Feature Interaction for Interpretability

Apr 22, 2020

Dr. Wei Wei, Prof. James Landay

CS 335: Fair, Accountable, and Transparent (FAccT) Deep Learning

Stanford University

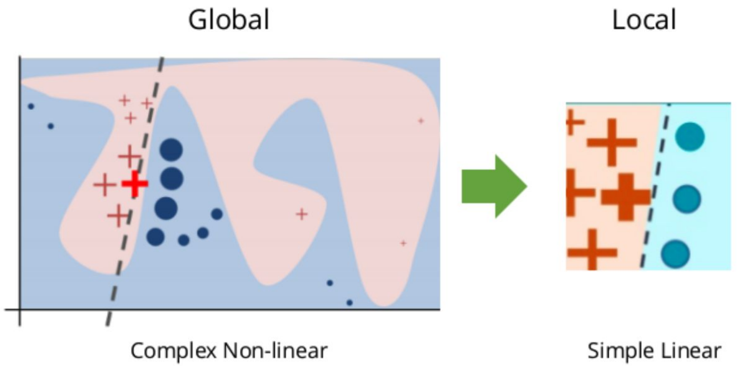
Announcements

- Project Proposal Due Apr 24
- Mid-way Presentation, May 13

Recap

- LIME
  - Optimizes Local Surrogate Loss between predictor f and explanation g

$$F(f, g, N_x) := \mathbb{E}_{x' \sim N_x}[(g(x') - f(x'))^2]$$



(a) Original Image (b) Explaining *Electric guitar*

Recap

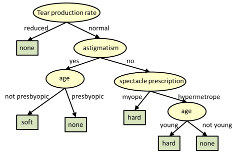
- Anchors are sets of feature predicates applied to the feature space
  - Optimize both Coverage and Precision

$$\max_{A \text{ s.t. } P(\text{prec}(A) \geq \tau) \geq 1-\delta} \text{cov}(A)$$

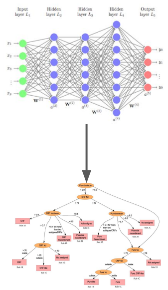
	If	Predict
adult	No capital gain or loss, never married	$\leq$ 50K
	Country is US, married, work hours > 45	$>$ 50K
rcdv	No priors, no prison violations and crime not against property	Not rearrested
	Male, black, 1 to 5 priors, not married, and crime not against property	Re-arrested
lending	FICO score ≤ 649	Bad Loan
	$649 \leq$ FICO score ≤ 699 and $\$5,400 \leq$ loan amount $\leq \$10,000$	Good Loan

Recap

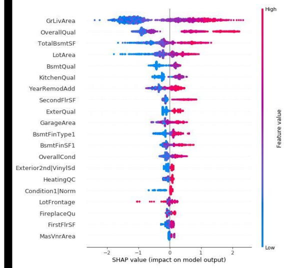
Lecture 3 Intrinsic Methods for Interpretability



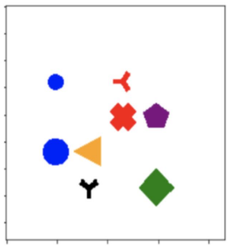
Post Hoc Methods for Interpretability



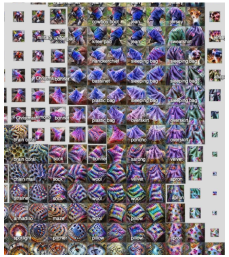
Lecture 4
Proxy Methods



Lecture 5
Feature Interaction



Lecture 6
Example Based
Methods

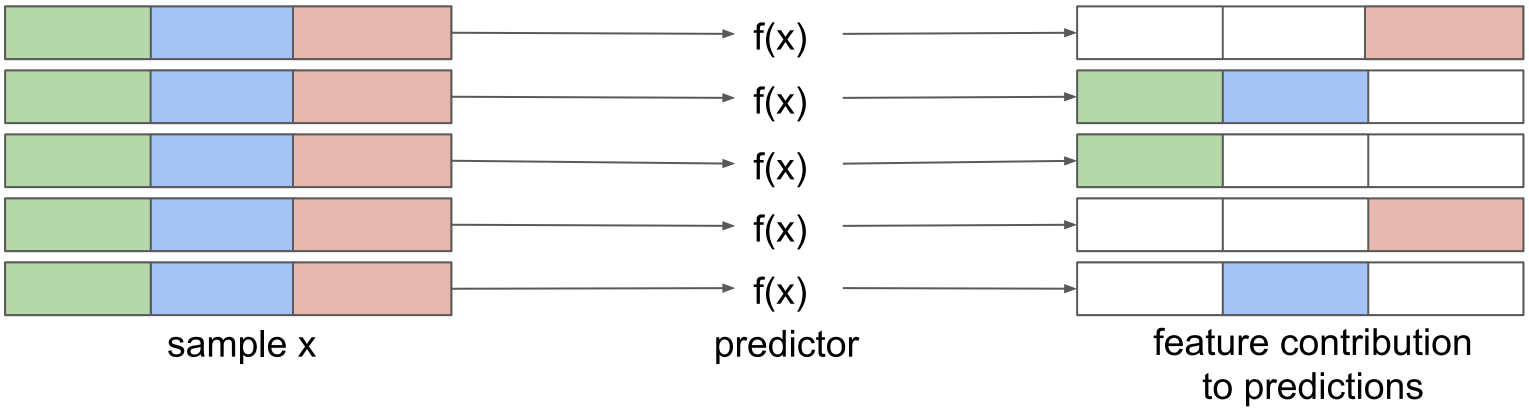


Lecture 7
Visualization Based
Methods

Outline

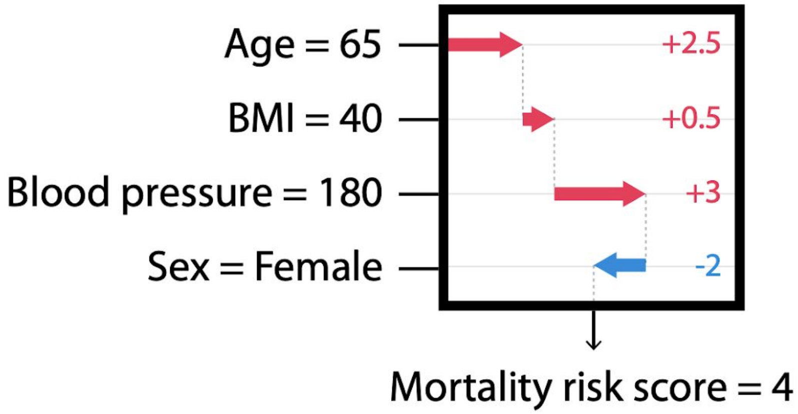
- Feature Interaction for Model Interpretability
- Layerwise Relevance Propagation
- DeepLift
- Shapley Additive Explanations (SHAP)
  - Coaliational Game and Shapley Values
  - Kernel SHAP
  - Deep SHAP
  - Tree SHAP
- Equatable Value of Data

Feature Interaction

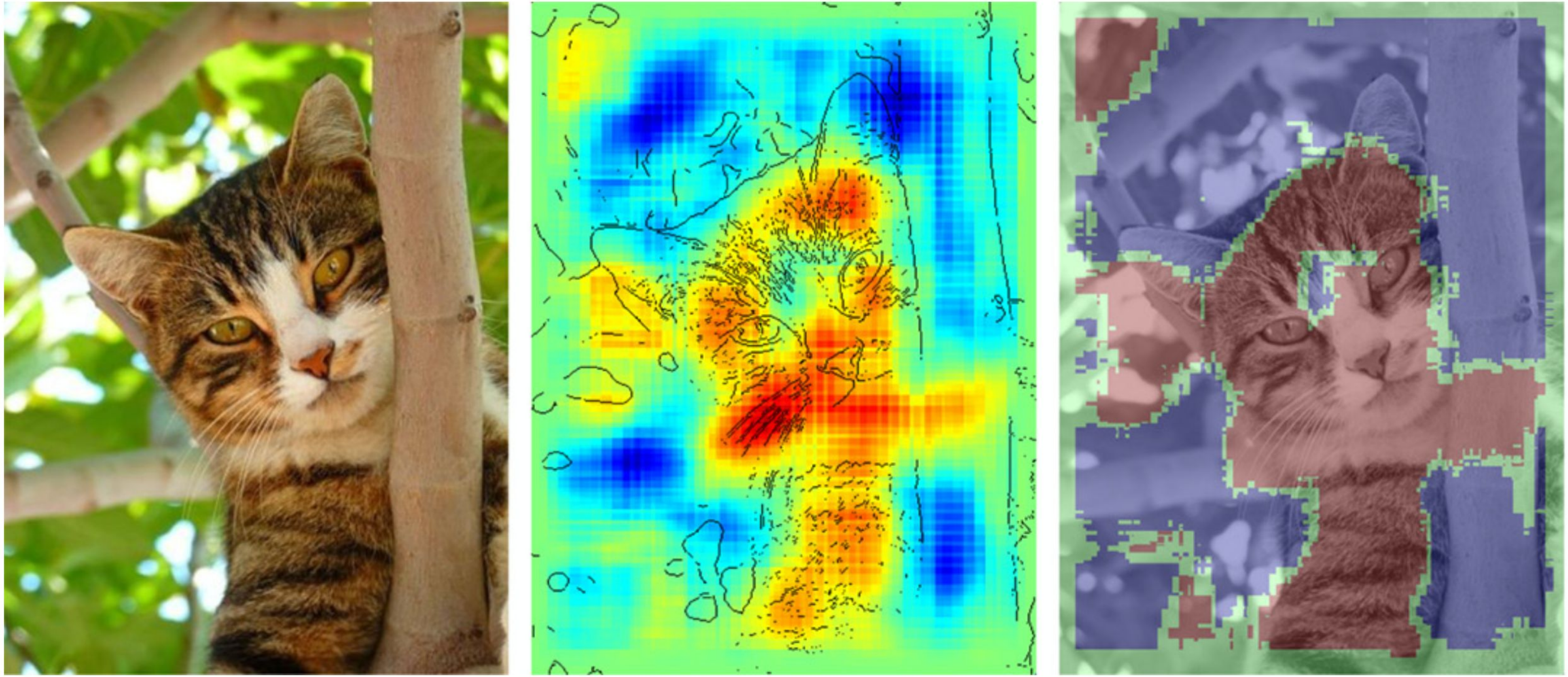


Feature Interaction

- Assign Importance Scores to Features
  - Each Feature i in the model will get a value Φ_i
  - Values explain how ML models make decisions



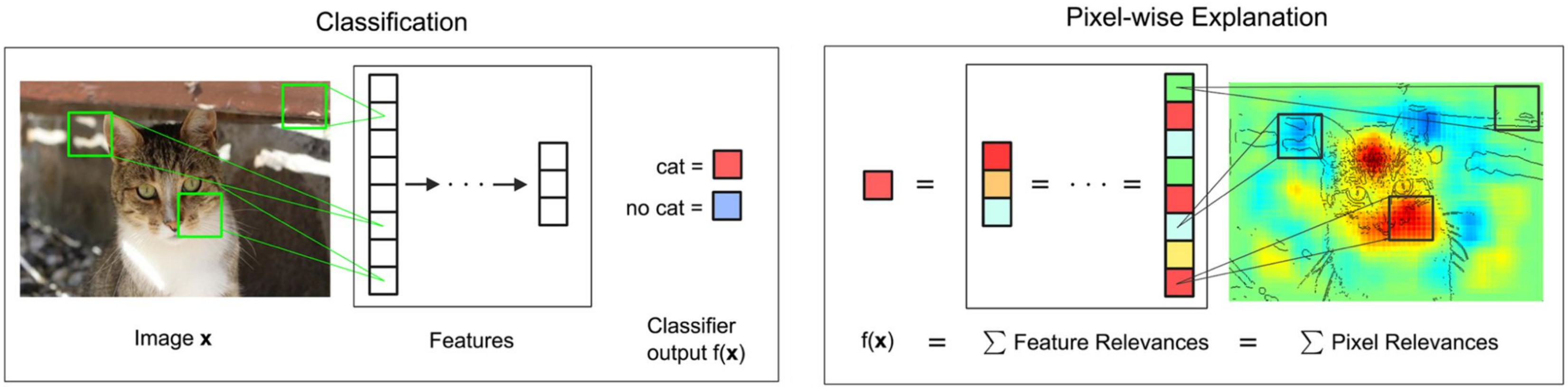
Feature Interaction



Outline

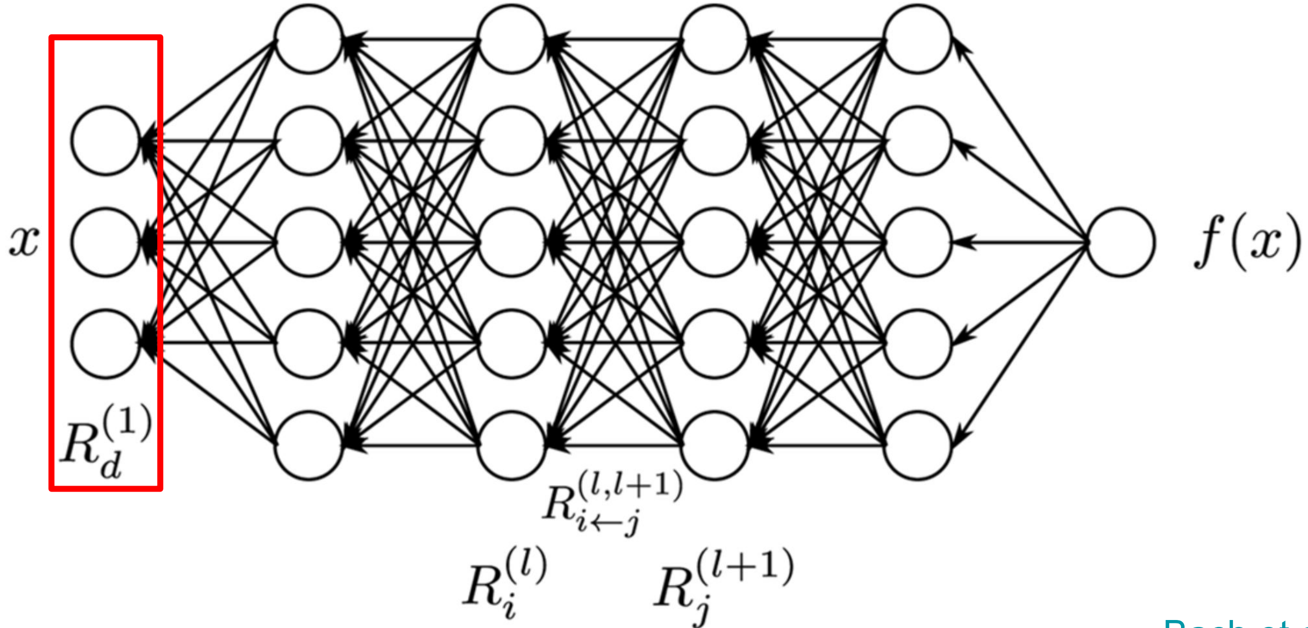
- Feature Interaction for Model Interpretability
- Layerwise Relevance Propagation
- DeepLift
- Shapley Additive Explanations (SHAP)
    - Coaliational Game and Shapley Values
    - Kernel SHAP
    - Deep SHAP
    - Tree SHAP
- Equatable Value of Data

Layerwise Relevance Propagation (LRP)



Bach et al, 2015

Layerwise Relevance Propagation (LRP)



Relevance Score R

Bach et al, 2015

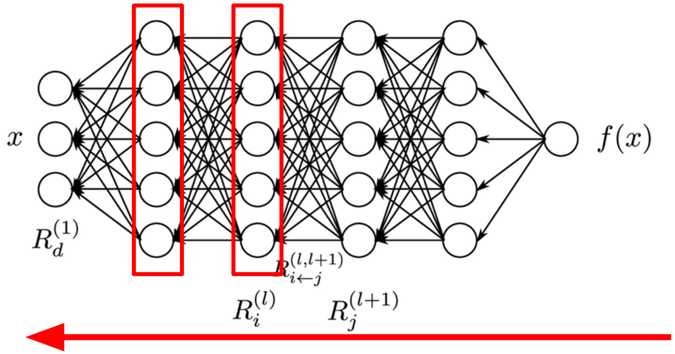
Relevance Scores

- x_i - output of neuron i $x_j = g(z_j)$
- g - activation function
- w_{ij} - weight of neural network connecting neuron x_i and x_j
- z_{ij} - linearly transformed neuron outputs

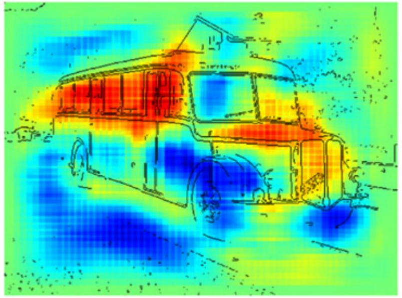
$$z_j = \sum_i z_{ij} + b_j \quad z_{ij} = x_i w_{ij}$$

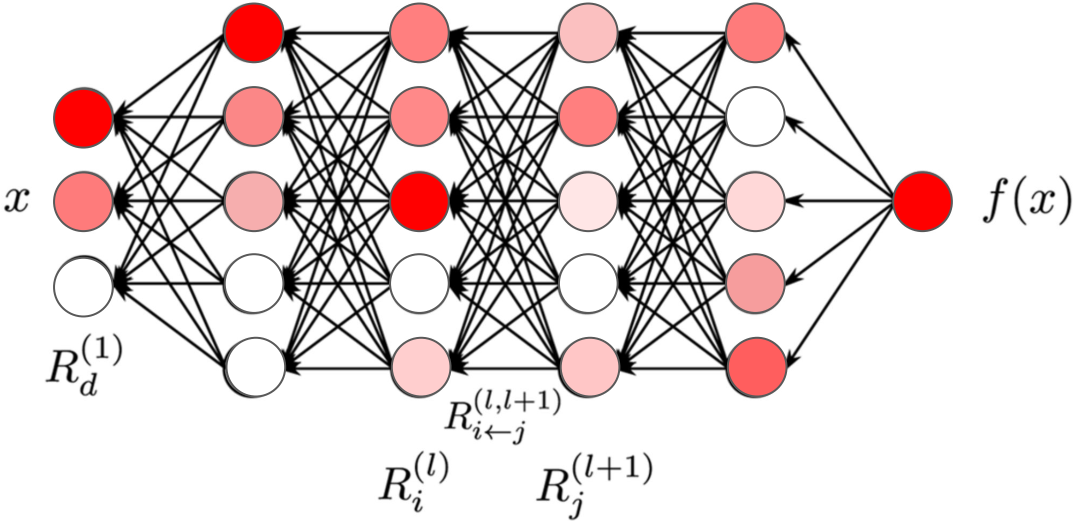
- Relevant Score $R_i^{(l)}$ of neuron i at level l

$$R_{i \leftarrow j}^{(l,l+1)} = \frac{z_{ij}}{z_j} \cdot R_j^{(l+1)} \qquad R_i^{(l)} = \sum_j R_{i \leftarrow j}^{(l,l+1)}$$



Relevance Score Propagation





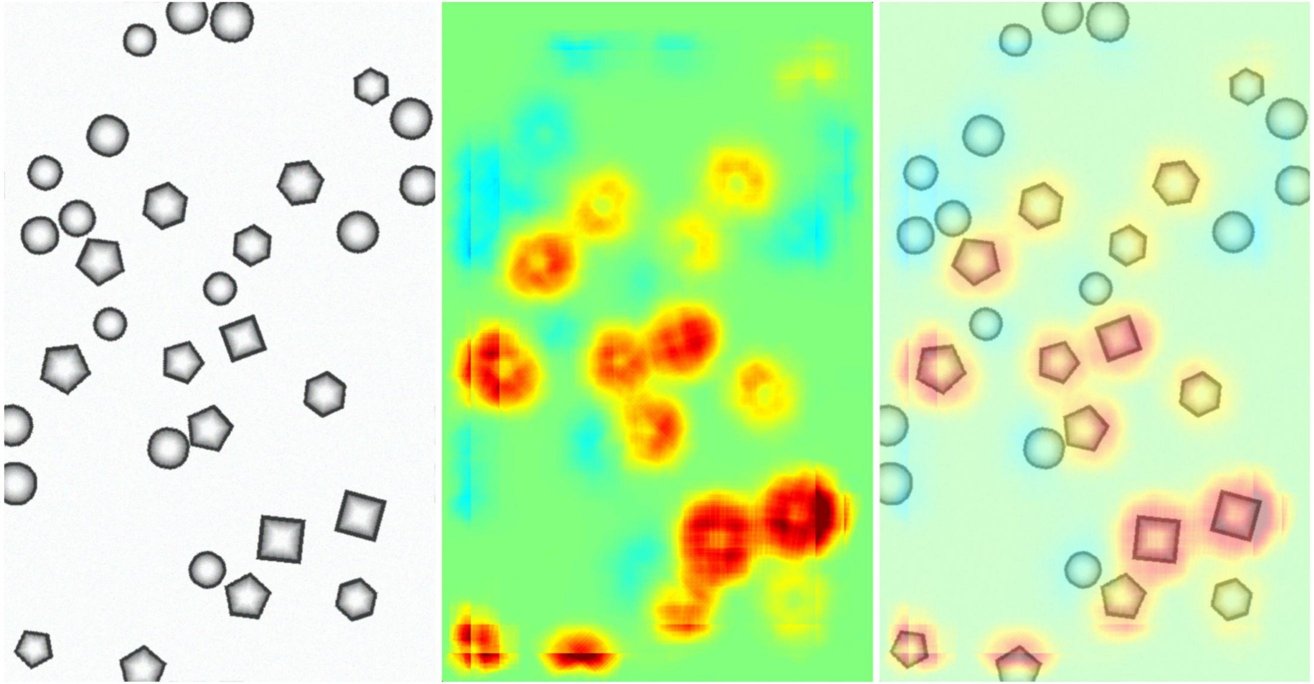
$$z_j = \sum_i z_{ij} + b_j$$

$$z_{ij} = x_i w_{ij}$$

$$R_{i\leftarrow j}^{(l,l+1)} = \frac{z_{ij}}{z_j} \cdot R_j^{(l+1)} \qquad R_i^{(l)} = \sum_j R_{i\leftarrow j}^{(l,l+1)}$$

[Bach et al, 2015]

Results on Synthetic Data

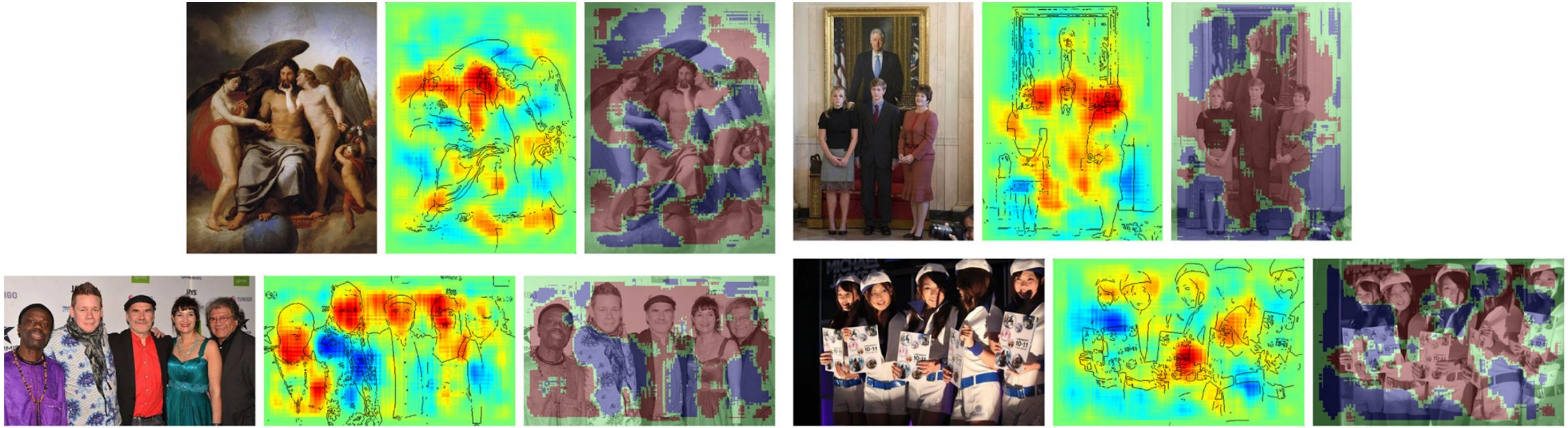


Bach et al, 2015

Resutls on Pascal Dataset

Bach et al, 2015

More Examples



Bach et al, 2015

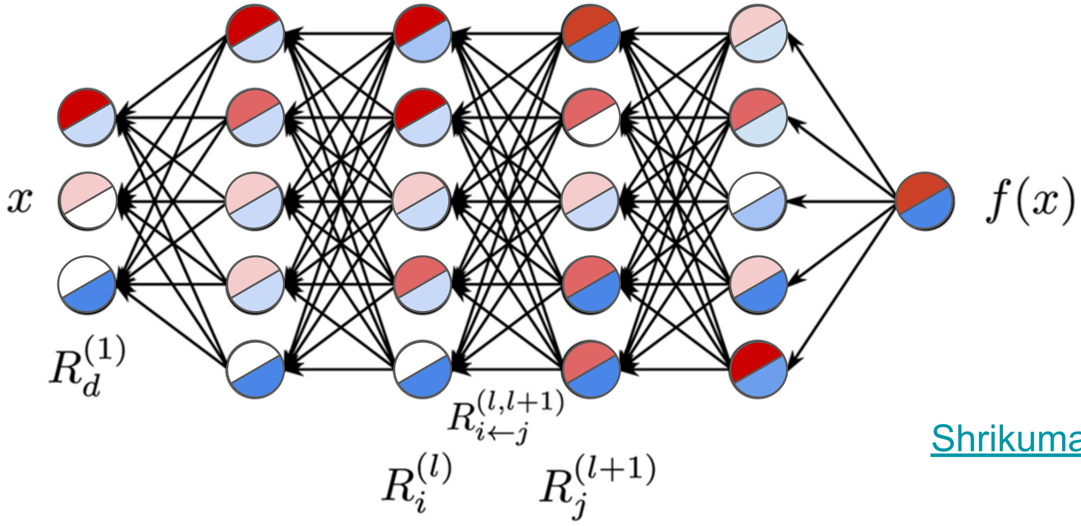
Outline

- Feature Interaction for Model Interpretability
- Layerwise Relevance Propagation
- DeepLift
- Shapley Additive Explanations (SHAP)
    - Coaliational Game and Shapley Values
    - Kernel SHAP
    - Deep SHAP
    - Tree SHAP
- Equatable Value of Data

DeepLift

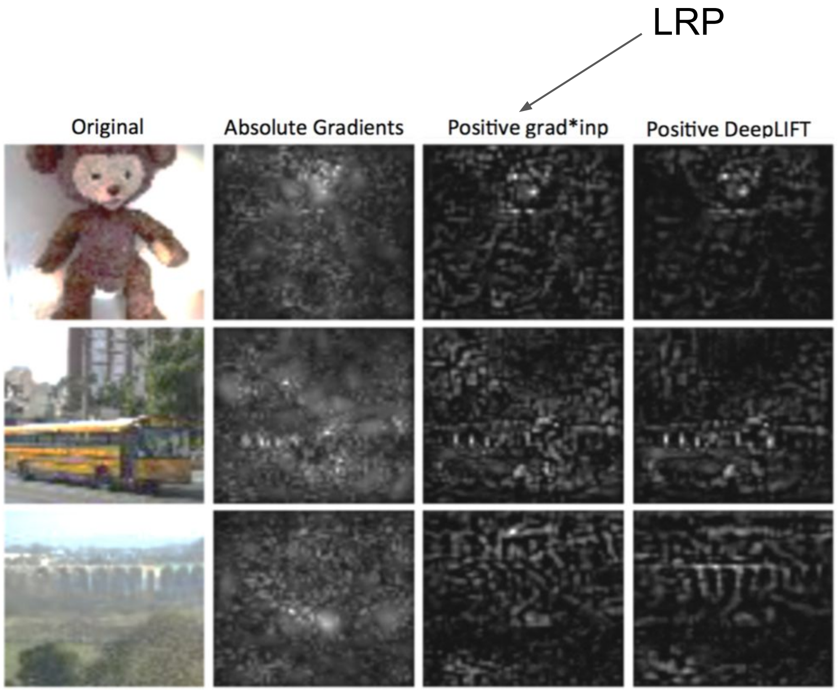
- DeepLift Allows Each Neuron A Reference Value for Activation Output x_i^0

$$\delta_i = x_i - x_i^0$$



Shrikumar et al, 2016

Results with VGG16 on Tiny Imagenet

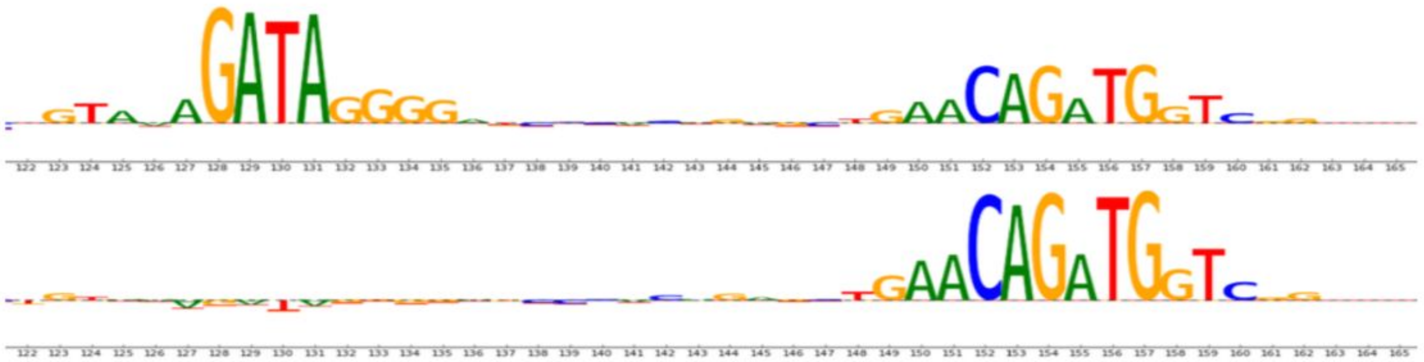


Shrikumar et al, 2016

Results with DNA Pattern Dataset

DeepLift

LRP



Shrikumar et al, 2016

Outline

- Feature Interaction for Model Interpretability
- Layerwise Relevance Propagation
- DeepLift
- Shapley Additive Explanations (SHAP)
  - Coaliational Game and Shapley Values
  - Kernel SHAP
  - Deep SHAP
  - Tree SHAP
- Equatable Value of Data

Shapley Additive Explanations (SHAP)

- Assigns Feature Importance Weights Based on Game Theory
  - Each feature ia a player
  - Probability $P(\hat{Y} \mid X)$ is the total payoff
  - Distribute the total payoff to players (features) "fairly"

Φ_0	Φ_1	Φ_2	Φ_3	Φ_4
0.6	0.05	0.03	0.01	0.01
0.1	0.2	0.3	-0.1	0
0.2	0.1	0.1	0.2	-0.1
0.05	0.10	0.05	0.1	-0.1

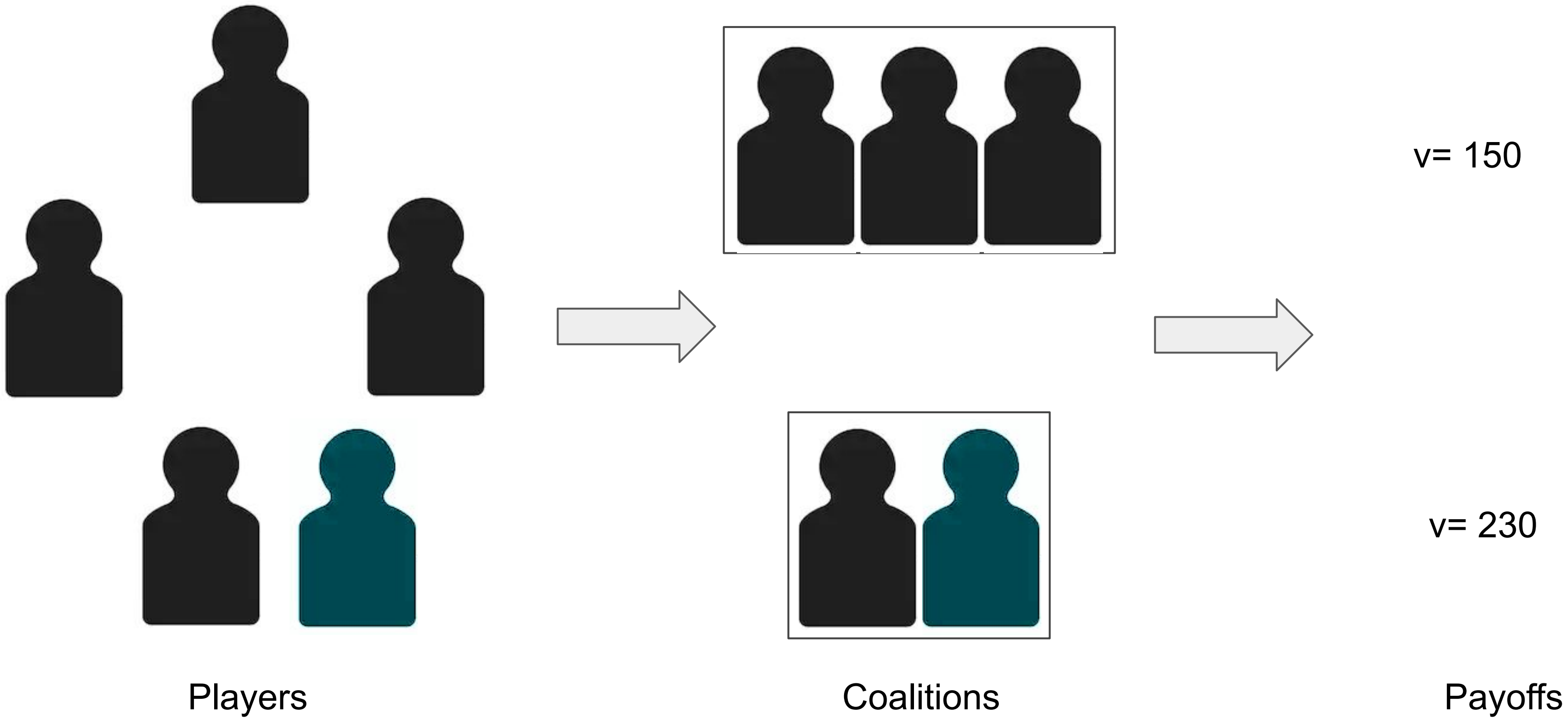
$\Sigma\Phi_i = P(\hat{Y} \mid X)$

$P(\hat{Y} \mid X)$
0.7
0.4
0.5
0.2

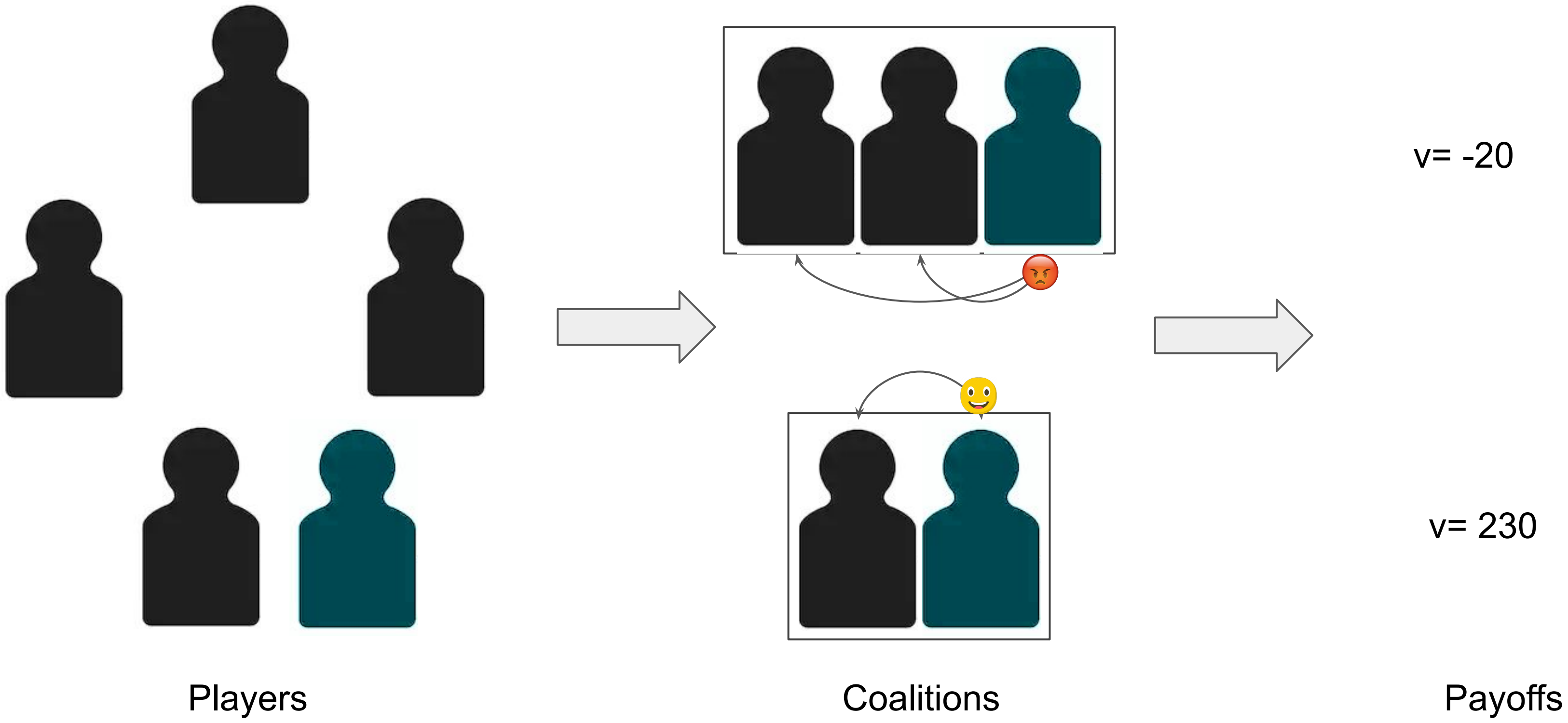
Shapley Values

- Developed by Lloyd Shapley
    - American mathematician
    - Nobel Prize-winning economist

Coalitional Game

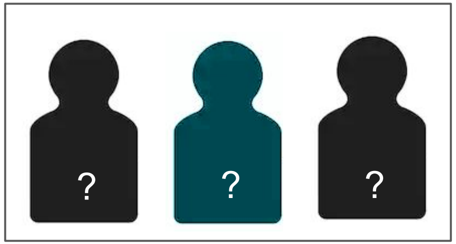


Coalitional Game



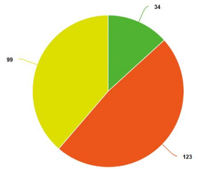
Coalitional Game

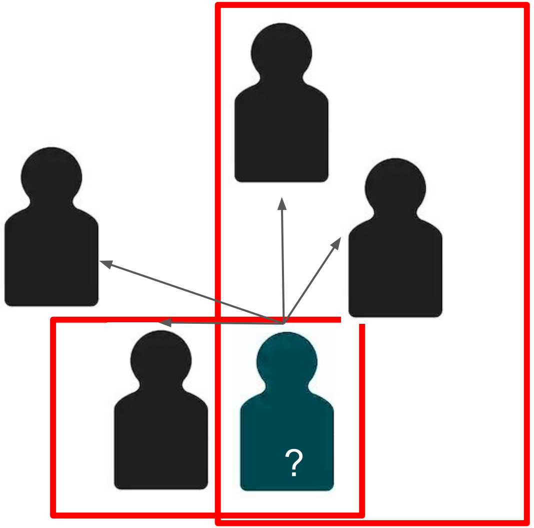
- How Do We Assign Importance Scores to Players?
    - Consider the *interactions* to all other players



How much value should we attribute to each player?

v= 100





How do we account for the interactions among players?

Shapley Values

- Design A Value Scheme Φi for player i
  - M-Player coalitional game
  - Payoff function v(S)
- Value Scheme Has to Follow Four Criteria
  - 1) Efficiency

$$\sum_{j=0}^{M} \phi_j = v(\mathcal{M})$$

  - 2) Symmetry

$$v(\mathcal{S} \cup \{i\}) = v(\mathcal{S} \cup \{j\}) \Longrightarrow \phi_i = \phi_j$$

  - 3) Dummy Player

$$v(\mathcal{S} \cup \{j\}) = v(\mathcal{S}) \Longrightarrow \phi_j = 0$$

  - 4) Linearity

$$\phi_i(v + w) = \phi_i(v) + \phi_i(w)$$

[Aas et al, 2019]()

Shapley Values

- Solution: Shparly Values
  - Unique solution that satisfies 1) - 4)
  - M - set of players
  - v (S) payoff function

$$\underbrace{\phi_j}_{\text{Shapley Value for player i}} = \frac{1}{M} \sum_{\mathcal{S} \subseteq \mathcal{M} \setminus \{j\}} \underbrace{\binom{M-1}{S}^{-1}}_{\text{normalizer}} \underbrace{(v(\mathcal{S} \cup \{j\}) - v(\mathcal{S}))}_{\text{impacts to to payoffs}}$$

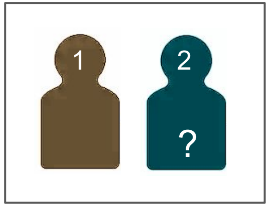
Shapley Values

- Calculate Shapley Value for Player 2

$$\phi_j = \frac{1}{M} \sum_{\mathcal{S} \subseteq \mathcal{M} \setminus \{j\}} \binom{M-1}{S}^{-1} \left(v(\mathcal{S} \cup \{j\}) - v(\mathcal{S})\right)$$

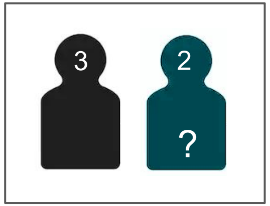
$$= \sum_{\mathcal{S} \subseteq \mathcal{M} \setminus \{j\}} \frac{S!(M-S-1)!}{M!} \left(v(\mathcal{S} \cup \{j\}) - v(\mathcal{S})\right)$$

$$\frac{1}{M}\binom{M-1}{S}^{-1} = \frac{1}{M}\left(\frac{(M-1)!}{S!(M-S-1)!}\right)^{-1} = \frac{S!(M-S-1)!}{M \cdot (M-1)!}$$



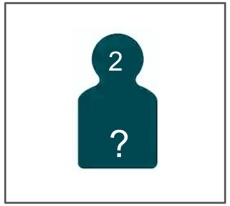
$$\binom{M-1}{S}^{-1} = \binom{2}{1}^{-1} = \frac{1}{2}$$

$$v(\{1,2\}) - v(\{1\})$$



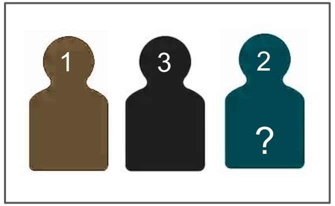
$$\binom{M-1}{S}^{-1} = \binom{2}{1}^{-1} = \frac{1}{2}$$

$$v(\{3,2\}) - v(\{3\})$$



$$\binom{M-1}{S}^{-1} = \binom{2}{0}^{-1} = 1$$

$$v(\{2\}) - v(\emptyset)$$

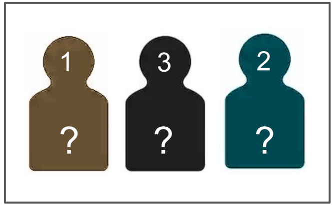


$$\binom{M-1}{S}^{-1} = \binom{2}{2}^{-1} = 1$$

$$v(\{1,2,3\}) - v(\{1,3\})$$

Shapley Values

$$\phi_j = \sum_{\mathcal{S} \subseteq \mathcal{M} \setminus \{j\}} \frac{S!(M - S - 1)!}{M!} (v(\mathcal{S} \cup \{j\}) - v(\mathcal{S}))$$



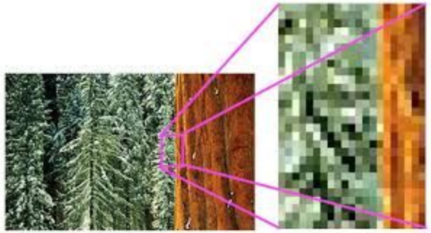
$$\begin{aligned}
\phi_1 &= \tfrac{1}{3}\Big(v(\{1,2,3\}) - v(\{2,3\})\Big) + \tfrac{1}{6}\Big(v(\{1,2\}) - v(\{2\})\Big) + \tfrac{1}{6}\Big(v(\{1,3\}) - v(\{3\})\Big) + \tfrac{1}{3}\Big(v(\{1\}) - v(\emptyset)\Big) \\
\phi_2 &= \tfrac{1}{3}\Big(v(\{1,2,3\}) - v(\{1,3\})\Big) + \tfrac{1}{6}\Big(v(\{1,2\}) - v(\{1\})\Big) + \tfrac{1}{6}\Big(v(\{2,3\}) - v(\{3\})\Big) + \tfrac{1}{3}\Big(v(\{2\}) - v(\emptyset)\Big) \\
\phi_3 &= \tfrac{1}{3}\Big(v(\{1,2,3\}) - v(\{1,2\})\Big) + \tfrac{1}{6}\Big(v(\{1,3\}) - v(\{1\})\Big) + \tfrac{1}{6}\Big(v(\{2,3\}) - v(\{2\})\Big) + \tfrac{1}{3}\Big(v(\{3\}) - v(\emptyset)\Big)
\end{aligned}$$

$$\phi_0 = v(\emptyset)$$

Aas et al, 2019

Back to ML Interpretability

- SHAP
  - Treat each feature i as a player as if we were in a coalitional game
  - Estimate the value of feature i by shapley values



predictor that uses feature $S \cup \{i\}$

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

value of feature i

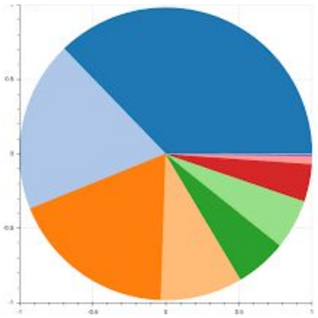
payoffs (probability)

Lundberg et al, 2017

Additive Feature Attribution

feature mask

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i x_i'$$



Efficiency of Shparly Values

$$\sum_{j=0}^{M} \phi_j = v(\mathcal{M})$$

Lundberg et al, 2017

Computational Challenges

- Terms Grow In the Order of 2^F
- Approximating Solutions
  - Shapley Sampling Values (Štrumbelj et al, 2013)
  - Tree SHAP (Lundberg et al, 2018)
  - Deep Approximate Shapley Propagation (Ancona et al, 2019)

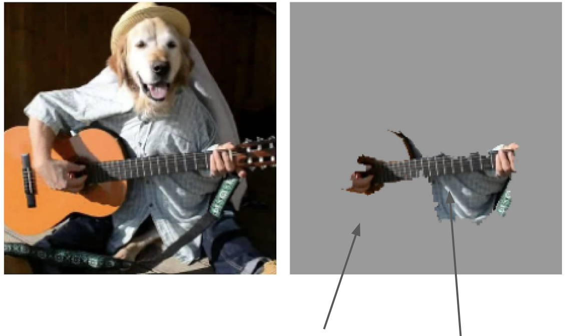
$$\phi_i = \underbrace{\sum_{S \subseteq F \setminus \{i\}}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

Computational Challenges

- Estimating Prediction Outcomes With Partial Features
  - Neural networks are not designed to use partial features
  - One solution is to use the expected value ([Lundberg et al, 2018](#))

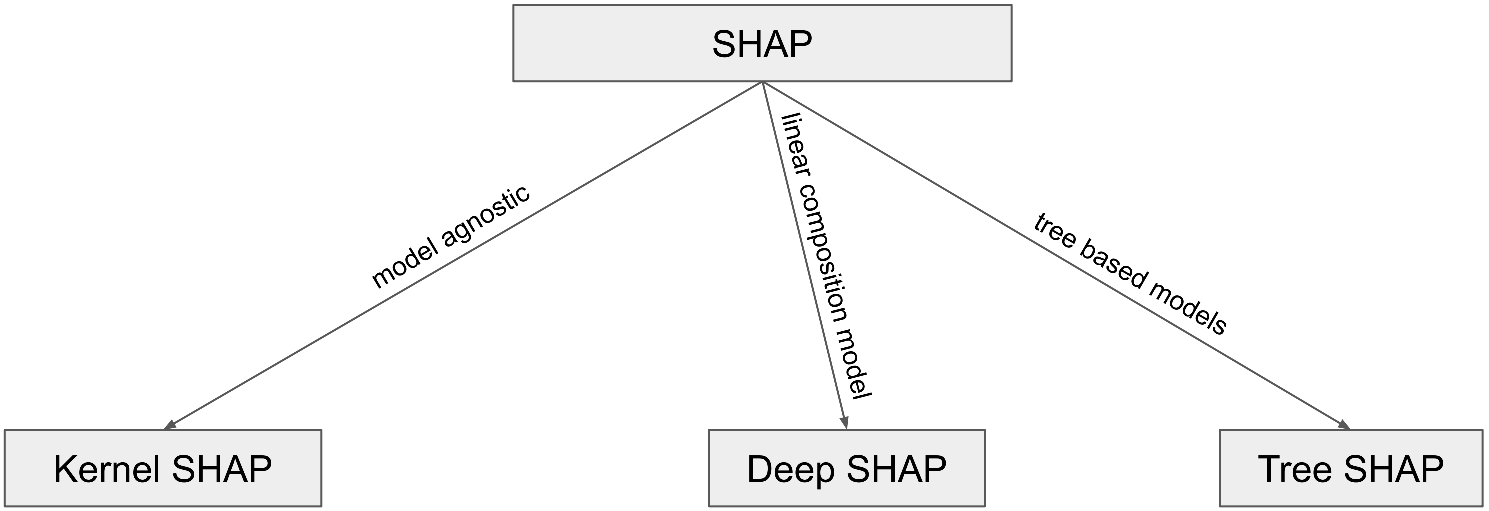
$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} \left[\underline{f_{S\cup\{i\}}(x_{S\cup\{i\}})} - \underline{f_S(x_S)} \right]$$

$$f_S(x_S) = \frac{1}{K} \sum_k f(x_{\bar{S}}^k, x_S^*)$$



$x_{\bar{S}}^k$ x_S^*

SHAP Based Methods



Equivalent to LIME

Equivalent to DeepLift, LRP

Kernel SHAP

- Remember the LIME Training Objective

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

$$L(f, g, \pi_{x'}) = \sum_{z' \in Z} \left[f(h_x(z')) - g(z') \right]^2 \pi_{x'}(z')$$

- SHAP Equivalent Objective

$$\Omega(g) = 0$$

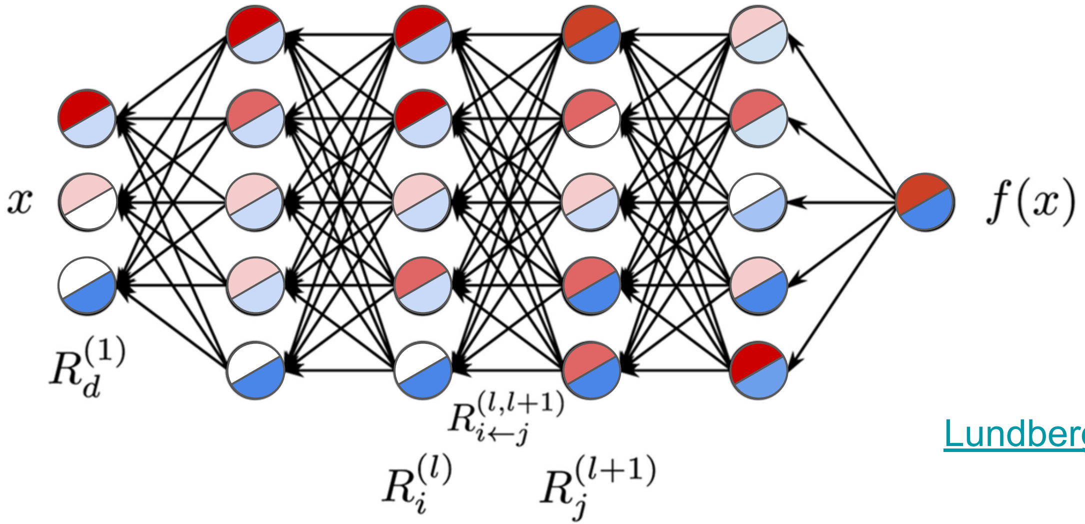
$$\pi_x(z') = \frac{(M-1)}{\binom{M}{|z'|} |z'| (M - |z'|)}$$

Lundberg et al, 2017

Deep SHAP

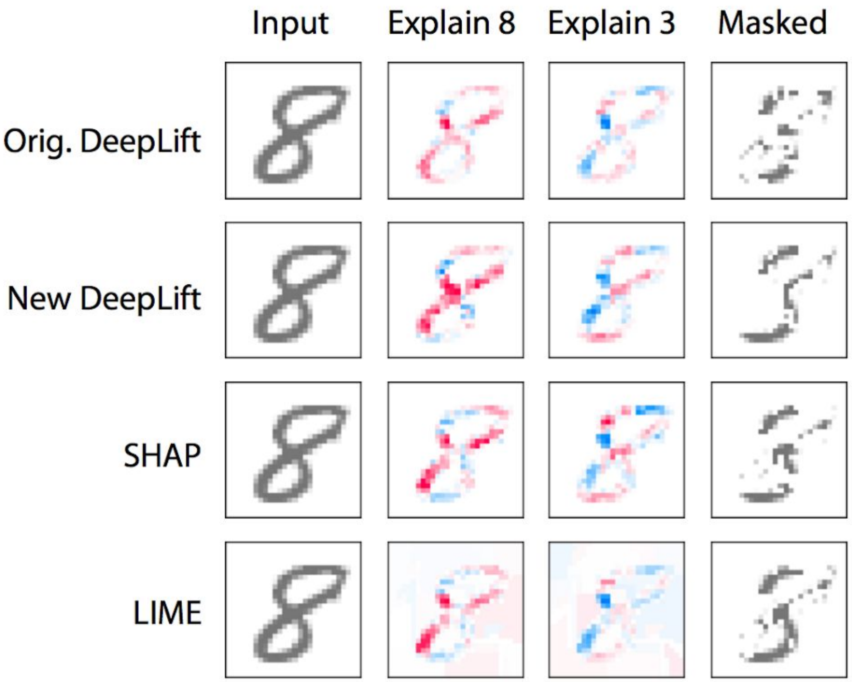
- Incorporate Shapley Values Into Linear Composition Model
- DeepLift approximates Deep SHAP when the reference value is taken to be $E[x]$

$$\delta_i = x_i - \mathbb{E}[x_i]$$



Lundberg et al, 2017

Feature Importance for MNIST

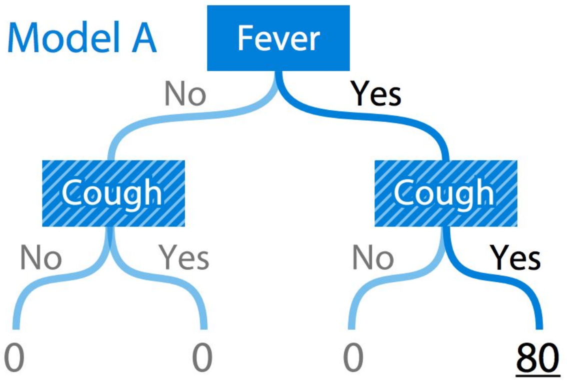


Outline

- Feature Interaction for Model Interpretability
- Layerwise Relevance Propagation
- DeepLift
- Shapley Additive Explanations (SHAP)
  - Coaliational Game and Shapley Values
  - Kernel SHAP
  - Deep SHAP
  - Tree SHAP
- Equatable Value of Data

Tree SHAP

- Incorporate Shapley Values to Tree Based Algorithms
  - Implemented into XGBoost and LightGBM
  - Reduced the complexity of estimating Shapley Values from $O(TL2^M)$ to $O(TLD^2)$
    - T - number of trees, L - maximum number of leaves
    - M - number of features, D - depth of tree



Lundberg et al, 2018

SHAP Interaction Values

- Pairwise Interactions of Shapley Values

$$\Phi_{i,j} = \sum_{S \subseteq N \setminus \{i,j\}} \frac{|S|!(M - |S| - 2)!}{2(M-1)!} \nabla_{ij}(S)$$

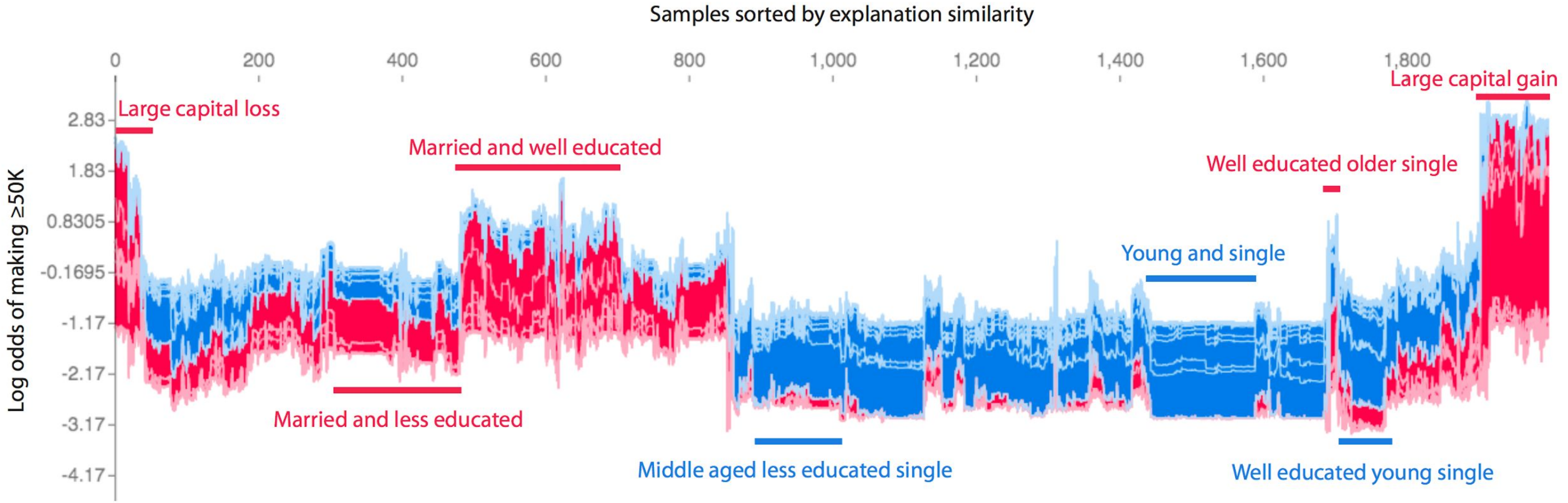
$$\nabla_{ij}(S) = f_x(S \cup \{i,j\}) - f_x(S \cup \{j\}) - [f_x(S \cup \{i\}) - f_x(S)]$$

$$\phi_j = \frac{1}{M} \sum_{\mathcal{S} \subseteq \mathcal{M} \setminus \{j\}} \binom{M-1}{S}^{-1} (v(\mathcal{S} \cup \{j\}) - v(\mathcal{S}))$$

Shapley Values

Adult Income

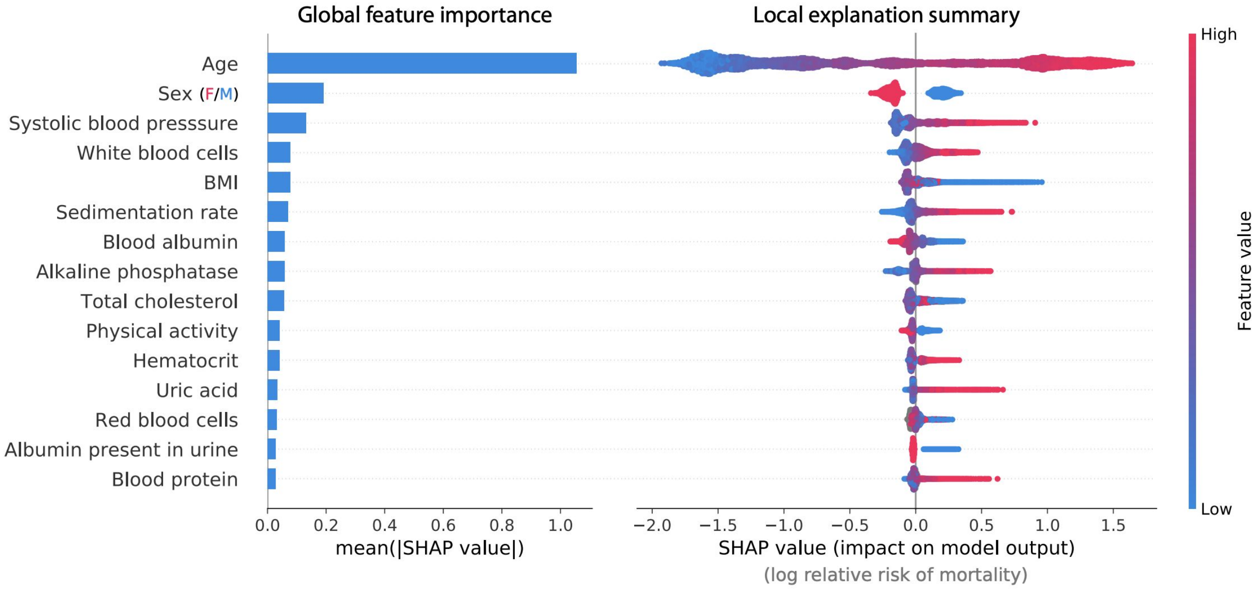
- Samples are clustered using the ordering in the leaf nodes



Lundberg et al, 2018

Mortality Data

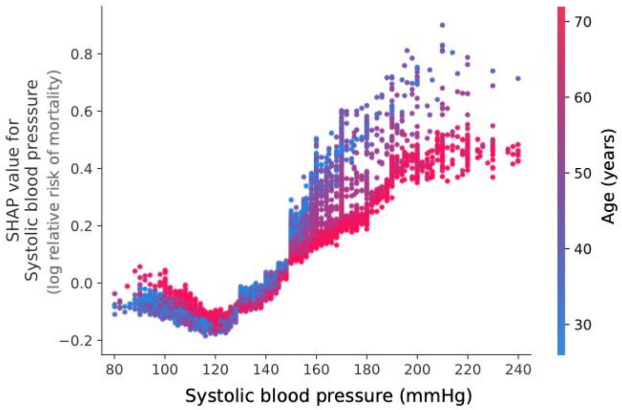
- Survival model on 20 year mortality followup data
    - 14,407 individuals and 79 features



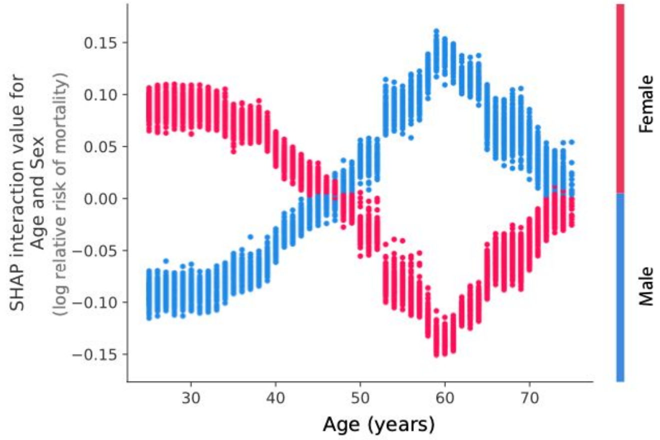
[Lundberg et al, 2019](#)

Mortality Data

Shapley Values



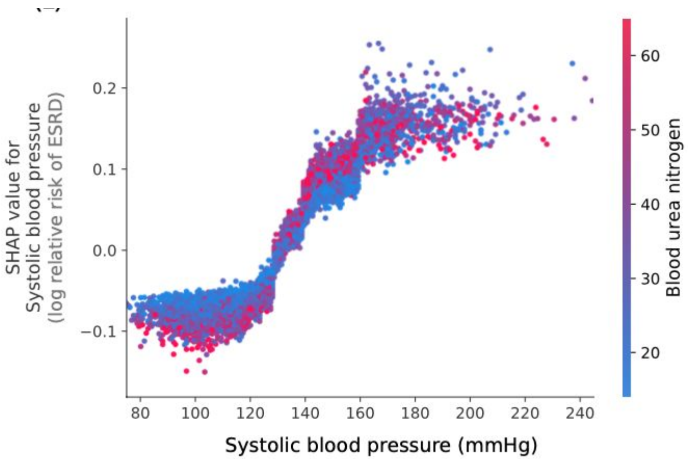
Shapley Interaction Values



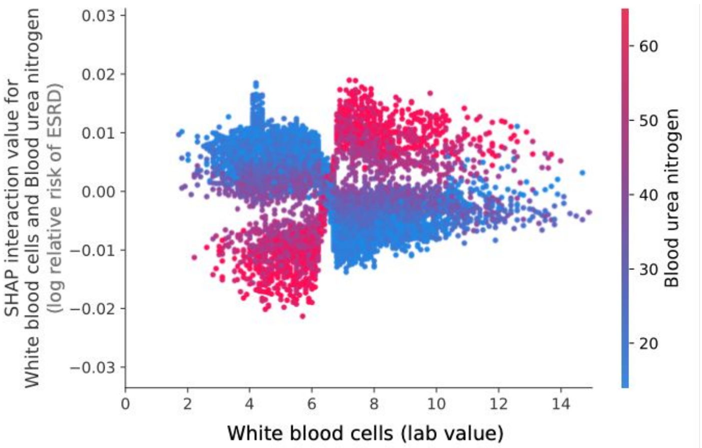
[Lundberg et al, 2019]

Chronic Kidney Disease

Shapley Values



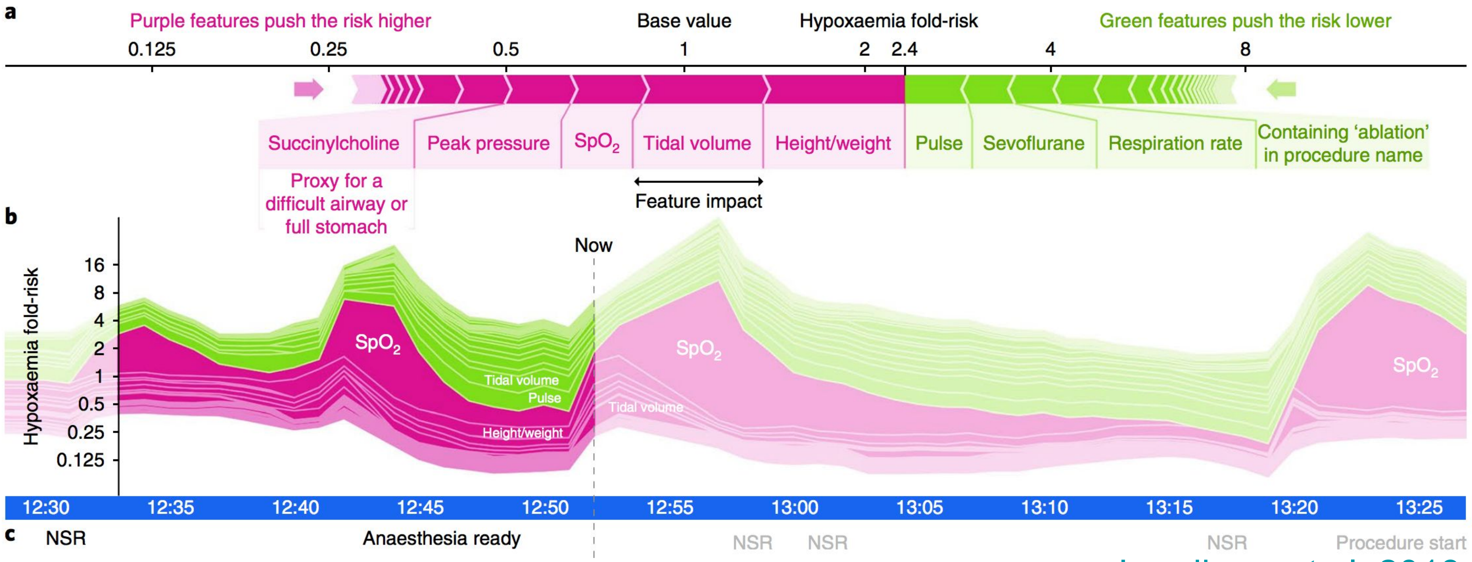
Shapley Interaction Values



Lundberg et al, 2019

Predicting Hypoxaemia During A Surgery

- Hypoxaemia
  - An abnormally low amount of oxygen in the blood



[Lundberg et al, 2018]

Outline

- Feature Interaction for Model Interpretability
- Layerwise Relevance Propagation
- DeepLift
- Shapley Additive Explanations (SHAP)
    - Coaliational Game and Shapley Values
    - Kernel SHAP
    - Deep SHAP
    - Tree SHAP
- Equatable Value of Data

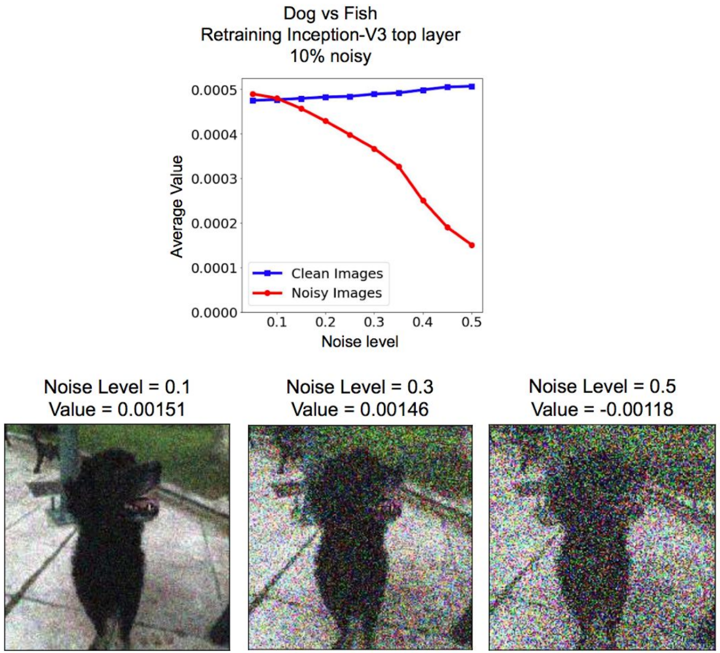
Equatable Value of Data Points

- Assign Shapley Values to Data Point for A Given Predictor
  - Each data point x_i receives a Shapley Value ϕ_i
  - V is a black-box predictor
  - C - a constant

$$\phi_i = C \sum_{S \subseteq D - \{i\}} \frac{V(S \cup \{i\}) - V(S)}{\binom{n-1}{|S|}}$$

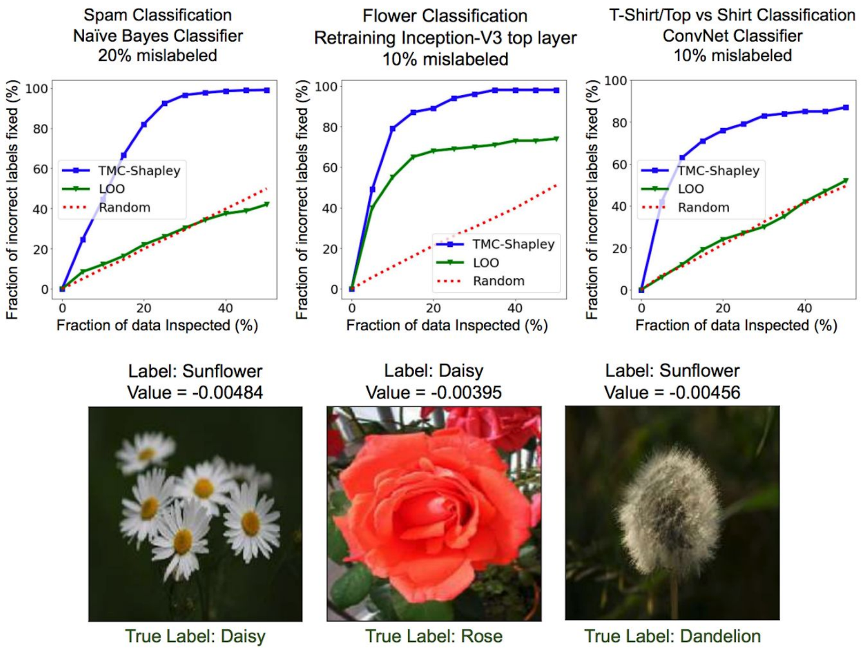
[Ghorbani et al, 2019]

Differentiating Noisy Data Using Shapley Values



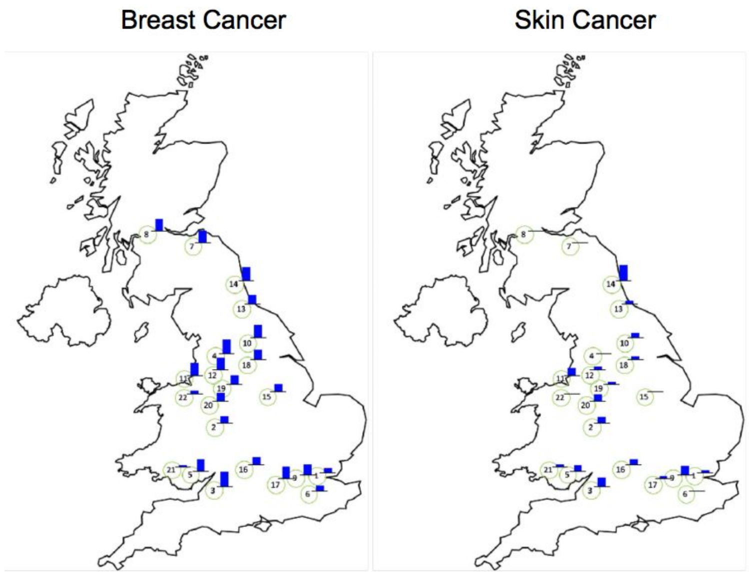
Ghorbani et al, 2019

Differentiating Mislabeled Data Using Shapley Values



[Ghorbani et al, 2019](#)

Value of Data for Each Cancer Type

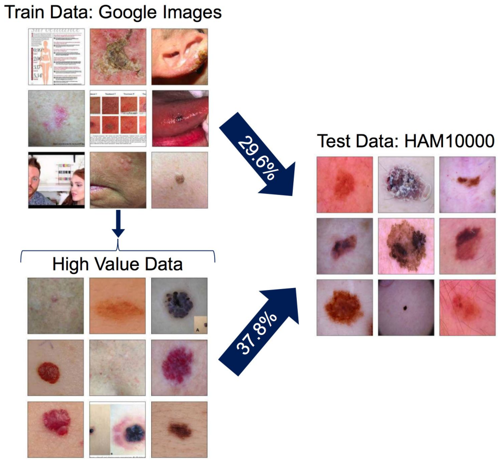


1- Barts
2- Birmingham
3- Bristol
4- Bury
5- Cardiff
6- Croydon
7- Edinburgh
8- Glasgow
9- Hounslow
10- Leeds
11- Liverpool
12- Manchester
13- Middlesborough
14- Newcastle
15- Nottingham
16- Oxford
17- Reading
18- Sheffield
19- Stockport
20- Stoke
21- Swansea
22- Wrexham

Ghorbani et al, 2019

Skin Pigmented Lesion Detection

- Search Online for Skin Pigmented Lesion Data Using Keyword Search
- Use Shapley Values to highlight high value data points
  - Performance improvements 29.6% -> 37.8%



Summary

- Feature Interaction
  - An importance score to explain how ML models make decisions
  - There exist interactions of features in the same ML model
- SHAP
  - Use Shapley Values as Feature Interaction Scores
    - Decomposes model prediction probabilities into an additive model

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i$$

    - Generalizes LIME, LRP and DeepLift
    - Tree SHAP implements SHAP efficiently into tree models
- Data Valuation
  - Estimate the value of data points based on Shapley Values

Required Reading

- Molnar: Ch 5.9, Ch 5.10

Reading Assignments

- Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, Costas J. Spanos, Towards Efficient Data Valuation Based on the Shapley Value, ICML 2019
- S Chang, Y Zhang, M Yu, T Jaakkola, A Game Theoretic Approach to Class-wise Selective Rationalization, NeurIPS 2019
- Schwab, Patrick, Djordje Miladinovic, and Walter Karlen. Granger-causal attentive mixtures of experts: Learning important features with neural networks, AAAI 2019
- Ying, Zhitao, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating explanations for graph neural networks, NeurIPS 2019
- Ancona, Marco, Cengiz Öztireli, and Markus Gross. Explaining deep neural networks with a polynomial time algorithm for shapley values approximation, ICML 2019

Next Lecture

Example Based Methods for Interpretability