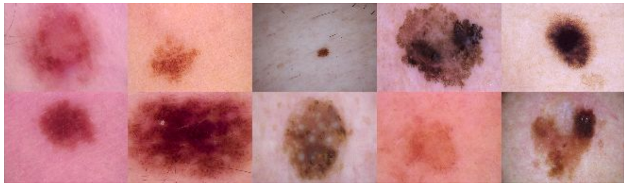
Equitable Valuation of Data

Amirata Ghorbani

A Simple Example

AI4H Company
Detecting Melanoma

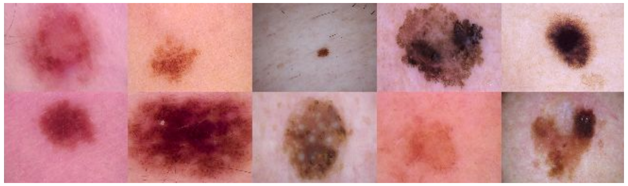


60% melanoma prediction accuracy

A Simple Example

Hospital

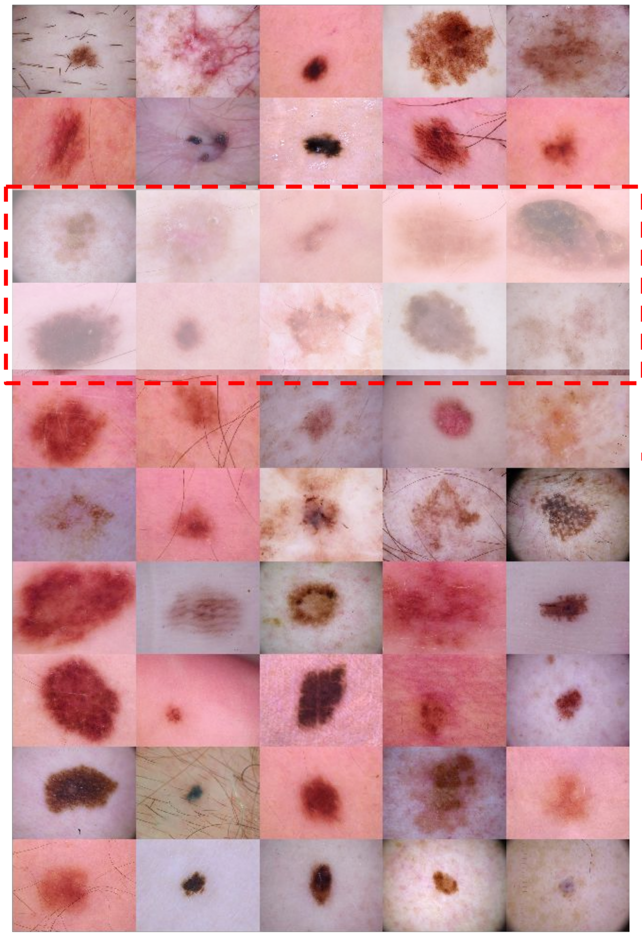
AI4H Company
Detecting Melanoma



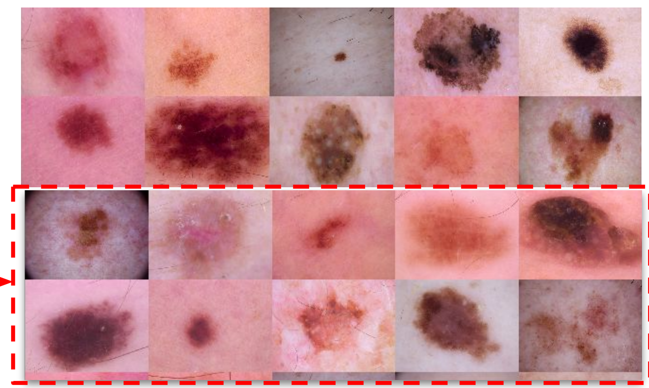
60% melanoma prediction accuracy

A Simple Example

Hospital



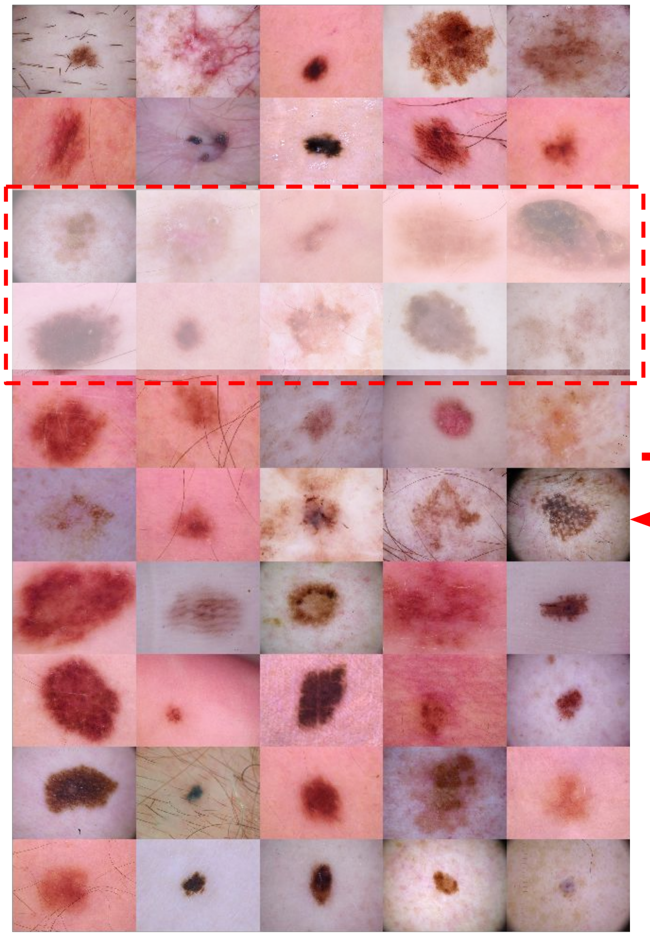
AI4H Company
Detecting Melanoma



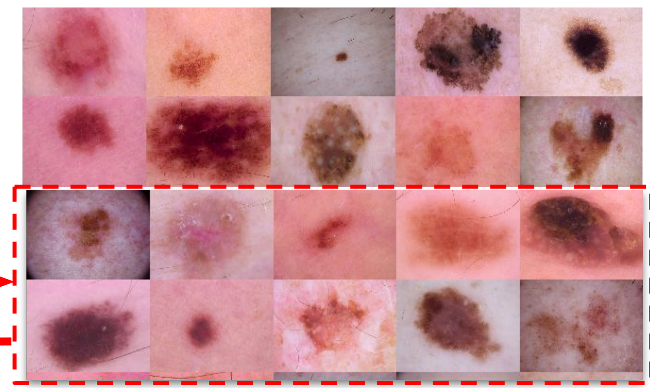
80% melanoma prediction accuracy

A Simple Example

Hospital



AI4H Company
Detecting Melanoma



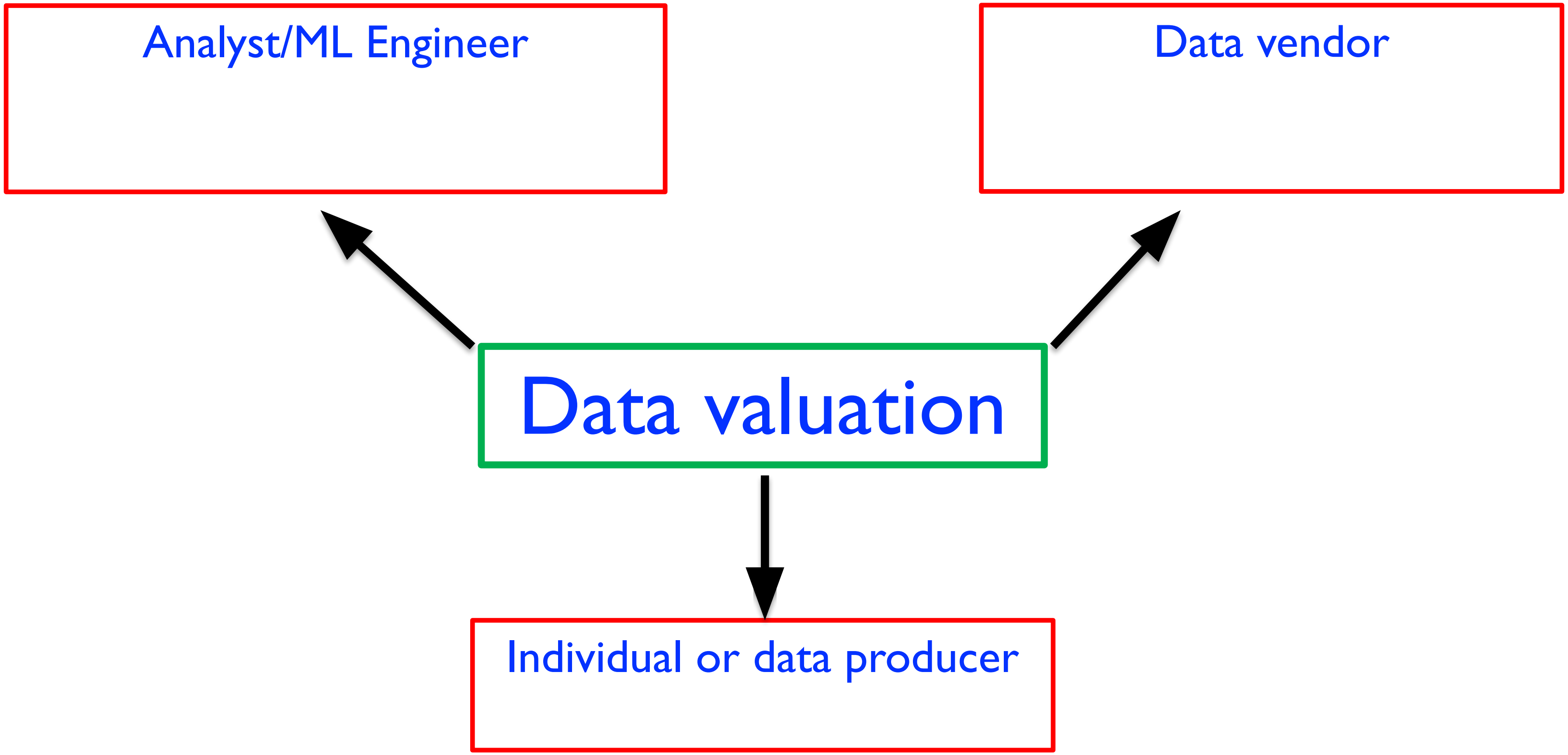
$$$
???

80% melanoma prediction accuracy

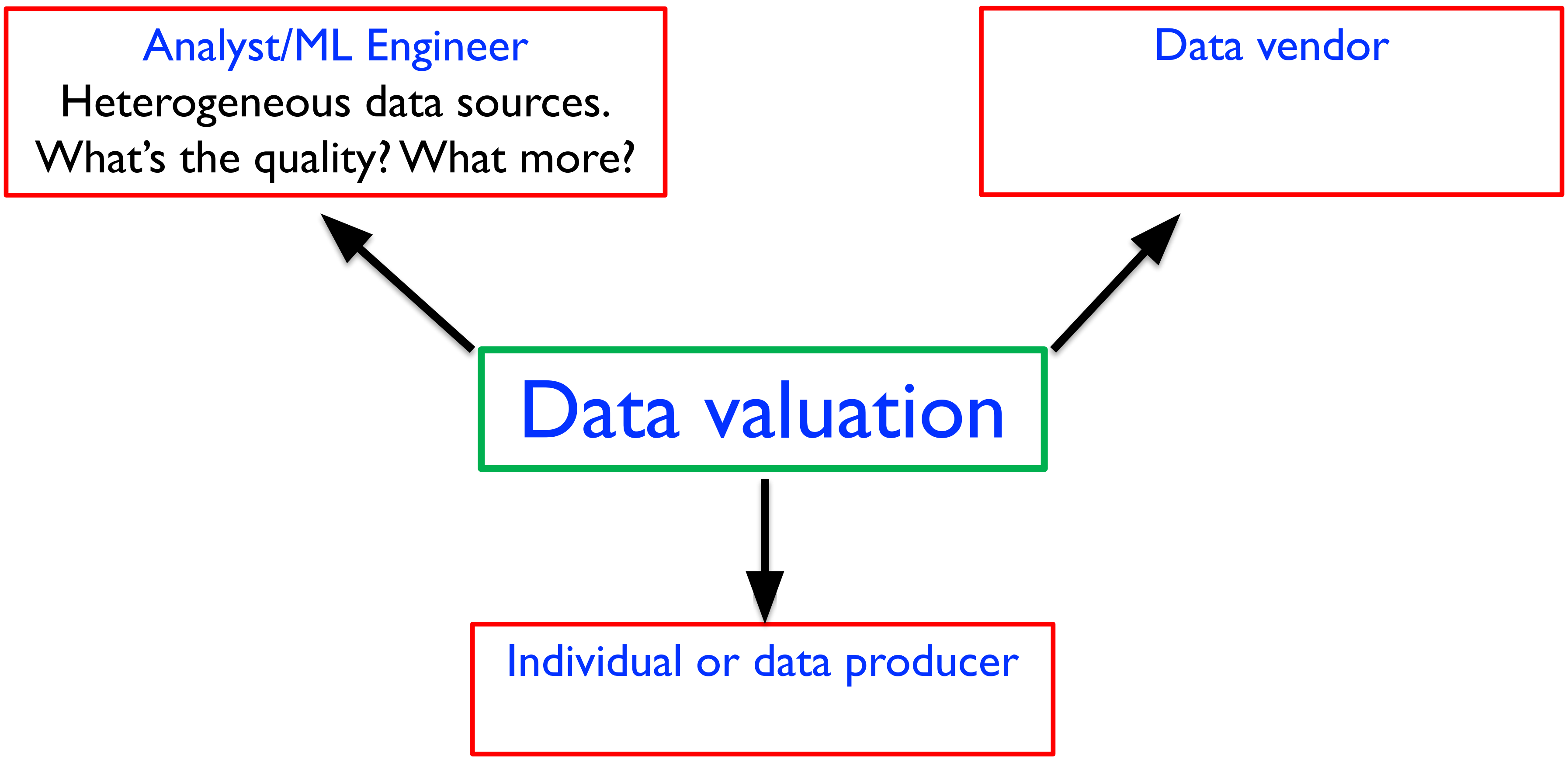
If data is fuel, then we need to measure its value

Data valuation

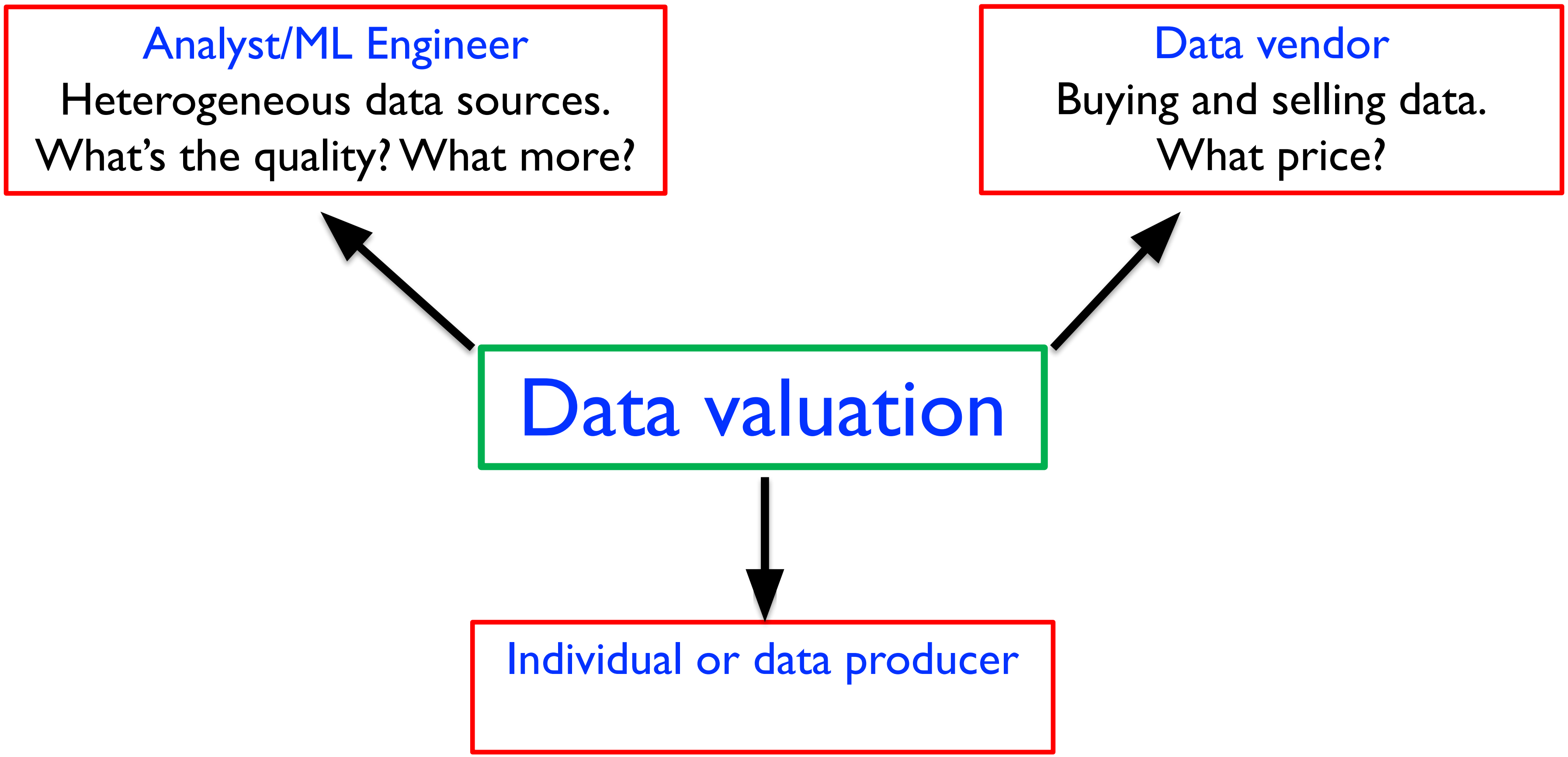
If data is fuel, then we need to measure its value



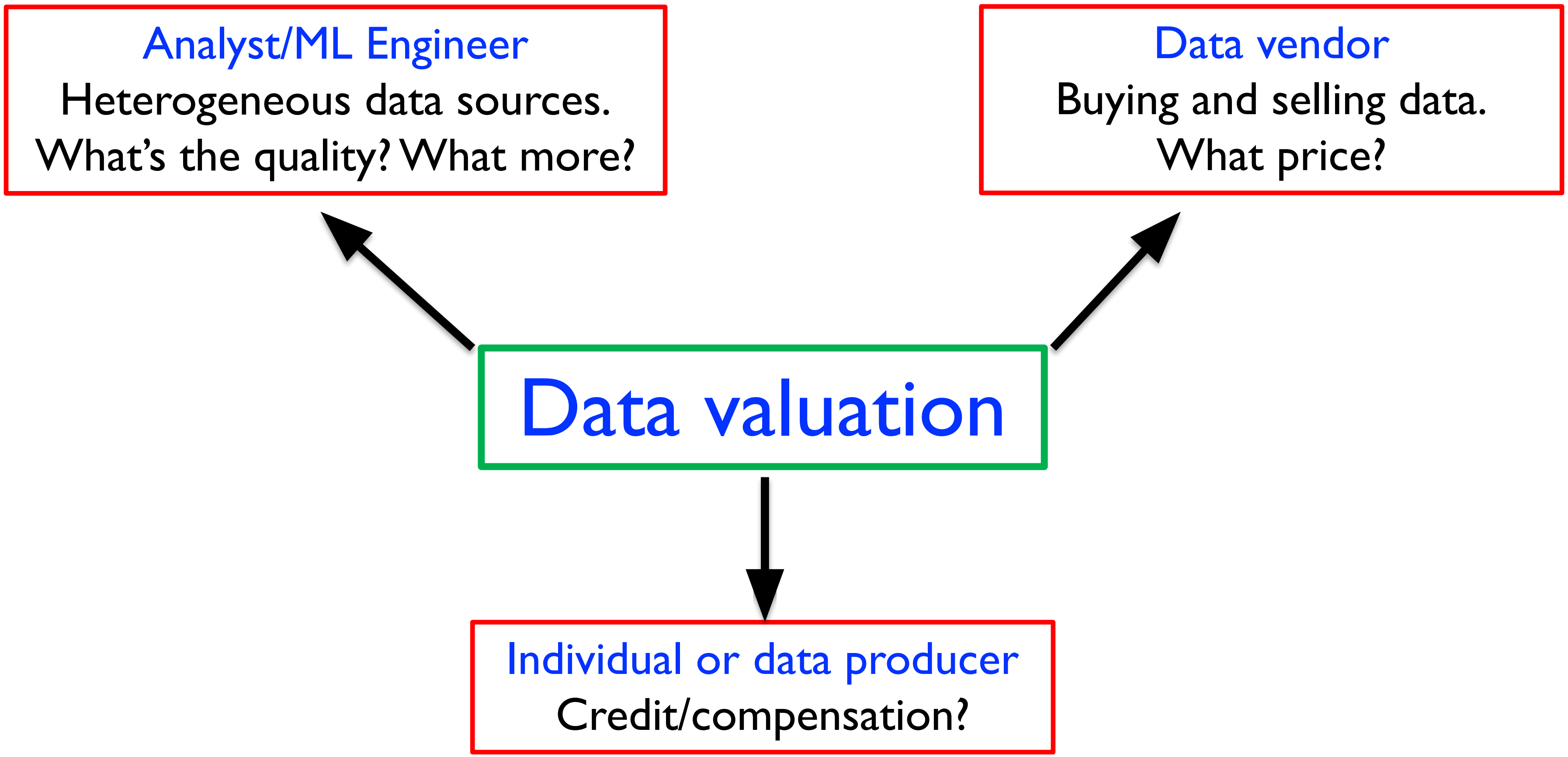
If data is fuel, then we need to measure its value



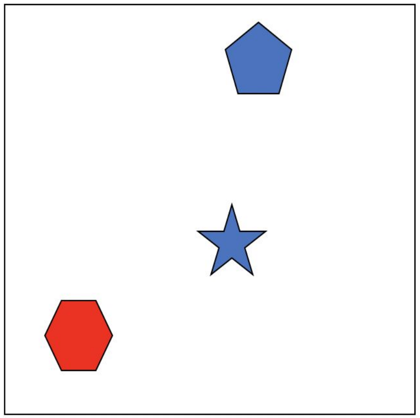
If data is fuel, then we need to measure its value

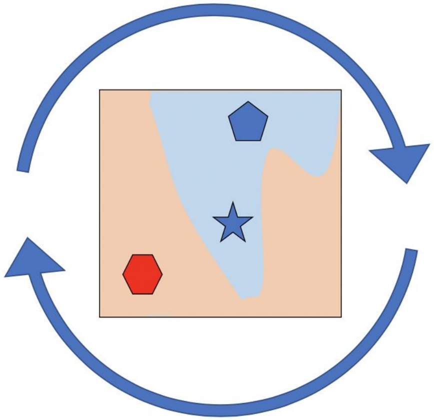


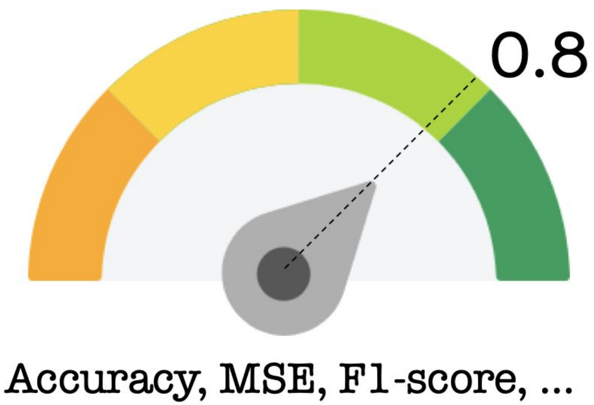
If data is fuel, then we need to measure its value



Ingredients of ML and Data Value

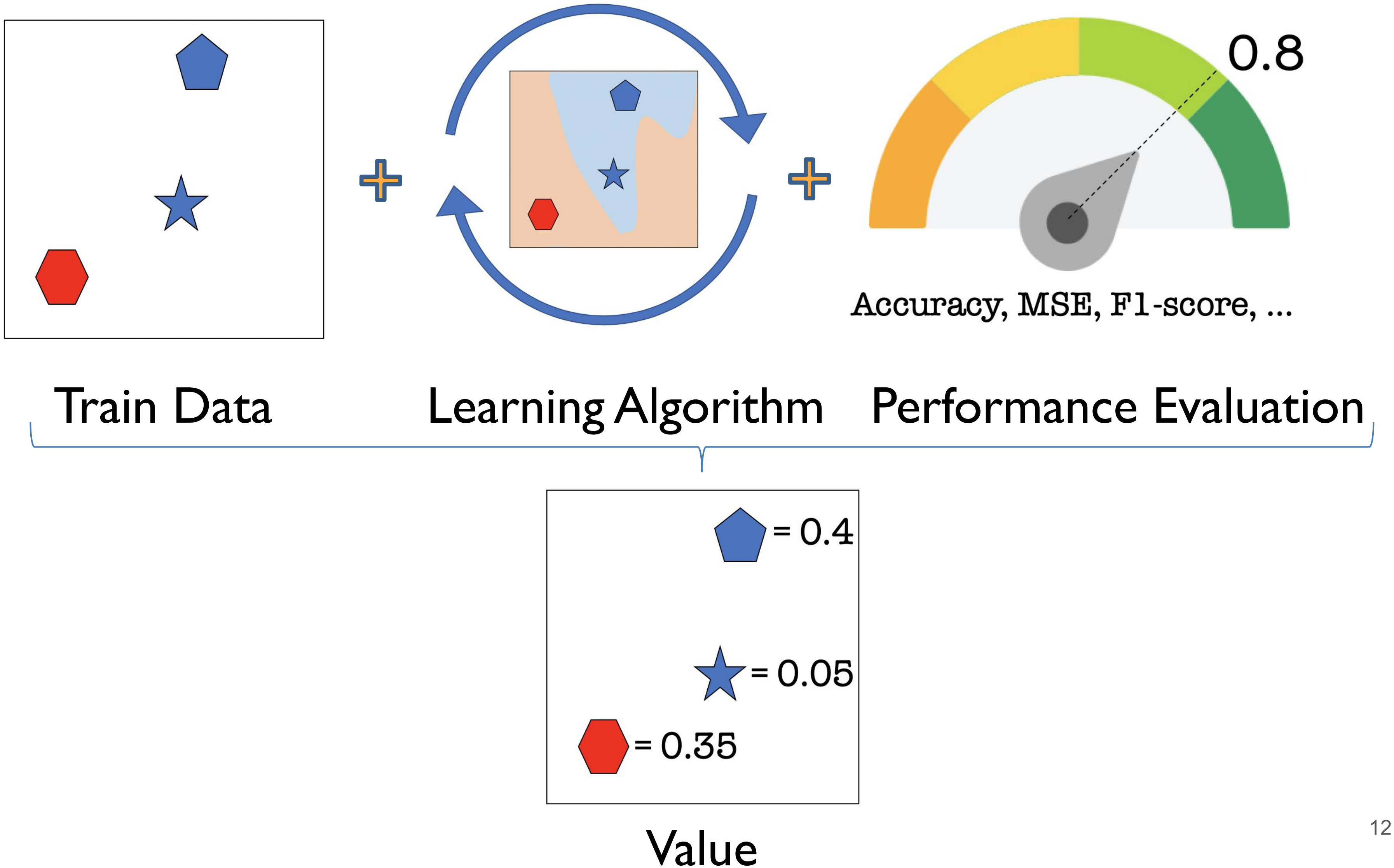




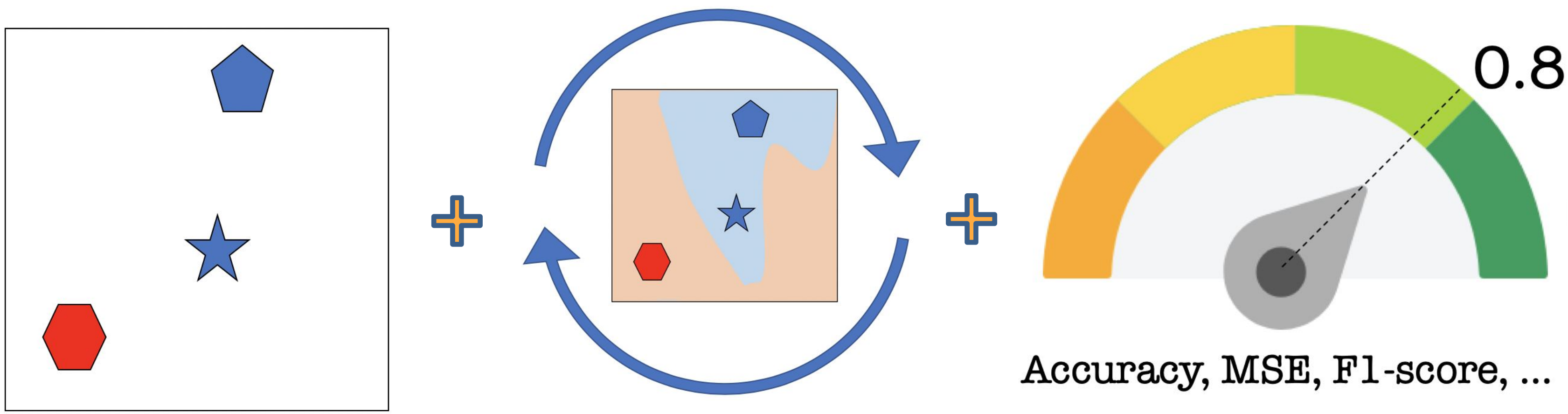


Train Data + Learning Algorithm + Performance Evaluation

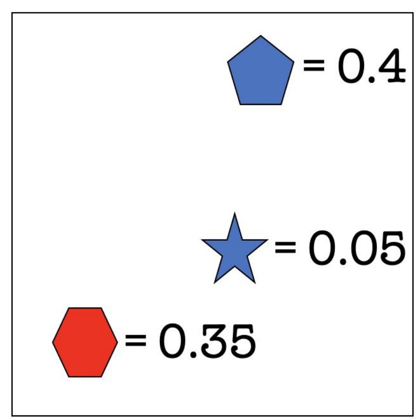
Ingredients of ML and Data Value



Ingredients of ML and Data Value



Train Data + Learning Algorithm + Performance Evaluation



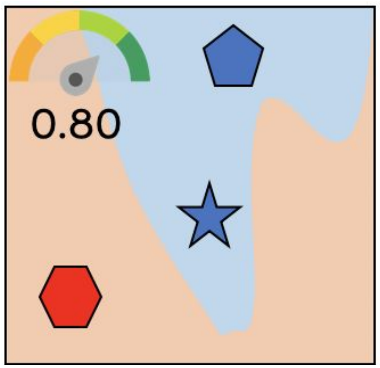
Value

Value depends on the learner, evaluation and dataset.

There are many ways to "value" data.
Is there one right way?

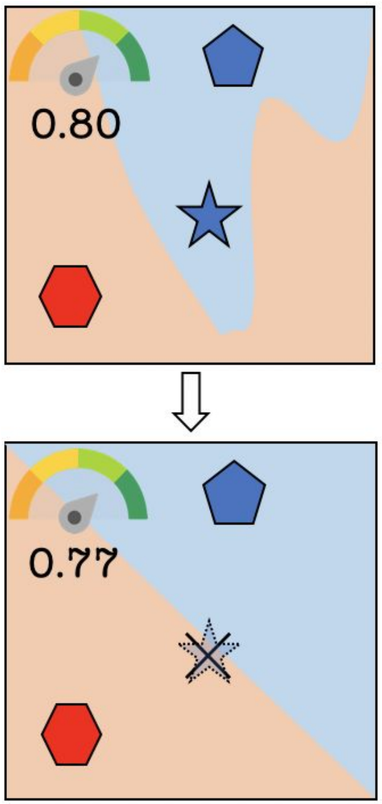
Leave One Out Score (LOO)

<u>Example</u>: value (★) = ?



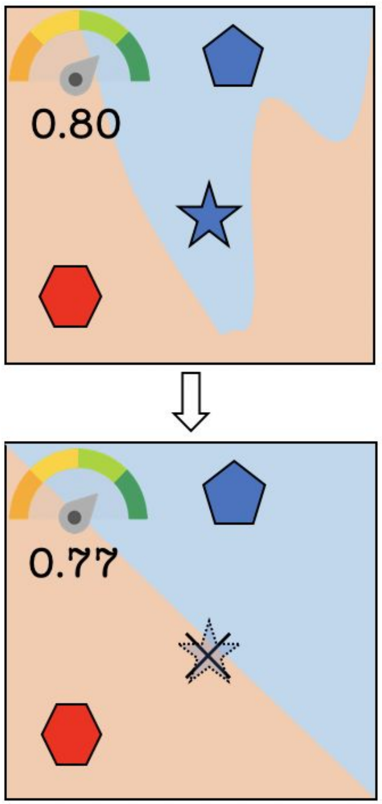
Leave One Out Score (LOO)

Example: value () = 0.80 – 0.77 = 0.03



Leave One Out Score (LOO)

Example: value (☆) = 0.80 – 0.77 = 0.03



Reasonable???

Desirable Properties of Valuation

Desirable Properties of Valuation

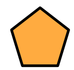
1. Null Element: If adding  to any subset of train data never changes the learned model's performance:

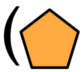
value() = 0

Desirable Properties of Valuation

1. Null Element: If adding  to any subset of train data never changes the learned model's performance:

value() = 0

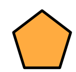
2. Symmetry: If adding  or  to any subset of train data always results in the same change in performance:

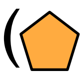
value () = value ()

Desirable Properties of Valuation

1. Null Element: If adding  to any subset of train data never changes the learned model's performance:

value() = 0

2. Symmetry: If adding  or  to any subset of train data always results in the same change in performance:

value () = value ()

3- Linearity: In ML, performance metric can be the sum of performance on individual tasks (e.g. individual test points)

$$\sum_i L(\text{classifier}(x_i^{\text{test}}), y_i^{\text{test}})$$

Desirable Properties of Valuation

1. Null Element: If adding ★ to any subset of train data never changes the learned model's performance:

value(★) = 0

2. Symmetry: If adding ★ or ⬟ to any subset of train data always results in the same change in performance:

value (⬟) = value (★)

3- Linearity: In ML, performance metric can be the sum of performance on individual tasks (e.g. individual test points)

$$\sum_i L(\text{classifier}(x_i^{\text{test}}), y_i^{\text{test}})$$

Add/remove one taks,… should correspond to add/remove value (★) for that task.

Data Shapley Value

Setting: A data point z in a dataset B containing n data points.

$$\text{value}(\text{z}) =$$

Data Shapley Value

Setting: A data point z in a dataset B containing n data points.

<u>*Theorem*</u> *(Ghorbani and Zou 19)* The only data value that satisfies these 3 properties is

$$\text{value(z)} = \sum_{S \subseteq \{\text{data points except } z\}} \frac{\text{Performance}(S \cup z) - \text{Performance}(S)}{\binom{|\{\text{data points except } z\}|}{|S|}}$$

Data Shapley Value

Setting: A data point z in a dataset B containing n data points.

Theorem (Ghorbani and Zou 19) The only data value that satisfies these 3 properties is

marginal contribution (LOO score with respect to S)

$$\underbrace{\mathrm{Performance}(S \cup z) - \mathrm{Performance}(S)}$$

Data Shapley Value

Setting: A data point z in a dataset B containing n data points.

<u>Theorem</u> (Ghorbani and Zou 19) The only data value that satisfies these 3 properties is

marginal contribution (LOO score with respect to S)

$$\frac{\text{Performance}(S \cup z) - \text{Performance}(S)}{\binom{|\{\text{data points except } z\}|}{|S|}}$$

Normalized by number of size |S| subsets

Data Shapley Value

Setting: A data point z in a dataset B containing n data points.

Theorem *(Ghorbani and Zou 19)* The only data value that satisfies these 3 properties is

marginal contribution (LOO score with respect to S)

$$\text{value(z)} = \sum_{S \subseteq \{\text{data points except } z\}} \frac{\text{Performance}(S \cup z) - \text{Performance}(S)}{\binom{|\{\text{data points except } z\}|}{|S|}}$$

Normalized by number of size |S| subsets

Data Shapley Value

Setting: A data point z in a dataset B containing n data points.

<u>*Theorem*</u> *(Ghorbani and Zou 19)* The only data value that satisfies these 3 properties is

marginal contribution (LOO score with respect to S)

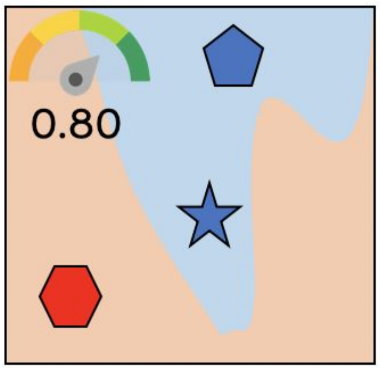
$$\text{value(z)} = \sum_{S \subseteq \{\text{data points except } z\}} \frac{\text{Performance}(S \cup z) - \text{Performance}(S)}{\binom{|\{\text{data points except } z\}|}{|S|}}$$

Normalized by number of size |S| subsets

Expected LLO scores with respect to all possible sizes of data

Data Shapley Value

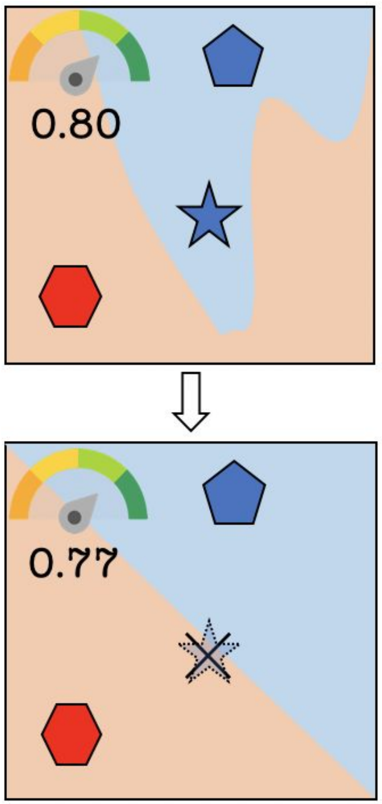
<u>Example</u>: value () = ?



Data Shapley Value

<u>Example</u>: value (☆) = ?

|S| = 2



$$\frac{0.80 - 0.77}{1}$$

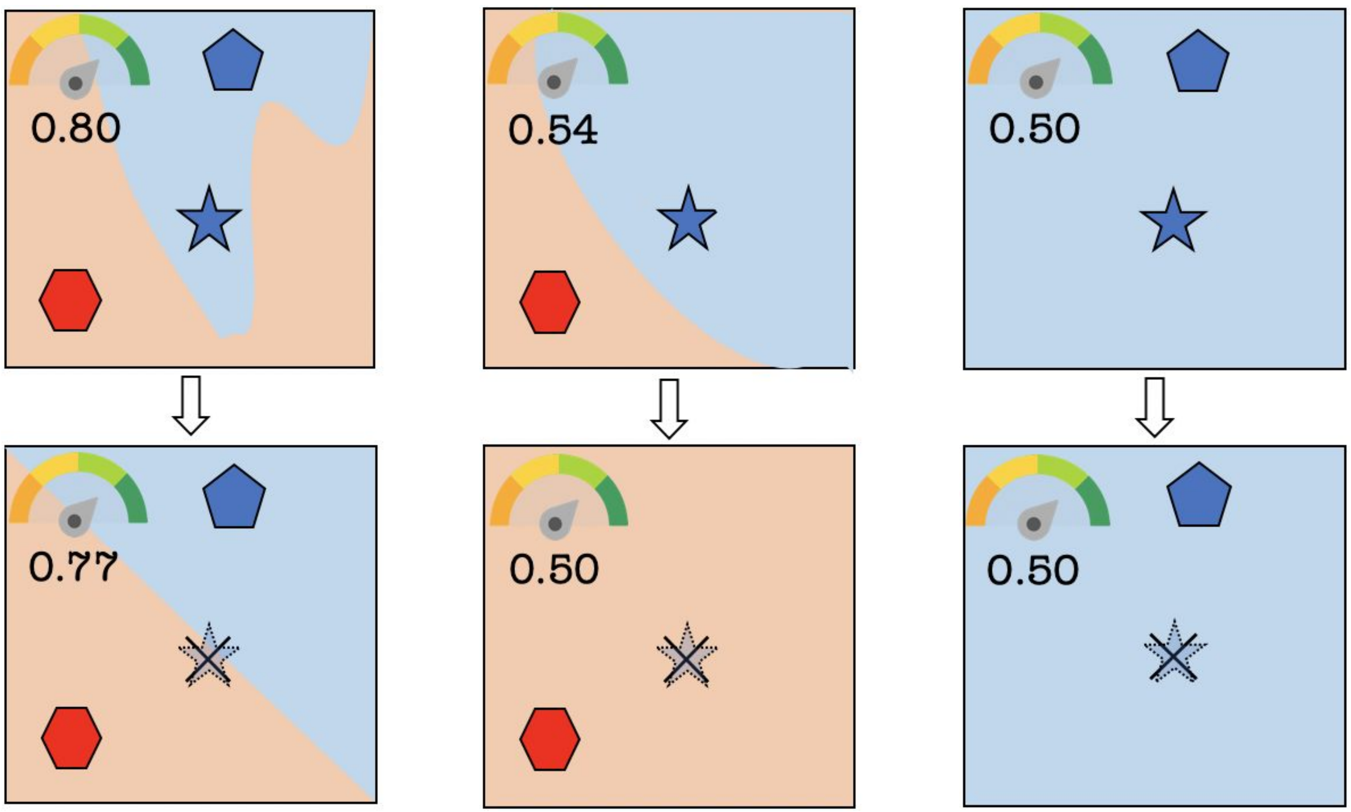
Two size one subset

Data Shapley Value

One size two subset

<u>Example</u>: value (★) = ?

|S| = 2 |S| = 1 |S| = 1



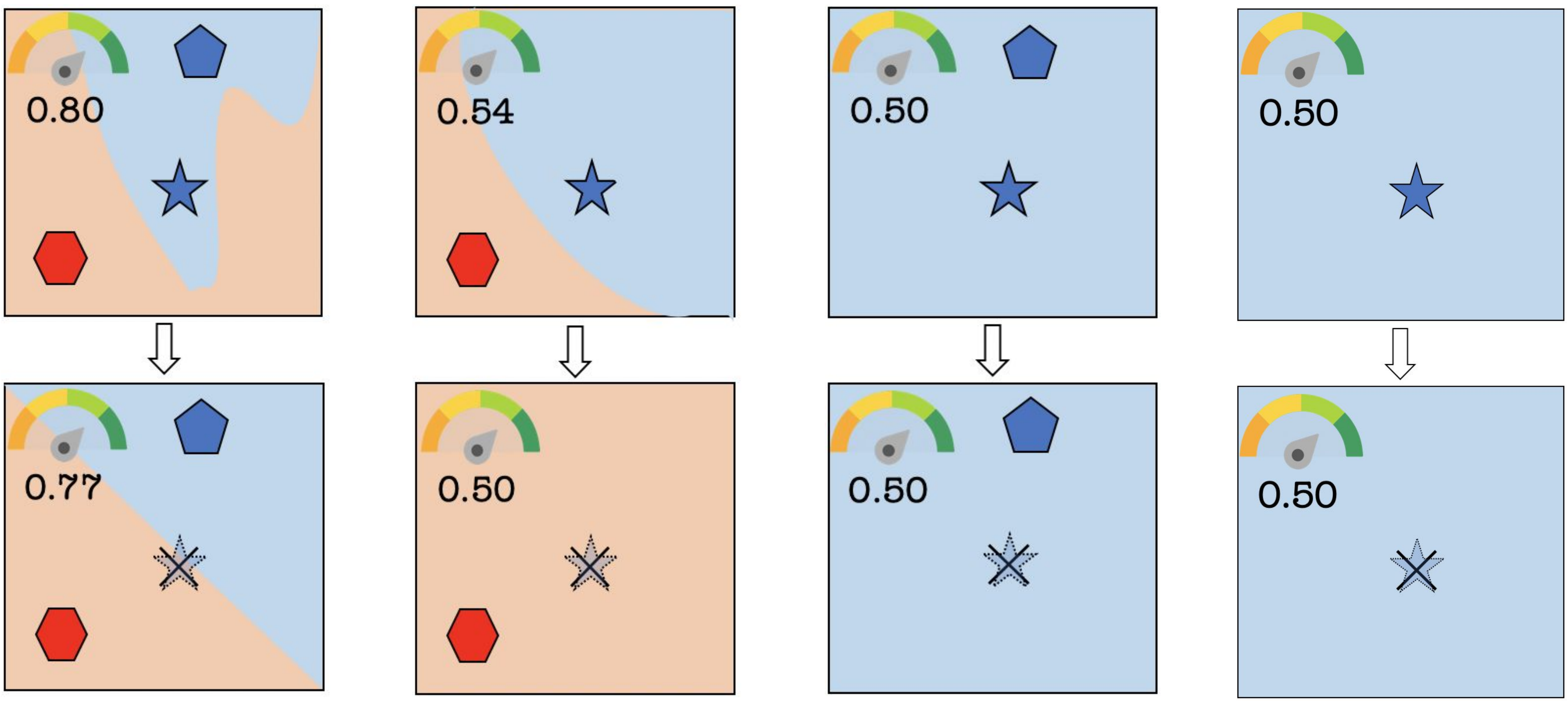
$$\frac{0.80 - 0.77}{1} + \frac{0.54 - 0.5}{2} + \frac{0.5 - 0.5}{2}$$

One size zero subset

Data Shapley Value

Example: value (★) = 0.05

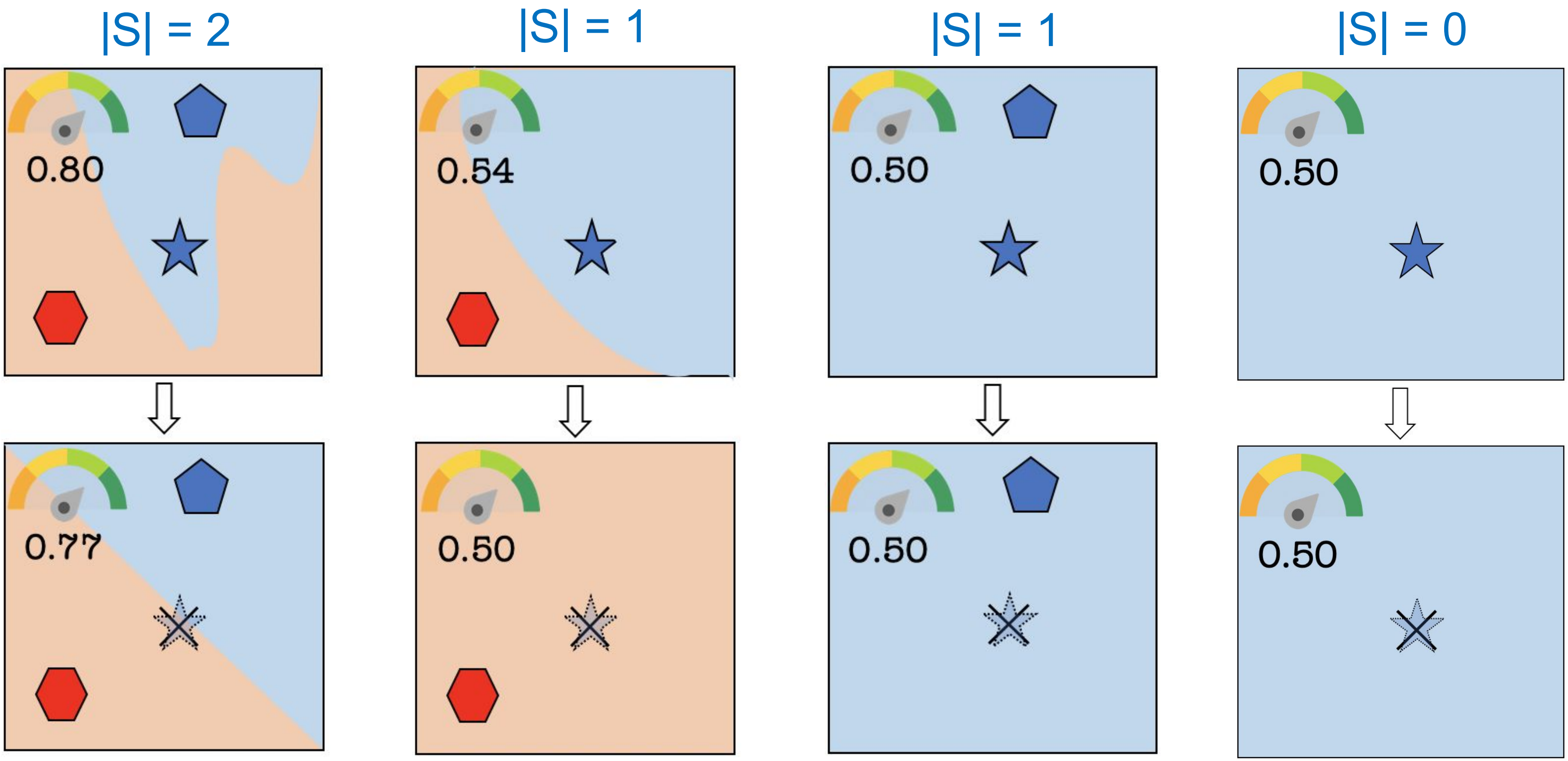
|S| = 2 | |S| = 1 | |S| = 1 | |S| = 0



$$\frac{0.80 - 0.77}{1} + \frac{0.54 - 0.5}{2} + \frac{0.5 - 0.5}{2} + \frac{0.5 - 0.5}{1}$$

Data Shapley Value

<u>Example</u>: value (★) = 0.05

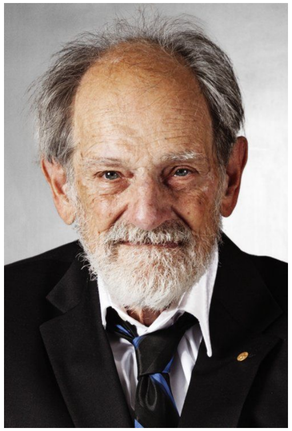


$$\frac{0.80 - 0.77}{1} + \frac{0.54 - 0.5}{2} + \frac{0.5 - 0.5}{2} + \frac{0.5 - 0.5}{1}$$

We developed efficient algorithms to estimate data Shapley for complex models.

Data Shapley Value

Lloyd Shapley



2012 Nobel Prize in Economics

Cooperative game

Applications

Data point value = expected contribution to performance

Applications

Data point value = expected contribution to performance

High value data

↓

Adds significant information

Applications

Data point value = expected contribution to performance

High value data

↓

Adds significant information
e.g. in-distribution clean data

Applications

Data point value = expected contribution to performance

High value data

↓

Adds significant information
e.g. in-distribution clean data

Low value data

↓

Adds low or harmful information

Applications

Data point value = expected contribution to performance

High value data

↓

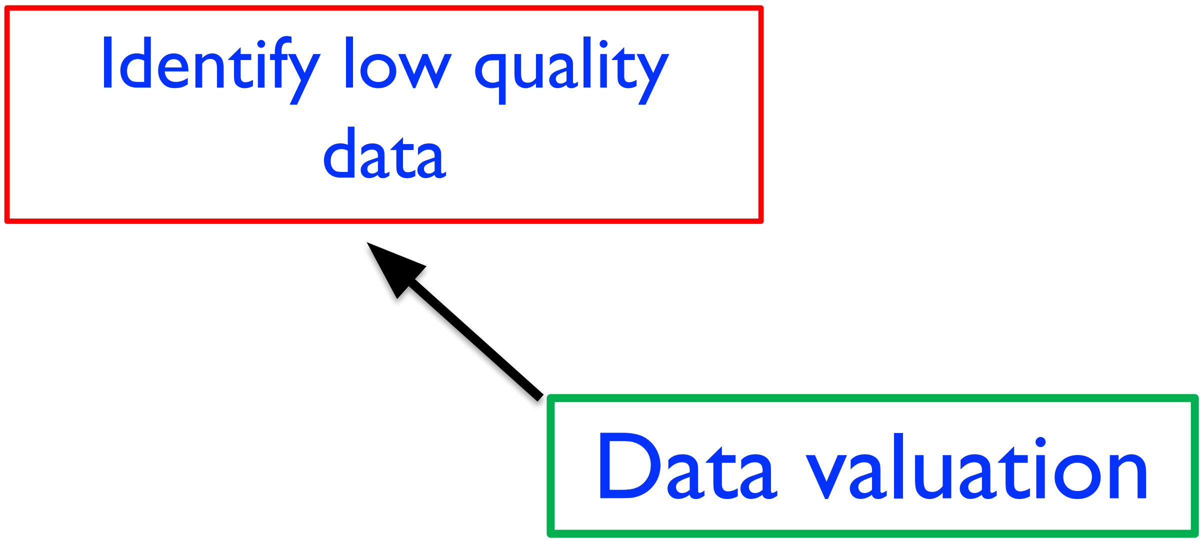
Adds significant information
e.g. in-distribution clean data

Low value data

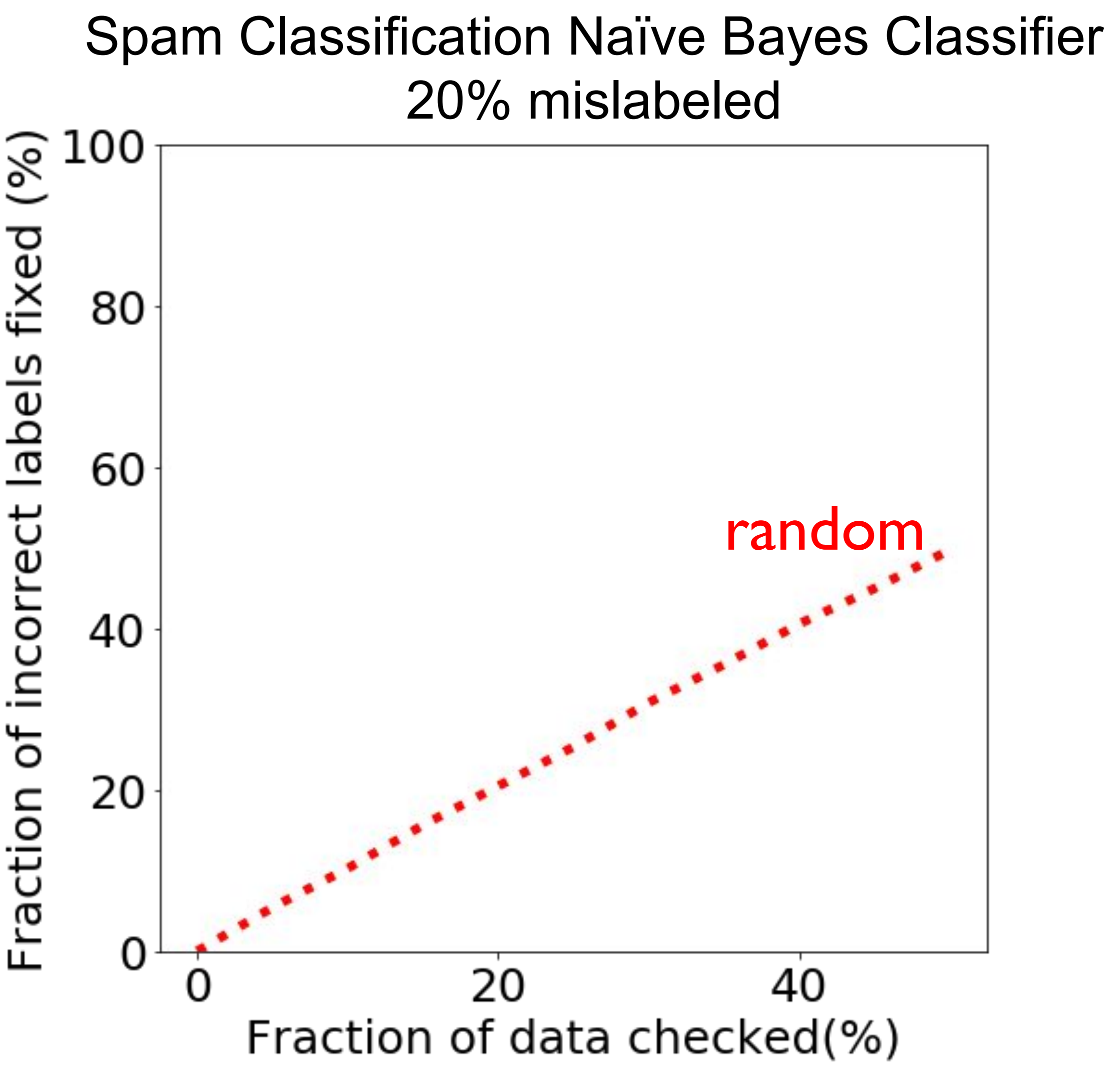
↓

Adds low or harmful information
e.g. noisy data, outliers, mislabeled data

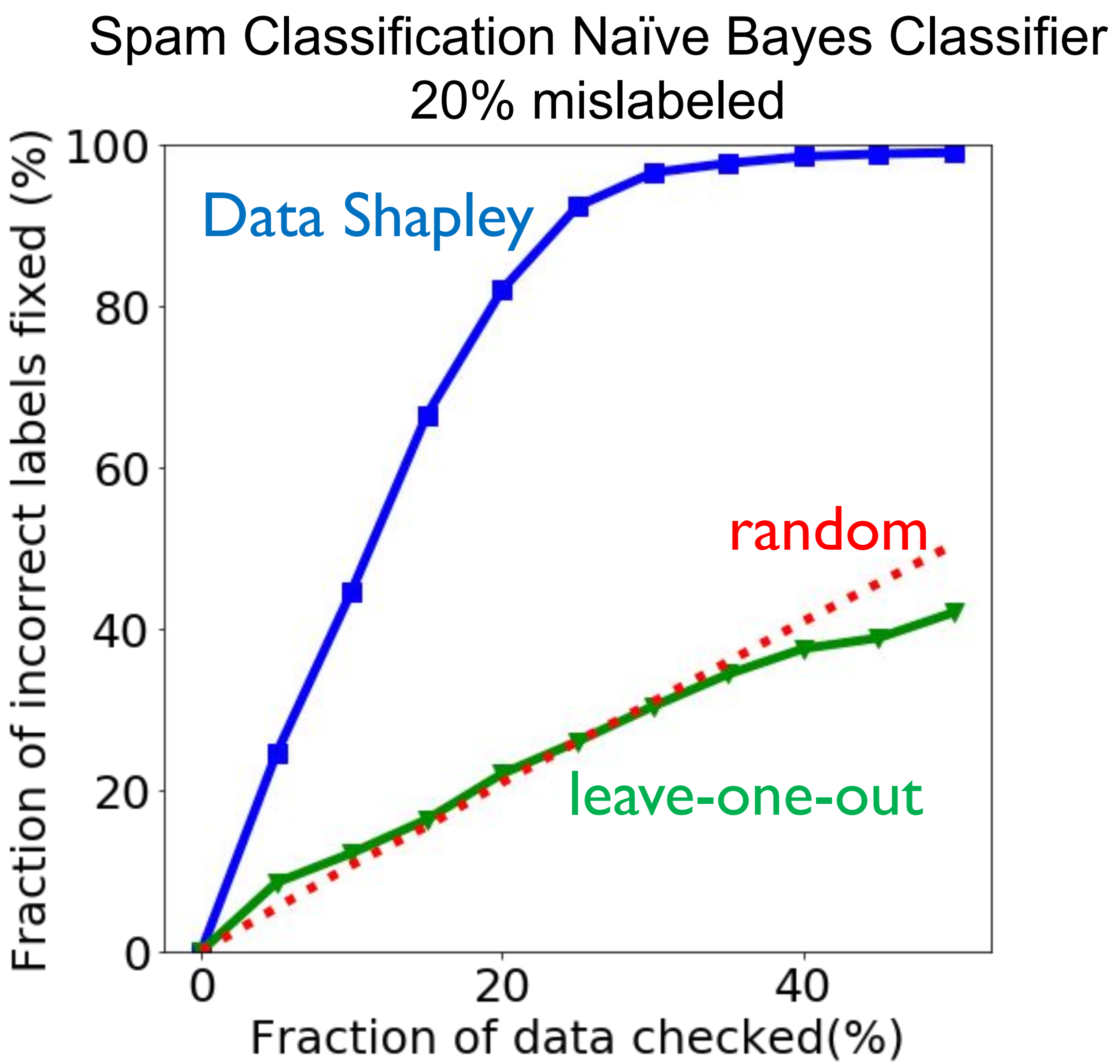
Applications



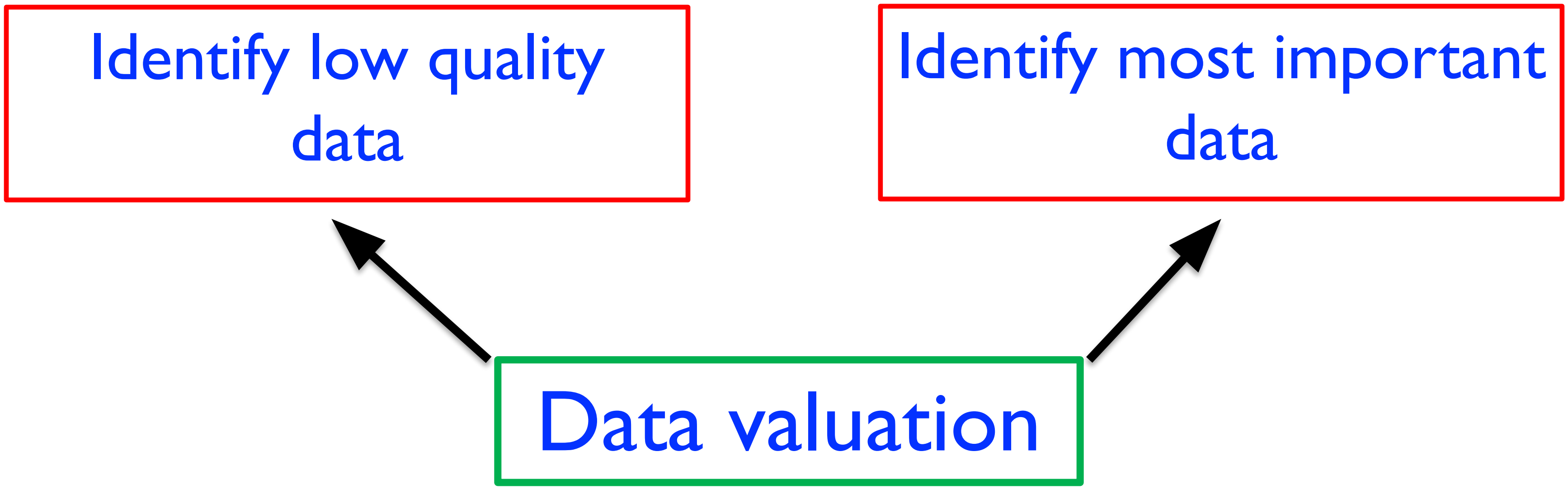
Applications: Idnetifying mislabeled data



Applications: Idnetifying mislabeled data



Applications



Applications: Identifying essential data

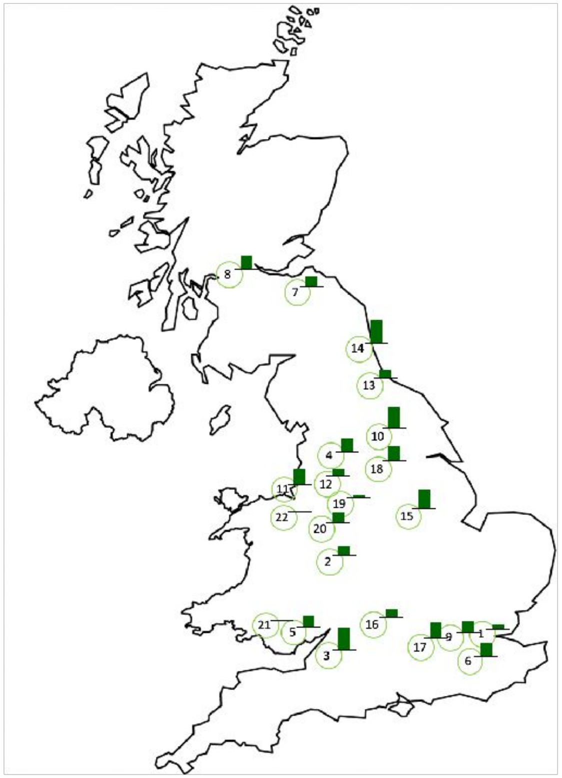
- UK Biobank Data set
- 500,000 individual in UK
- Phenotype, Genotype
- Gathered from 22 centers in UK

Applications: Identifying essential data

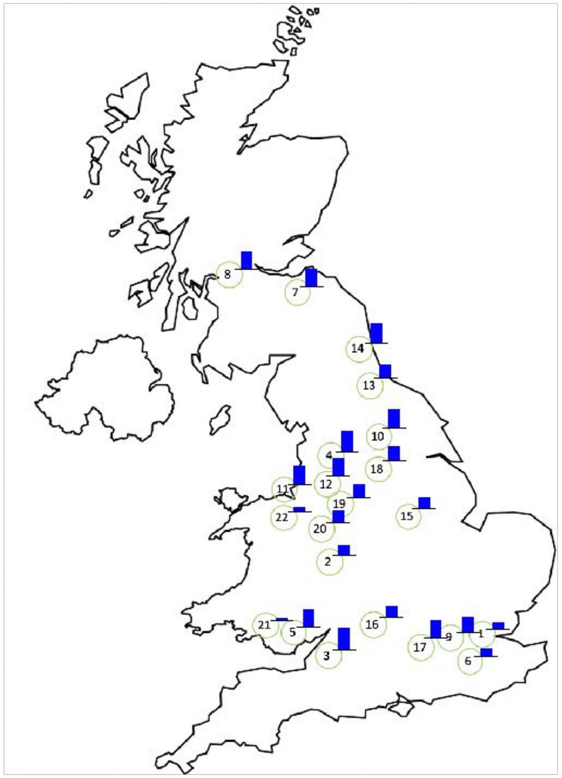
- UK Biobank Data set
- 500,000 individual in UK
- Phenotype, Genotype
- Gathered from 22 centers in UK = 22 data sources
- We create binary-balanced disease prediction datasets
- Let's look at each center as source of data...

Applications: Identifying essential data

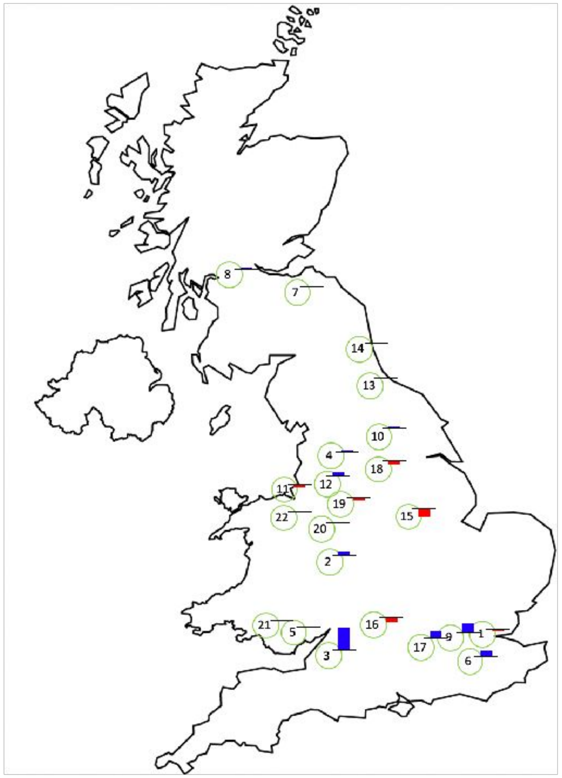
of patients



Breast Cancer

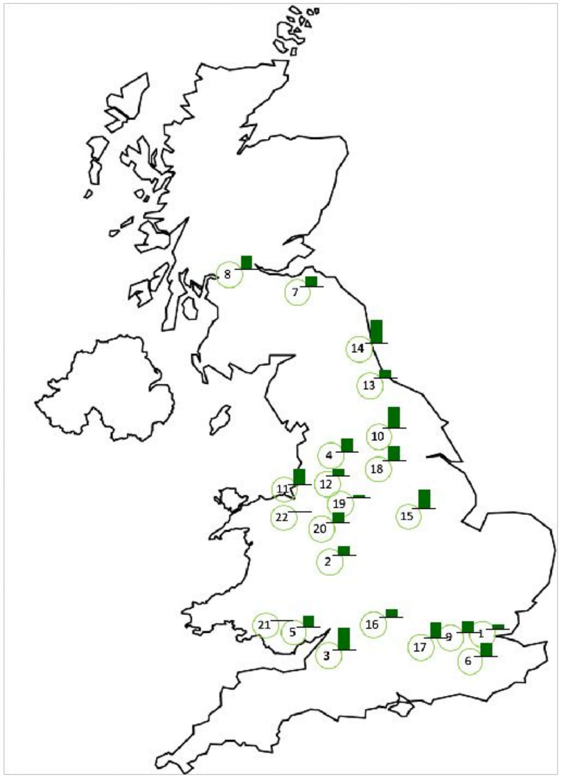


Colon Cancer



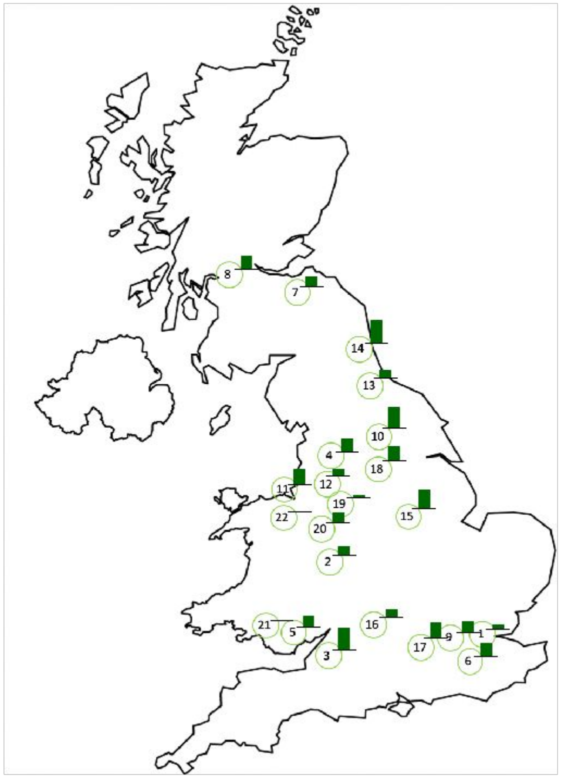
Applications: Identifying essential data

of patients

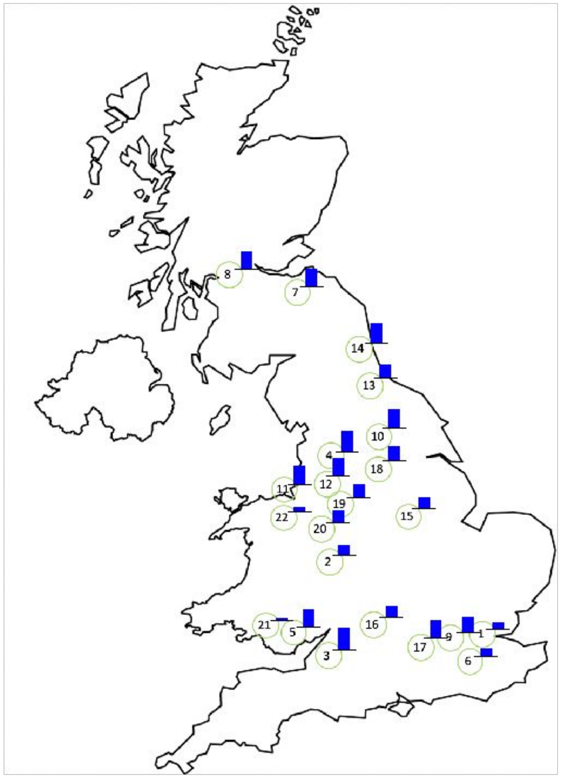


Applications: Identifying essential data

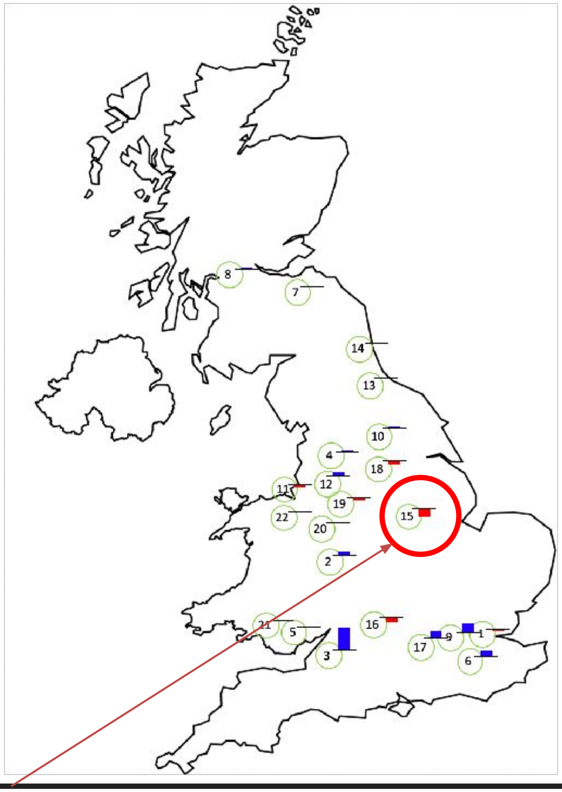
of patients



Breast Cancer



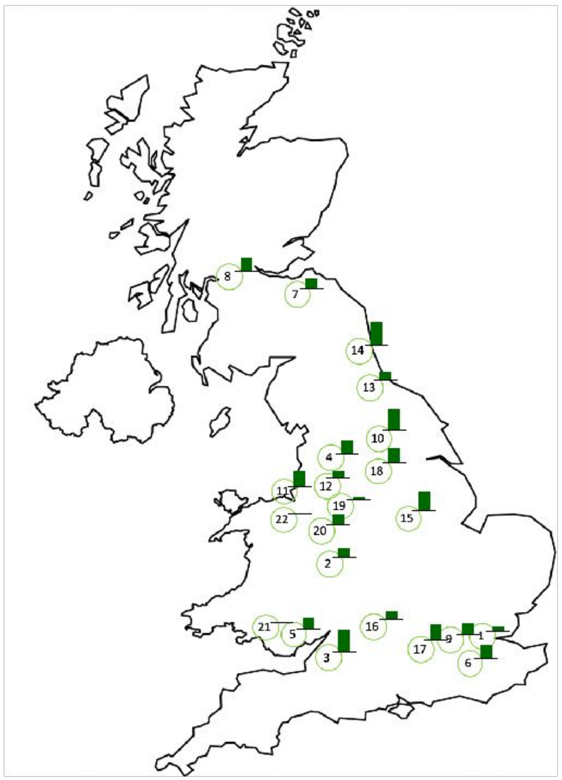
Colon Cancer



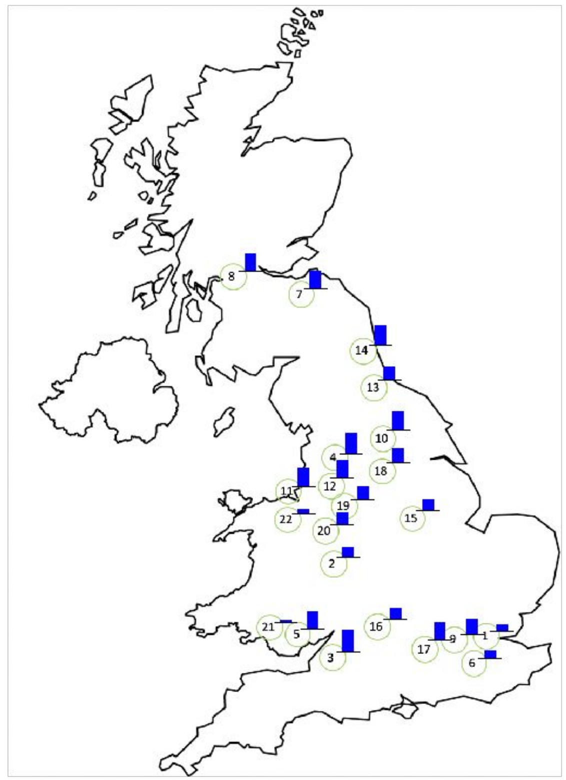
Most predictive feature: Age ⬆ □ Colon Cancer ⬆ (p=1.5e-6)

Applications: Identifying essential data

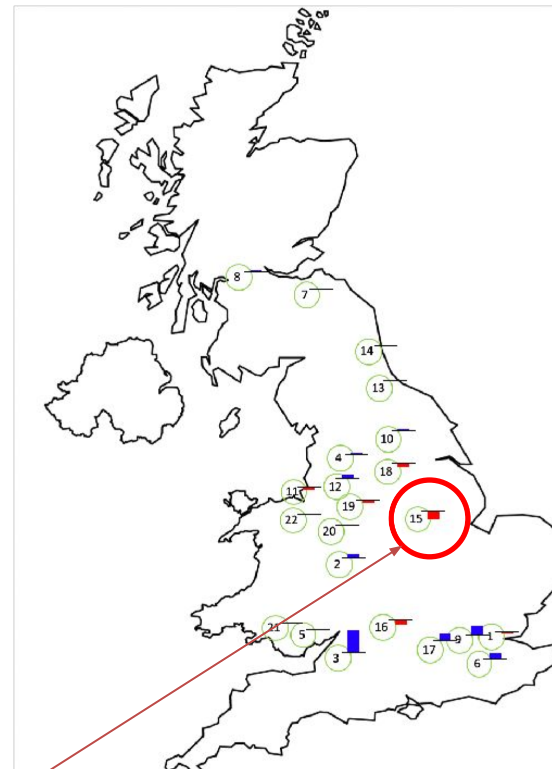
of patients



Breast Cancer



Colon Cancer



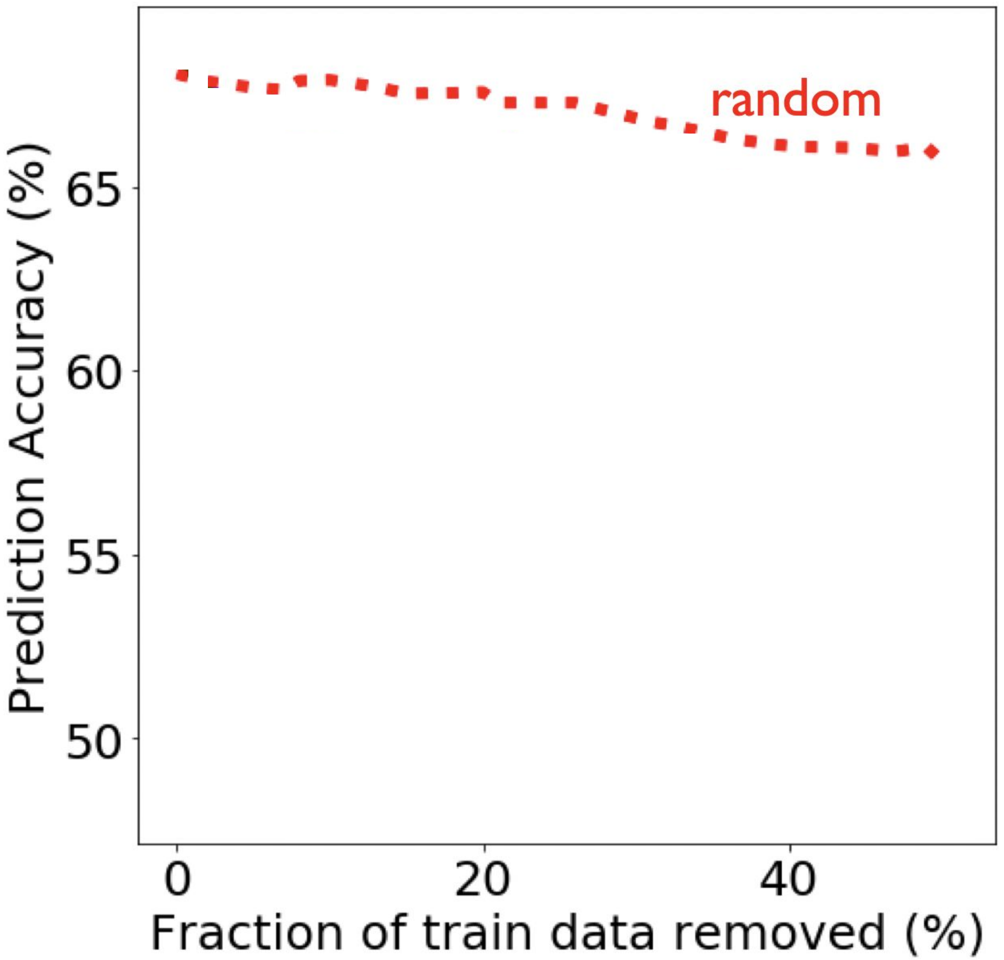
Most predictive feature: Age↑ □ Colon Cancer ↑(p=1.5e-6)
Center-15: cancer unrelated to age (p=0.14)

Applications: Identifying essential data

- UK Biobank Data set
- 500,000 individual in UK
- Phenotype, Genotype
- Gathered from 22 centers in UK = 22 data sources
- We create binary-balanced disease prediction datasets
- Let's look at individual data points as sources of data...

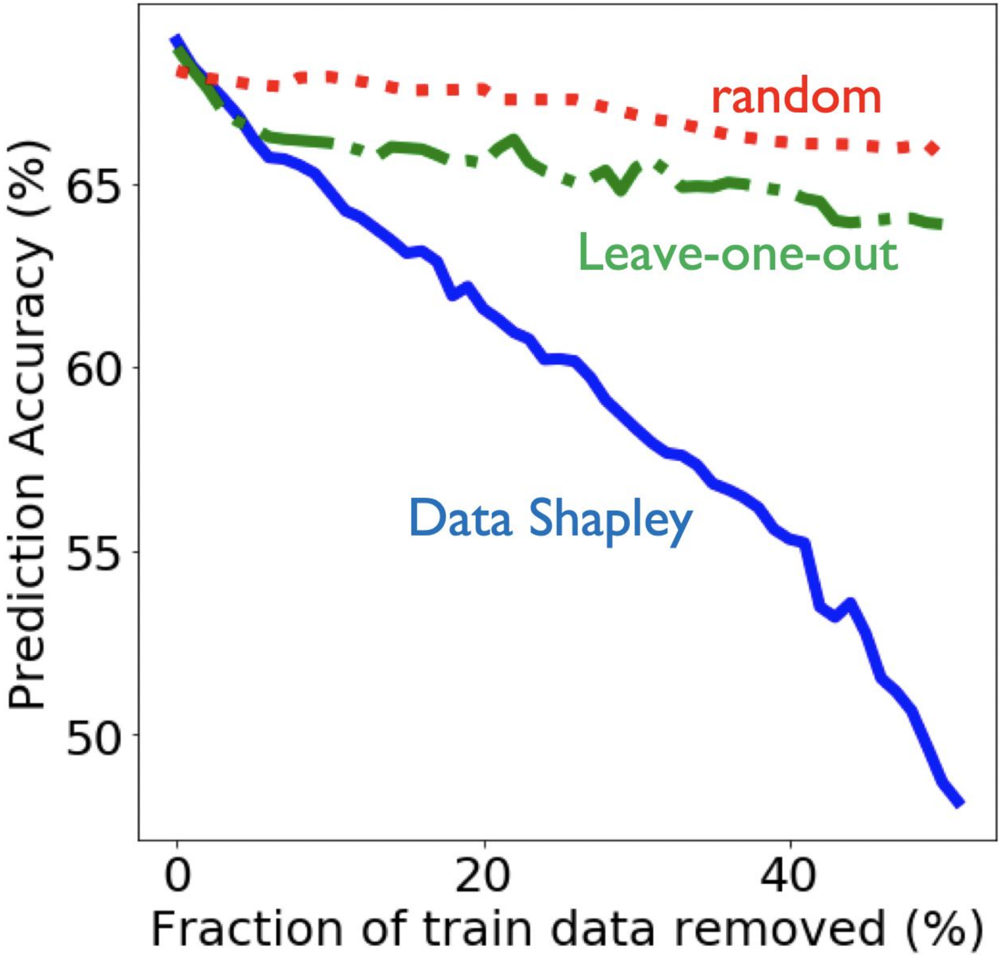
Applications: Identifying essential data

Breast Cancer



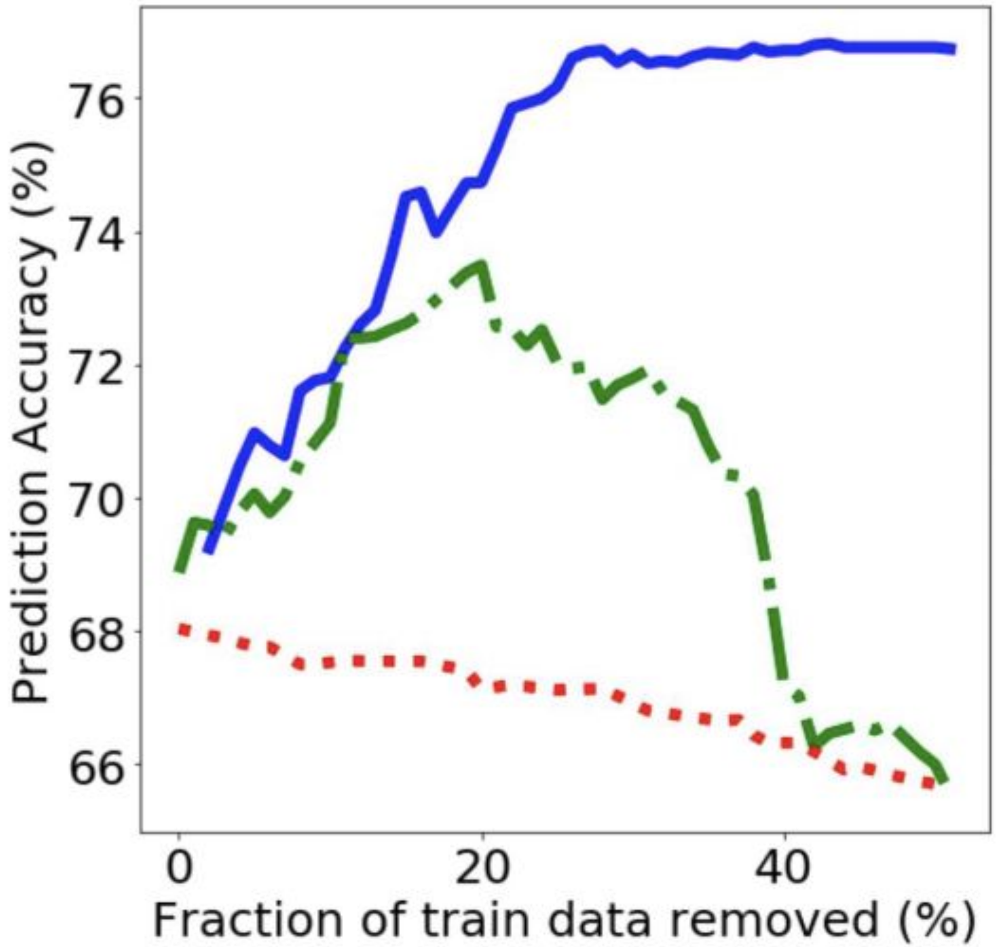
Applications: Identifying essential data

Breast Cancer



Applications: Identifying bad data

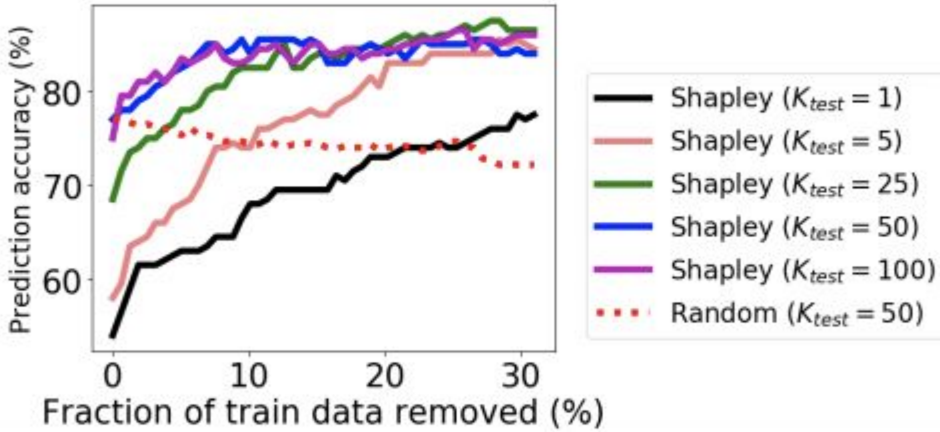
Breast Cancer



Applications: Identifying bad data

#acl2020nlp #acl2020en

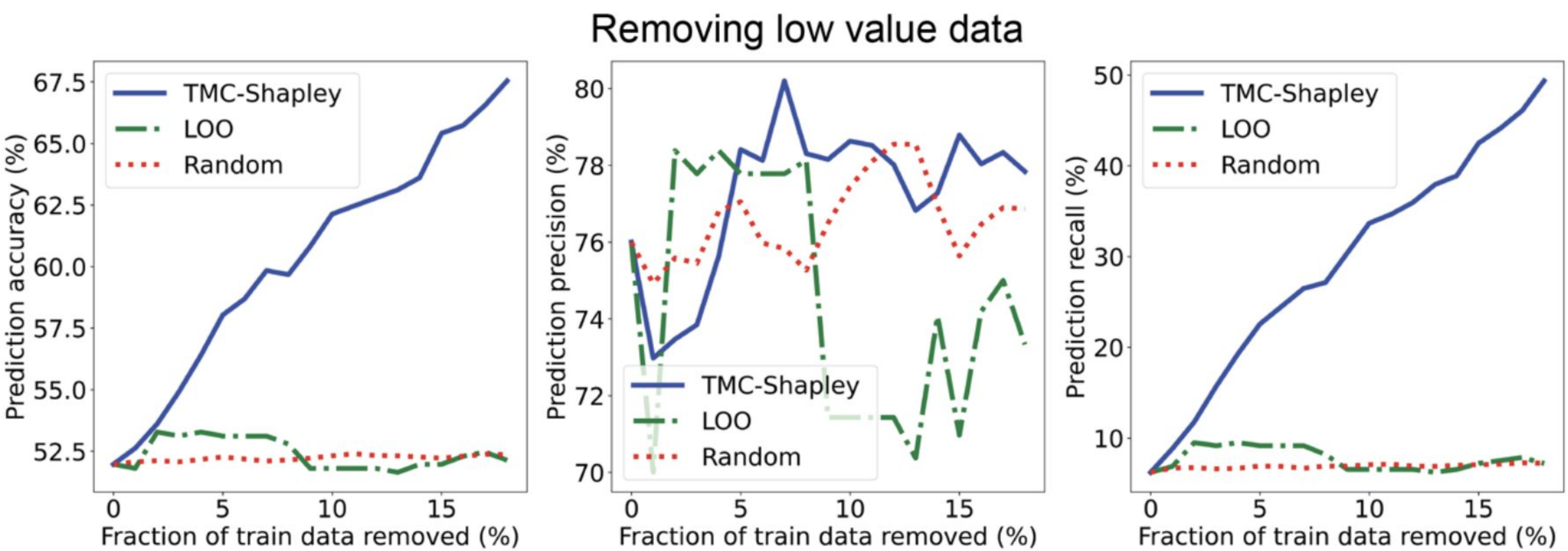
"Beyond User Self-Reported Likert Scale Ratings: A Comparison Model for Automatic Dialog Evaluation"



No.	Model	Test Acc.	Kappa	
			κ	SE
(1)	BERT-Classification	0.581	0.161	0.049
(2)	BERT-Regression	0.640	0.280	0.048
(3)	BERT-Pairwise	0.730	0.459	0.044
(4)	BERT-Pairwise+Dev	0.749	0.499	0.043
(5)	Stage 2	0.755	0.509	0.043
(6)	Stage 2 + 3	0.764	0.529	0.042
(7)	Stage 3	0.714	0.429	0.045
(8)	Stage 1	0.620	0.241	0.048
(9)	Stage 1 + 3	0.788	0.628	0.039
(10)	Stage 1 + 2	0.837	0.673	0.037
(11)	**CMADE**	**0.892**	**0.787**	**0.031**

Applications: Identifying bad data

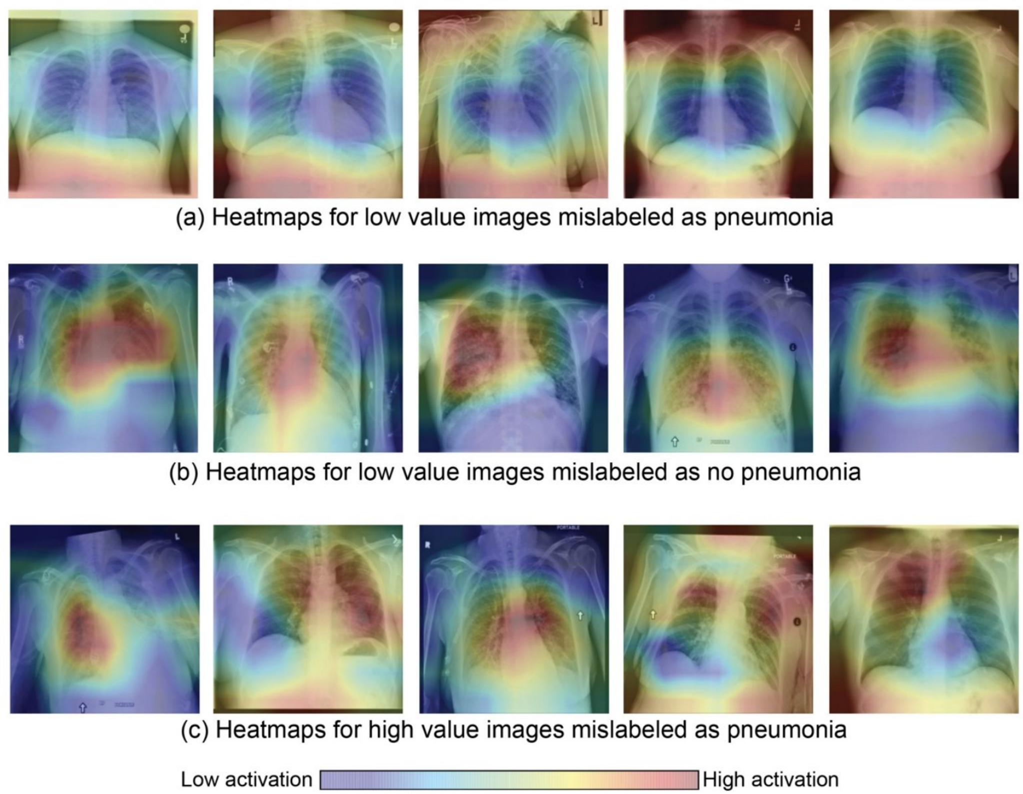
Data Valuation for Medical Imaging Using Shapley Value: Application on A Large-scale Chest X-ray Dataset



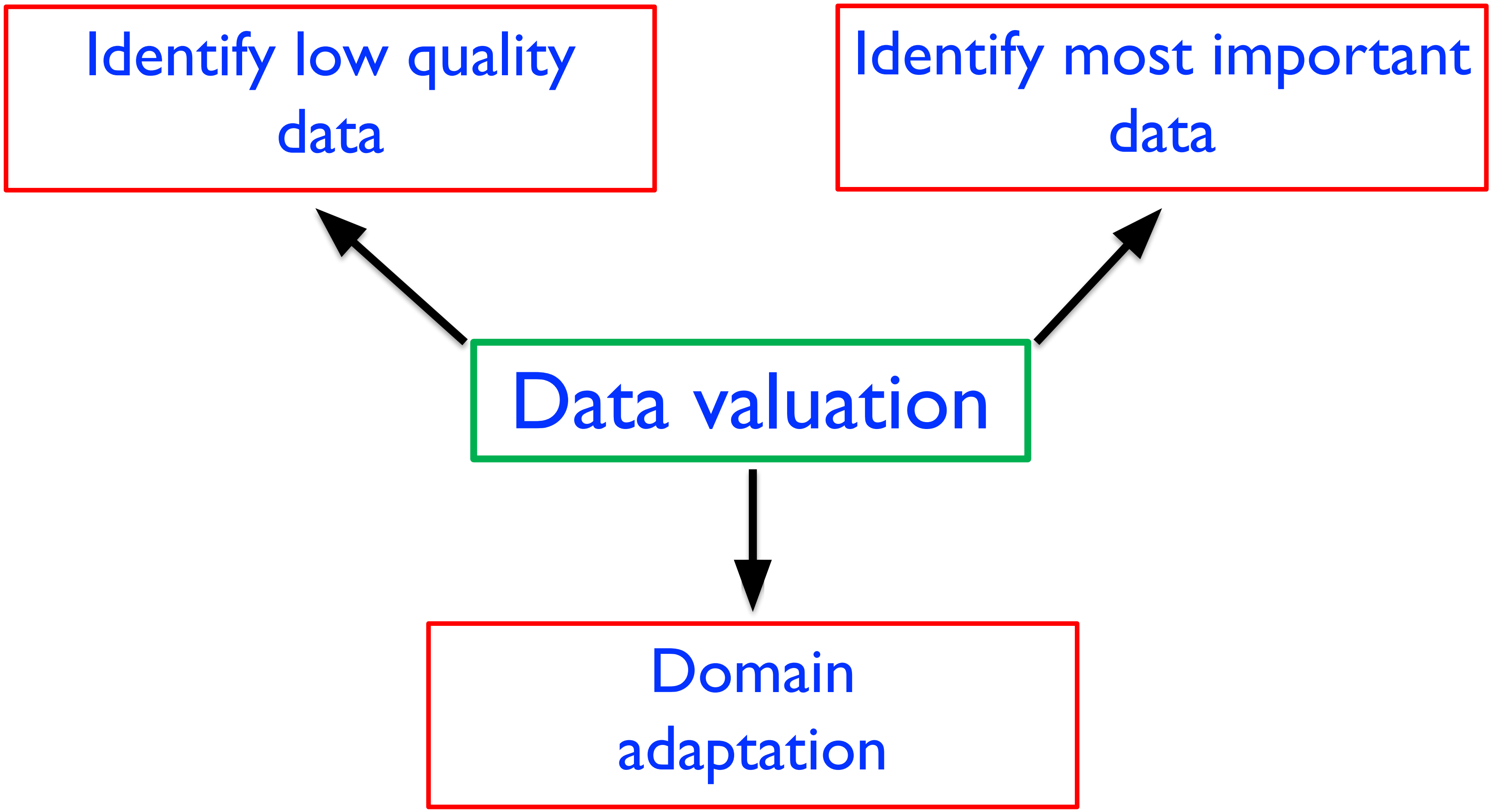
(d) Accuracy (e) Precision (f) Recall

Applications: Identifying bad data

Data Valuation for Medical Imaging Using Shapley Value: Application on A Large-scale Chest X-ray Dataset



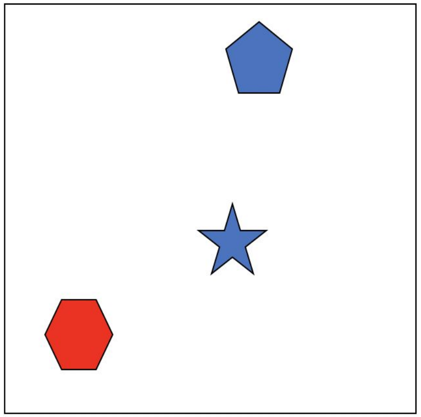
If data is fuel, then we need to measure its value



Domain adaptation: face recognition

Training data

Trained model



Test data

Performance Valution

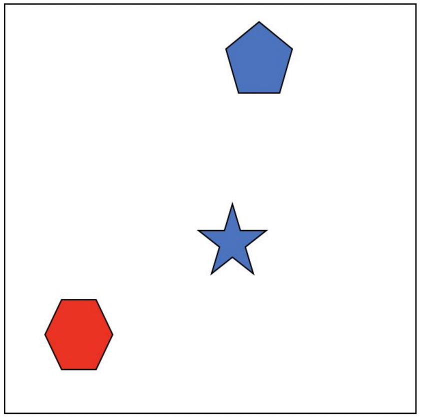
Accuracy, MSE, F1-score, ...

Domain adaptation: face recognition

Are there train data points that are harmful/helpful for adaptation?

Training data

Trained model



Different in quality, distribution, class balance, etc.

Test data

Performance Valution

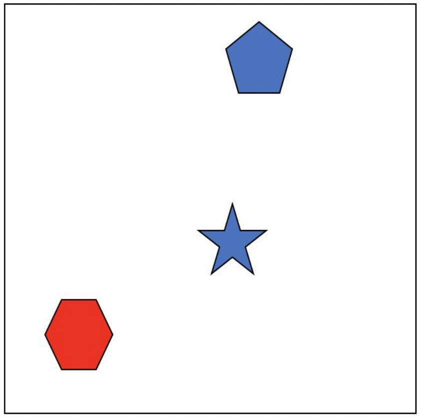
Accuracy, MSE, F1-score, ...

Domain adaptation: face recognition

1-Remove data with negative value

Training data

Trained model



Different in quality, distribution, class balance, etc.

Test data

Performance Valution

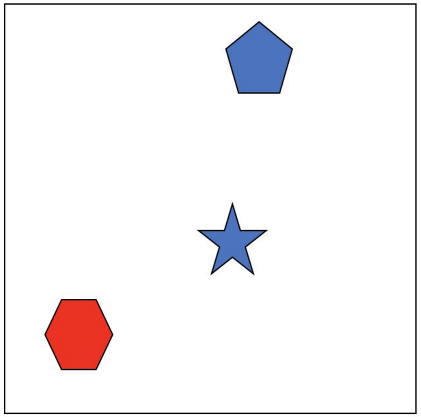
Accuracy, MSE, F1-score, ...

Domain adaptation: face recognition

I-Remove data with negative value
II-Reweight rest of the data with relative weight

Training data

Trained model



Different in quality, distribution, class balance, etc.

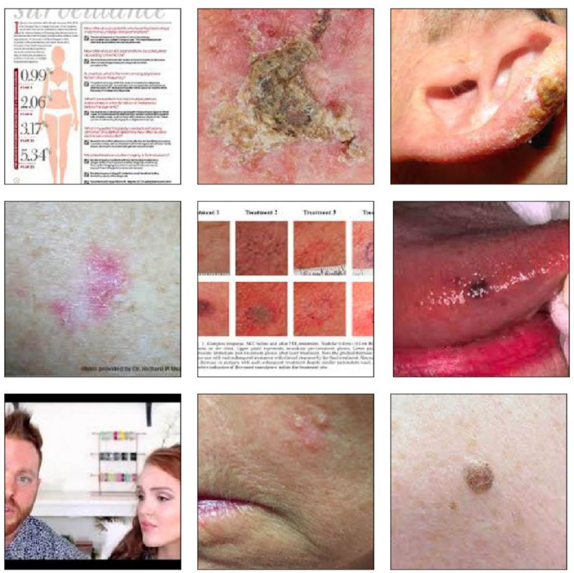
Test data

Performance Valution

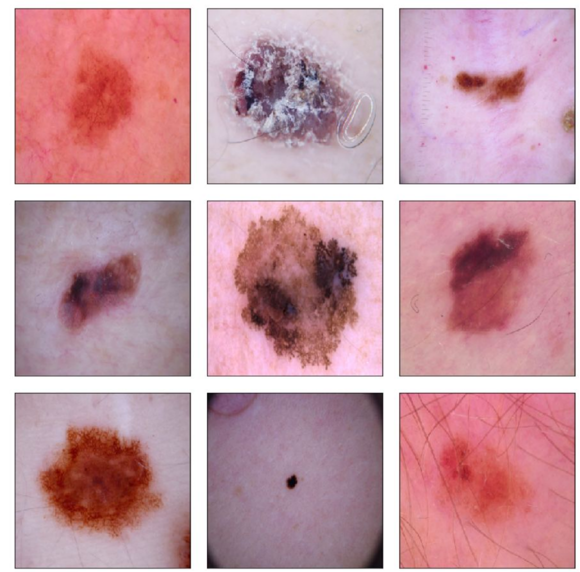
Accuracy, MSE, F1-score, ...

Skin lesion classification

Train data
google image search

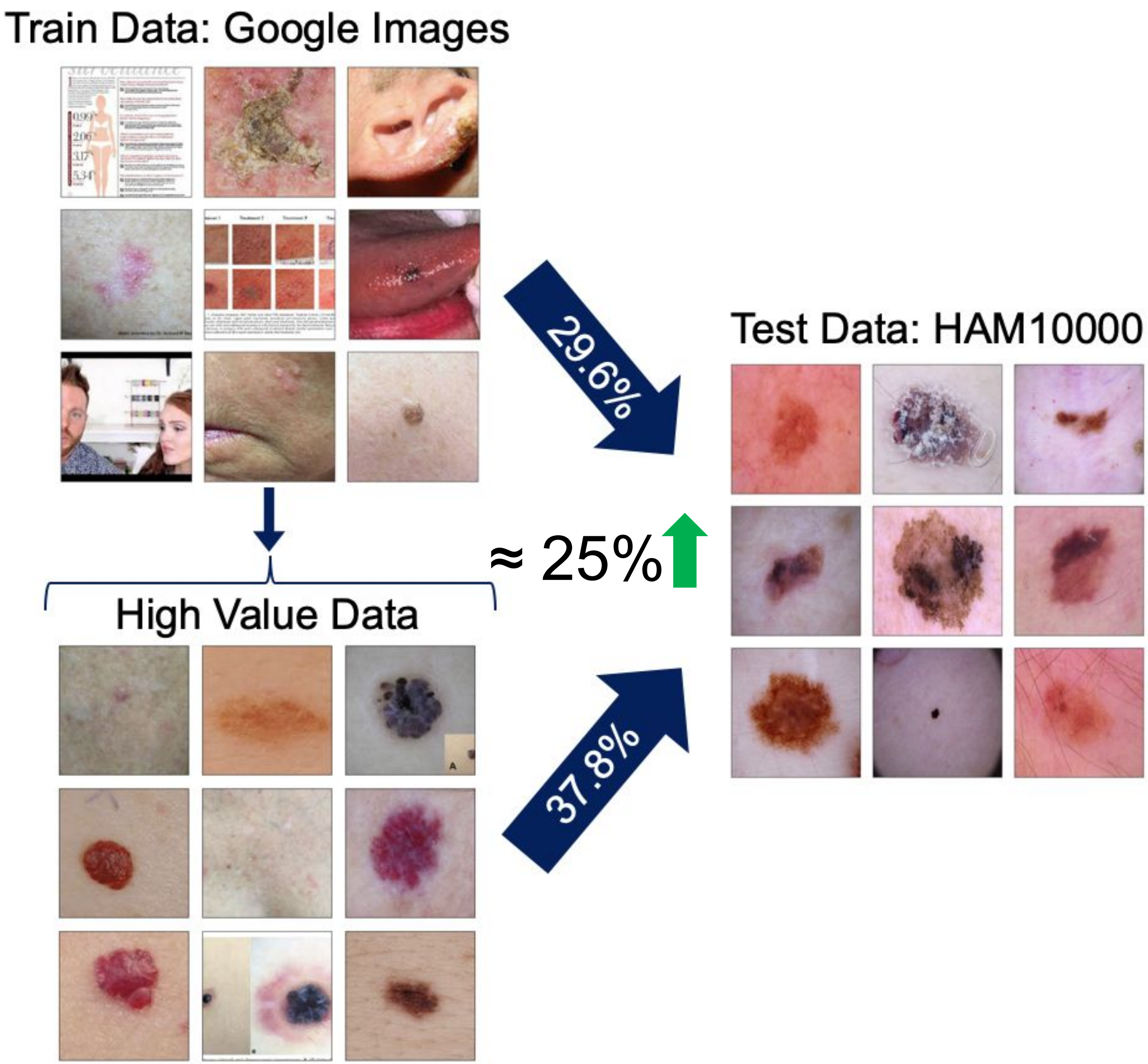


Target data
Clinical examples



accuracy

Skin lesion classification



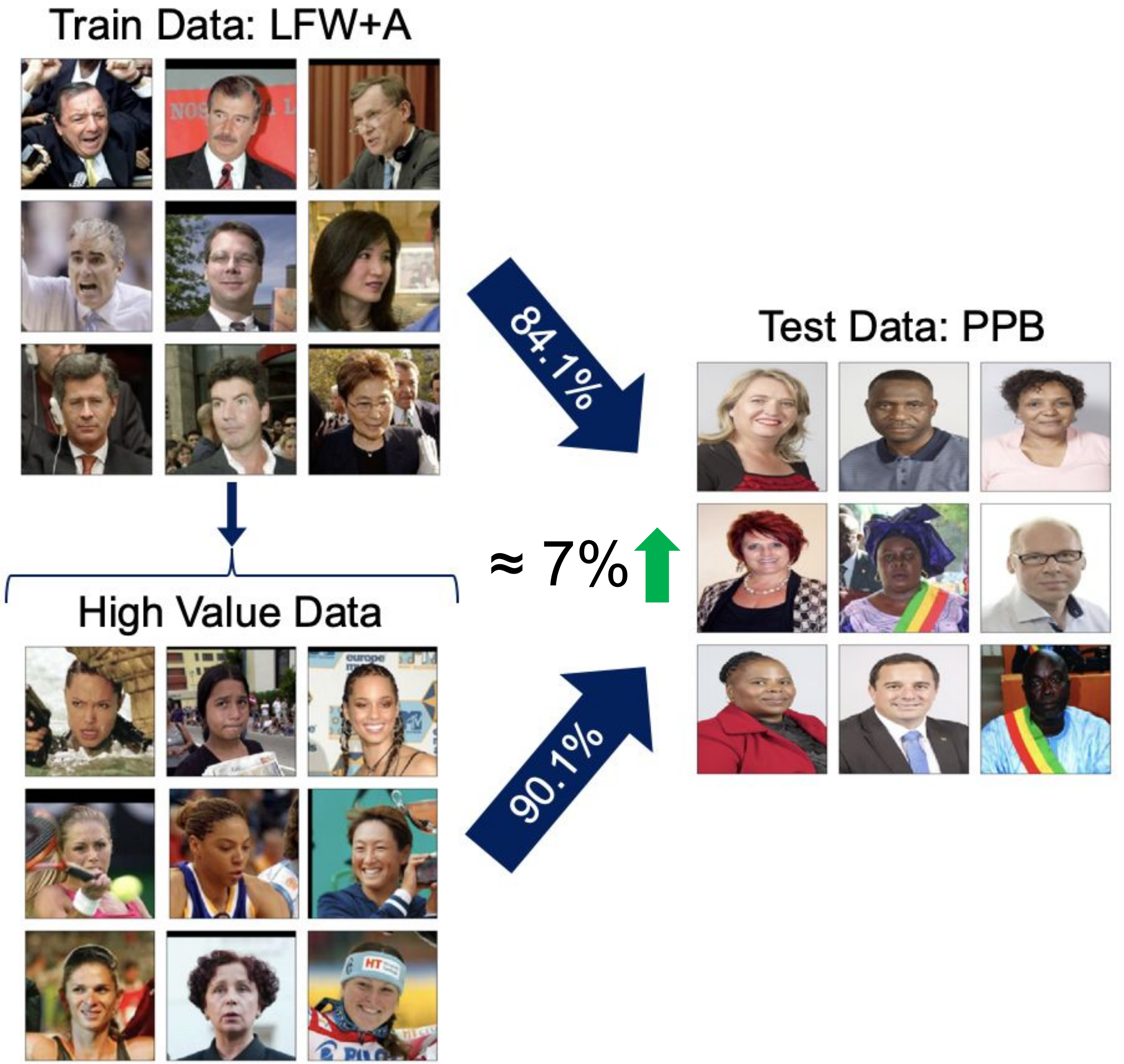
Domain adaptation: gender detection

Train Data: LFW+A

accuracy ↓
esp. for minoriti

Test Data: PPB

Domain adaptation: gender detection



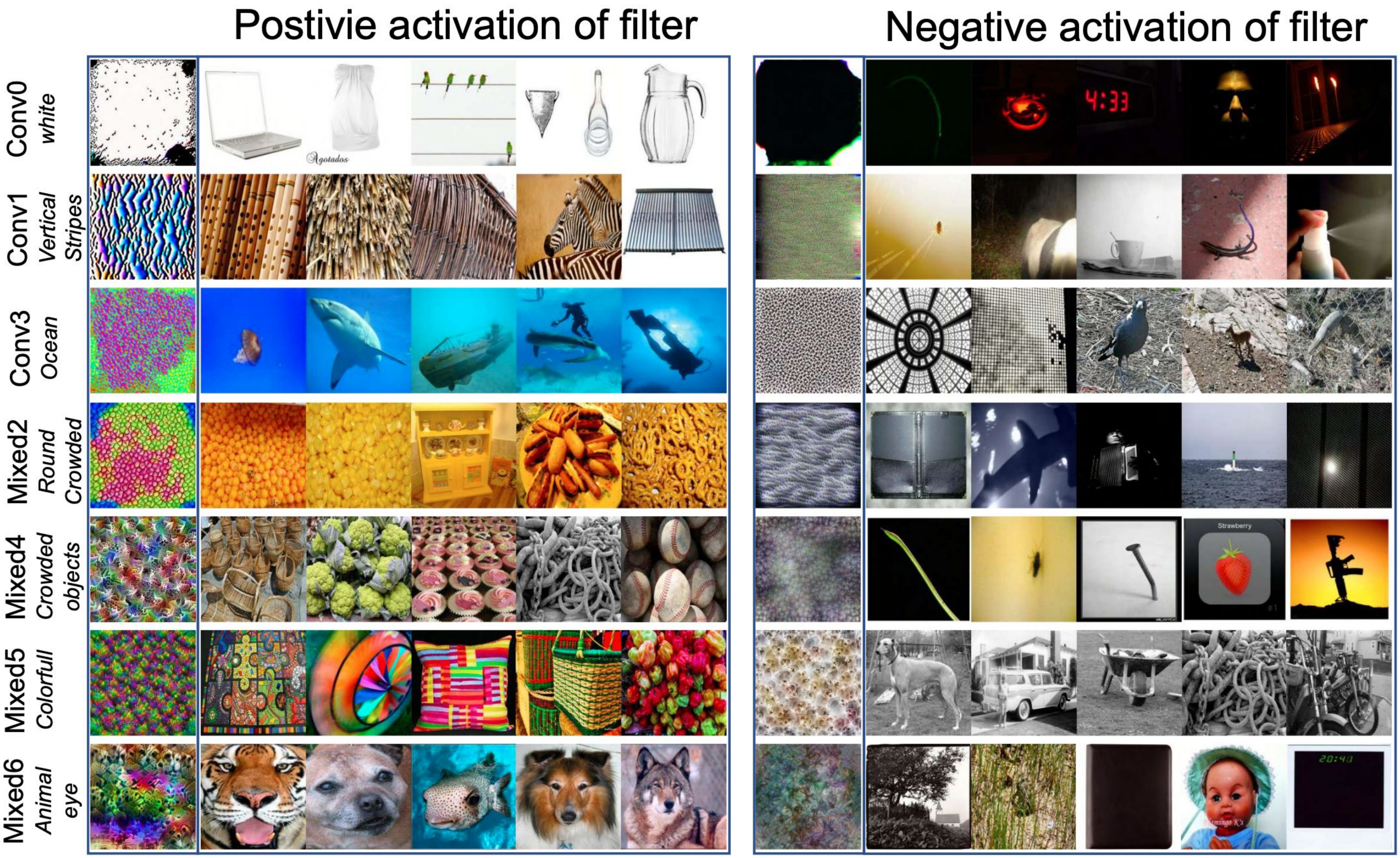
Neuron Shapley: Similar idea

Neuron Shapley: Discovering the Responsible Neurons

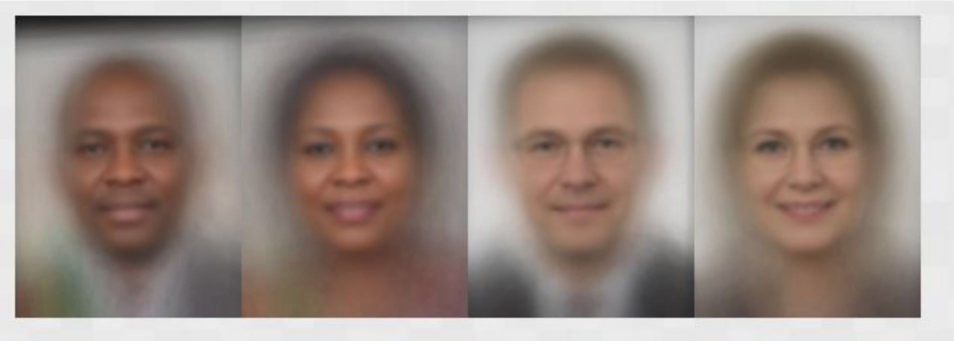
Algorithm 1 Truncated Multi Armed Bandit Shapley

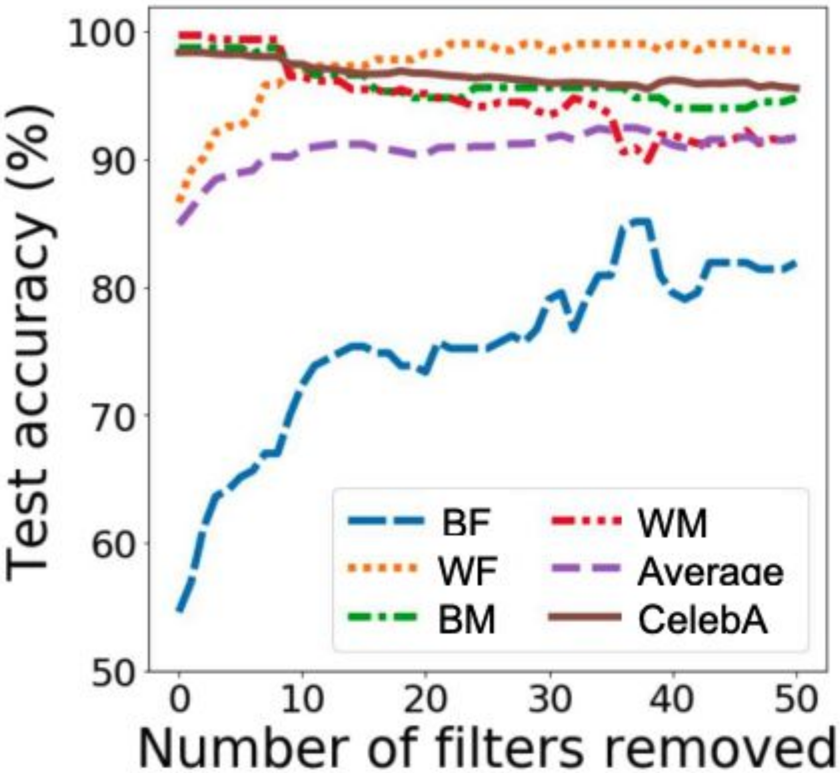
Input: Network's elements $N = \{1, \ldots, n\}$; performance metric $V(.)$; failure probability δ, tolerance ϵ, number of important elements k, Early truncation performance v_T
Output: Shapley value of elements: $\{\phi_i\}_{i=1}^n$
Initializations: $\{\phi_i\}_{i=1}^n = 0$, $\{\sigma_i\}_{i=1}^n = 0$, $\mathscr{U} = N$, $t = 0$
while $\mathscr{U} \neq \emptyset$ **do**
  $t \leftarrow t + 1$
  Random permutation of network's elements: $\pi^t = \{\pi^t[1], \ldots, \pi^t[n]\}$
  $v_0^t \leftarrow V(N)$
  for $j \in \{1, \ldots, N\}$ **do**
    if $j \in \mathscr{U}$ **then**
      if $v_{j-1}^t < v_T$ **then**
        $v_j^t \leftarrow v_{j-1}^t$
      else
        $v_j^t \leftarrow v(\{\pi^t[j+1], \ldots, \pi^t[n]\})$
      $\phi_{\pi^t[j]}, \sigma_{\pi^t[j]} \leftarrow$ Moving Average$(v_{j-1}^t - v_j^t, \phi_{\pi^t[j]})$, Moving Variance$(v_{j-1}^t - v_j^t, \phi_{\pi^t[j]})$
      $\phi_{\pi^t[j]}^{ub}, \phi_{\pi^t[j]}^{lb} \leftarrow$ Confidence Bounds$(\phi_{\pi^t[j]}, \sigma_{\pi^t[j]}, t)$
  $\mathscr{U} \leftarrow \{i : \phi_i^{lb} + \epsilon <$ k'th largest $\{\phi_i\}_i = 1^n < \phi_i^{ub} - \epsilon\}$

Neuron Shapley: Important ImageNet filters

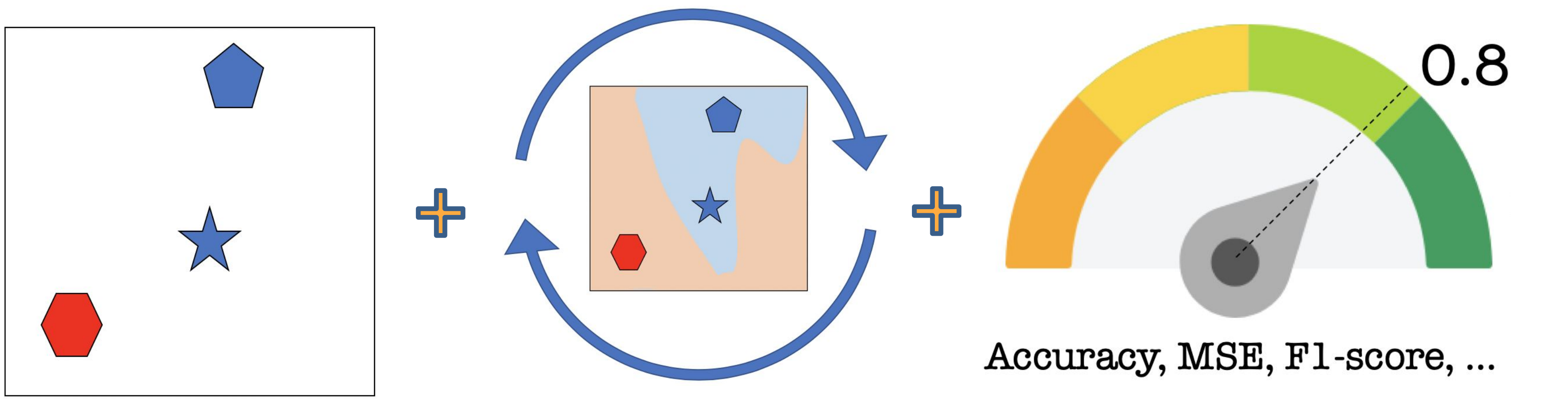


Neuron Shapley: Removing unfair filters





Distributional Shapley Value



Train Data | Learning Algorithm | Performance Evaluation

Distribution

Distributional value (★) = ???

A distributional framework tailored to ML applications

Distributional Shapley Value

Setting: A data point z with respect to a data distribution D

Definition *(GKZ20)*
For a data point z, its distributional shapley value for size m datasets coming from distribution D:

$$\text{value of data } z = \underset{\substack{B \sim \mathcal{D}^{m-1} \\ (m-1 \text{ points sampled from } \mathcal{D})}}{\mathbf{E}} \left[\begin{array}{c} \text{Data Shapley value of } z \\ \text{in dataset } B \cup \{z\} \end{array} \right]$$

Distributional Shapley Value

Definition (GKZ20)
For a data point z, its distributional shapley value for size m datasets coming from distribution D:

$$\text{value of data } z = \underset{\substack{B\sim\mathcal{D}^{m-1} \\ (m-1 \text{ points sampled from } \mathcal{D})}}{\mathbf{E}} \begin{bmatrix} \text{Data Shapley value of } z \\ \text{in dataset } B\cup\{z\} \end{bmatrix}$$

A random variable

Distributional Shapley Value

Definition (GKZ20)
For a data point z, its distributional shapley value for size m datasets coming from distribution D:

$$\text{value of data } z = \underset{\substack{B \sim \mathcal{D}^{m-1} \\ (m-1 \text{ points sampled from } \mathcal{D})}}{\mathbf{E}} \begin{bmatrix} \text{Data Shapley value of } z \\ \text{in dataset } B \cup \{z\} \end{bmatrix}$$

A random variable

Problem solved:

No dependance on a specific dataset!

Distributional Shapley Value

<u>Definition</u> (GKZ20)

For a data point z, its distributional shapley value for size m datasets coming from distribution D:

$$\text{value of data } z = \mathop{\mathbf{E}}_{\substack{k \sim \text{unif}[1,\ldots,m-1] \\ S \sim \mathcal{D}^{k-1}}} [\text{Performance}(S \cup z) - \text{Performance}(S)]$$

Expectation over leave-one-out scores

Distributional Shapley Value

<u>Definition</u> *(GKZ20)*
For a data point z, its distributional shapley value for size m datasets coming from distribution D:

$$\text{value of data } z = \mathop{\mathbf{E}}_{\substack{k \sim \text{unif}[1,\ldots,m-1] \\ S \sim \mathcal{D}^{k-1}}} [\text{Performance}(S \cup z) - \text{Performance}(S)]$$

Expectation over leave-one-out scores

Good news:
It satisfies (statistical variant of) Shapley axioms

Distributional Shapley Value

Definition (GKZ20)
For a data point z, its distributional shapley value for size m datasets coming from distribution D:

$$\text{value of data } z = \mathop{\mathbf{E}}_{\substack{k\sim\text{unif}[1,\ldots,m-1] \\ S\sim\mathcal{D}^{k-1}}} [\text{Performance}(S \cup z) - \text{Performance}(S)]$$

Expectation over leave-one-out scores

Good news:
It satisfies (statistical variant of) Shapley axioms
Efficient monte-carlo approximation

Distributional Shapley Value

<u>Definition</u> (GKZ20)
For a data point z, its distributional shapley value for size m datasets coming from distribution D:

$$\text{value of data } z = \mathop{\mathbf{E}}_{\substack{k \sim \text{unif}[1,\ldots,m-1] \\ S \sim \mathcal{D}^{k-1}}} [\text{Performance}(S \cup z) - \text{Performance}(S)]$$

Expectation over leave-one-out scores

Good news:

It satisfies (statistical variant of) Shapley axioms

Efficient monte-carlo approximation

Value is not dependent on a particular dataset ⇒ Intrinsic

Distributional Shapley Value

<u>Definition</u> (GKZ20)
For a data point z, its distributional shapley value for size m datasets coming from distribution D:

$$\text{value of data } z = \mathop{\mathbf{E}}_{\substack{k\sim\text{unif}[1,\ldots,m-1] \\ S\sim\mathcal{D}^{k-1}}} [\text{Performance}(S\cup z) - \text{Performance}(S)]$$

Expectation over leave-one-out scores

Good news:

It satisfies (statistical variant of) Shapley axioms

Efficient monte-carlo approximation

Value is not dependent on a particular dataset ⇒ Intrinsic

We can apply existing ML knowledge to value

Thank you!