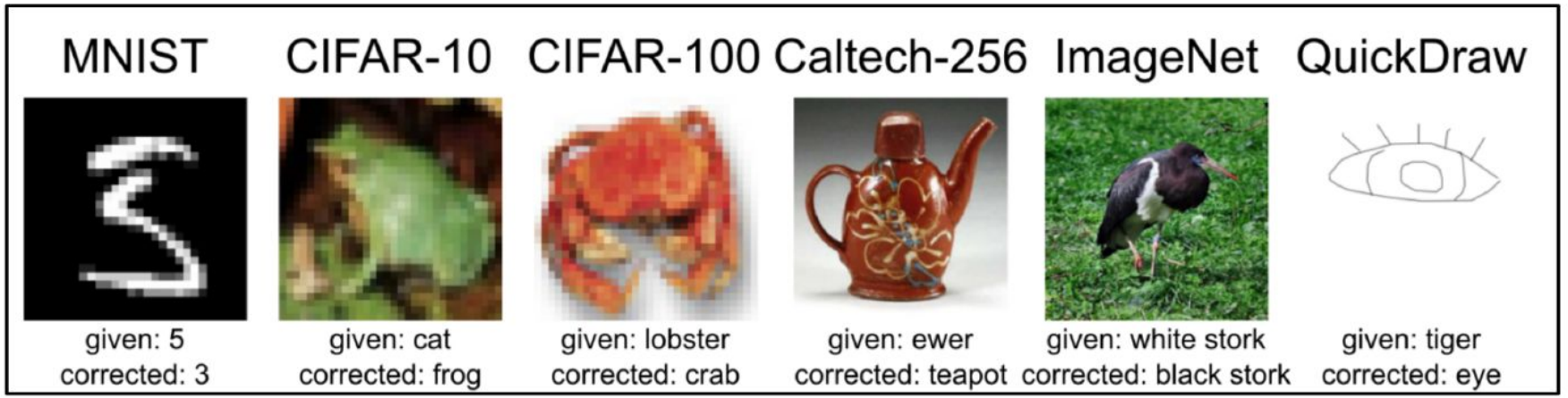


Figure: Example of mislabeled samples in commonly used image datasets.

At least 6% of the ImageNet data are mislabeled
A smaller model performs better after fixing label errors!

Source: Northcutt et al. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks.

Data value in politics

Figure: The DASHBOARD Act proposed in 2019.

Main contributions

- We propose a **noise-reduced** data valuation method called **Beta Shapley**.
- Beta Shapley is powerful at **many downstream ML tasks** such as (i) detecting mislabeled training data or (ii) identifying the most positively impactful data points.

Data value as a change in prediction after removing one data point

TECHNOMETRICS©, VOL. 19, NO. 1, FEBRUARY 1977

Detection of Influential Observation in Linear Regression

R. Dennis Cook

Department of Applied Statistics
University of Minnesota
St. Paul, Minnesota 55108

A new measure based on confidence ellipsoids is developed for judging the contribution of each data point to the determination of the least squares estimate of the parameter vector in full rank linear regression models. It is shown that the measure combines information from the studentized residuals and the variances of the residuals and predicted values. Two examples are presented.

Data Shapley and Marginal contributions

$$\text{Data Shapley}(z^*) := \frac{1}{n}\sum_{j=1}^{n}\Delta_j(z^*)$$

$$\Delta_{|S|}(z^*) := \text{Average of } (\text{Accuracy}(S \cup \{z^*\}) - \text{Accuracy}(S))$$

- Marginal contribution considers all possible subsets S with the same cardinality.
- Data Shapley is **a simple average** of the marginal contributions.

Question: Is this uniform weight best?

Which marginal contribution should we focus on?

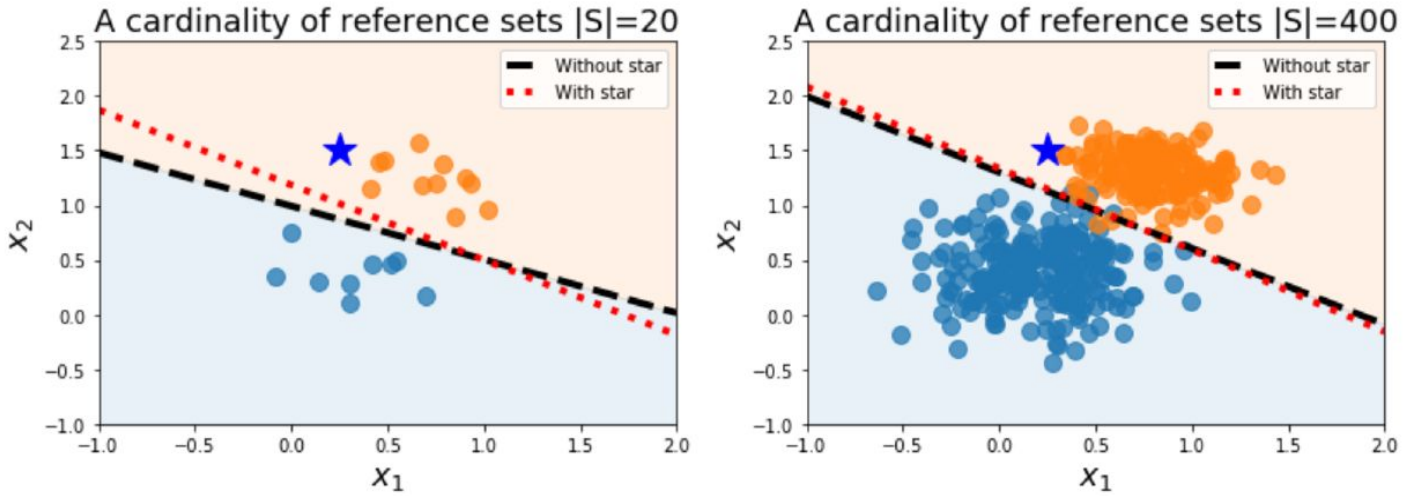


Figure: Impact of a blue star point based on (left) 20 and (right) 400 circle data points.

Which marginal contribution should we focus on?

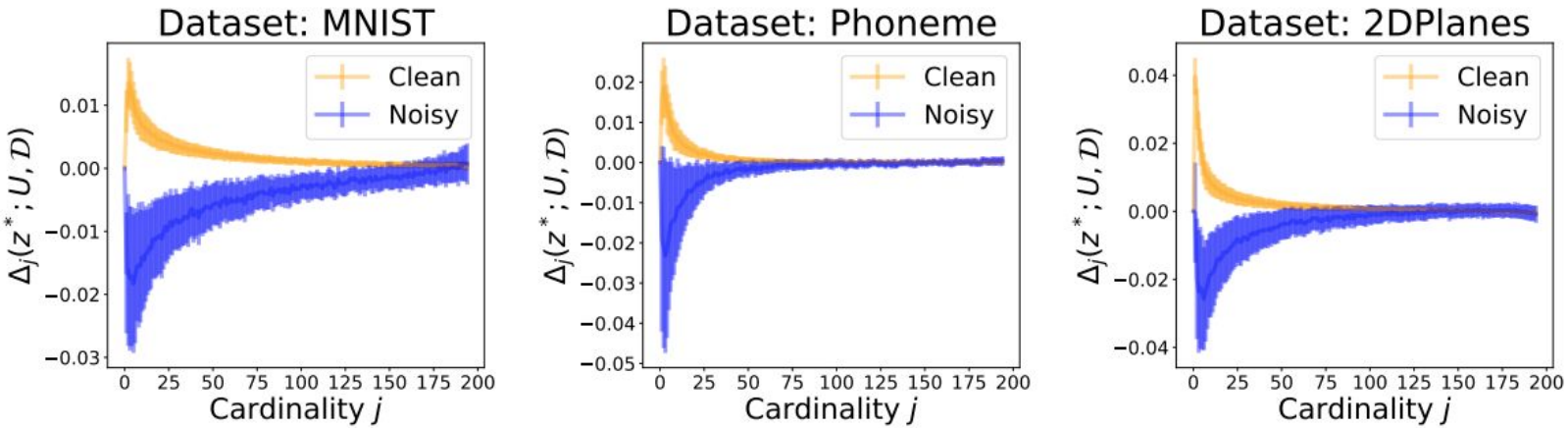


Figure: Marginal contribution as a function of the cardinality for **clean** and **noisy** data points.

A marginal contribution based on **small cardinality** is more effective to detect changes!

Detecting REAL noisy labels in CIFAR100 dataset

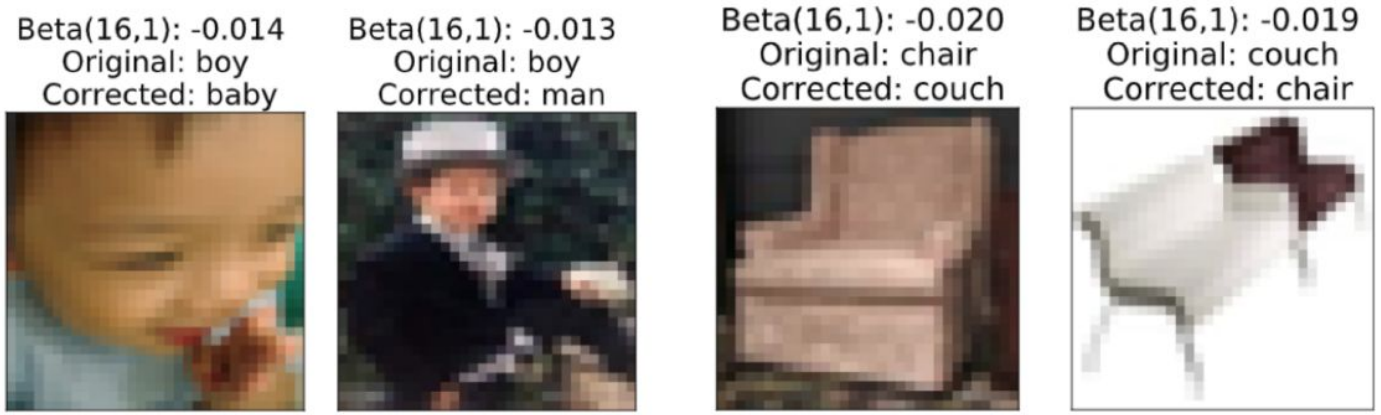
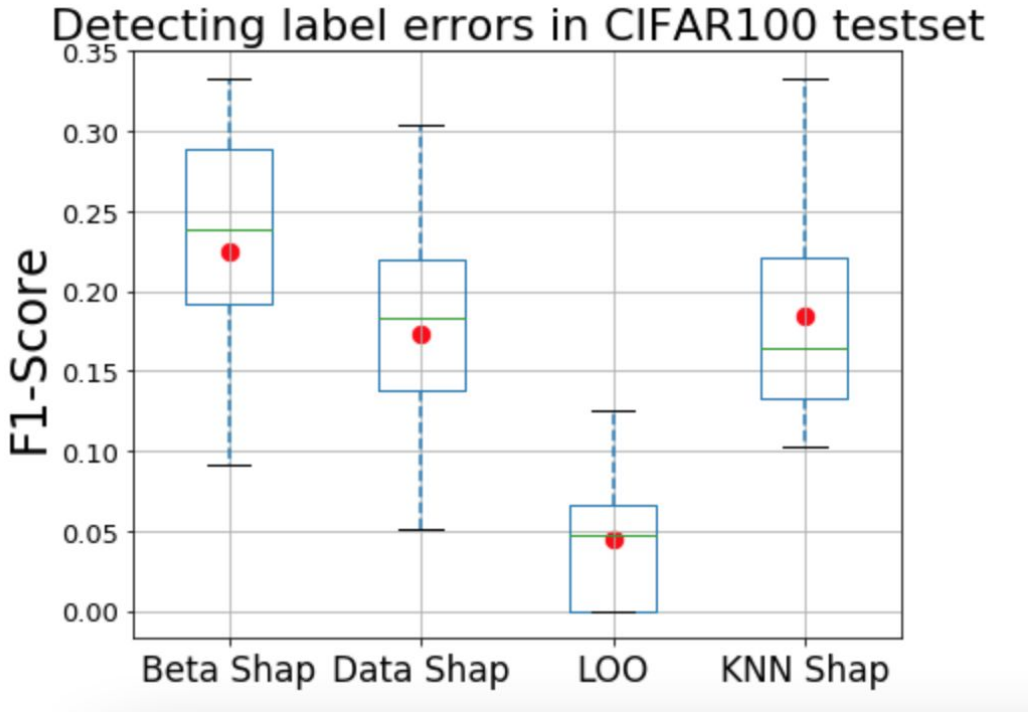
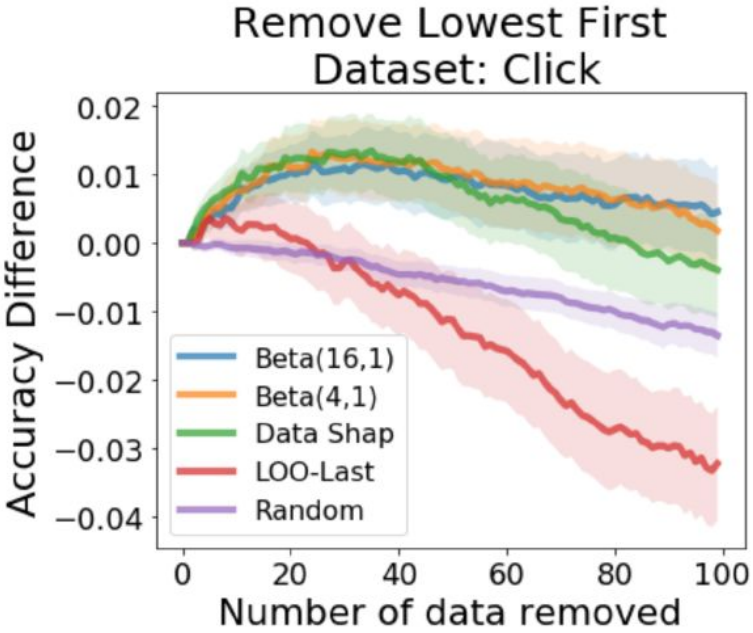


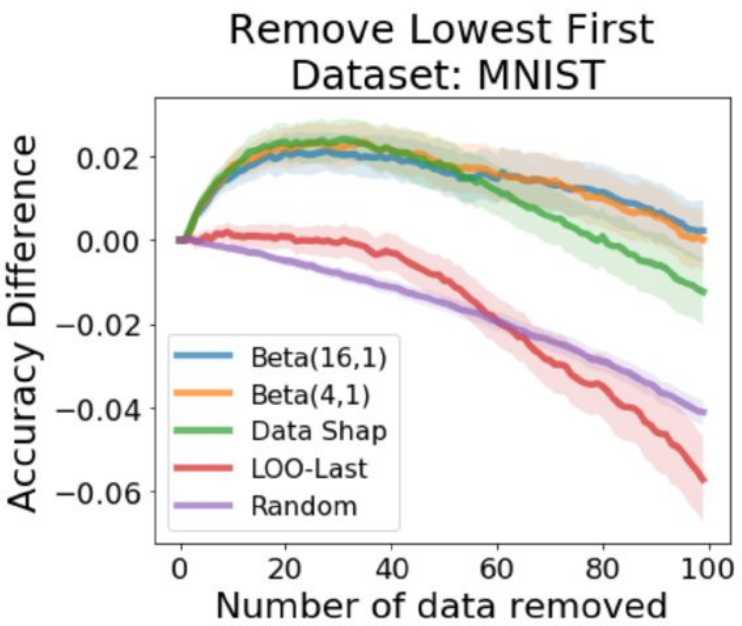
Figure: Example of CIFAR100 images with low data values. The original labels are mislabeled. The corrected labels are given in Northcutt et al. (2021).

Beta Shapley is powerful at **identifying mislabeled samples**

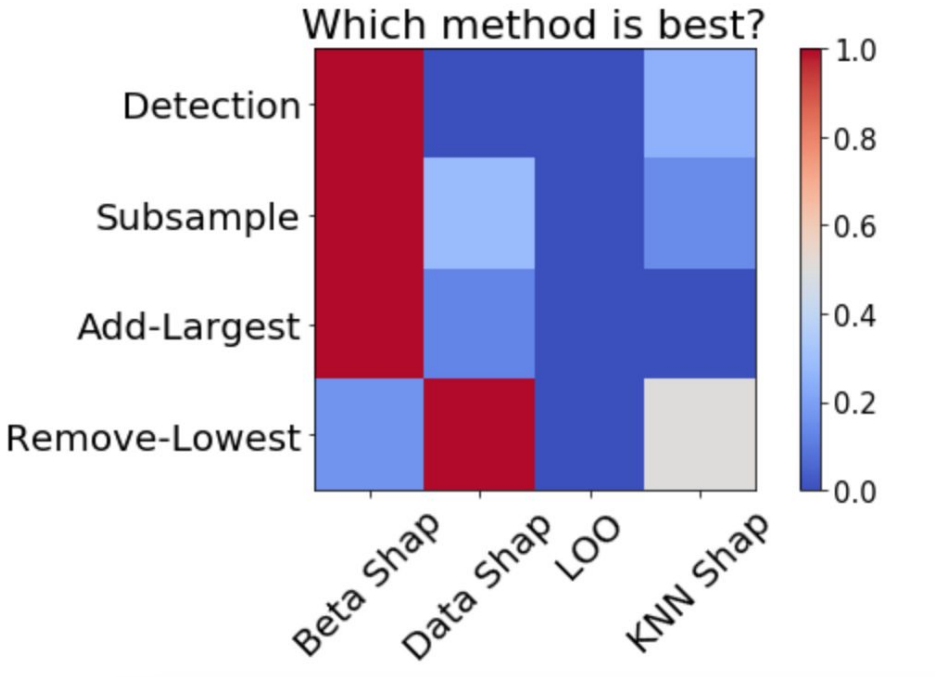


Beta Shapley is powerful at **identifying harmful data**





Beta Shapley on downstream ML tasks



Beta Shapley ***outperforms***
other valuation methods in many ML tasks

Thank you for listening!

How to compute Beta Shapley?

Check out our **easy-to-start Jupyter notebook!**