Explaining by ~~Removing:~~
A Unified Framework for Model Explanation*

Authors: Ian C. Covert, Scott Lundberg, Su-In Lee
Presenters: Hongjin Lin, Jason Wang, Oam Patel

Monday 03/27/2023

* Journal of Machine Learning Research, 2021

Motivation

There is a need for a unifying framework

Minimal Image Representation approach, Zhou et al. (2014)

LIME (Ribeiro et al., 2016)

Masking Model approach (Dabkowski and Gal, 2017)

SHAP (Lundberg and Lee, 2017)

Prediction Difference Analysis (Zintgraf et al., 2017)

Meaningful Perturbations (Fong and Vedaldi, 2017)

RISE (Petsiuk et al., 2018)

?

FIDO-CA (Chang et al., 2018)

INVASE (Yoon et al., 2018)

Extremal Perturbations (Fong et al., 2019)

REAL-X (Jethani et al., 2021a)

...

Contributions

1. A unified framework that characterizes **26** existing explanation methods → **removal-based explanations**.

2. **Mathematical tools** to represent different approaches for **removing features** from ML models.

3. Removal-based explanations are implicitly tied to **cooperative game theory** → advantages of the **Shapley value** over alternative allocation strategies.

4. Feature removal is a simple application of **subtractive counterfactual reasoning**.

Previous Unifying Efforts

2017

A Unified Approach to Interpreting Model Predictions

2017

TOWARDS BETTER UNDERSTANDING OF GRADIENT-BASED ATTRIBUTION METHODS FOR DEEP NEURAL NETWORKS

Marco Ancona
Department of Computer Science
ETH Zurich, Switzerland

Enea Ceolini
Institute of Neuroinformatics
University Zürich and ETH Zürich

2020

Understanding Global Feature Contributions With Additive Importance Measures

Ian C. Covert
University of Washington
Seattle, WA
icovert@uw.edu

Scott Lundberg
Microsoft Research
Redmond, WA
scott.lundberg@microsoft.com

Su-In Lee
University of Washington
Seattle, WA
suinlee@uw.edu

LIME, DeepLIFT, LRP, QII, …

SHAP

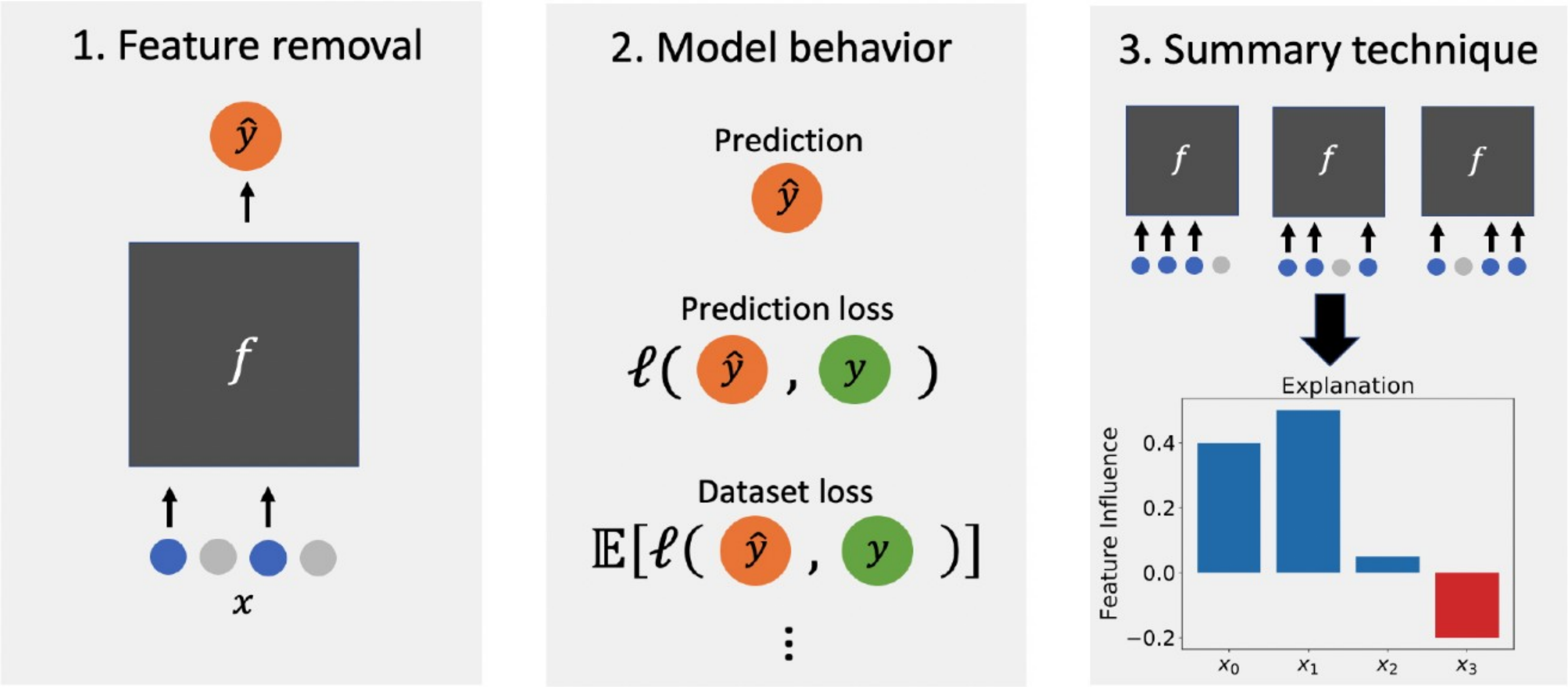
Grad * Input, DeepLIFT, LRP and Integrated Gradients

modified gradient back propagations

permutation tests, Shapley Net Effects, feature ablation, SAGE…

additive importance measures

The Removal-based Explanations Framework



Methods Survey

Key insights

1. Common
   a. marginalize out removed features with their conditional distribution
   b. Shapley values

2. Relatively unique and spatially isolated
   a. RISE; LIME for tabular data; INVASE

3. Many combinations left unexplored!

3

Summary technique

1

Feature removal	Feature attribution: Remove individual	Include individual	Mean when included	Shapley value	Additive model	Feature selection: High value subset	Low value subset	Partitioned subsets
Zeros	Occlusion CXPlain		RISE					MM
Default values					LIME (images)			
Extend pixels						MIR		
Blurring						EP	MP	
Generative model						FIDO-CA		
Marginalize (replacement distribution)					LIME (tabular)			
Marginalize (uniform)				IME 2010				
Marginalize (marginals product)				QII				
Marginalize (marginal)	Permutation test			SHAP KernelSHAP				
Marginalize (conditional)	PredDiff Conditional perm. test			SHAP SAGE LossSHAP Shapley Effects				
Tree distribution				TreeSHAP				
Surrogate model						L2X REAL-X		
Missingness during training						INVASE		
Separate models	Feature ablation	Univariate predictors		Shapley Net Effects IME 2009 SPVIM				

2

Model behavior: ■ Prediction ■ Prediction loss ■ Prediction mean loss ■ Dataset loss ■ Prediction loss (output) ■ Dataset loss (output)

Mathematical Formulation

Overview

1. Feature removal

$$F: \mathcal{X} \times \mathcal{P}(D) \mapsto \mathcal{Y}$$

2. Model behavior

$$u: \mathcal{P}(D) \mapsto \mathbb{R}$$

3. Summary technique

$$E: U \mapsto \mathbb{R}^d$$

or

$$E: U \mapsto \mathcal{P}(D)$$

$$F\colon \mathcal{X} \times \mathcal{P}(D) \mapsto \mathcal{Y}$$

Defining Feature Removal

$$F(x) = f(\text{features\_to\_keep}, \text{features\_to\_remove})$$

- Zero ablation
- Default values
- Generative model
- Train separate models
  - Train surrogate models
- “Missingness” during training

$$F(x_S) = f(x_S, 0).$$

$$F(x_S) = f(x_S, r_{\bar{S}}).$$

$$F(x_S) = f(x_S, \tilde{x}_{\bar{S}}).$$

$F\colon \mathcal{X} \times \mathcal{P}(D) \mapsto \mathcal{Y}$

Defining Feature Removal (cont)

Marginalization

- Marginalize with conditional
  - Tree distribution
- Marginalize with marginal
- Marginalize with product of marginals
- Marginalize with uniform
- Marginalize with replacement distribu

$$F(x_S) = \mathbb{E}\big[f(X) \mid X_S = x_S\big].$$

$$F(x_S) = \mathbb{E}\big[f(x_S, X_{\bar{S}})\big].$$

$$F(x_S) = \mathbb{E}_{\prod_{i \in D} p(X_i)}\big[f(x_S, X_{\bar{S}})\big].$$

$$F(x_S) = \mathbb{E}_{\prod_{i \in D} u_i(X_i)}\big[f(x_S, X_{\bar{S}})\big].$$

$$F(x, S) = \mathbb{E}_{\prod_{i \in D} q_{x_i}(X_i)}\big[f(x_S, X_{\bar{S}})\big].$$

$u: \mathcal{P}(D) \mapsto \mathbb{R}$

Explaining Model Behaviors

- Given the newly defined model F(x_S), we need a metric to assess how important the x_S features are
- Options
  - Prediction* $F(x_S).$
  - Prediction loss* $-\ell\big(F(x_S), y\big).$
  - Prediction mean loss $-\mathbb{E}_{p(Y|X=x)}\Big[\ell\big(F(x_S), Y\big)\Big].$
  - Dataset loss* $-\mathbb{E}_{XY}\Big[\ell\big(F(X_S), Y\big)\Big].$
  - Prediction loss w.r.t. output $-\ell\big(F(x_S), F(x)\big).$
  - Dataset loss w.r.t. output $-\mathbb{E}_X\Big[\ell\big(F(X_S), F(X)\big)\Big].$

Summarizing Feature Influence

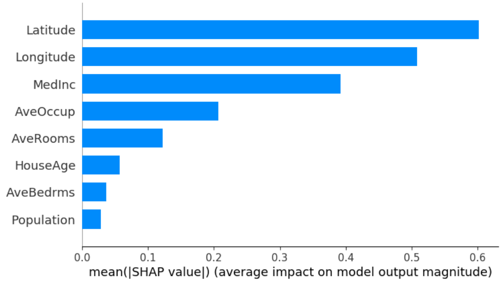
Two related approaches

- Map feature to real number value (feature attribution)*
- Map dataset to a set of important features (feature selection)

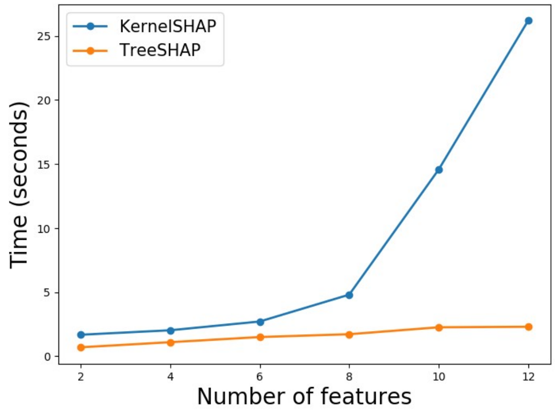
$$E : \mathcal{U} \rightarrow \mathbb{R}^d$$

$$E : \mathcal{U} \rightarrow \mathcal{P}(D)$$

Both are related, often the second is just a threshold applied to the first



Computational Complexity



How to isolate features?

- Consider every subset including the feature in question
  - Exact solutions are O(2^d / d) in the worst case
  - SHAP, RISE summarization, LIME additive model, etc
- Only consider the subset where you remove the feature
  - Polynomial in d
  - Occlusion, PRedDiff, CXPlain, permutation tests, etc

Approximations? (much faster, but lose worst-case guarantees)

- Sampling
- TreeSHAP (dynamic programming solution for tree models)
- Solve a continuous relaxation of the subset problem (i.e. learn a mask)
- Greedy search (MIR)
- Learn a model to do the explanation task (map from dataset to feature importance vector)

Connections to Other Theoretical Frameworks

Game Theory

Same setup as what the SHAP people said

- Cooperative games where we solve for allocations

Makes SHAP unique; however, other methods can be viewed as fitting (linear, additive) models to cooperative games

- Proofs in appendix, not critical

Summarization	Methods	Related To
Shapley value	Shapley Net Effects, IME, QII, SHAP, TreeSHAP, KernelSHAP, LossSHAP, Shapley Effects, SAGE, SPVIM	Shapley value, probabilistic values, modeling cooperative games
Mean value when included	RISE	Banzhaf value, probabilistic values, modeling cooperative games
Remove/include individual players	Occlusion, PredDiff, CXPlain, permutation tests, univariate predictors, feature ablation (LOCO)	Probabilistic values, modeling cooperative games
Fit additive model	LIME	Shapley value, Banzhaf value, modeling cooperative games
High/low value coalitions	MP, EP, MIR, MM, L2X, INVASE, REAL-X, FIDO-CA	Maximum/minimum excess

Information Theory

- f approximates response variable's conditional distribution $f(x) \approx p(Y|X=x)$
  - Classification: $f(x) \approx \mathbb{E}[Y|X=x]$
  - Regression:

Information Theory

- f approximates response variable's conditional distribution $f(x) \approx p(Y|X=x)$
  - Classification: $f(x) \approx \mathbb{E}[Y|X=x]$
  - Regression:
- F can also approximate $\{q(Y|X_S) : S \subseteq D\}$ conditional distribution (over subsets!)
  - Set of conditional distributions:

Information Theory

- f approximates response variable's conditional distribution
  - Classification: $f(x) \approx p(Y|X = x)$
  - Regression: $f(x) \approx \mathbb{E}[Y|X = x]$
- F can also approximate conditional distribution (over subsets!)
  - Set of conditional distributions: $\{q(Y|X_S) : S \subseteq D\}$
- Want this to be probabilistically valid (called consistent)
  - Countable Additivity: $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$
  - Bayes' Rule: $P(A \mid B) = \frac{P(A, B)}{P(B)}$
  - Occurs only when average over conditional distribution

Approximations of SHAP

Intractable
!

Conditional Distribution is Intracta

$$\mathbb{E}[f(X) \mid X_S = x_S] = \mathbb{E}_{p(X_{\bar{S}} \mid X_S = x_S)}[f(x_S, X_{\bar{S}})]$$

- Assume feature independence
- Assume model linearity

$$\approx \mathbb{E}_{p(X_{\bar{S}})}[f(x_S, X_{\bar{S}})]$$

$$\approx f\left(x_S, \mathbb{E}[X_{\bar{S}}]\right)$$

Approximations of SHAP

Intractable!

Conditional Distribution is Intracta

$$\mathbb{E}[f(X) \mid X_S = x_S] = \mathbb{E}_{p(X_{\bar{S}} \mid X_S = x_S)}[f(x_S, X_{\bar{S}})]$$

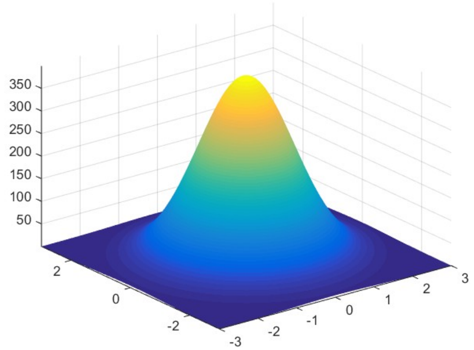
$$\approx \mathbb{E}_{p(X_{\bar{S}})}[f(x_S, X_{\bar{S}})]$$

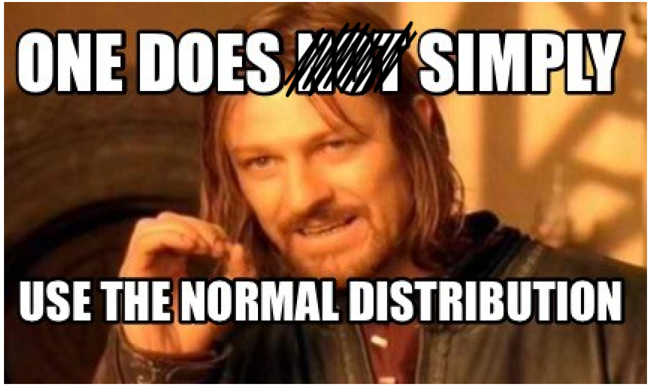
$$\approx f(x_S, \mathbb{E}[X_{\bar{S}}])$$

- Assume feature independence
- Assume model linearity

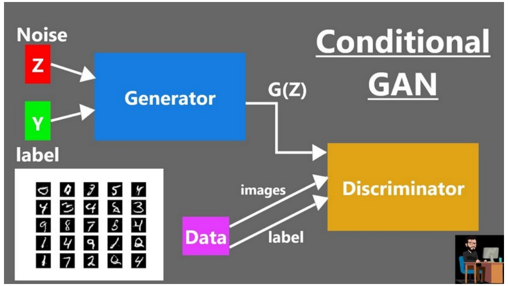
Or:

- Assume norma



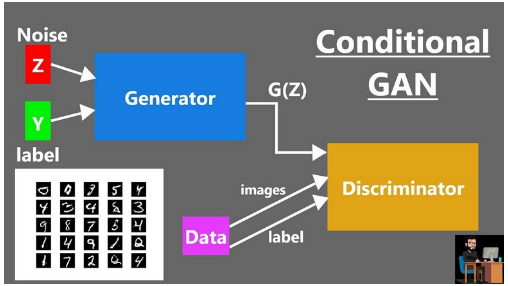


Approximations of SHAP



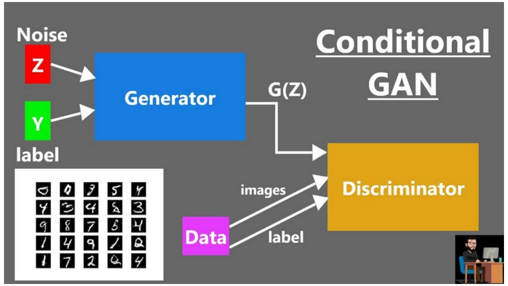
- Generative Model
  - Draw samples from cGAN
  - Single-Sample Monte Carlo Approximation

Approximations of SHAP



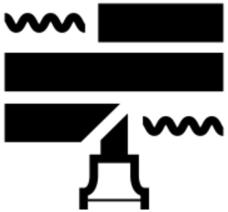
- Generative Model
  - Draw samples from cGAN
  - Single-Sample Monte Carlo Approximation
- Surrogate Model
  - Train a model to match model predictions
  - Objective $\min_{F} \mathbb{E}_X \mathbb{E}_S \left[\ell\big(F(X_S), f(X)\big) \right]$

Approximations of SHAP



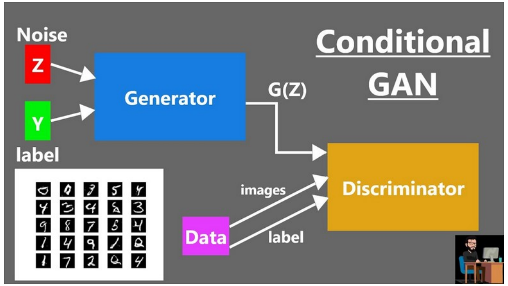
- Generative Model
  - Draw samples from cGAN
  - Single-Sample Monte Carlo Approximation

- Surrogate Model
  - Train a model to match model predictions
  - Objective $\min_{F} \mathbb{E}_X \mathbb{E}_S \Big[\ell\big(F(X_S), f(X)\big)\Big]$



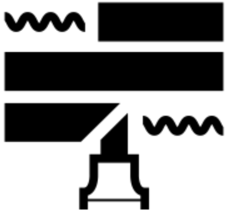
- Training with Missing
  - Train your original model with missing features
  - Objective $\min_{F} \mathbb{E}_{XY} \mathbb{E}_S \Big[\ell\big(F(X_S), Y\big)\Big]$

Approximations of SHAP



- Generative Model
  - Draw samples from cGAN
  - Single-Sample Monte Carlo Approximation

- Surrogate Model
  - Train a model to match model predictions
  - Objective $\min_{F} \mathbb{E}_X \mathbb{E}_S \Big[\ell\big(F(X_S), f(X)\big) \Big]$



- Training with Missing
  - Train your original model with missing features
  - Objective $\min_{F} \mathbb{E}_{XY} \mathbb{E}_S \Big[\ell\big(F(X_S), Y\big) \Big]$

- Separate Model
  - Train separate model for each possible subset

Information Theory Quantities

Model Behavior	Set Function	Methods	Related To
Prediction	u_x	Occlusion, MIR, MM, IME, QII, LIME, MP, EP, FIDO-CA, RISE, SHAP, KernelSHAP, TreeSHAP	Conditional probability, conditional expectation
Prediction loss	v_{xy}	LossSHAP, CXPlain	Pointwise mutual information
Prediction mean loss	v_x	INVASE	KL divergence with conditional distribution
Dataset loss	v	Permutation tests, univariate predictors, feature ablation (LOCO), Shapley Net Effects, SAGE, SPVIM	Mutual information (with label)
Prediction loss (output)	w_x	L2X, REAL-X	KL divergence with full model output
Dataset loss (output)	w	Shapley Effects	Mutual information (with output)

Cognition Theory

- Subtractive Counterfactual (Epstude and Roese, 2008) and Method of Difference (Mill, 1884)
- Norm Theory and the downhill rule
- Trade-off between simplicity and completeness

Experiments

	Summary technique							
	Feature attribution					Feature selection		
Feature removal	Remove individual	Include individual	Mean when included	Shapley value	Additive model	High value subset	Low value subset	Partitioned subsets
Zeros	Occlusion CXPlain		RISE					MM
Default values					LIME (images)			
Extend pixels						MIR		
Blurring						EP	MP	
Generative model						FIDO-CA		
Marginalize (replacement distribution)					LIME (tabular)			
Marginalize (uniform)				IME 2010				
Marginalize (marginals product)				QII				
Marginalize (marginal)	Permutation test			SHAP KernelSHAP				
Marginalize (conditional)	PredDiff Conditional perm. test			SHAP SAGE LossSHAP Shapley Effects				
Tree distribution				TreeSHAP				
Surrogate model						L2X REAL-X		
Missingness during training						INVASE		
Separate models	Feature ablation	Univariate predictors		Shapley Net Effects IME 2009 SPVIM				

Model behavior ■ Prediction ■ Prediction loss ■ Prediction mean loss ■ Dataset loss ■ Prediction loss (output) ■ Dataset loss (output)

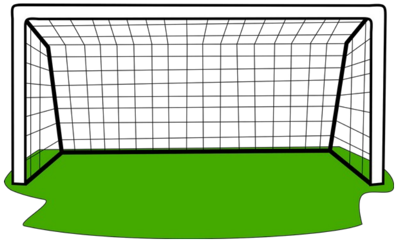
Summary technique

Feature removal	Feature attribution					Feature selection		
	Remove individual	Include individual	Mean when included	Shapley value	Additive model	High value subset	Low value subset	Partitioned subsets
Zeros	Occlusion CXPlain		RISE					MM
Default values					LIME (images)			
Extend pixels						MIR		
Blurring						EP	MP	
Generative model						FIDO-CA		
Marginalize (replacement distribution)					LIME (tabular)			
Marginalize (uniform)				IME 2010				
Marginalize (marginals product)				QII				
Marginalize (marginal)	Permutation test			SHAP KernelSHAP				
Marginalize (conditional)	PredDiff Conditional perm. test			SHAP SAGE LossSHAP Shapley Effects				
Tree distribution				TreeSHAP				
Surrogate model						L2X REAL-X		
Missingness during training						INVASE		
Separate models	Feature ablation	Univariate predictors		Shapley Net Effects IME 2009 SPVIM				

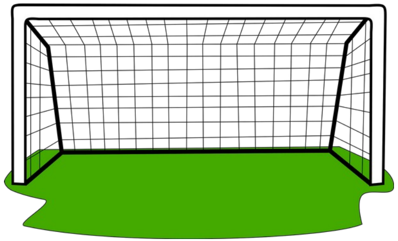
Model behavior ■ Prediction ■ Prediction loss ■ Prediction mean loss ■ Dataset loss ■ Prediction loss (output) ■ Dataset loss (output)

So many unexplored combinations!!!

Experimental Goals

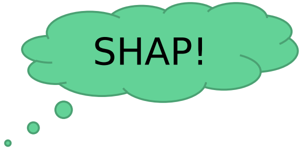


1. Fill in the gaps of removal-based method combinations



Experimental Goals

1. Fill in the gaps of removal-based method combinations



2. Verify that the information-theoretic model works best

3. Verify other theorized relationships between existing methods

Removal-Based Model Combinations

Feature Removal

- Default Values
- Marginalization
  - Uniform
  - Product
  - Joint

Model Behavior

- Prediction
- Prediction Loss
- Dataset Loss

Summary Technique

- Removing
- Including
- Mean when Included
- Banzhaf
- Shapley

Removal-Based Model Combinations

Feature Removal

- Default Values
- Marginalization
  - Uniform
  - Product
  - Joint

Model Behavior

- Prediction
- Prediction Loss
- Dataset Loss

Summary Technique

- Removing
- Including
- Mean when Included
- Banzhaf
- Shapley

Over 80 Combinations Tested!

Removal-Based Model Combinations

Feature Removal

- Default Values
- Marginalization
  - Uniform
  - Product
  - Joint

Model Behavior

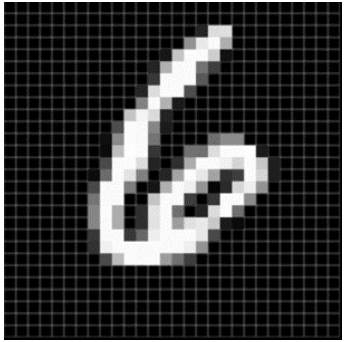
- Prediction
- Prediction Loss
- Dataset Loss

Summary Technique

- Removing
- Including
- Mean when Included
- Banzhaf
- Shapley

Experimental Domains

- Census Income
  - 48,842 Individuals
  - 14 Socioeconomic Features, >$50k income or not
  - Light-GBM (gradient boosted tree)



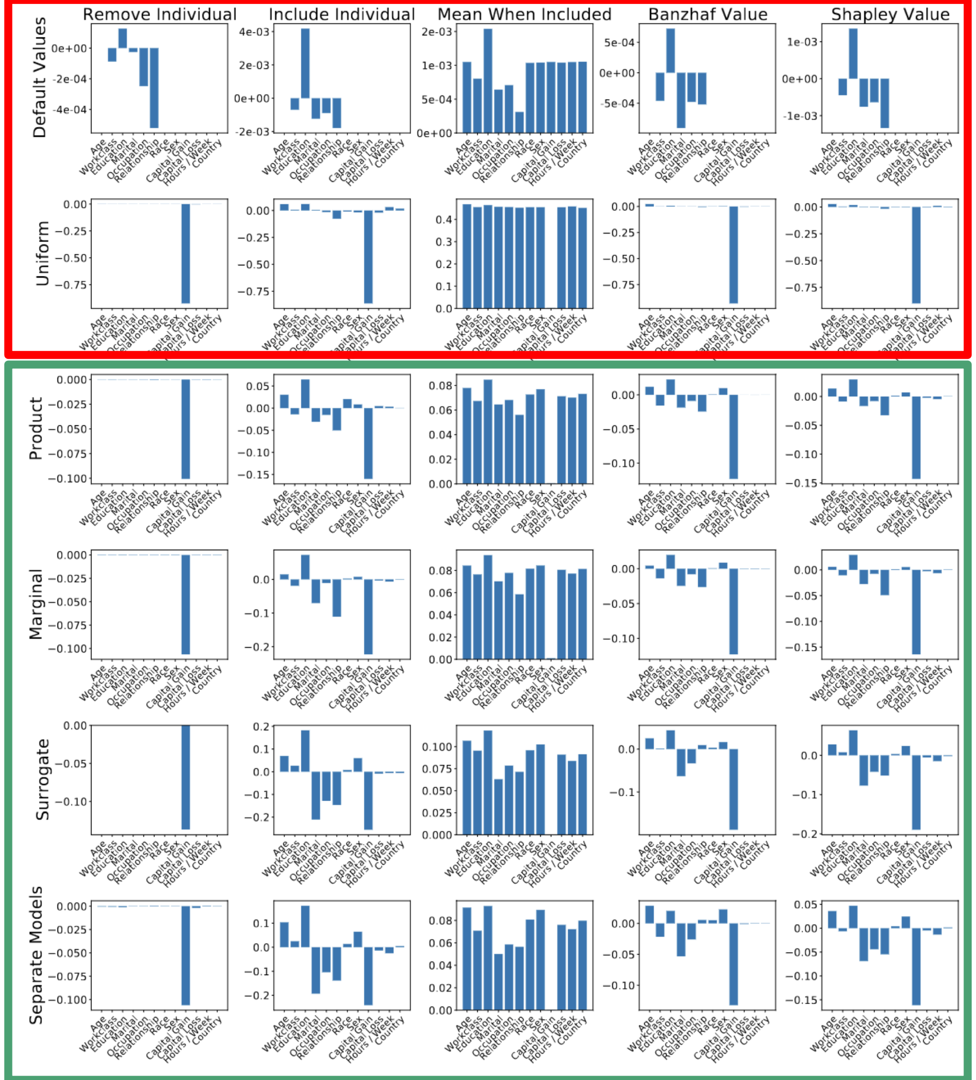
- MNIST
  - 70,000 Handwritten Digits
  - 32x32 Grayscale Pixels
  - CNN (14 Layers)

- Breast Cancer Subtypes
  - 510 Patients
  - Took a random subset of 100/17,814 genes
  - Logistic Regression

1. Census Income

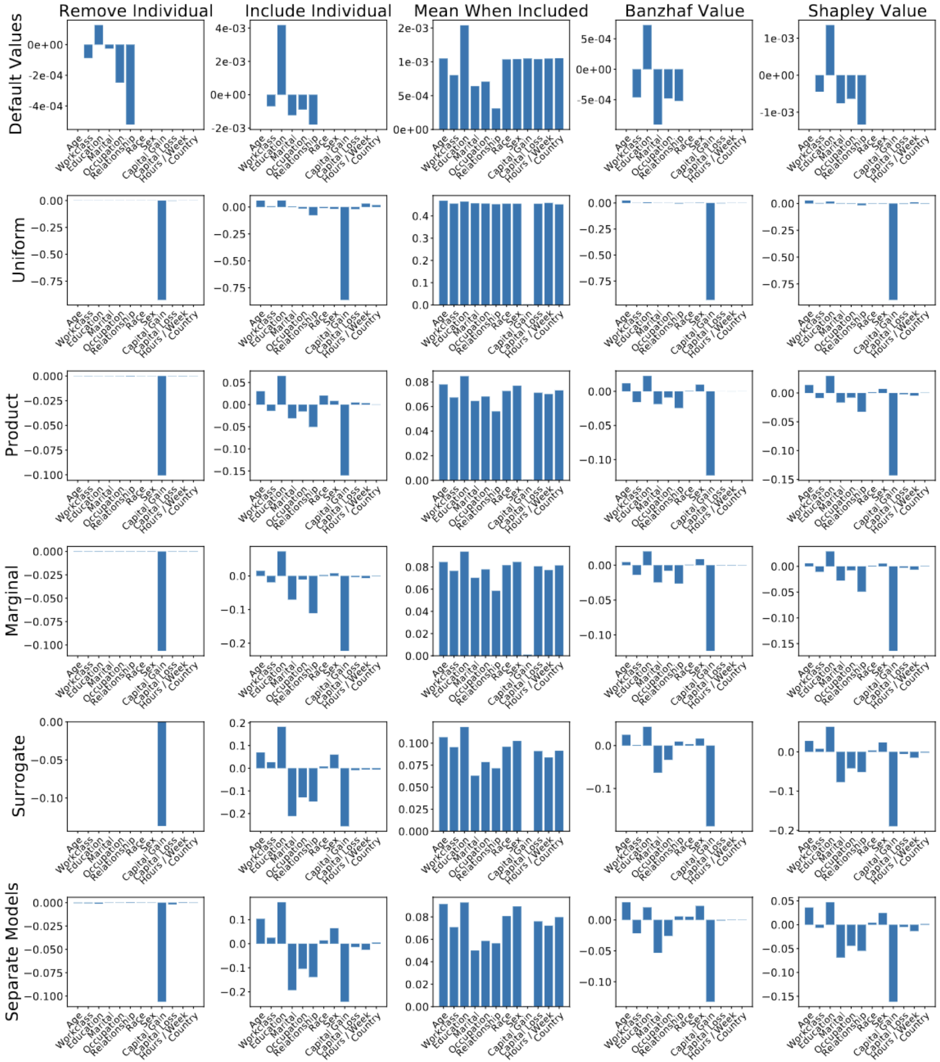
Qualitatively:

- Bottom four the same
  - Approximate conditional distribution



1. Census Income

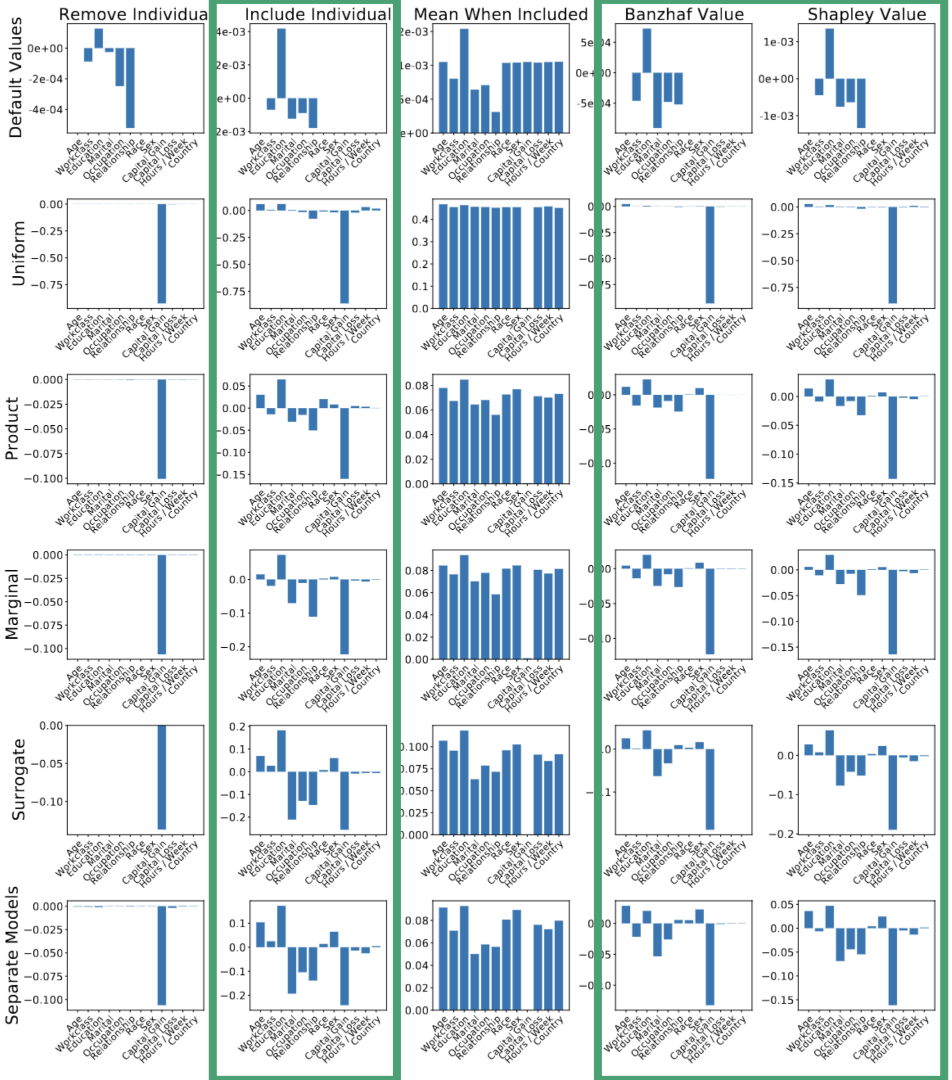
Qualitatively:



1. Census Income

Qualitatively:

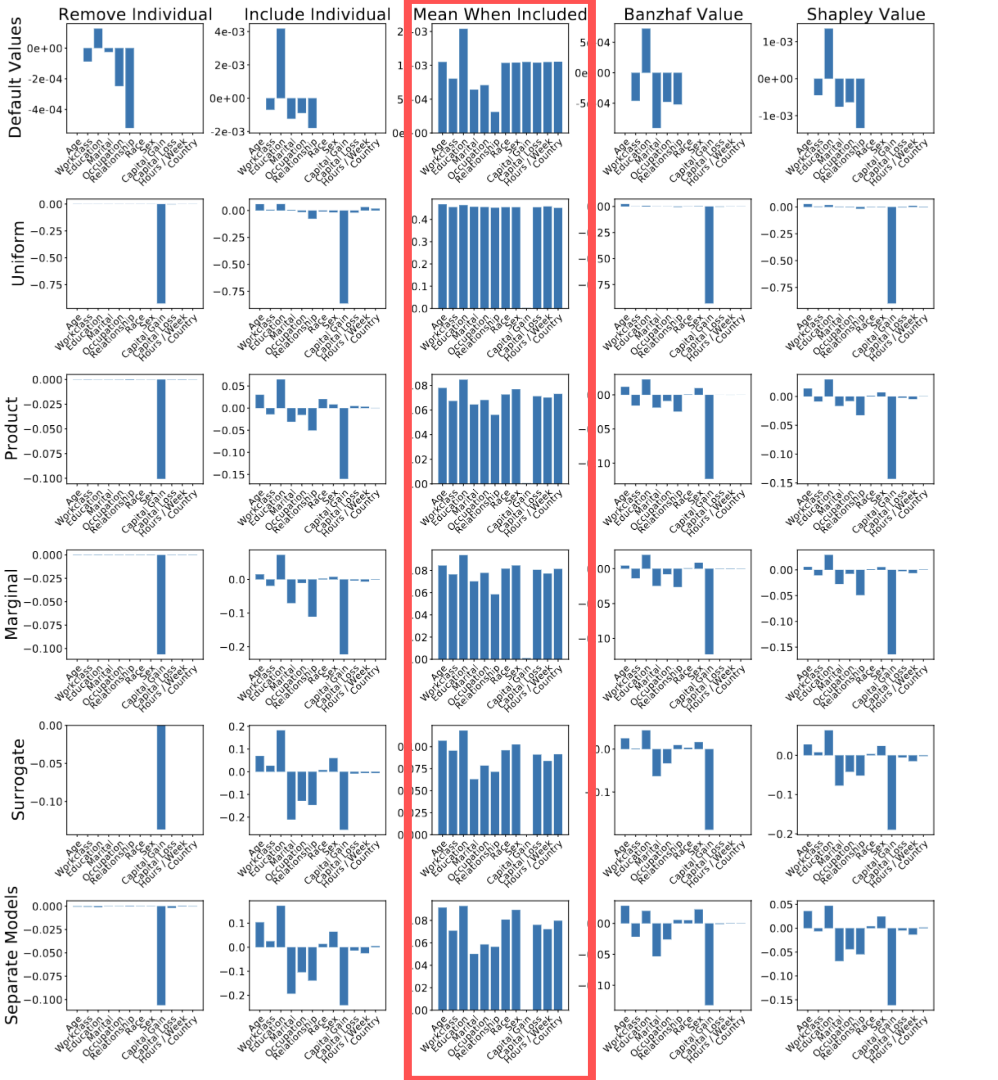
- Bottom four the same
  - Approximate conditional distribution
- Include and Banzhaf/Shapley the same
  - Similar formulations



1. Census Income

Qualitatively:

- Bottom four the same
  - Approximate conditional distribution
- Include and Banzhaf/Shapley the same
  - Similar formulations
- Mean When Included is

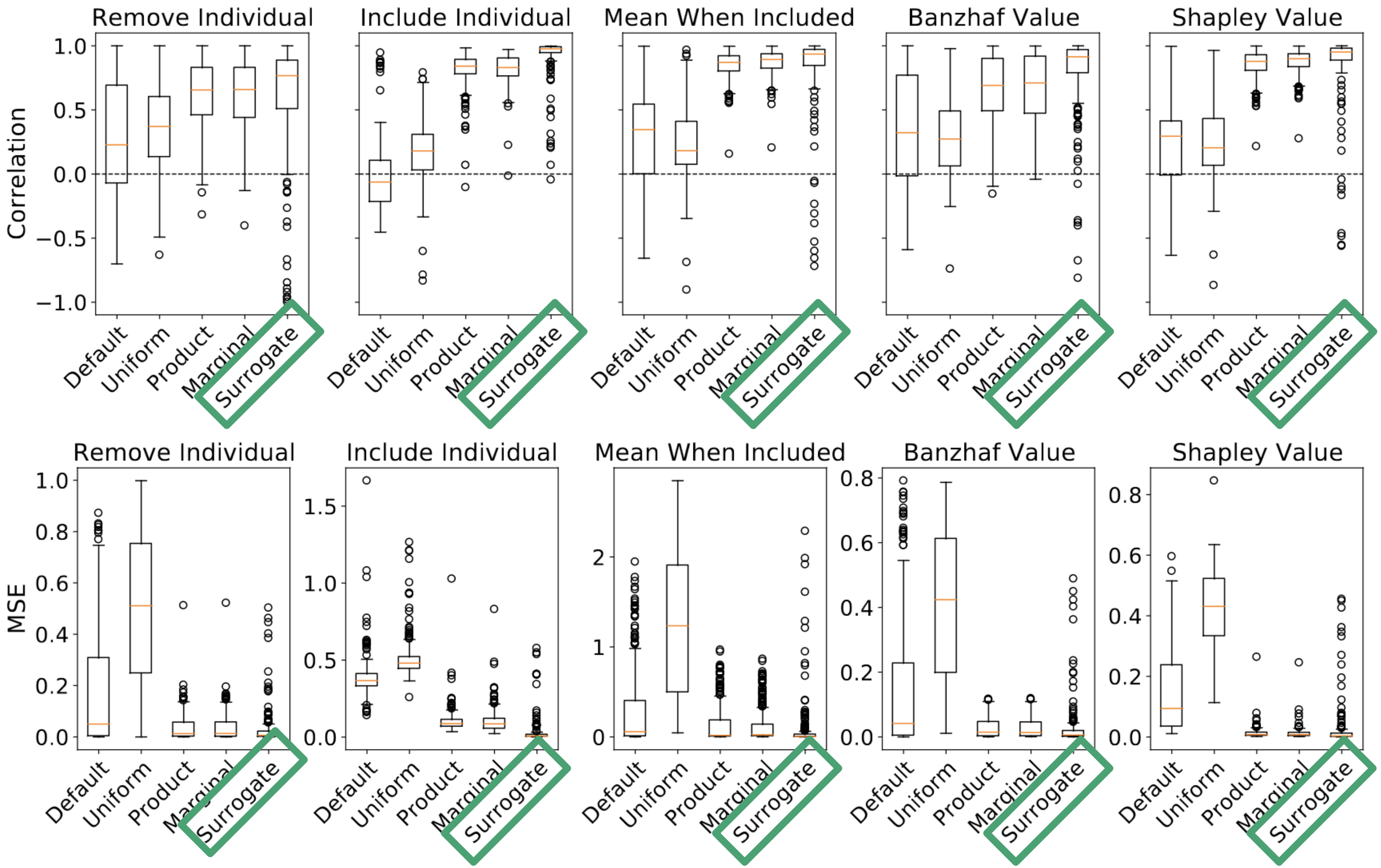


1. Census Income

Quantitatively:

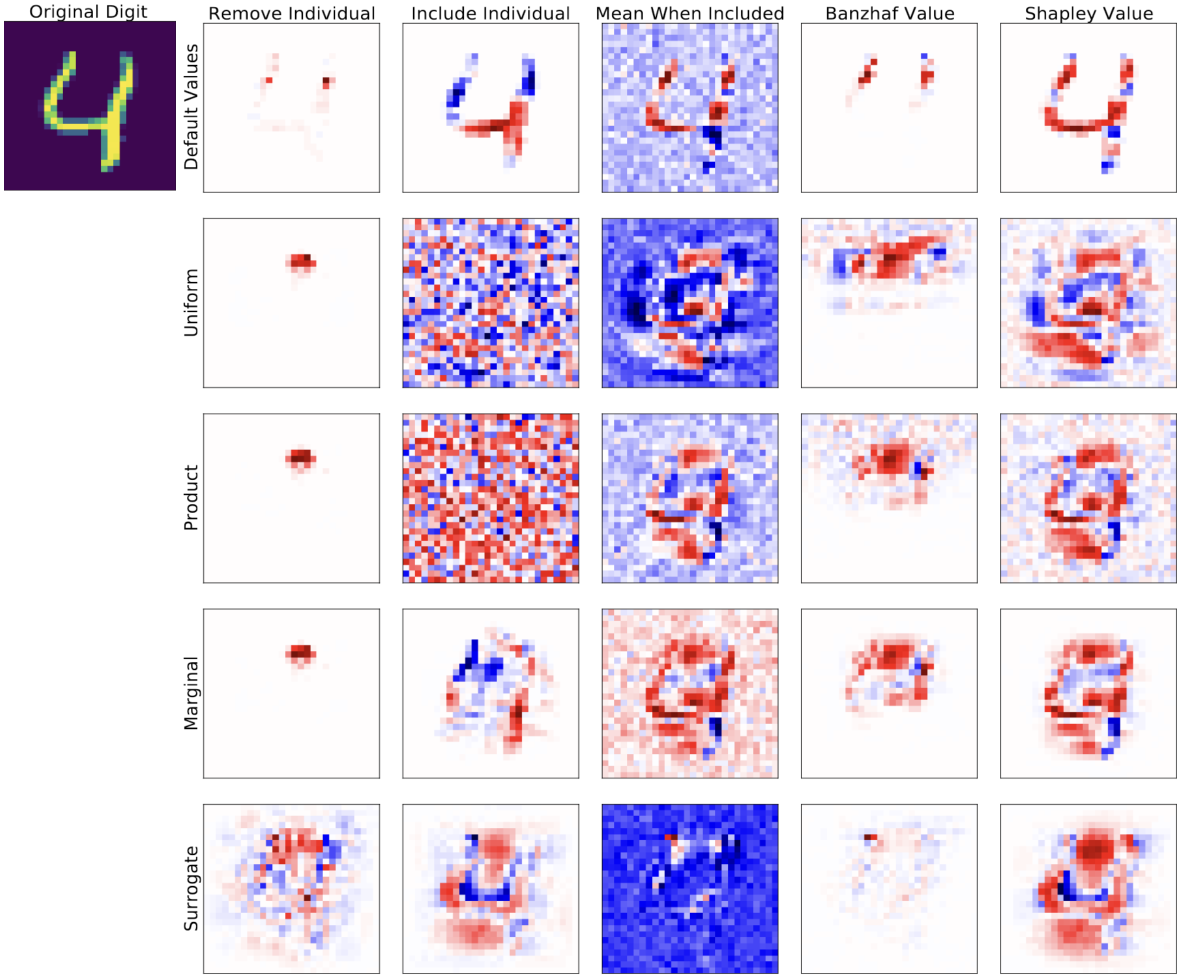
How good is the conditional distribution approximation?

- Intractable, so treat separate models as ground truth
- Surrogate Removal Method does best


Remove Individual
Include Individual
Mean When Included
Banzhaf Value
Shapley Value
Correlation
MSE
Default
Uniform
Product
Marginal
Surrogate

2. MNIST

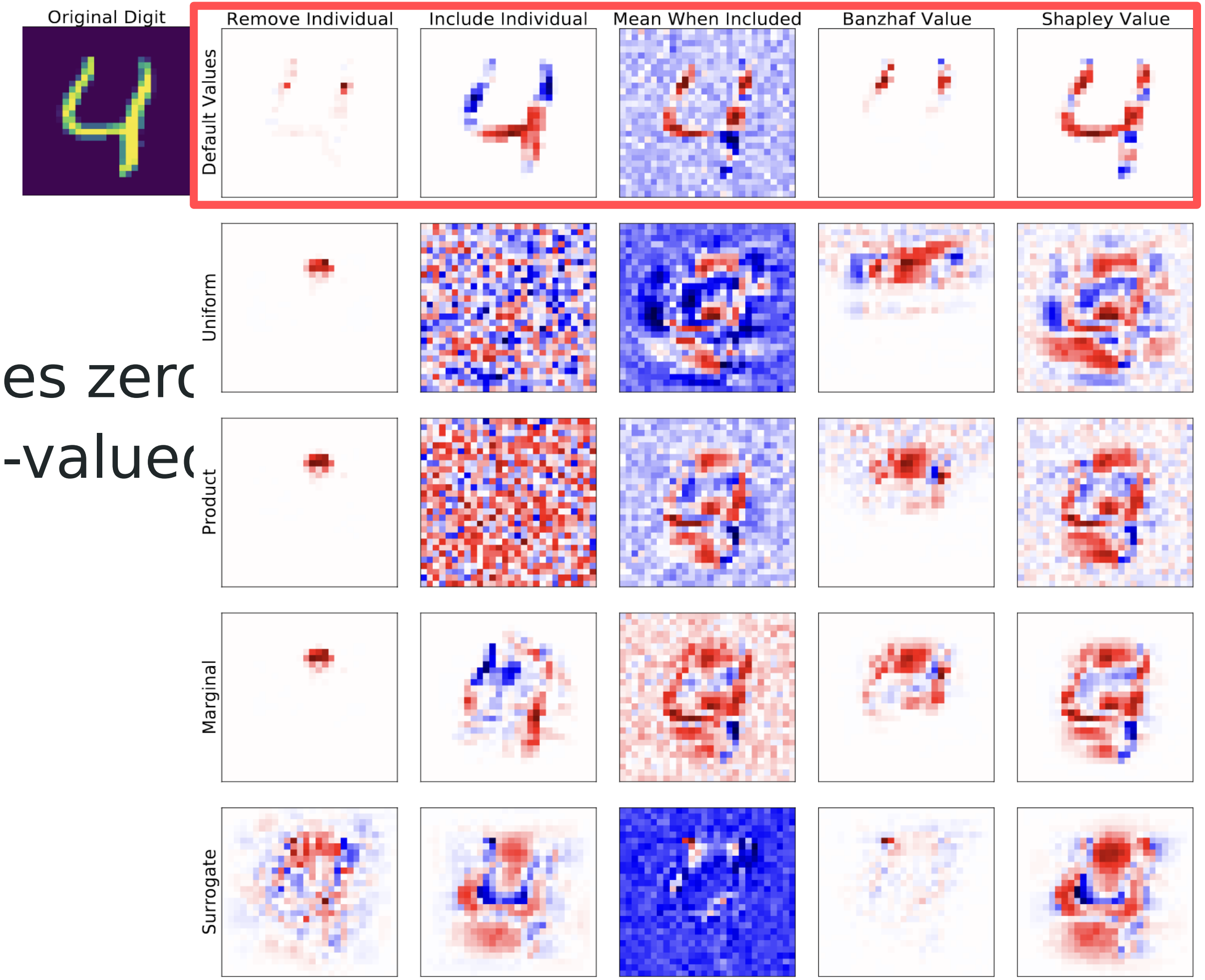
Qualitatively:



2. MNIST

Qualitatively:

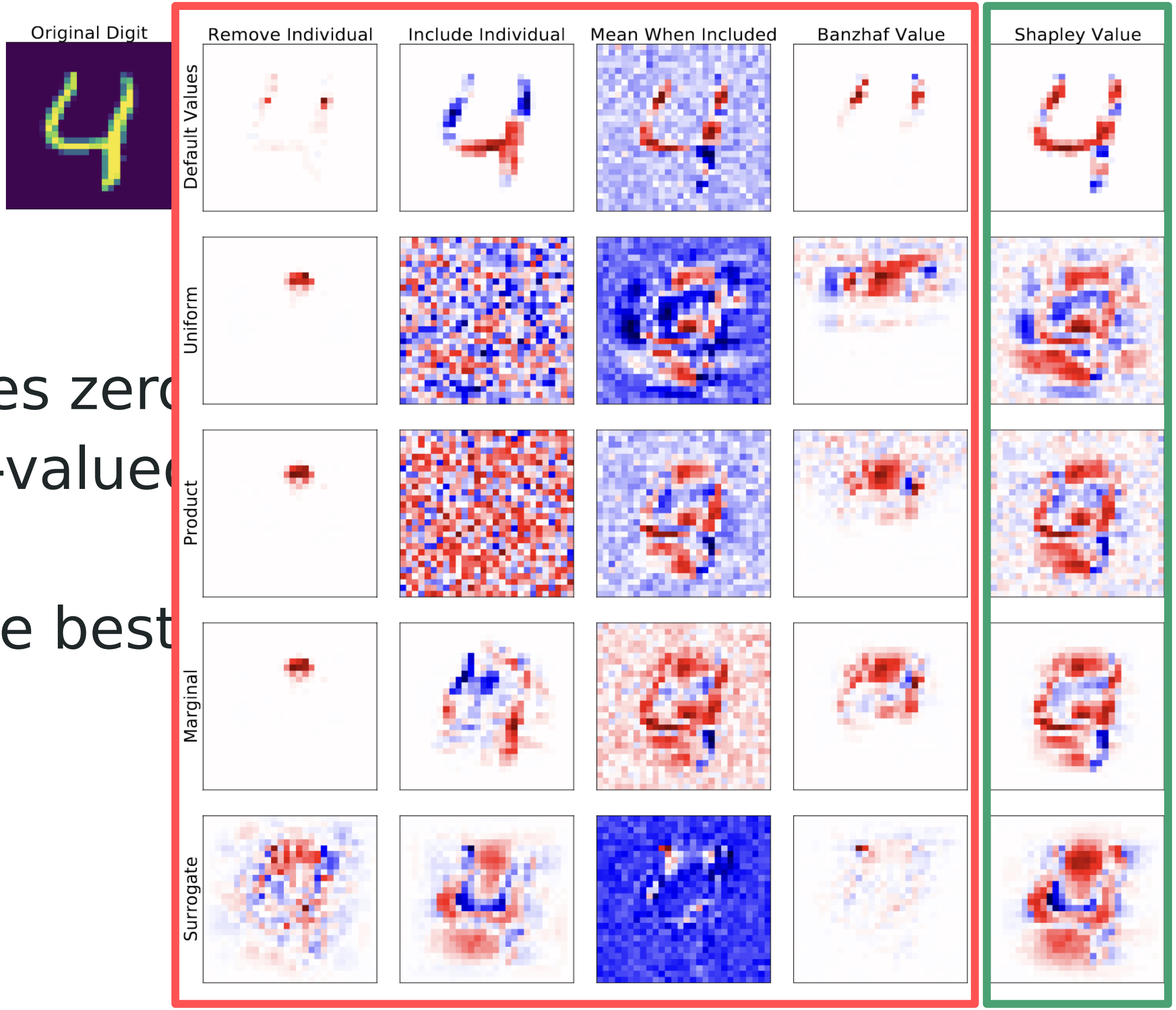
- Default Values gives zerc attribution to zero-valuec features



2. MNIST

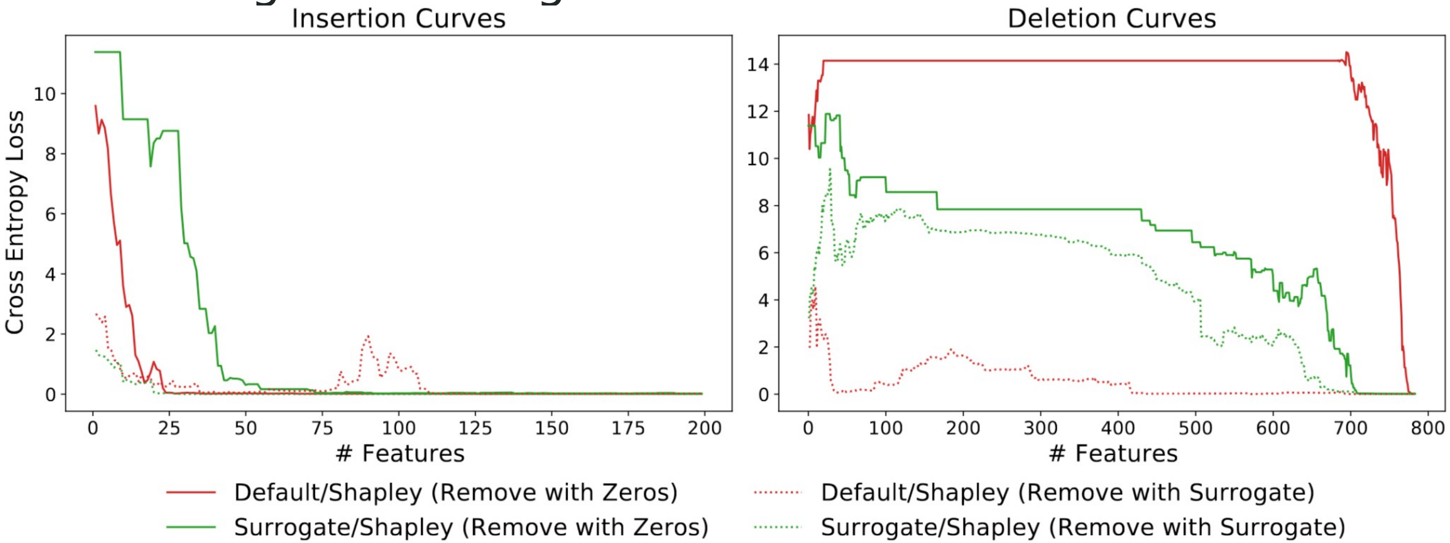
Qualitatively:

- Default Values gives zer
  attribution to zero-value
  features
- Shapley “looks” the best



2. MNIST

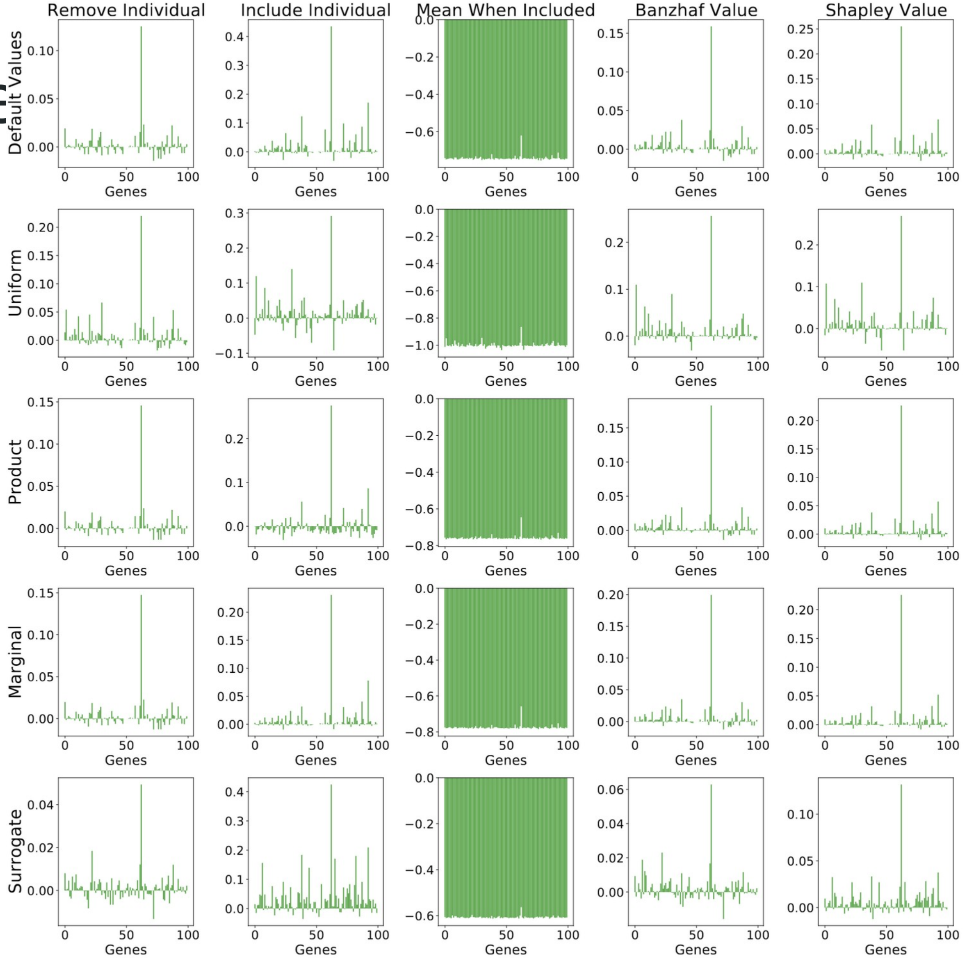
- Quantitative Analysis:
  - Insertion: Remove the bottom k important features
  - Deletion: Remove the top k important features
  - Removing with surrogate is better!



3. Breast Cancer Subtype

Qualitatively:

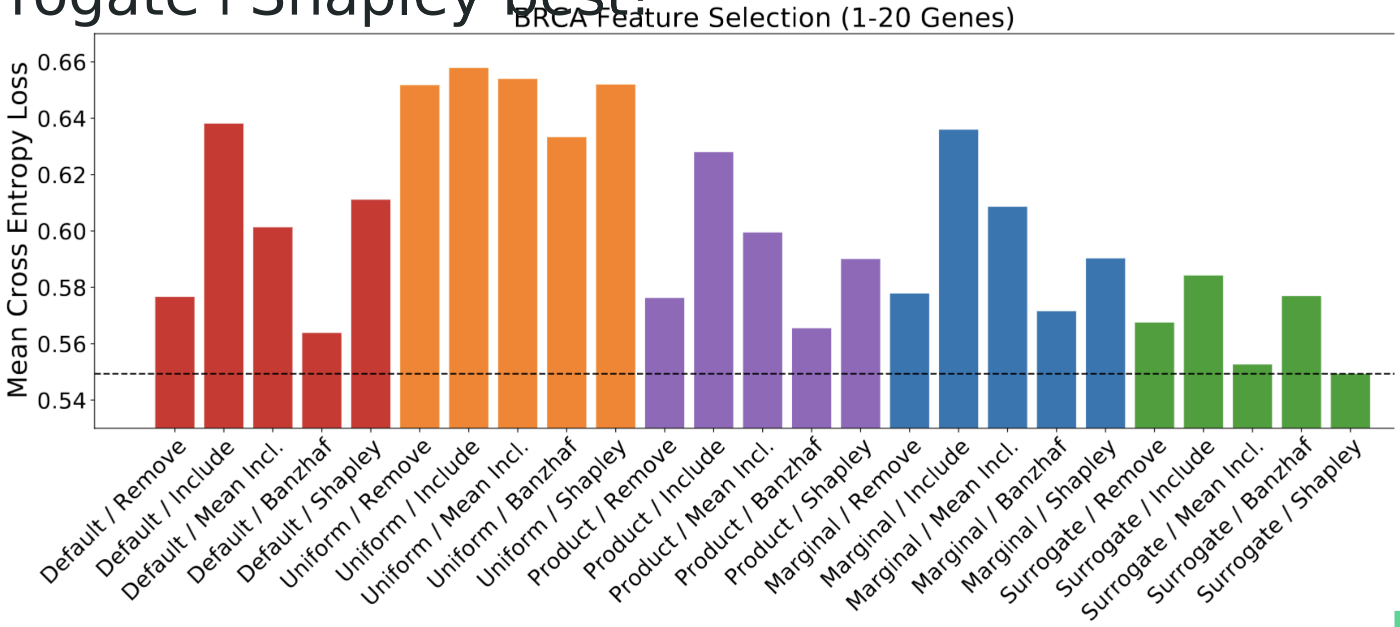
- ESR1 gene is most important
- Hard to compare qualitatively
- Verify that “mean when included” is much different from the rest



3. Breast Cancer Subtypes

Quantitatively:

- Select only 20 most important genes
- Surrogate+Shapley best!



Conclusion

- SHAP+Surrogate Model is the best
  - In line with information-theoretic connection
- The authors’ unified framework helps us reason abstractly about these techniques
  - Able to identify and fill so many gaps in the combination of removal-based models

Limitations

- No limitations section
- No comparisons across model behavior
- Quantifying approximation quality with an approximation
- Evaluation metrics can be aligned with explanation method
  - Doesn’t bode well for a universal unbiased metric
- Slightly suspect analysis for Breast Cancer dataset

Discussion

- Are removal-based the right framework to go in XAI?
- How do we reconcile with the fact that removing features may result in out-of-distribution behavior?
- What other unifying frameworks can you think of?
- Do you agree with the assessment that SHAP provides the best explanation?
- What methods do not fall under removal-based methods?