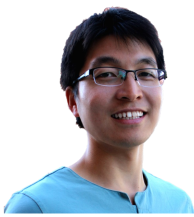
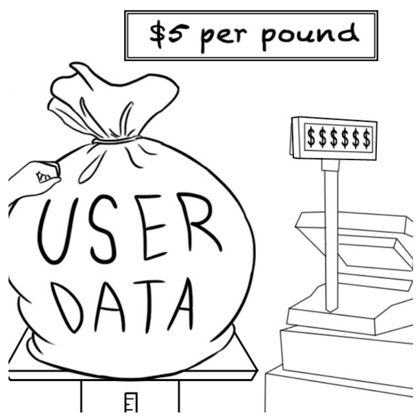
AAAI 2023 Tutorial: Economics of Data and ML



Haifeng Xu (Chicago) Shuran Zheng (CMU) James Zou (Stanford)



2/8/2023

Tutorial outline: economics of data and ML

Part I: Data buyer's perspective.

- What data is the most useful? Statistical data valuation
- How to quantify the value of information.

Short break

Part II: Data seller's perspective.

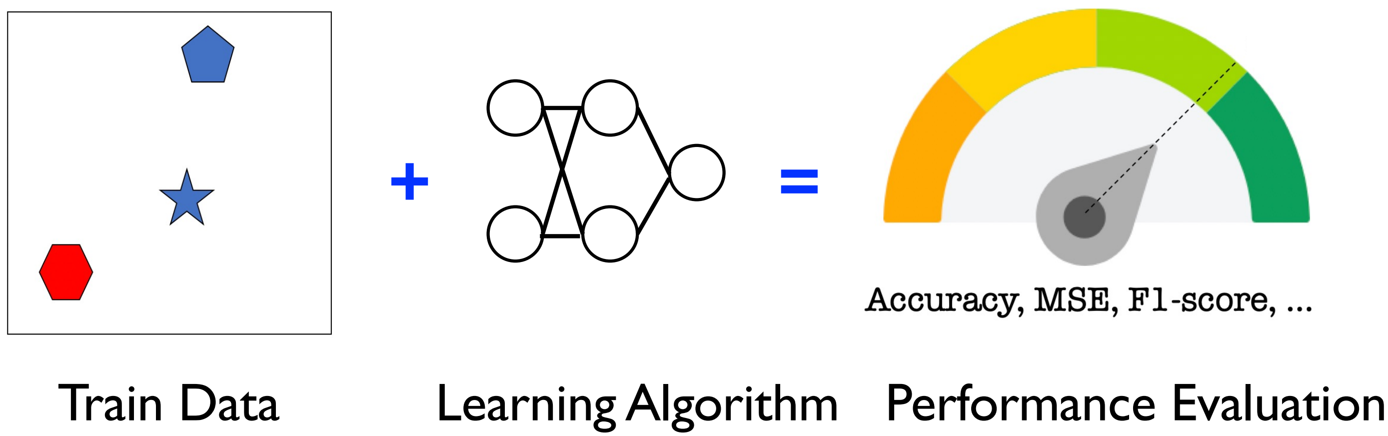
- How to price information.
- How to collect truthful data.

Short break

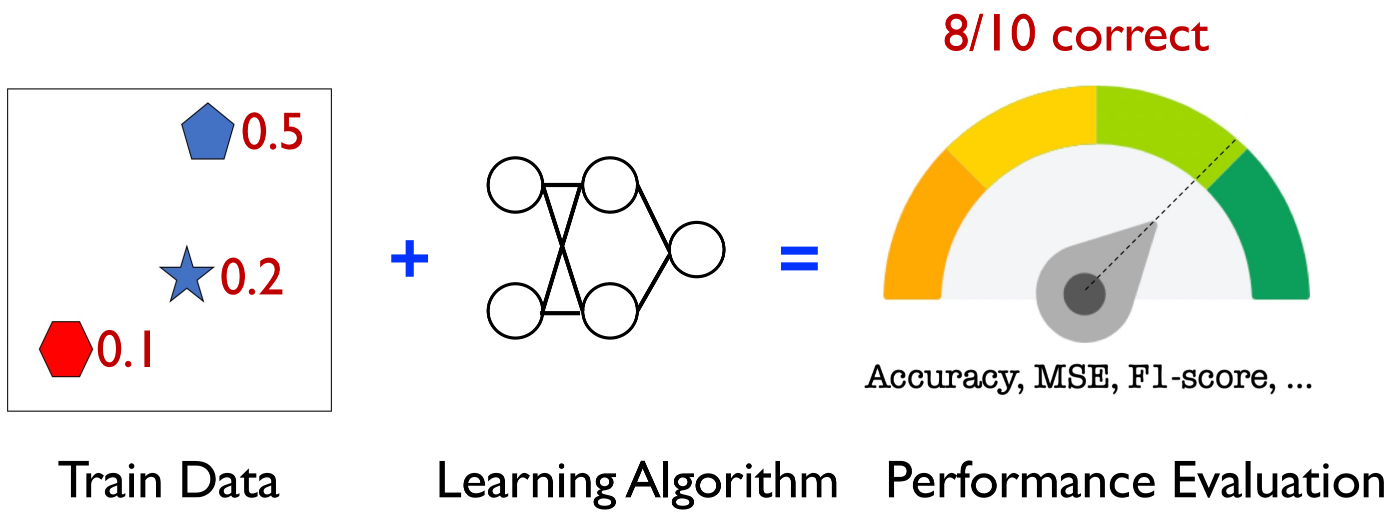
Part III: economics of ML

- Market for ML-as-a-service

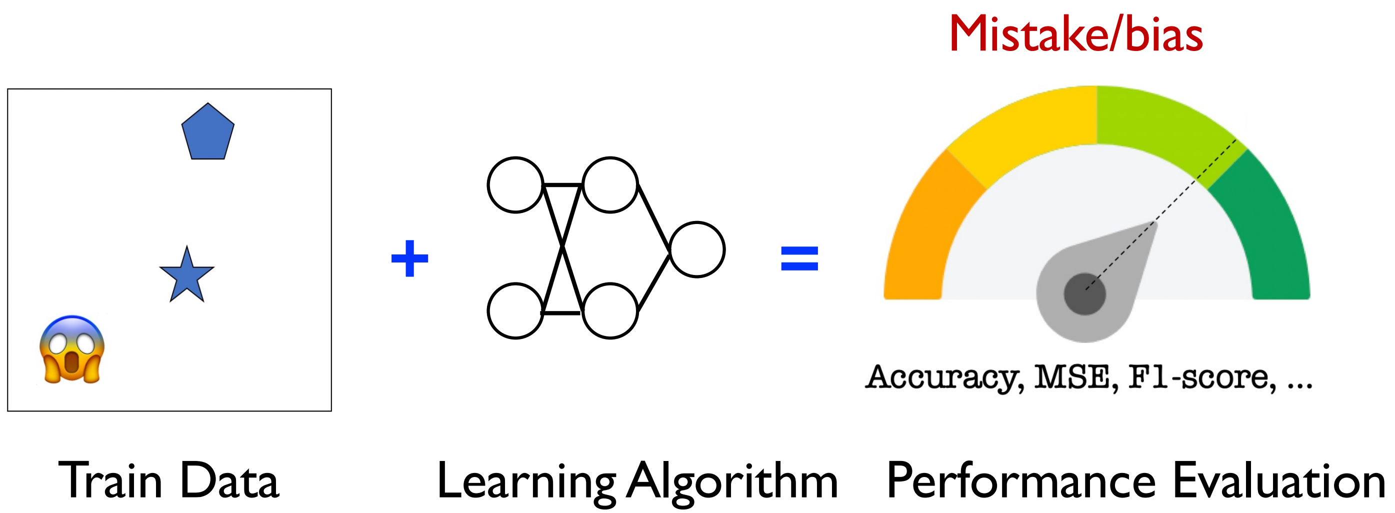
Data valuation for machine learning + statistics



Data valuation for machine learning + statistics

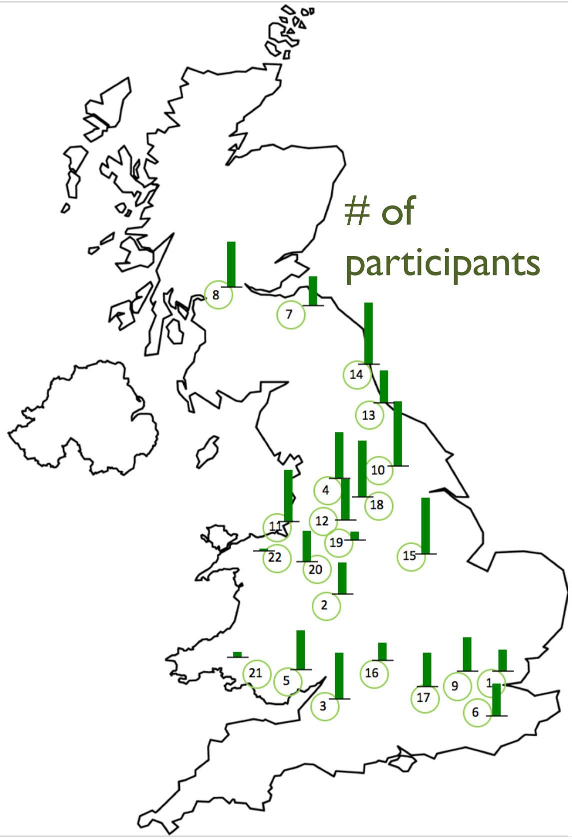


Data valuation for machine learning + statistics



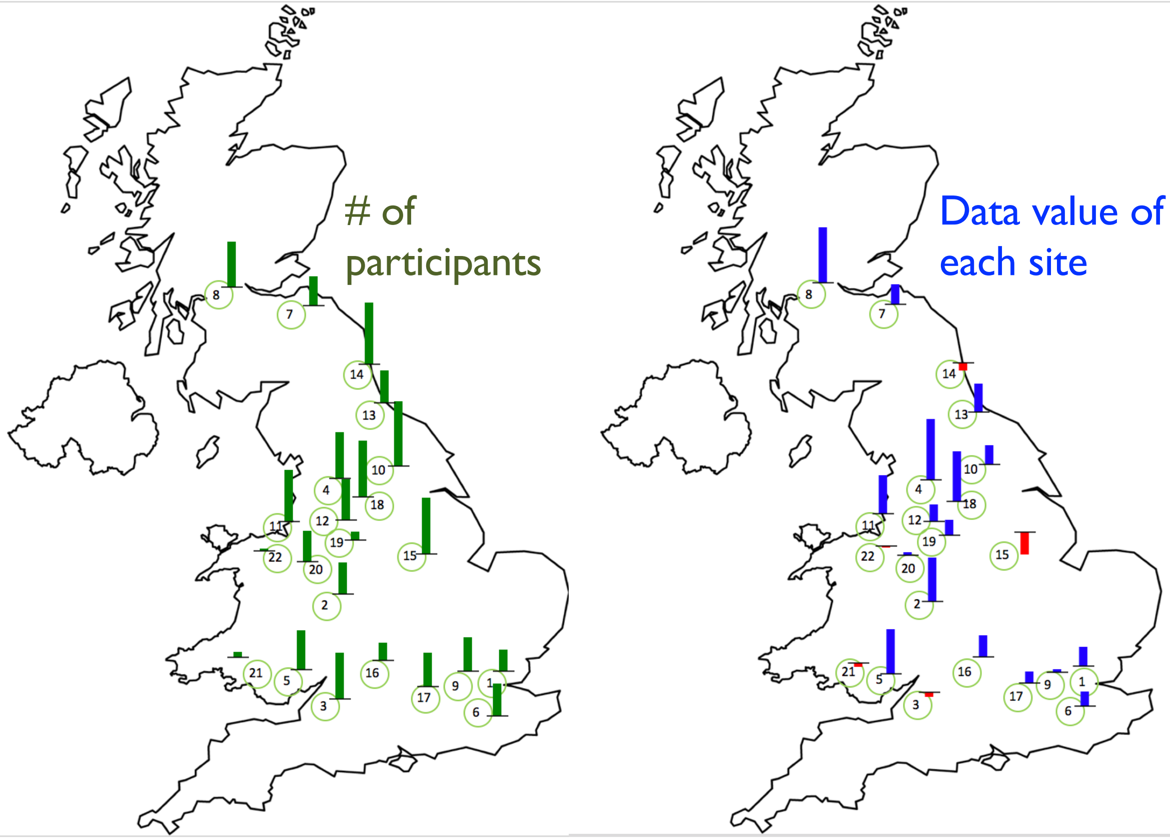
Goal is to address:
- What data to collect?
- Which data sources are high quality?
- How to audit, clean and filter data?

UK Biobank: 500k participants with genotypes and EHR



1- Barts
2- Birmingham
3- Bristol
4- Bury
5- Cardiff
6- Croydon
7- Edinburgh
8- Glasgow
9- Hounslow
10- Leeds
11- Liverpool
12- Manchester
13- Middlesborough
14- Newcastle
15- Nottingham
16- Oxford
17- Reading
18- Sheffield
19- Stockport
20- Stoke
21- Swansea
22- Wrexham

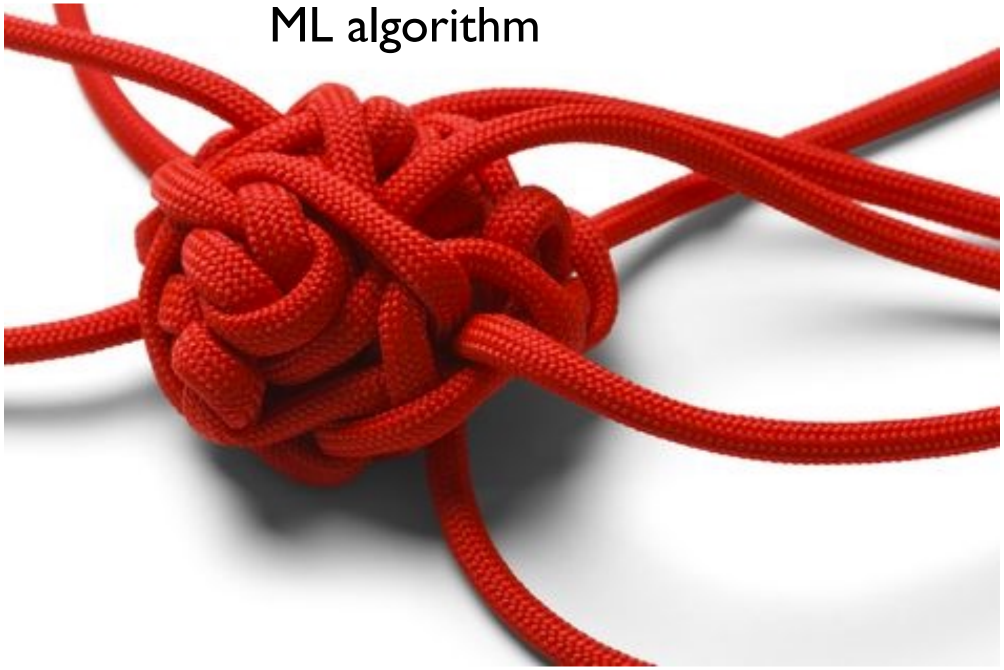
UK Biobank Lung Cancer prediction



1- Barts
2- Birmingham
3- Bristol
4- Bury
5- Cardiff
6- Croydon
7- Edinburgh
8- Glasgow
9- Hounslow
10- Leeds
11- Liverpool
12- Manchester
13- Middlesborough
14- Newcastle
15- Nottingham
16- Oxford
17- Reading
18- Sheffield
19- Stockport
20- Stoke
21- Swansea
22- Wrexham

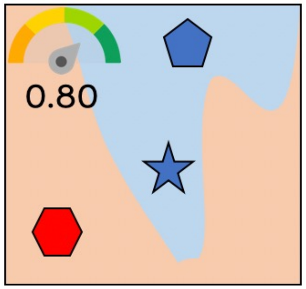
Removing negative valued centers improves performance.

How do we do it?



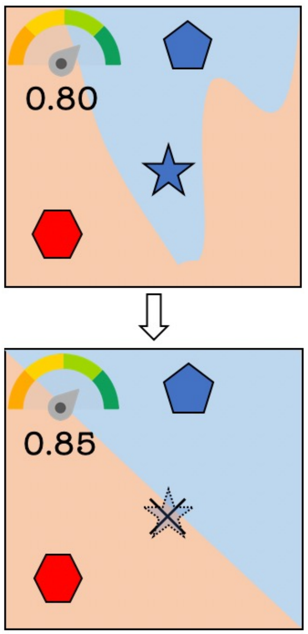
Leave-one-out valuation

Example: value(★) = ?



Leave-one-out valuation

Example: value() = ?



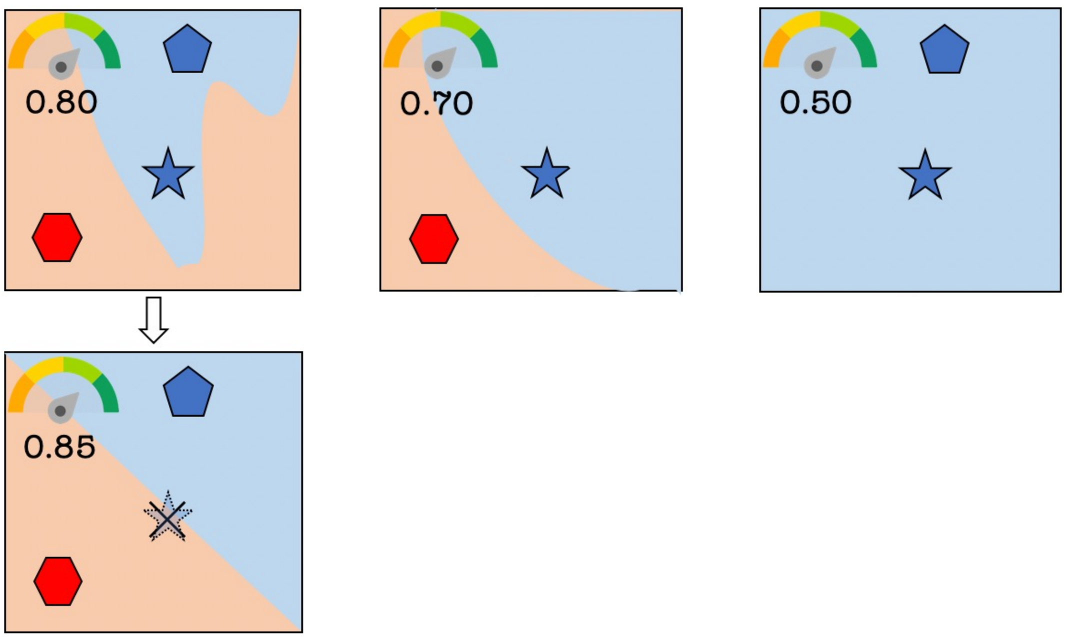
0.8 – 0.85

delete  and see how the model performance changes.

Problem: LOO too noisy

Data Shapley Value

<u>Example</u>: value(★) = ?

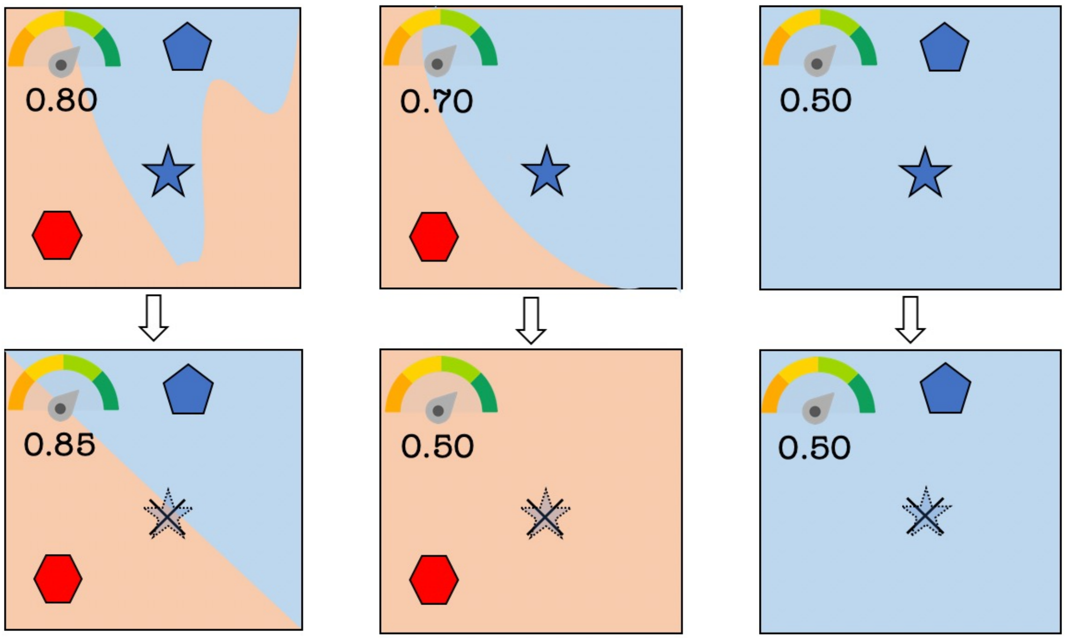


0.8 – 0.85

Ghorbani and Zou. *ICML* 2019; Jia et al *AISTATS* 2019; Agarwal, Dahleh, Sarkar *EC* 2018.

Data Shapley Value

<u>Example</u>: value(★) = 0.05



0.8 – 0.85 0.7 – 0.5 0.5 – 0.5

Unique way to aggregate these scores into data Shapley.

Ghorbani and Zou. *ICML* 2019; Jia et al *AISTATS* 2019; Agarwal, Dahleh, Sarkar *EC* 2018.

Data Shapley Value

$$\textit{Value}(\text{data k}) = \sum_{\text{subsets S not containing k}} \frac{\overbrace{performance(S \cup k) - performance(S)}^{\text{marginal contribution}}}{\underbrace{\binom{n-1}{|S|}}_{\text{\# of size |S| subsets}}}$$

Expected contribution to all possible sizes of train data samples.

Uniquely satisfies Shapley axioms: null, symmetry, efficiency, linearity

Data Shapley Value

Lloyd Shapley

2012 Nobel Prize in Economics

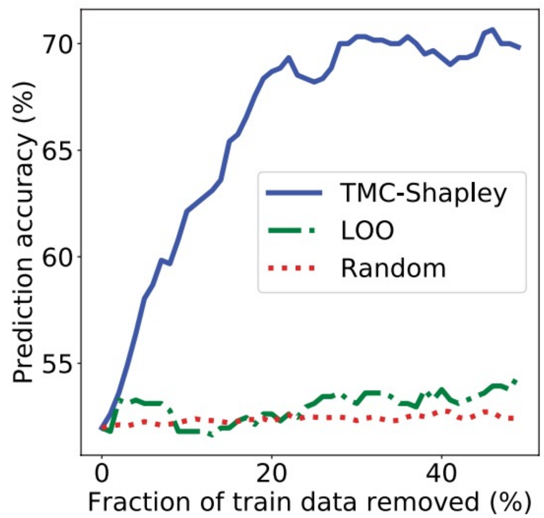
Cooperative game

Application 1: Shapley value identifies mis-annotations

CheXpert database

60% images with low Shapley were mis-annotated in the database

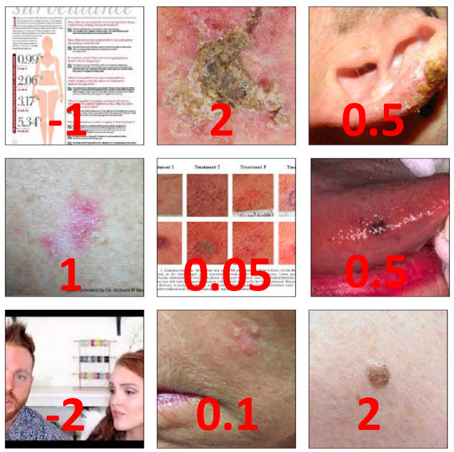
Removing training data with low Shapley values



Tang et al. *Scientific Reports* 2021

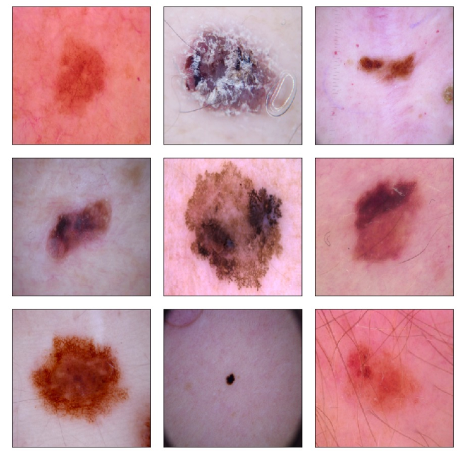
Application 2: improves model via data weighting

Training data



accuracy ↓

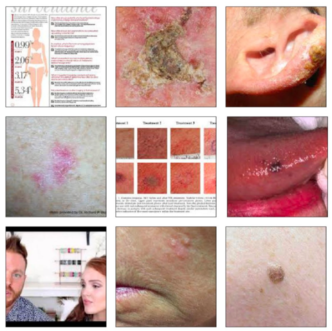
Clinical examples



Ghorbani and Zou. *ICML* 2019; Ghorbani, Kim and Zou *ICML* 2020; Kwon and Zou *AISTATS* 2021, 2022

Application 2: improves model via data weighting

Training data

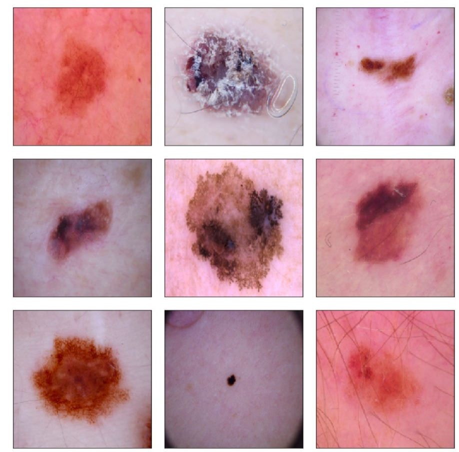


High value data

weight train data
by Shapley value

accuracy ⬆ 11%.

Clinical examples



Application 3: data Shapley improves fairness

Training data

Deployment examples

accuracy ⬇
esp. for minorities

Ghorbani and Zou. *ICML* 2019; Ghorbani, Kim and Zou *ICML* 2020; Kwon and Zou *AISTATS* 2021, 2022

Application 3: data Shapley improves fairness

Data Shapley

accuracy
esp. for minorities

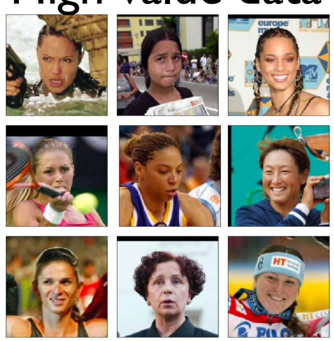
Deployment examples

Ghorbani and Zou. *ICML* 2019; Ghorbani, Kim and Zou *ICML* 2020; Kwon and Zou *AISTATS* 2021, 2022

Application 3: data Shapley improves fairness

Training data

High value data



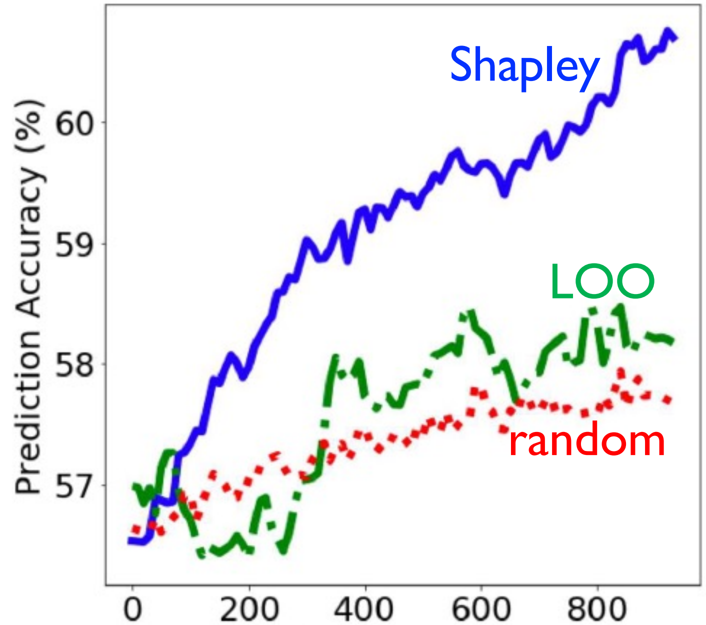
weight train data
by Shapley value

accuracy 8%.

Deployment examples

Application 4: active learning

Improving lung cancer prediction in the UK Biobank



# of additional participants recruited

How to efficiently estimate data Shapley values

$$\textit{Value}(\text{data } k) = \sum_{\text{subsets S not containing k}} \frac{\textit{performance}(S \cup k) - \textit{performance}(S)}{\binom{n-1}{|S|}}$$

Analytic forms of data Shapley available for specific ML models.

- KNN predictor: recursive formula for Shapley (Jia et al 2019)
- Linear regression: modified least squares (Kwon, Rivas, Zou 2021)
- Logistic classifier: lower bound for Shapley (Kwon, Rivas, Zou 2021)

Approach 1: fix the encoder and compute data Shapley using the last layer of NN. Scales to $>10^6$ data.

How to efficiently estimate data Shapley values

$$\textit{Value}(\text{data k}) = \sum_{\text{subsets S not containing k}} \frac{\textit{performance}(S \cup k) - \textit{performance}(S)}{\binom{n-1}{|S|}}$$

Approach 2: Monte Carlo approximations for general ML models

- Sample coalitions until convergence (Ghorbani, Kim, Zou 2020)

Can scale to compute data Shapley for CNN on ~100k data points.

Unifying approach to data valuation

$$\textit{Value}(\text{data k}) = \sum_{\text{subsets S not containing k}} \boxed{?} \; \overbrace{\frac{\textit{performance}(S \cup k) - \textit{performance}(S)}{\binom{n-1}{|S|}}}^{\text{marginal contribution}}$$

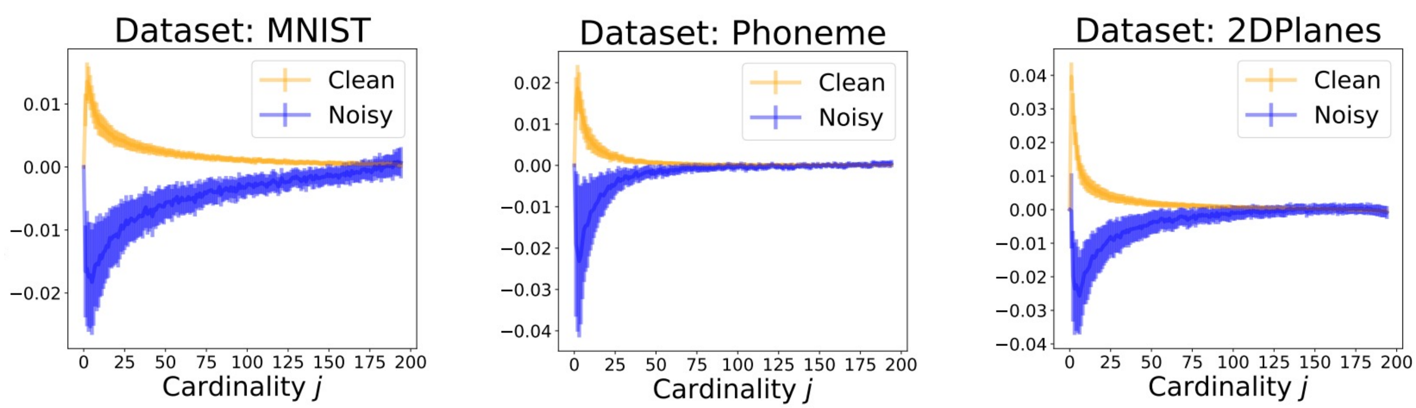
(Almost) all of data valuation developments: different ways of combining marginal contributions into a data value.

- Beta-Shapley (Kwon and Zou *AISTATS* 2022)
- AME (Lin et al. *ICML* 2022)
- Data model (Ilyas et al. *ICML* 2022)

Data Shapley is statistically suboptimal

$$\text{Data Shapley of } x = \frac{1}{n}\sum_{\text{cardinality}=j} \underbrace{E[\text{perf}(x \cup j \text{ pts}) - \text{perf}(j \text{ pts})]}_{\text{marginal contribution of cardinality } j}$$

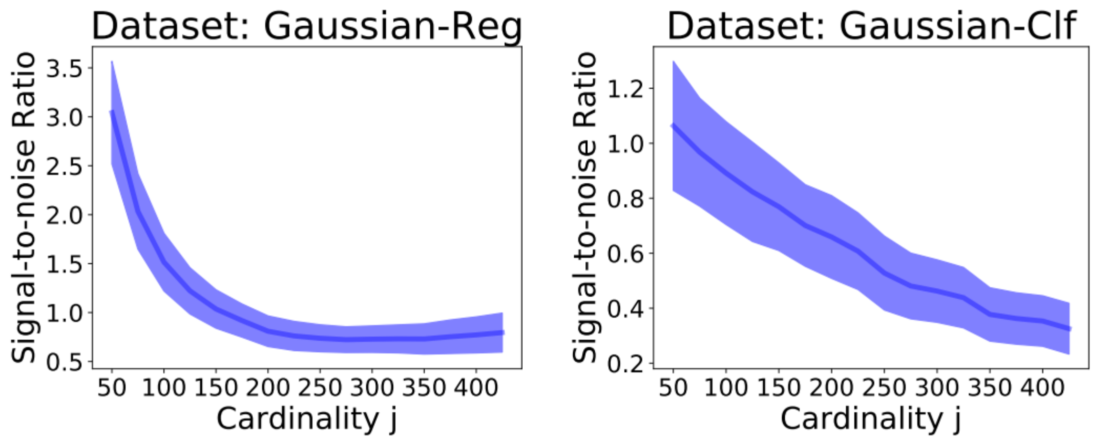
marginal contribution



Large cardinalities are noisier

Kwon and Zou. *AISTATS* 2022

Large cardinalities are less informative

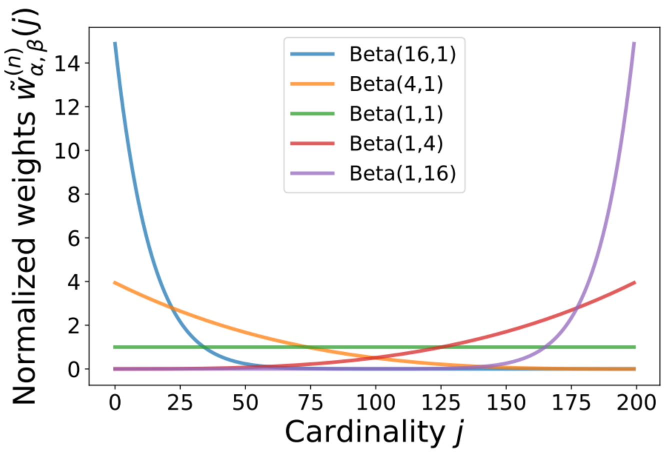


Proposition (informal): The signal-to-noise ratio (i.e. marginal contribution divided by its standard deviation) decreases as cardinality increases.

Kwon and Zou. *AISTATS* 2022

Beta-Shapley extends Shapley value

Weight coalitions of different cardinalities w/ Beta distribution

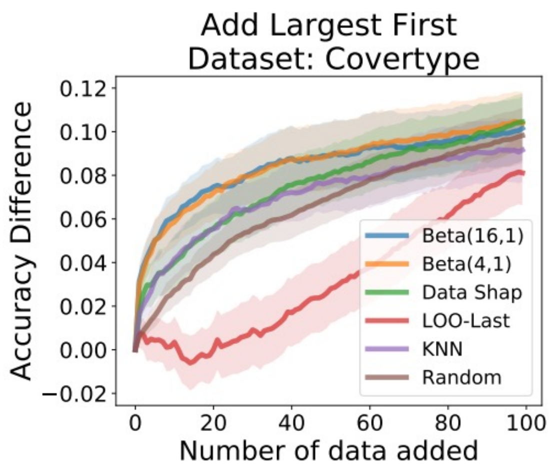


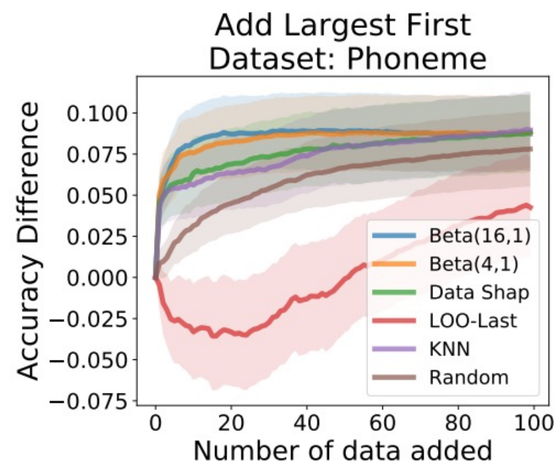
$$\text{Beta-Shapley}(x) = \sum_{\text{cardinality}=j} \text{Beta}(j+\beta-1, n-j+\alpha) E[\text{perf}(x \cup j \text{ pts}) - \text{perf}(j \text{ pts})]$$

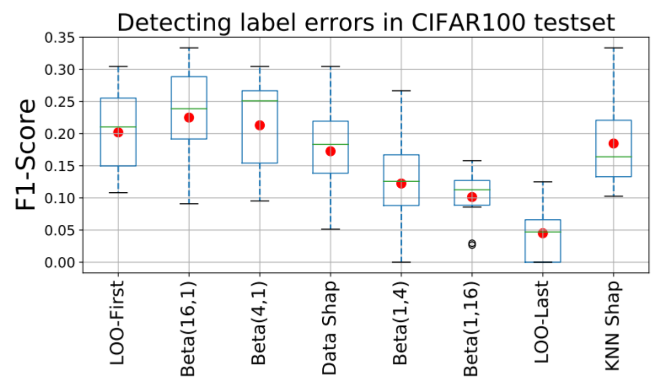
Beta distribution is flexible and computationally efficient.

* Does not satisfy the efficiency axiom.

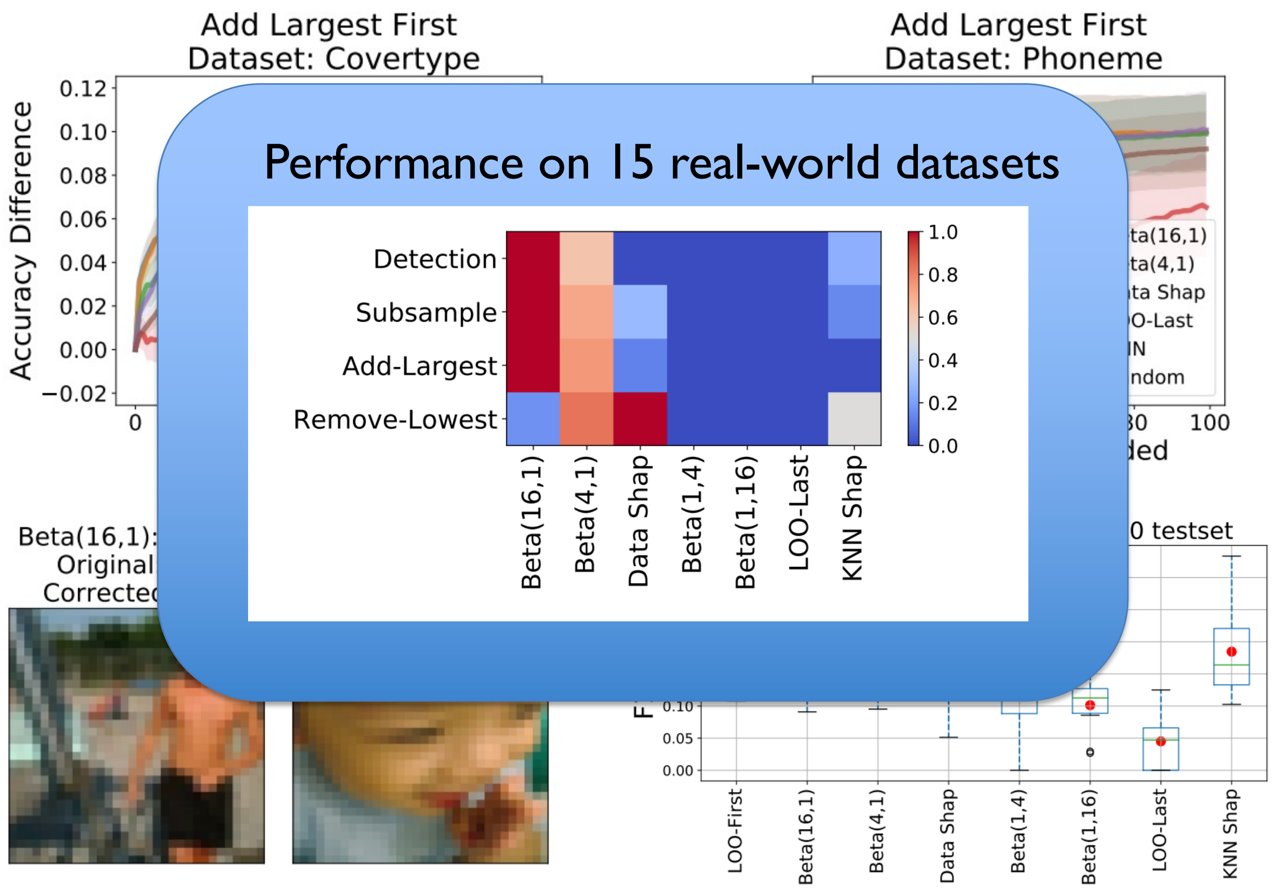
Beta-Shapley is more informative than Data Shapley







Beta-Shapley is more informative than Data Shapley



Learning weights for data valuation

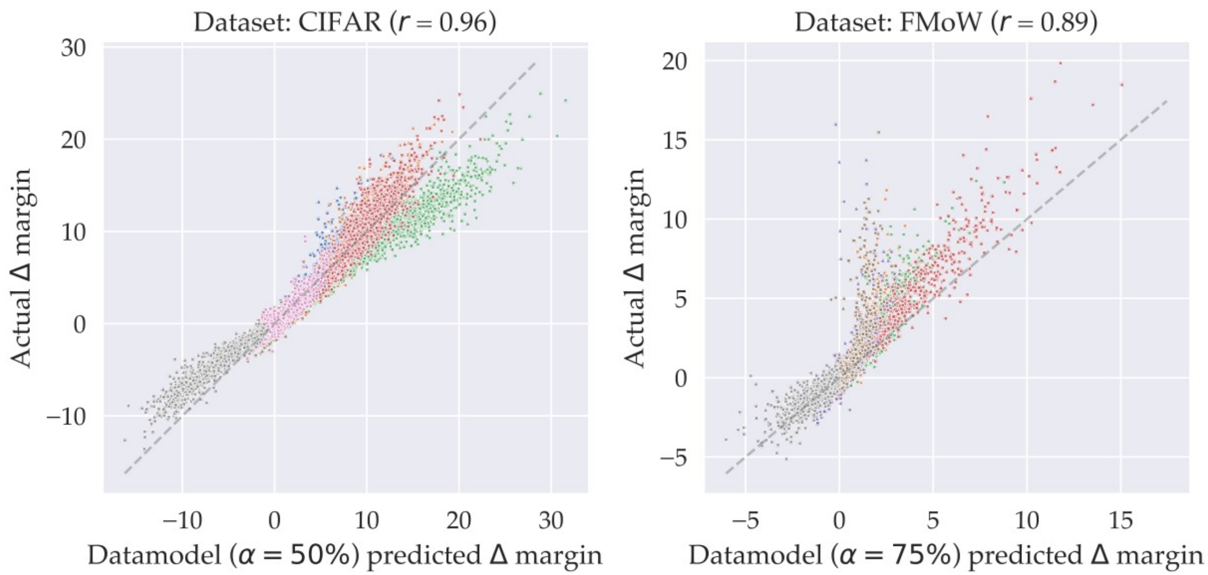
	Data 1	Data 2	…	Data N	Test acc
Model 1	1	0		1	0.7
Model 2	1	1		0	0.8
…					
Model M	0	1		1	0.85

Which data points are used to train each model

Fit a linear regression to predict test acc. Data value = regression coeff.

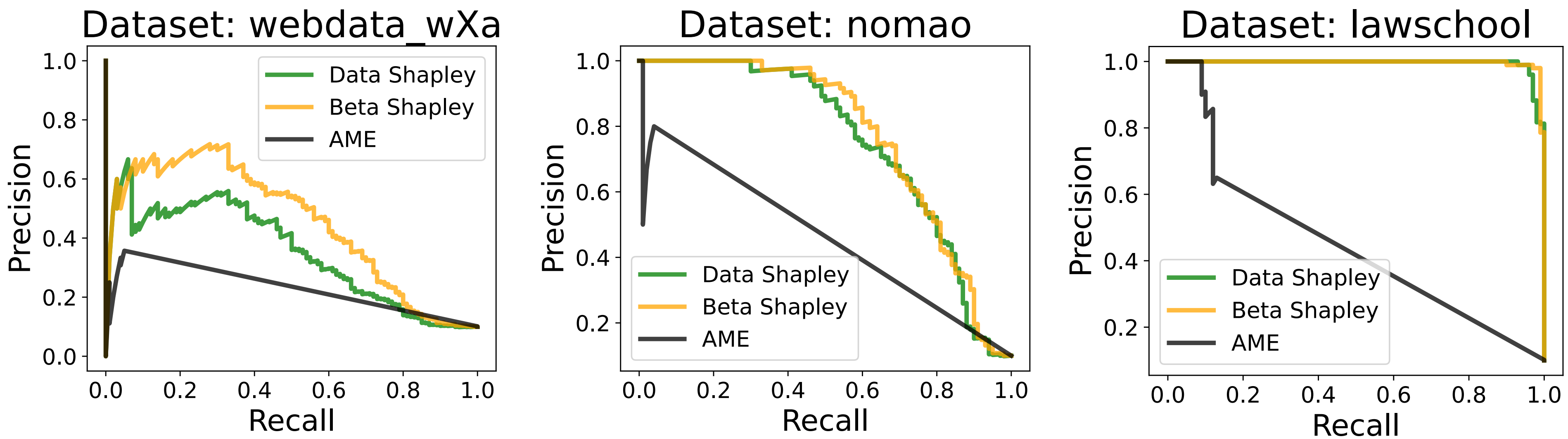
Lin et al. *ICML* 2022; Ilyas et al. *ICML* 2022

Predicting the impact of each point on model accuracy



Ilyas et al. *ICML* 2022

Comparison of data valuation methods for detecting noisy data



Yongchan Kwon

Takeaways

Data valuation depends on the context (model, performance metric).

Applications to data cleaning, curation, active learning, fairness.

Most data valuation approaches aggregate marginal contributions of a data point. They differ in the aggregation weights.

Open challenge: current data valuation requires access to data.
How to estimate data value w/o seeing the data?

References

Data Shapley value: Ghorbani and Zou. *ICML* 2019; Jia et al *AISTATS* 2019; Agarwal, Dahleh, Sarkar *EC* 2018.

Statistical properties of data valuation: Ghorbani, Kim and Zou *ICML* 2020.

Efficient computations of data valuation: Jia et al *AISTATS* 2019; Kwon, Rivas and Zou *AISTATS* 2021.

Alternative weights for data valuation: Kwon and Zou AISTATS 2022; Lin et al. *ICML* 2022; Ilyas et al. *ICML* 2022.

Applications of data valuation: Liang et al. *Nature Machine Intelligence* 2022.