

# Visualization Methods for Explainable Machine Learning

## Statistics PhD Oral Prelim

Katherine Goode  
Iowa State University  
May 6, 2020

# Personal Background

## Education

B.A. in Mathematics

- Lawrence University (Appleton, WI)
- Graduated in June 2013

M.S. in Statistics

- University of Wisconsin, Madison
- Graduated in May 2015

Ph.D. in Statistics (in progress)

- Iowa State University
- Started in January 2016

## Teaching

- Teaching assistant at UW Madison
- Lecturer at Lawrence University
- Lecturer at ISU

## Consulting

- AES Statistical Consultant
- NREM Research Assistant

## Internship

- Sandia National Labs: Statistical Sciences Research and Development Intern

# Overview of Talk

1. Background and Overview of Thesis

2. Detailed explanation of Chapter 1

- Visual Diagnostics of a Model Explainer -- Tools for the Assessment of LIME Explanations

3. Plan for Chapter 2

- Explaining Random Forests using Clustering of Trees

4. Ideas for Chapter 3

- Extensions of Neural Network Explanation Tools to Tabular Data Applications

5. Timeline for Completion

6. Discussion Points

# Background and Overview of Thesis

# Explainable Machine Learning

- Machine learning models:
  - Good in prediction problems
  - Many considered "black-boxes" since too complex to directly interpret
- Led to research area of explainable machine learning
  - Goal: provide methods to explain predictions made by black-box models
  - Method overview papers/books: (Gilpin, Bau, Yuan, Bajwa, Specter, and Kagal, 2018; Guidotti, Monreale, Ruggieri, Turini, Pedreschi, and Giannotti, 2018; Ming, 2017; Mohseni, Zarei, and Ragan, 2018; Molnar, 2019).
- European General Data Protection Regulation (GDPR) implemented in 2018 includes a "right to explanation"

Goodman and Flaxman (2016): "It is reasonable to suppose that any adequate explanation would, at a minimum, provide an account of how input features relate to predictions, allowing one to answer questions such as: Is the model more or less likely to recommend a loan if the applicant is a minority?"

# Explainability versus Interpretability

No accepted definitions for explainability and interpretability (Gilpin, Bau, Yuan, et al., 2018; Lipton, 2016; Molnar, 2019; Montavon, Samek, and Müller, 2017; Murdoch, Singh, Kumbier, Abbasi-Asl, and Yu, 2019)

My definitions (implicitly used by Rudin (2018) and Ribeiro, Singh, and Guestrin (2016)):

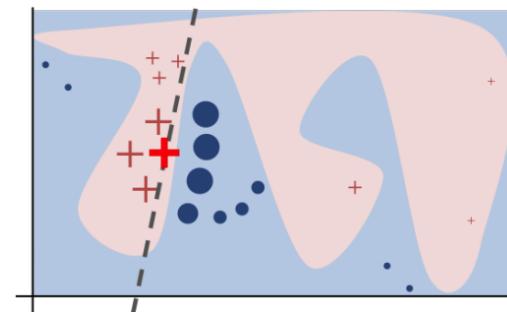
**Interpretability** is the ability to **directly use model parameters** to understand the mechanism of how the model **makes predictions**.

- Linear regression model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

**Explainability** is the ability to **use the model in an indirect manner** to understand the relationships in the data captured by the model.

- LIME: local interpretable model-agnostic explanations (Ribeiro, Singh, and Guestrin, 2016)



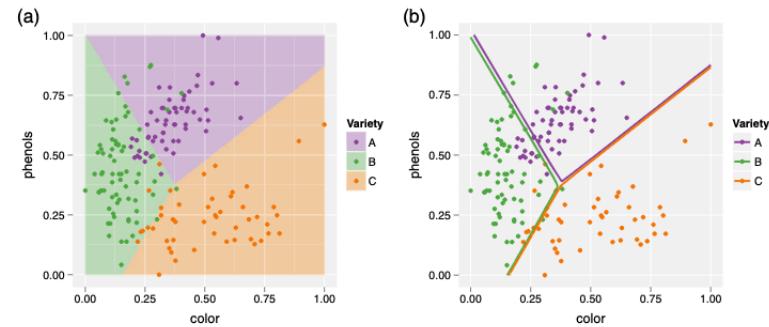
# Model Agnostic Methods

## General Model Visualizations: Strategies for understanding any model

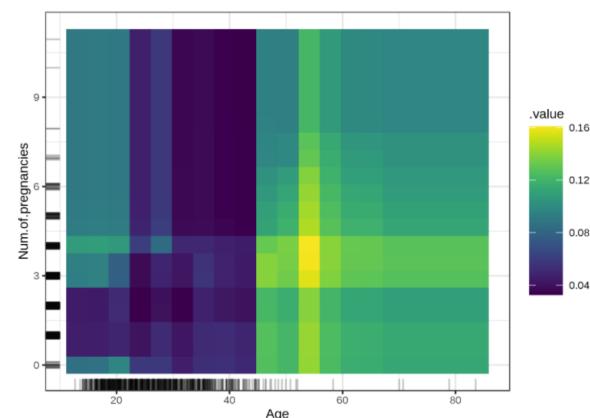
- Removing the blindfold (Wickham, Cook, and Hofmann, 2015)

## Global Methods: Explanation for model as a whole

- Partial dependence plots (Friedman, 2001) and extensions:
  - Interactive partial dependence plots (Krause, Perer, and Ng, 2016)
  - Individual conditional expectation plots (Goldstein, Kapelner, Bleich, and Pitkin, 2013)
  - Accumulated local effect plots (Apley and Zhu, 2016)
  - Feature interaction plots (Friedman and Popescu, 2008; Greenwell, Boehmke, and McCarthy, 2018; Hooker, 2004)
- Global feature importance plots (Fisher, Rudin, and Dominici, 2018; Altmann, Tološi, Sander, and Lengauer, 2010; Casalicchio, Molnar, and Bischl, 2019)
- Global surrogate models (Molnar, 2019)



Model in data-space from Wickham, Cook, and Hofmann (2015).



Partial dependence plot from Molnar (2019).

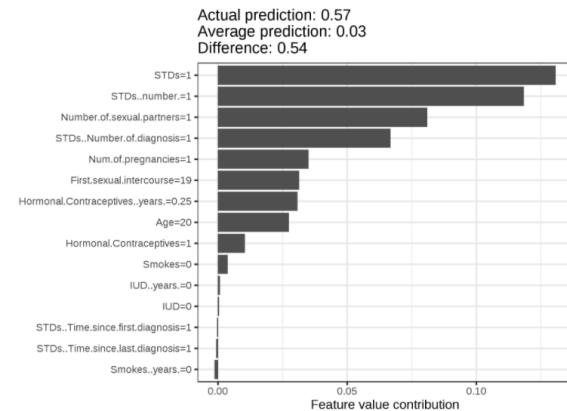
# Model Agnostic Methods

## Local Methods: Explanation for an individual prediction

- Individual conditional importance plots (Casalicchio, Molnar, and Bischl, 2019)
- LIME (Ribeiro, Singh, and Guestrin, 2016)
- Anchors (scoped rules) (Ribeiro, Singh, and Guestrin, 2018)
- Shapely values
- SHAP (Lundberg and Lee, 2017)
- breakDown (Staniak and Biecek, 2018)

## Example Based: Explanations based on examples from the data

- Counterfactual examples (Wachter, Mittelstadt, and Russell, 2017; Martens and Provost, 2014; Looveren and Klaise, 2019; Laugel, Lesot, Marsala, Renard, and Deryniecki, 2017)
- Adversarial examples (Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, and Fergus, 2013; Goodfellow, Shlens, and Szegedy, 2014; Biggio and Roli, 2018; Su, Vargas, and Sakurai, 2019)
- Prototypes and criticisms
- Influential instances (Koh and Liang, 2017)



Bar chart of shapley values from Molnar (2019).



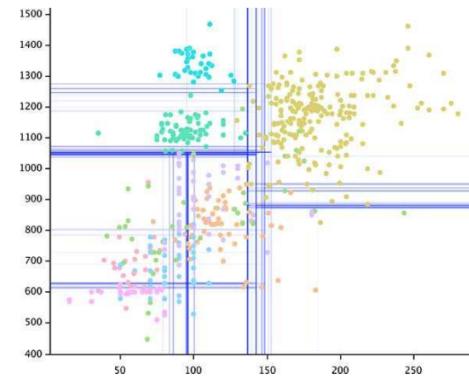
Adversarial example where one pixel change affects prediction from Su (2019).

# Model Specific Methods

## Random Forests

- Random forest impurity based feature importance (Breiman, 2001)
- Sectioned scatterplots (Urbanek, 2008)
- Trace plots of trees (Urbanek, 2008)
- Simplified model (Hara and Hayashi, 2016a)
- Forest floor visualizations (Welling, Refsgaard, Brockhoff, and Clemmensen, 2016)
- Interactive visualizations

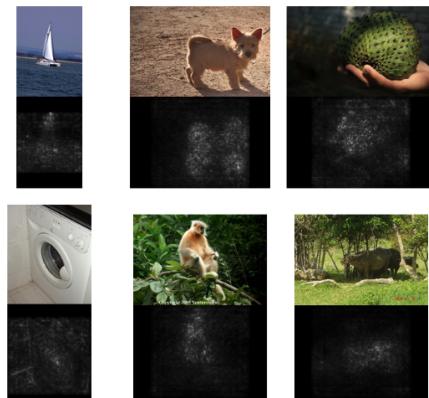
(Beckett, 2018; da Silva, Cook, and Lee, 2017)



Sectioned scatterplot from Urbanek (2008).

## Neural Networks

- Extracting tree structures (Craven and Shavlik, 1996)
- Saliency maps (Simonyan, Vedaldi, and Zisserman, 2013)
- Feature visualization (Olah, Mordvintsev, and Schubert, 2017)
- Grand tours (Li, Zhao, and Scheidegger, 2020)
- Flows (Halnaut, Giot, Bourqui, and Auber, 2020)



Saliency maps from Simonyan, Vedaldi, and Zisserman (2013).

# Assessments of Explainable Machine Learning

## General

Argument against black box model explanations (Rudin, 2018):

- "Explanations must be wrong."
- Explanations may not:
  - be faithful to the original model
  - make sense
  - be detailed enough to understand the "black-box" model
- Debunks accuracy and interpretability trade-off myth
- Use interpretable models for high-stakes decisions

## Method Specific

Assessment of LIME (Laugel, Renard, Lesot, Marsala, and Detyniecki, 2018):

- How to choose a local region?

Assessment of counterfactual examples (Laugel, Renard, Lesot, et al., 2018):

- Issues with unjustified counterfactual examples

Assessment of saliency maps (Kindermans, Hooker, Adebayo, Alber, Schütt, Dähne, Erhan, and Kim, 2017):

- Transformation to input data affects saliency map but not model

# Overview of Dissertation Chapters

## Chapter 1: Visual Diagnostics for LIME

- Discuss importance of assessing LIME
- Suggest the use of visualizations for assessment and provide example visualizations

## Chapter 2: Visualizations for Explaining Random Forests

- Use clustering to identify key tree structures within the random forest
- Improve visualizations for use of trees as global surrogate models

## Chapter 3: Visualizations for Explaining Neural Networks

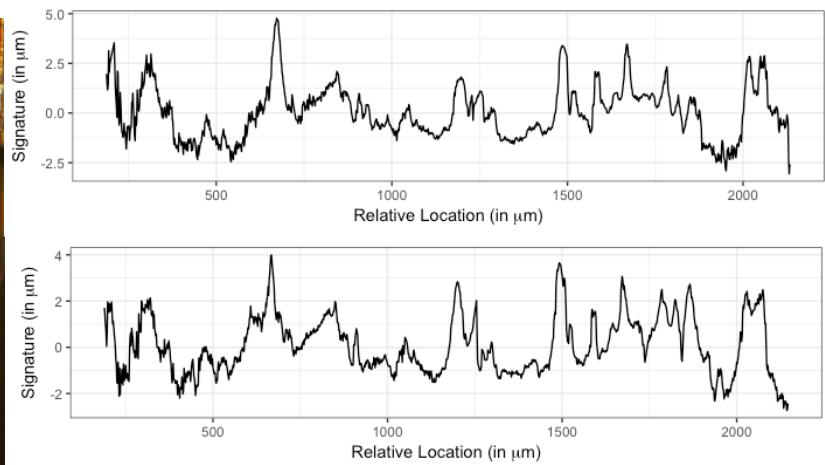
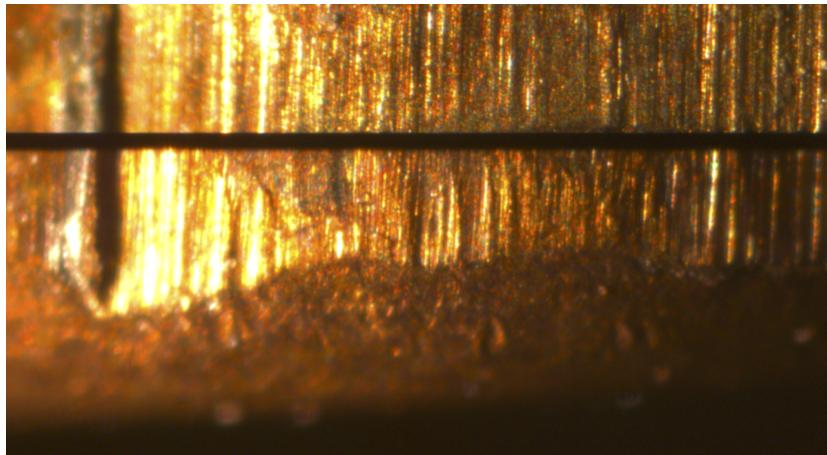
- Project for my internship with Sandia National Labs
- Application with functional data
- Visualizations for the explaining and understanding the models

# Chapter 1: Visual Diagnostics of a Model Explainer -- Tools for the Assessment of LIME Explanations

# Motivation

Hare, Hofmann, and Carriquiry (2016):

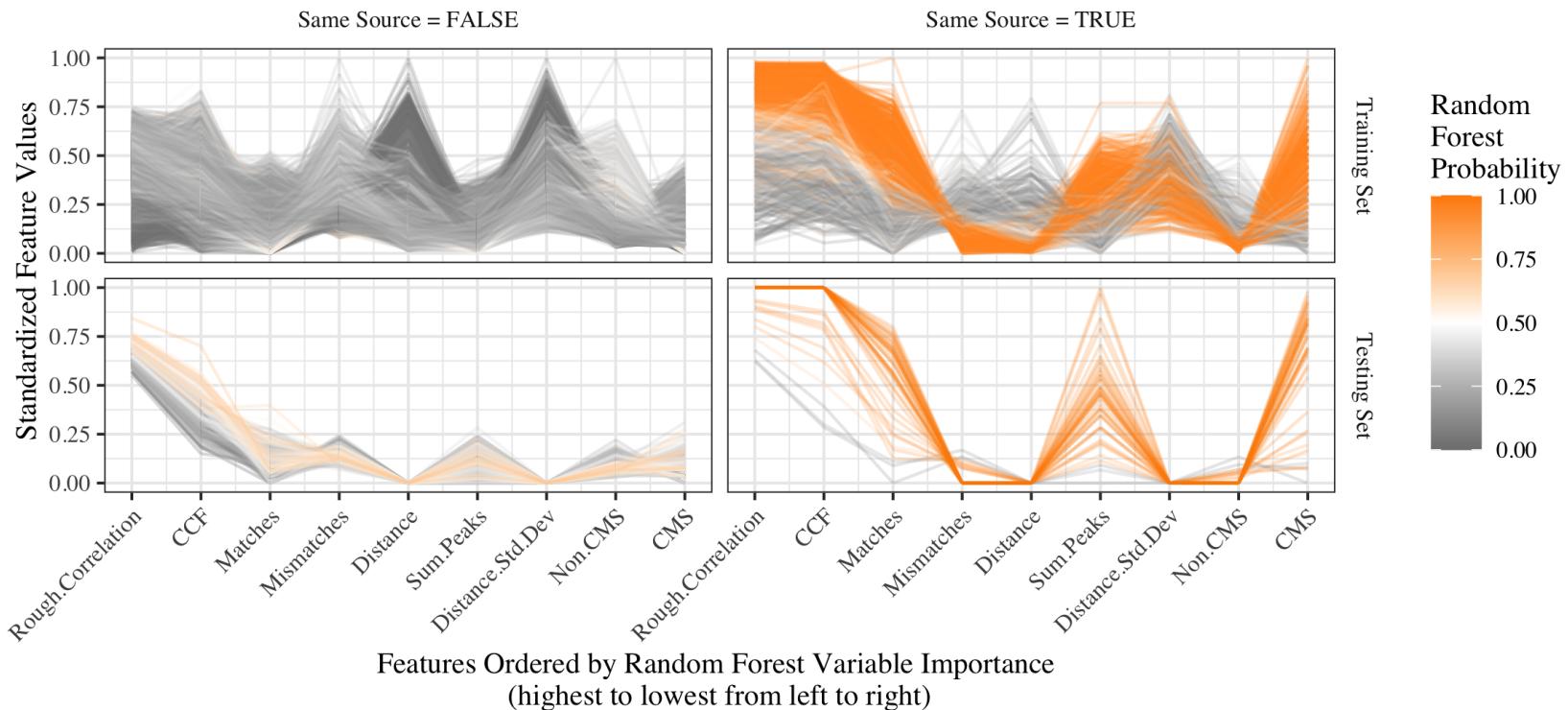
- Want to provide quantitative evidence for whether two bullets were fired from the same gun
- Use high definition scans of striations on bullet lands to extract "signatures"
- Compute similarity features to compare two signatures



# Motivation

Hare, Hofmann, and Carriquiry (2016) approach:

- Random forest model
- 9 signature similarity features
- Returns a score for the comparison of two lands



# Motivation

**Our Original Goal:** Provide explanations for specific signatures comparisons

**Attempt:** Applied LIME

**Result:** Unreasonable explanations (e.g., LIME explanation does not agree with random forest prediction)

**Example:** Known non-match

# Conceptual Depiction of LIME

LIME (Ribeiro, Singh, and Guestrin, 2016):

- Local
- Interpretable
- Model-agnostic
- Explanation

**Concept:** For *one prediction of interest*

- Focus on a neighborhood around the prediction of interest
- Use an inherently interpretable model
- Understand the complex model
- Capture the relationship between the complex model predictions and predictor variables

# Importance of Assessing LIME

Additional layer of complexity:

- Start with a complex model
- LIME uses another model
- End with two models to assess

Questions raised:

- Explainer model a good approximation?
- Appropriate local region?
- Relationship linear in the local region?
- Which tuning parameter settings to use when applying LIME?

# Visualizations for Model Assessment

Claims made about LIME (Ribeiro, Singh, and Guestrin, 2016):

- **Interpretability:** Easy to interpret the explainer model to provide meaningful explanations
- **Faithfulness:** Explainer model sufficiently captures the relationship between the complex model predictions and the features in the local region around a prediction of interest
- **Linearity:** Using a ridge regression as the explainer model assumes a linear relationship between complex model predictions and the features
- **Localness:** Explanations are local in regards to a prediction of interest

We suggest visual diagnostics to assess these claims:

- **Diagnostics for individual explanations**
- **Diagnostics for sets of explanations**
- **Diagnostics for comparisons of tuning parameters**

# Sine Example Data

Training data: 500 observations

Testing data: 100 observations

Black-box model: random forest

$$x_1 \sim \text{Unif}(-10, 10)$$

$$x_2 \sim \text{Unif}(-10, 10)$$

$$x_3 \sim \mathcal{N}(0, 1)$$

$$y = \begin{cases} \text{blue} & \text{if } x'_2 > \sin(x'_1) \\ \text{red} & \text{if } x'_2 \leq \sin(x'_1). \end{cases}$$

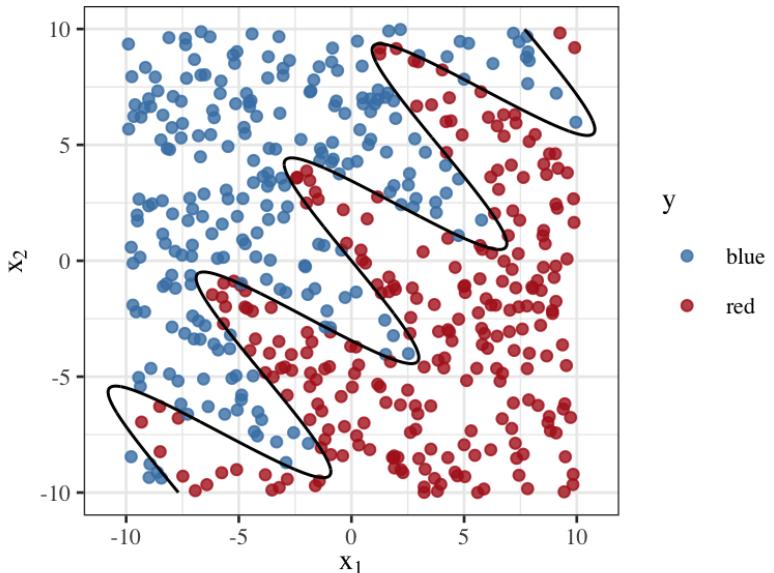
where

$$x'_1 = x_1 \cos(\theta) - x_2 \sin(\theta)$$

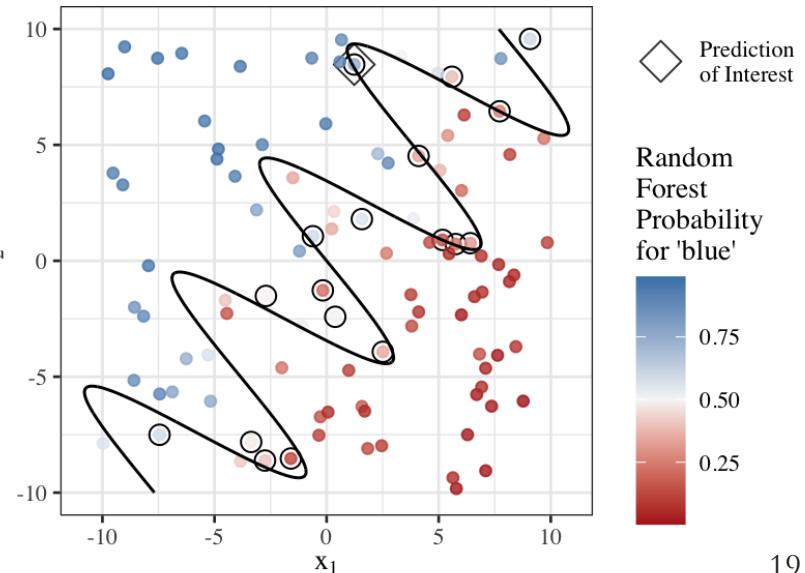
$$x'_2 = x_1 \sin(\theta) + x_2 \cos(\theta)$$

$$\theta = -0.9$$

Training Data



Testing Data



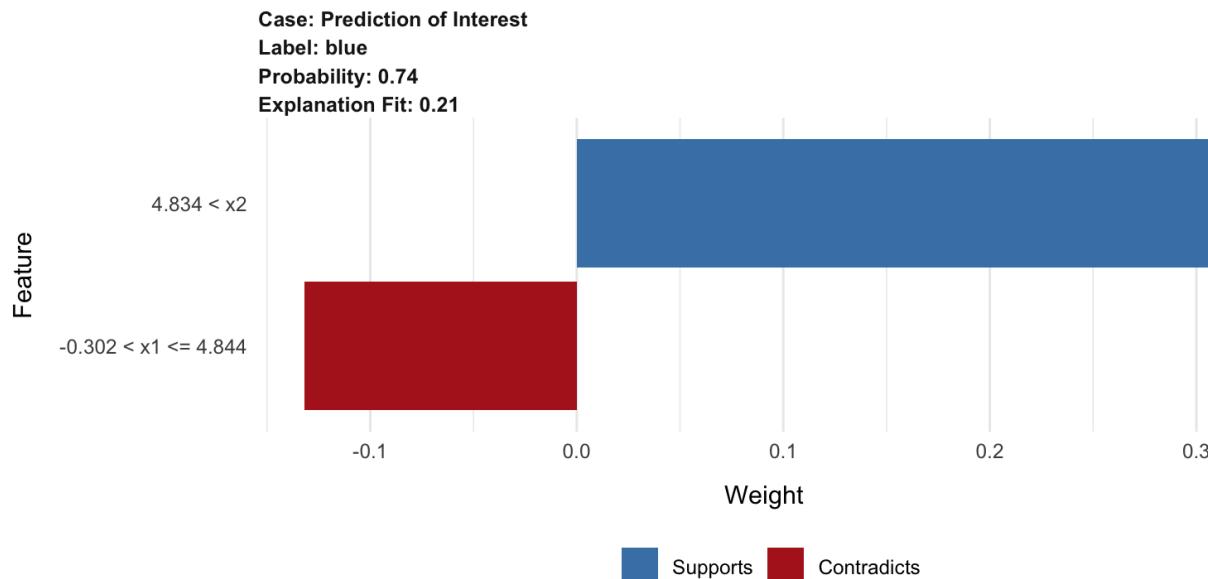
# Set 1 of Visualizations: Diagnostics for Individual Explanations

# Diagnostics for Individual Explanations

Applied LIME using *lime* R package (Pedersen and Benesty, 2020)

- To prediction of interest
- Number of features to return in explanation: 2
- Default tuning parameters settings

*lime* R package visualization of the explanation:



# Diagnostics for Individual Explanations

## Step 1a: Data Simulation

- Sample 4999 observations uniformly from 4 quantile bins for each feature in the training data

## Corresponding Diagnostic:

### *Training Data Plot*

- **Axes**: two features selected by LIME
- **Points**: training data
- **Color**: observed response
- **Lines**: 4 quantile bins boundaries



# Diagnostics for Individual Explanations

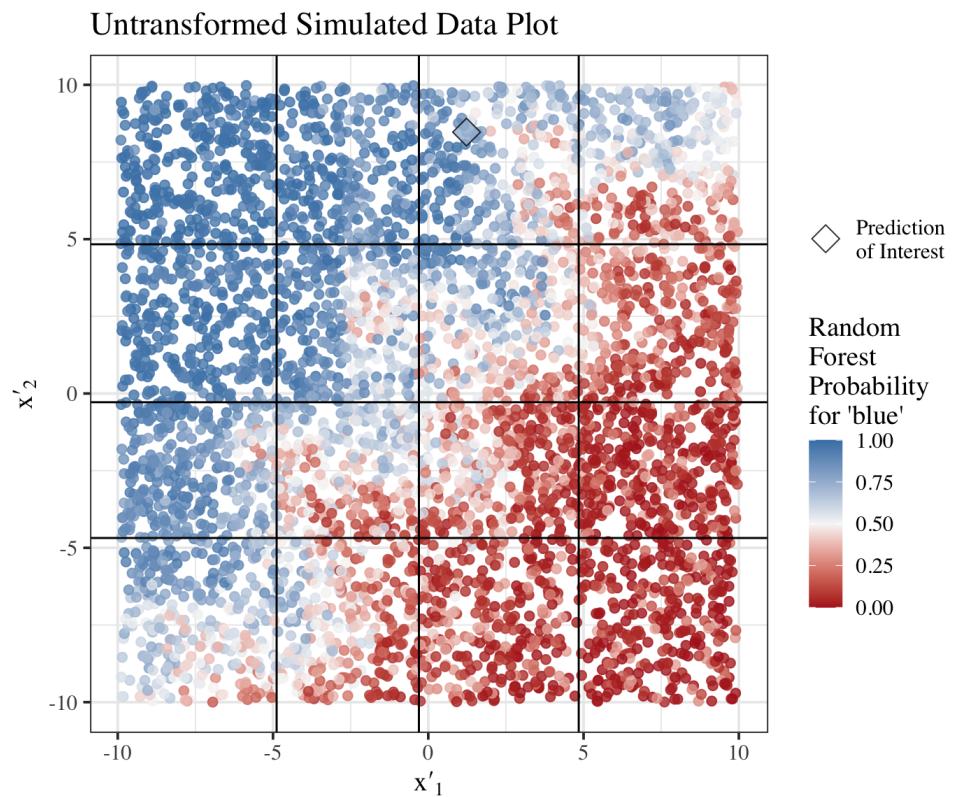
## Step 1b: Complex Model Predictions

- Apply complex model to simulated data to obtain predictions

### Corresponding Diagnostic:

#### *Untransformed Simulated Data Plot*

- **Axes:** simulated data features
  - denoted as  $x'_1$  and  $x'_2$ )
- **Points:** simulated data
- **Diamond:** prediction of interest
- **Color:** random forest prediction (for 'blue')
- **Lines:** 4 quantile bins boundaries



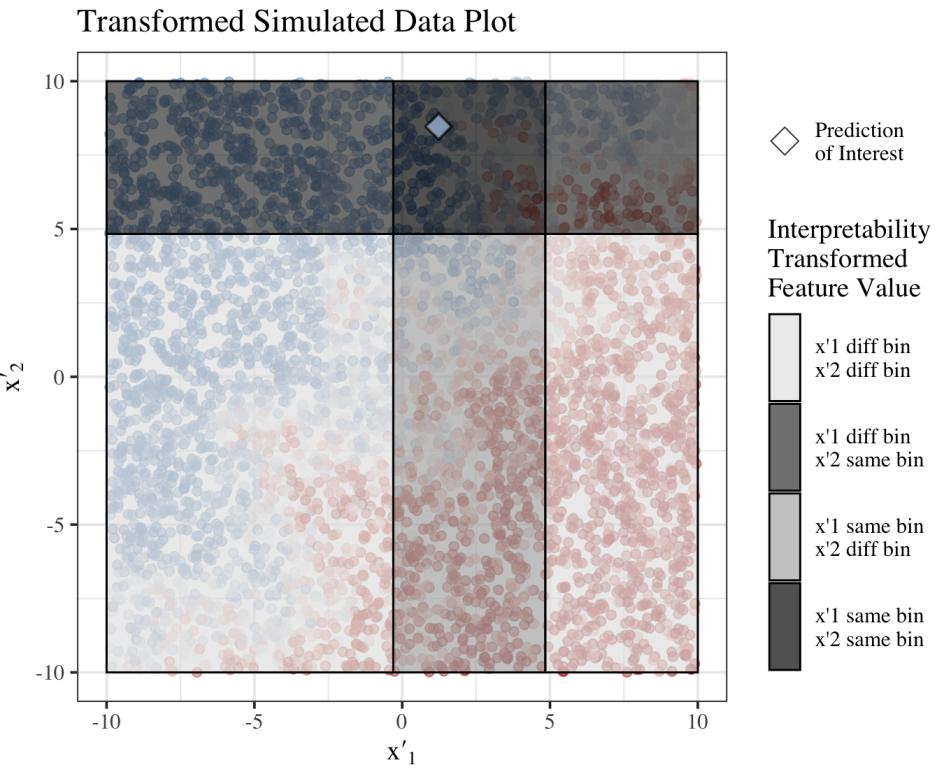
# Diagnostics for Individual Explanations

## Step 1c: Interpretability Transformation

- Convert continuous features to binary variables based on whether the observation falls in the same quantile bin as the prediction of interest or not

Corresponding Diagnostic: *Transformed Simulated Data Plot*

- Axes:** Simulated data features
- Points:** simulated data
- Diamond:** prediction of interest
- Color:** random forest prediction (for 'blue')
- Rectangle Shades:** interpretability transformed feature regions



# Diagnostics for Individual Explanations

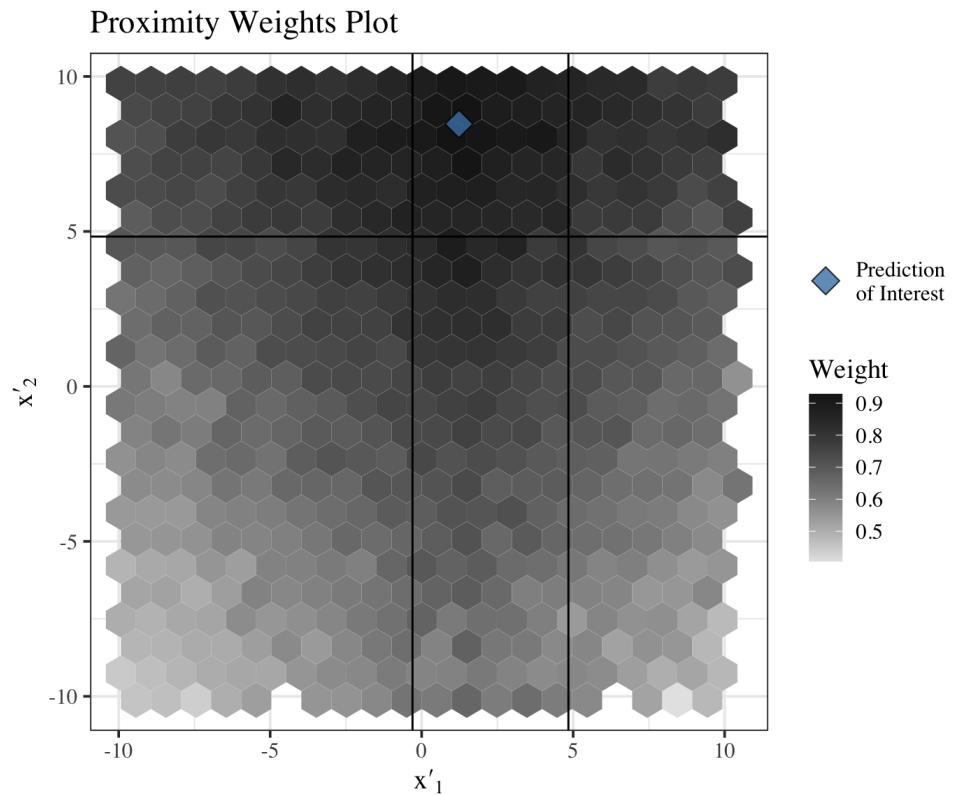
## Step 2a: Assign Weights

- Assign weights to simulated data based on proximity to the prediction of interest (using untransformed feature values)
- Gower distance metric

### Corresponding Diagnostic:

#### *Proximity Weights Plot*

- **Axes:** Simulated data features
- **Rectangle Color:** average weight within hexagon region
- **Lines:** interpretability transformation boundaries
- **Diamond:** prediction of interest



# Diagnostics for Individual Explanations

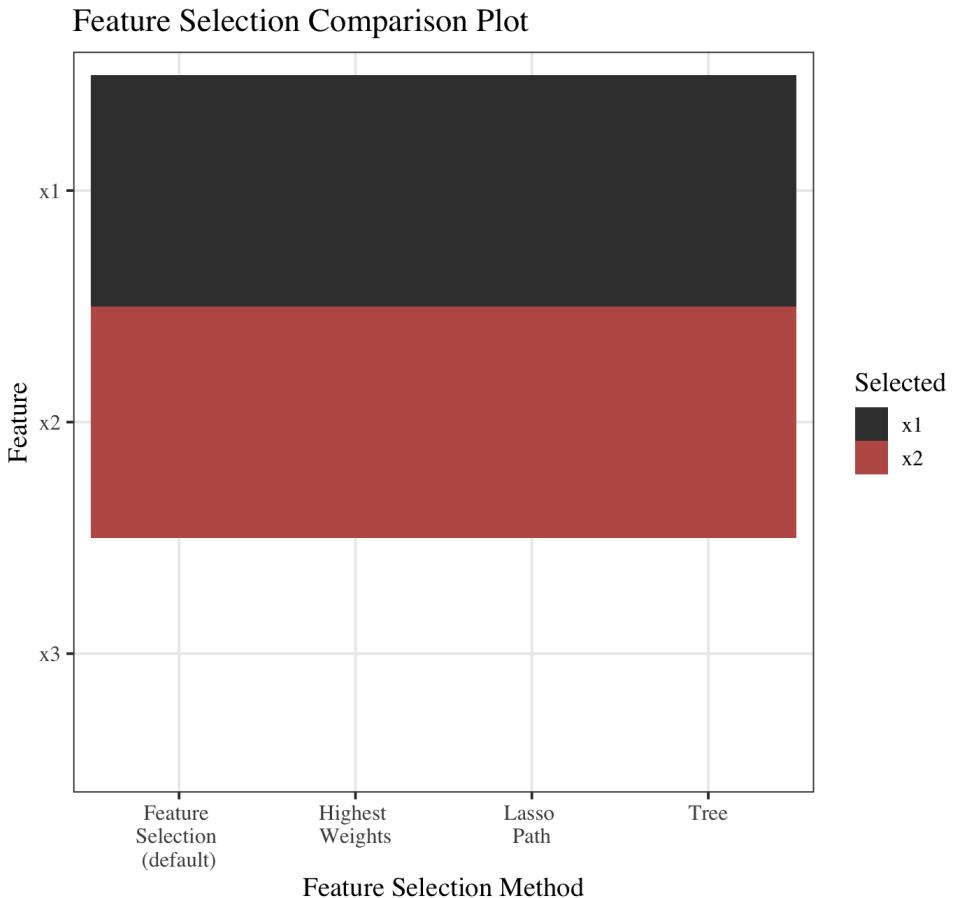
## Step 2b: Feature Selection

- Ridge regression model fit to simulated data
  - Response: complex model predictions
  - Features: interpretability transformed features
- Forward selection (if less than 6 features specified)

### Corresponding Diagnostic:

*Feature Selection Comparison Plot*

- **Axes:** Training data features versus feature selection method
- **Tile Color:** indicates if feature selected



# Diagnostics for Individual Explanations

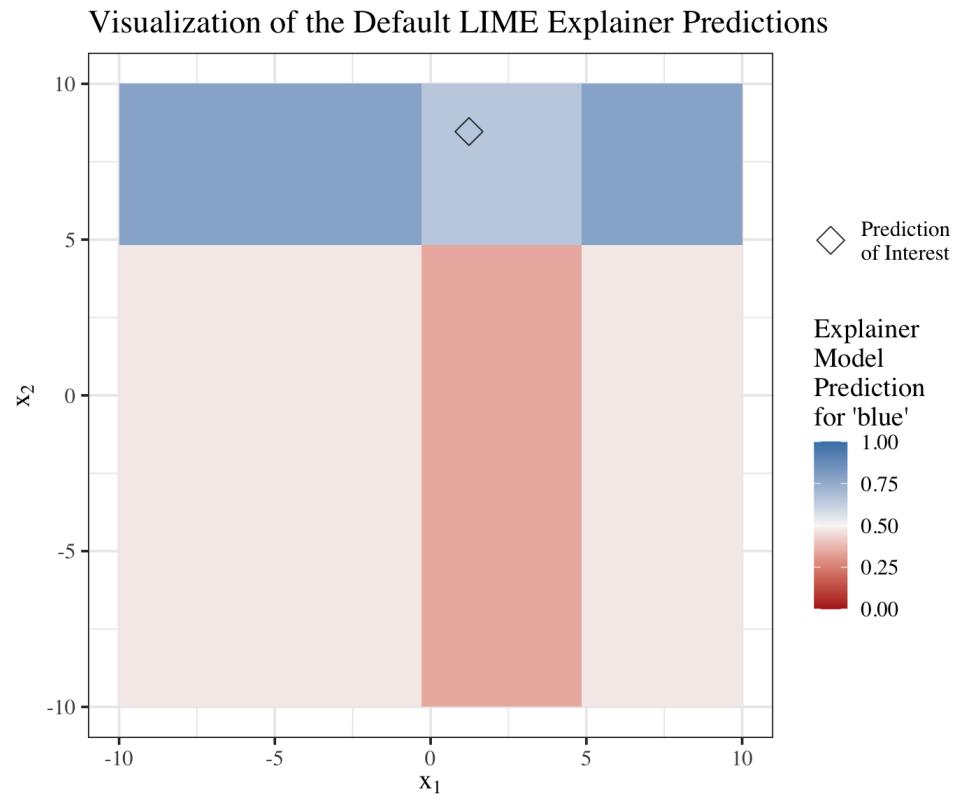
## Step 2c: Fit Explainer Model

- Ridge regression model fit to simulated data
  - Response: complex model predictions
  - Features: selected interpretability transformed features

### Corresponding Diagnostic:

#### *Explainer Model Prediction Plot*

- **Axes**: simulated data features
- **Diamond**: prediction of interest
- **Rectangle**: Interpretability transformed regions
- **Color**: explainer model prediction



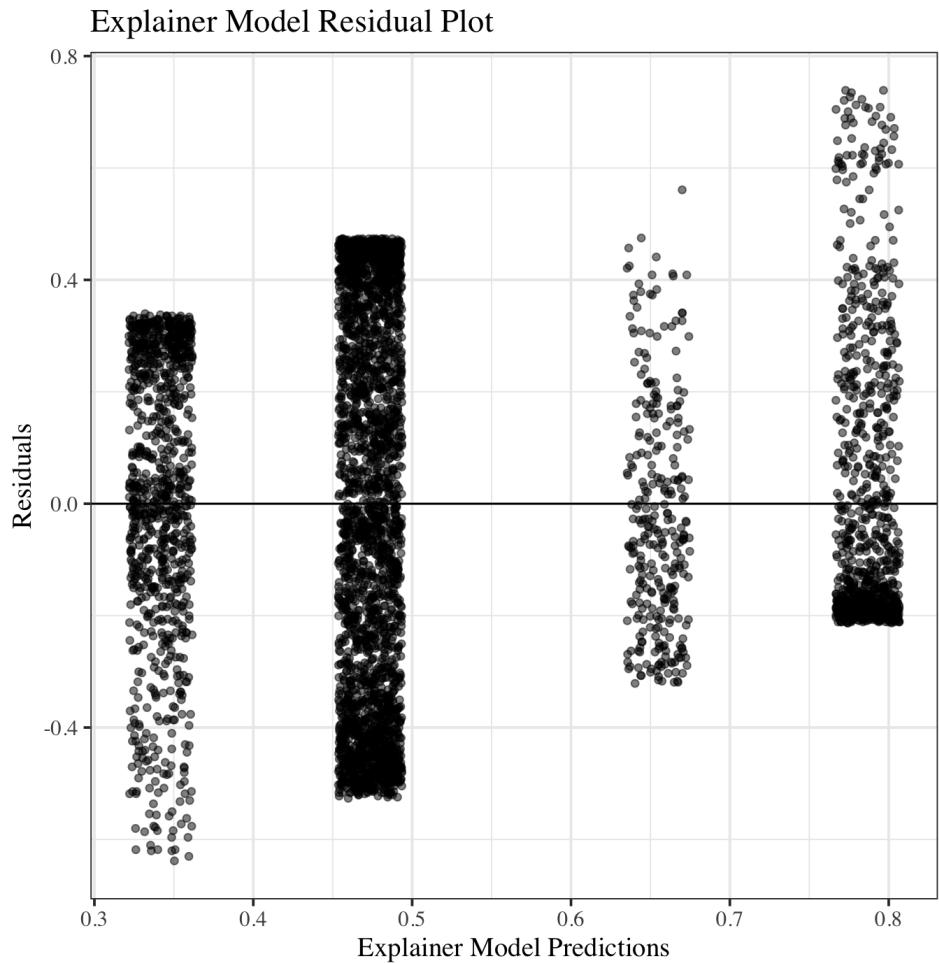
# Diagnostics for Individual Explanations

Step 2c: Fit Explainer Model (continued)

Corresponding Diagnostic:

*Explainer Model Residual Plot*

- Residual plot for the explainer model
- Points have been jittered in the x-direction



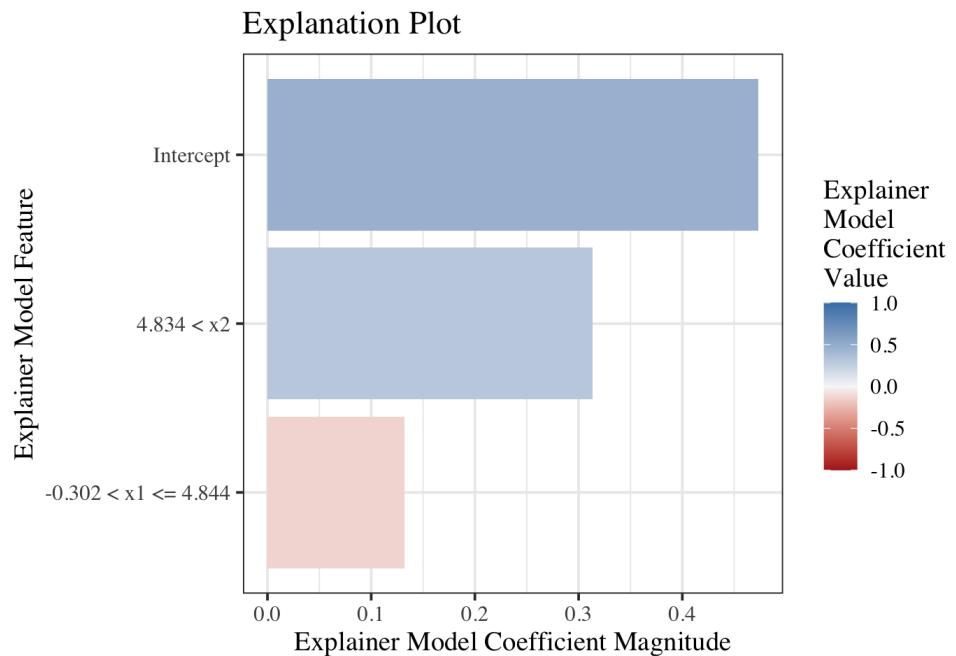
# Diagnostics for Individual Explanations

## Step 3a: Explainer Model Interpretation

- Interpret coefficients of explainer model
- Explains complex model prediction

### Corresponding Diagnostic: *Explanation Plot*

- Adaptation of plot from *lime* R package
- **Axes:** Explainer model (interpretability transformed) feature versus explainer model coefficient
- **Bar Length:** magnitude of explainer model coefficients
- **Bar Color:** value of the explainer model coefficients

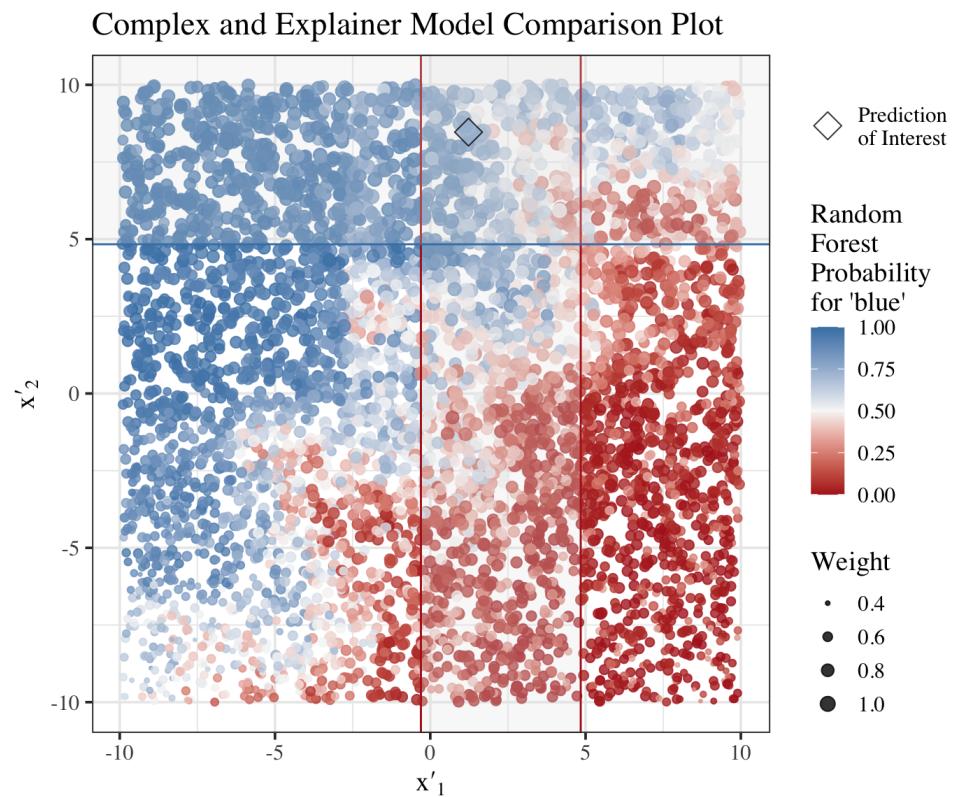


# Diagnostics for Individual Explanations

## Step 3b: Explainer Model Interpretation (continued)

### Corresponding Diagnostic: *Complex and Explainer Model Comparison Plot*

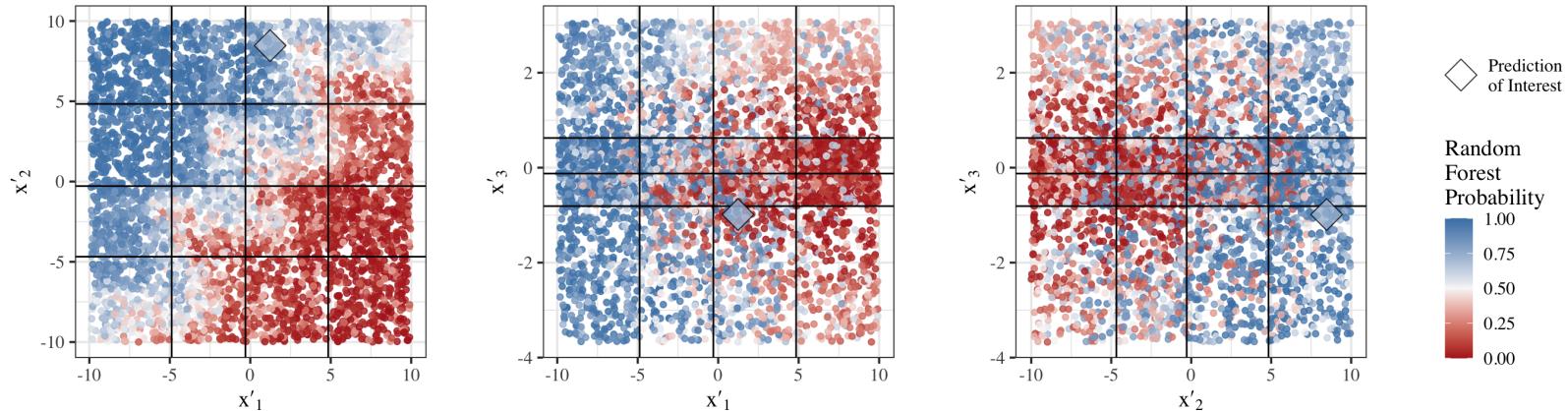
- **Axes:** simulated data features
- **Points:** simulated data
- **Diamond:** prediction of interest
- **Point Color:** random forest prediction (for 'blue')
- **Point Size:** proximity weight
- **Lines:** interpretability transformation boundaries
- **Line Color:** explainer coefficient supports a random forest prediction of 'blue' (blue) or not (red)



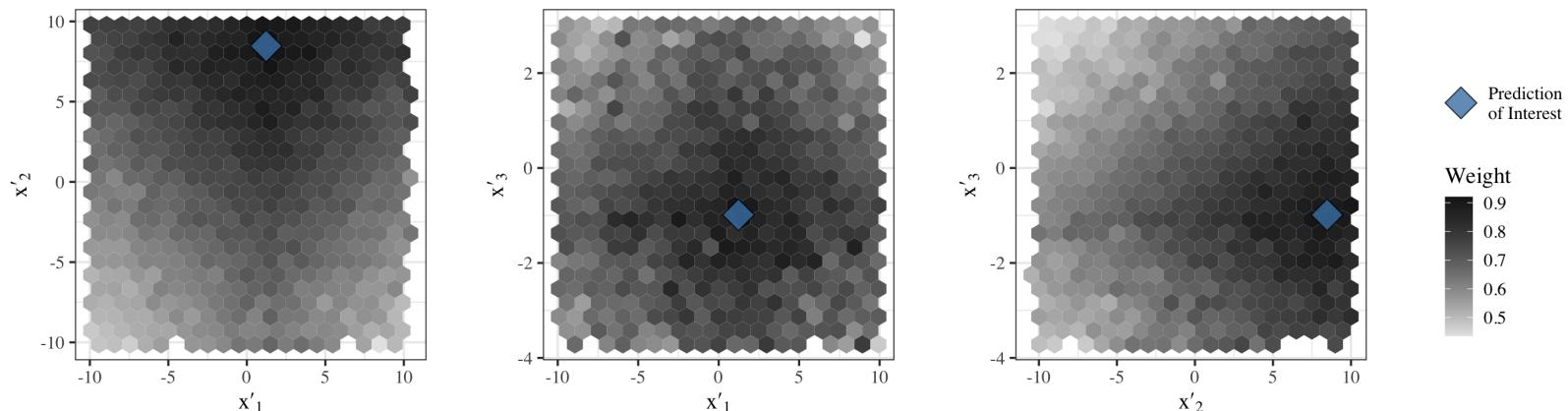
# Diagnostics for Individual Explanations

## Extensions to Higher Dimensions

Untransformed Simulated Data Plots



Proximity Weights Plots



# Set 2 of Visualizations: Diagnostics for Sets of Explanations

# Diagnostics for Sets of Explanations

Applied LIME using *lime* R package (Pedersen and Benesty, 2020)

- All test data observations
- Number of features to return in explanation: 2
- Default tuning parameters settings

Plot of all explanations from *lime* R package:

# Diagnostics for Sets of Explanations

## Explanation Set Plot

Provides an overview of groupings of LIME explanations

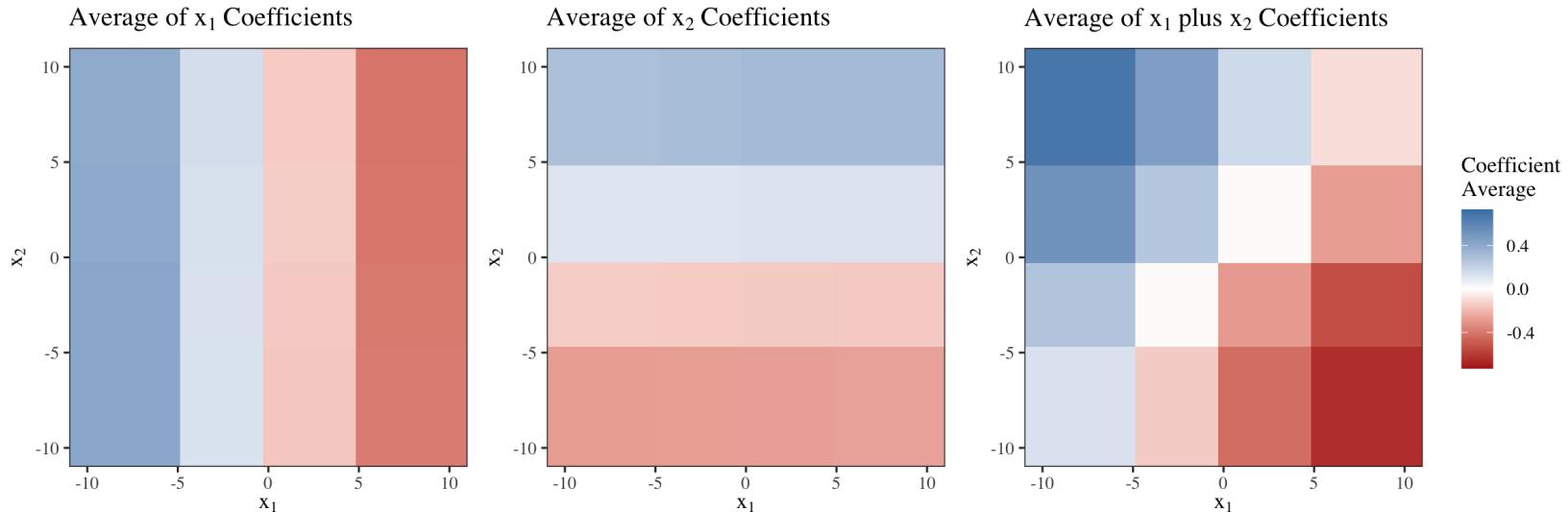
- Adaptation to the *lime* plot of all explanations
- **Y-axis:** Interpretability transformed features
- **X-axis:** Observation in the data set
- **Tile Color:** Ridge regression coefficient from the corresponding model

# Diagnostics for Sets of Explanations

## Average Coefficient Plots

Provides a summary of explainer model coefficients across the set of explanations within quantile bin regions

- **Axes:** Test data features
- **Cells:** Intersections of quantile bins
- **Color:** Average of ridge regression coefficients (or sum of ridge regression coefficients) for observations within a cell

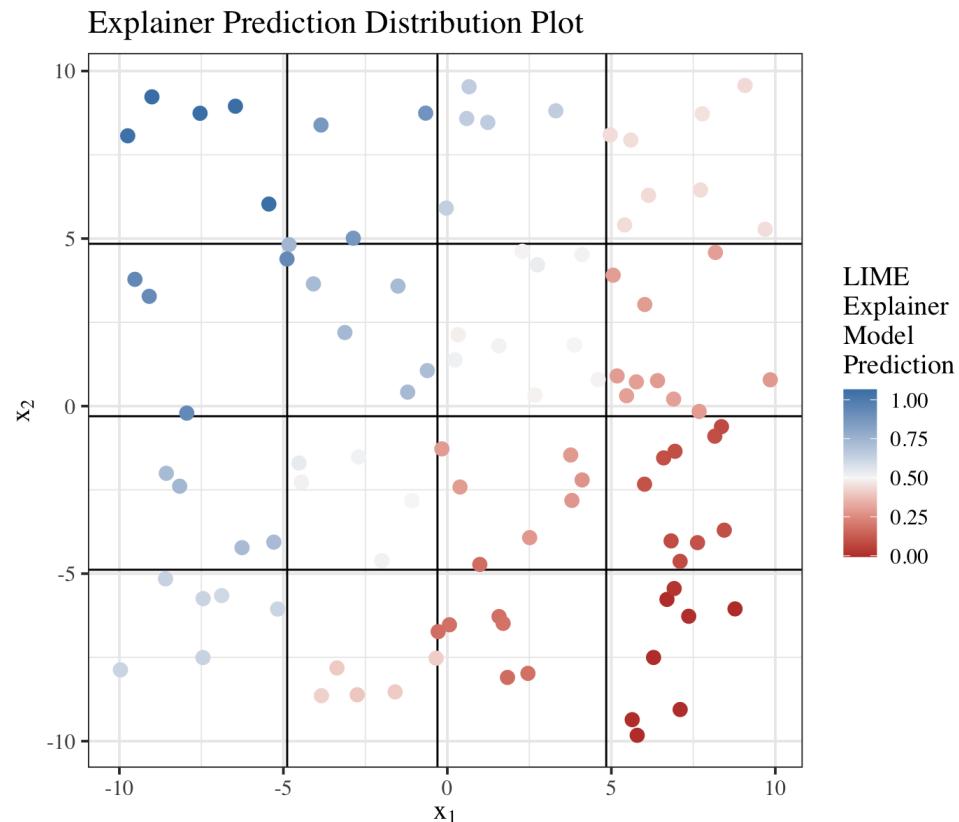


# Diagnostics for Sets of Explanations

## Explainer Prediction Distribution Plot

Shows the relationship between the explainer model predictions and the quantile bins

- **Axes:** Test data features
- **Points:** Test data observations
- **Color:** Explainer model prediction
- **Lines:** Quantile bin boundaries

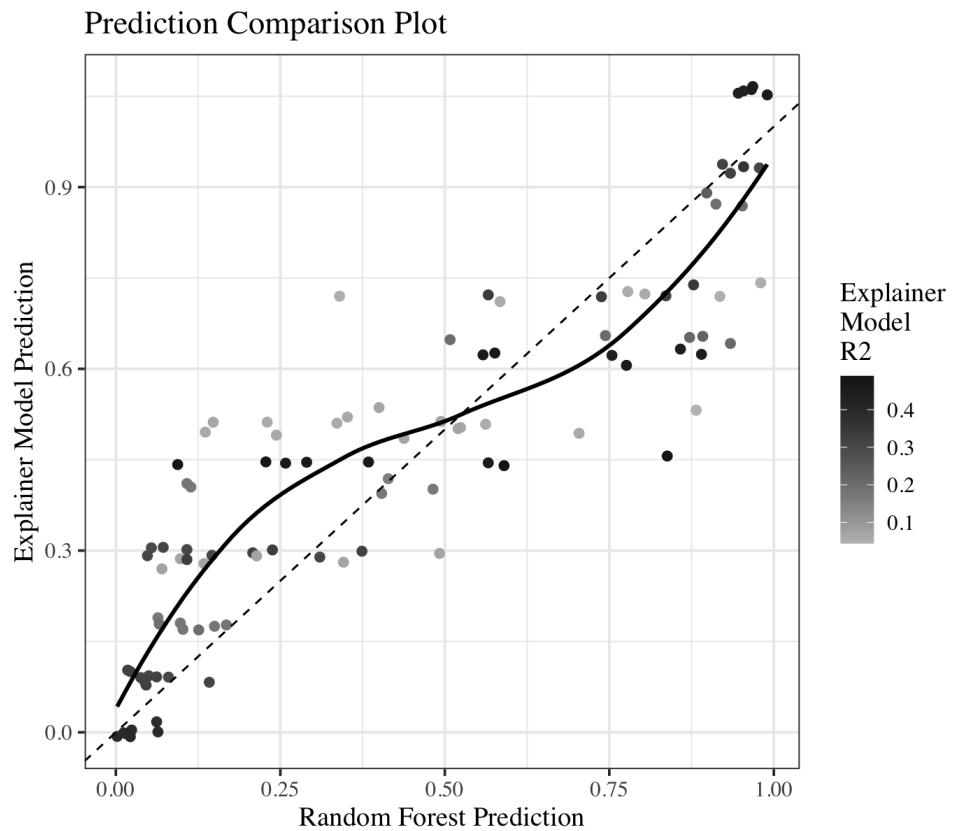


# Diagnostics for Sets of Explanations

## Prediction Comparison Plot

Shows the relationship between the explainer model and complex model predictions

- Y-Axis: Explainer model predictions
- X-Axis: Complex model predictions
- Points: Observations from the test data
- Color: Corresponding explainer model  $R^2$
- Dashed Line: 1-1 line
- .red.[Solid Line]: Loess smoother fit to the points

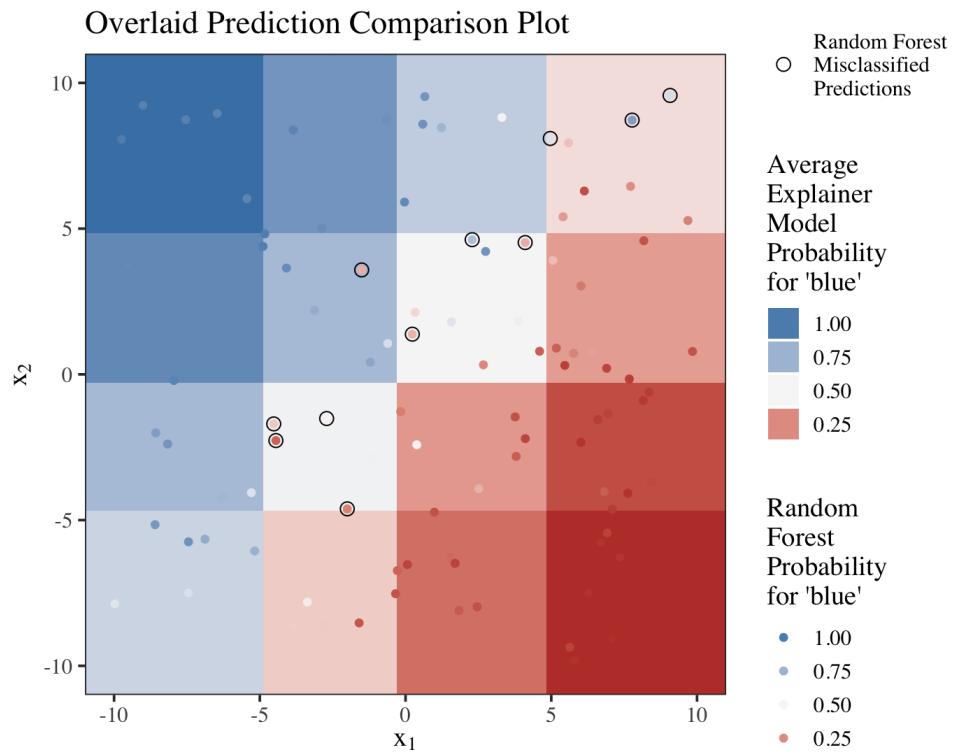


# Diagnostics for Sets of Explanations

## Overlaid Prediction Comparison Plot

Shows the relationship between the average explainer model predictions within a quantile bins cell and the complex model predictions

- **Axes:** Test data features
- **Points:** Test data observations
- **Point Color:** Explainer model prediction
- **Cells:** Intersections of quantile bins
- **Cell Color:** Average explainer model prediction
- **Black Circles:** Identify observations misclassified by the complex model



# Set 3 of Visualizations: Diagnostics for Comparisons of Tuning Parameters

# Diagnostics for Comparisons of Tuning Parameters

Applied LIME using *lime* R package (Pedersen and Benesty, 2020)

- All test data observations
- Number of features to return in explanation: 2
- Multiple applications of LIME for each observation
  - 2 to 6 quantile bins
  - Kernel density simulation

Plot of all explanations from *lime* R package:

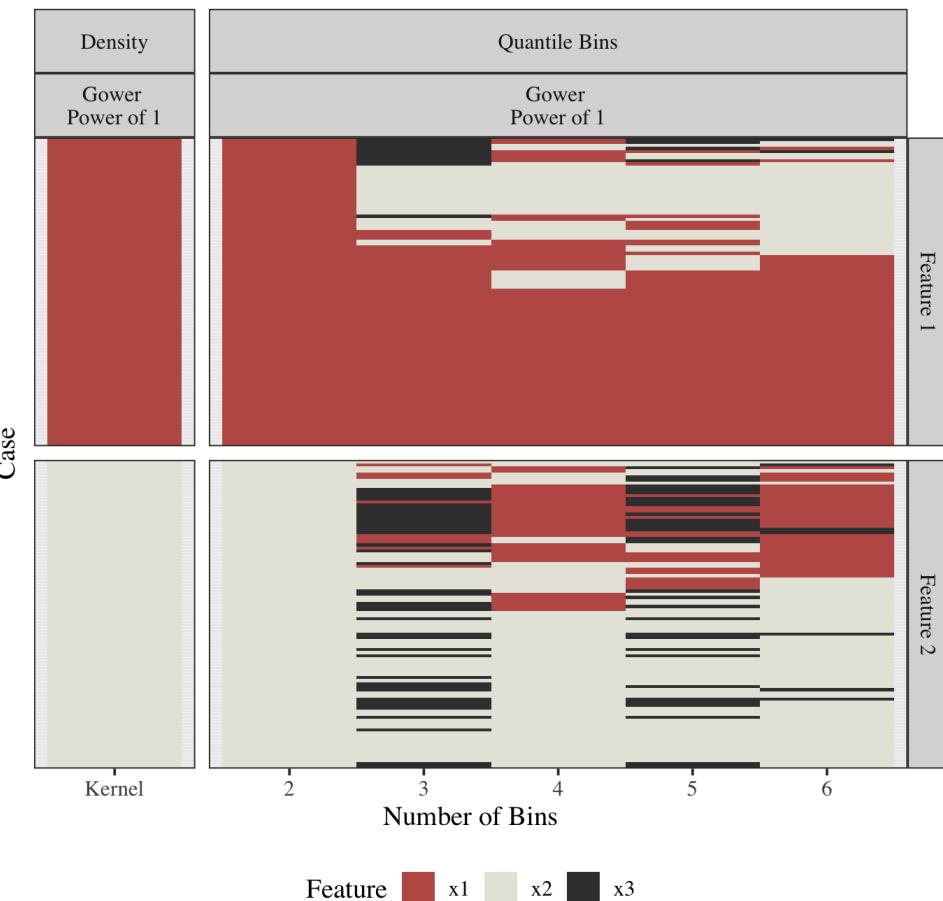
# Diagnostics for Comparisons of Tuning Parameters

## Feature Heatmap Plot

Provides an overview of the features selected by LIME across observations and tuning parameters

- **Y-Axis:** Test data observation
- **X-Axis:** Data simulation method
- **Y-Facet:** LIME feature order selection (first, second, etc.)
- **X-Facet:** Density based or bin based simulation method
- **Color:** Feature selected

## Feature Heatmap Plot



# Diagnostics for Comparisons of Tuning Parameters

## Metrics for the Assessment of LIME

Define the following:

**Average Fidelity** (Ribeiro, Singh, and Guestrin, 2016):

**Average  $R^2$** :

**MSEE (Mean Squared Explanation Error)**:

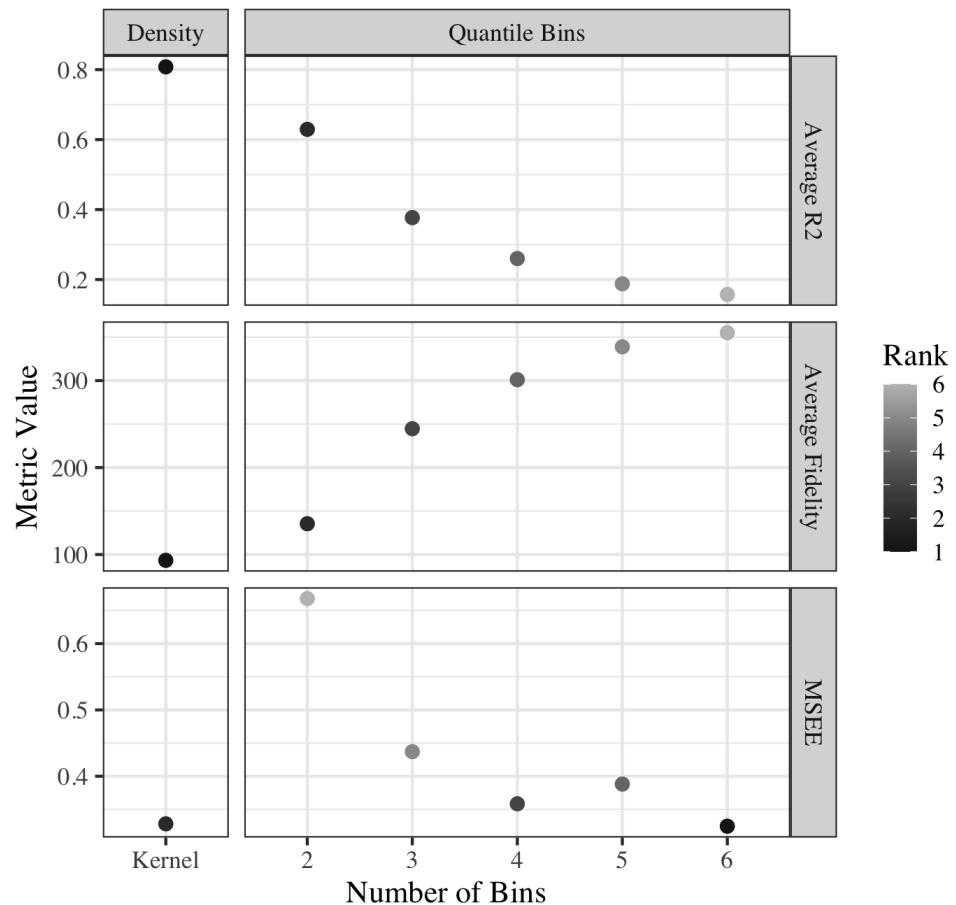
# Diagnostics for Comparisons of Tuning Parameters

## Assessment Metric Plot

Visualization of assessment metrics for comparing tuning parameter performance

- **Y-Axis:** Metric value
- **X-Axis:** Data simulation method
- **Y-Facet:** Metric type
- **X-Facet:** Density based or bin based simulation method
- **Point:** Represents application of LIME to a set of observations
- **Color:** Rank of the application based on metric value (within a metric)

## Assessment Metric Plot



# R package: limeaid

# Application of Visual Diagnostics to Bullet

# Chapter 2: Explaining Random Forests using Clustering of Trees

# Motivation

- Since LIME did not provide reasonable explanations for the bullet matching random forest from the study by (Hare, Hofmann, and Carriquiry, 2016), it is still of interest to visualize the random forest to gain insight into the predictions.
- It is particularly of interest to provide explanations for incorrect predictions.

# Current Ideas

## Classification Tree as a Global Explainer Model

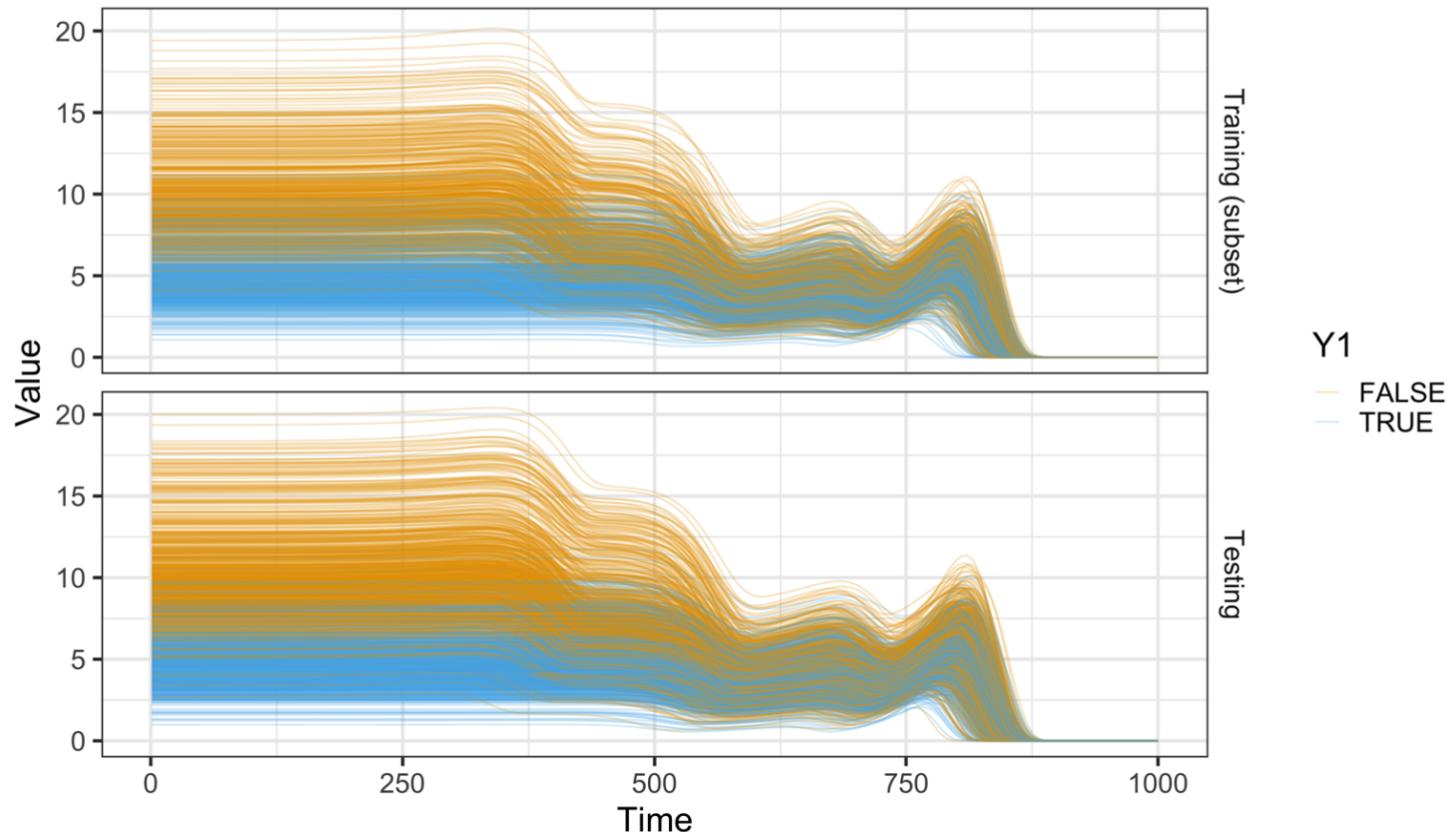
# Current Ideas

Clustering to Identify Key Tree Paths in the Random Forest

# Chapter 3: Extensions of Neural Network Explanation Tools to Functional Data

# Application from Sandia National Labs

Plots of the training (randomly selected subset) and testing data



Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

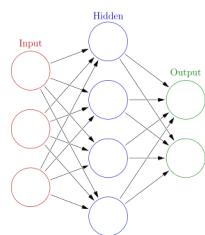
# Current Approach: Feature Visualization

## Concept

- Focus on a "location" in the neural network
- Determine example observation that maximize the activation function

## Process

1. Fit a model
2. Fix estimated parameter values
3. Determine values that maximized activation function at desired "location"



## Example

- Response:  $y \in \{0, 1\}$
- Features:

$$PC = (PC_1, PC_2, PC_3, PC_4, PC_5)$$

- Model: 1 hidden layer, 5 neurons
- Neuron  $i$  coefficients ( $i = 1, \dots, 5$ ):

$$(\beta_{0,i}, \beta_{1,i}, \beta_{2,i}, \beta_{3,i}, \beta_{4,i}, \beta_{5,i})$$

- Activation Function: logistic

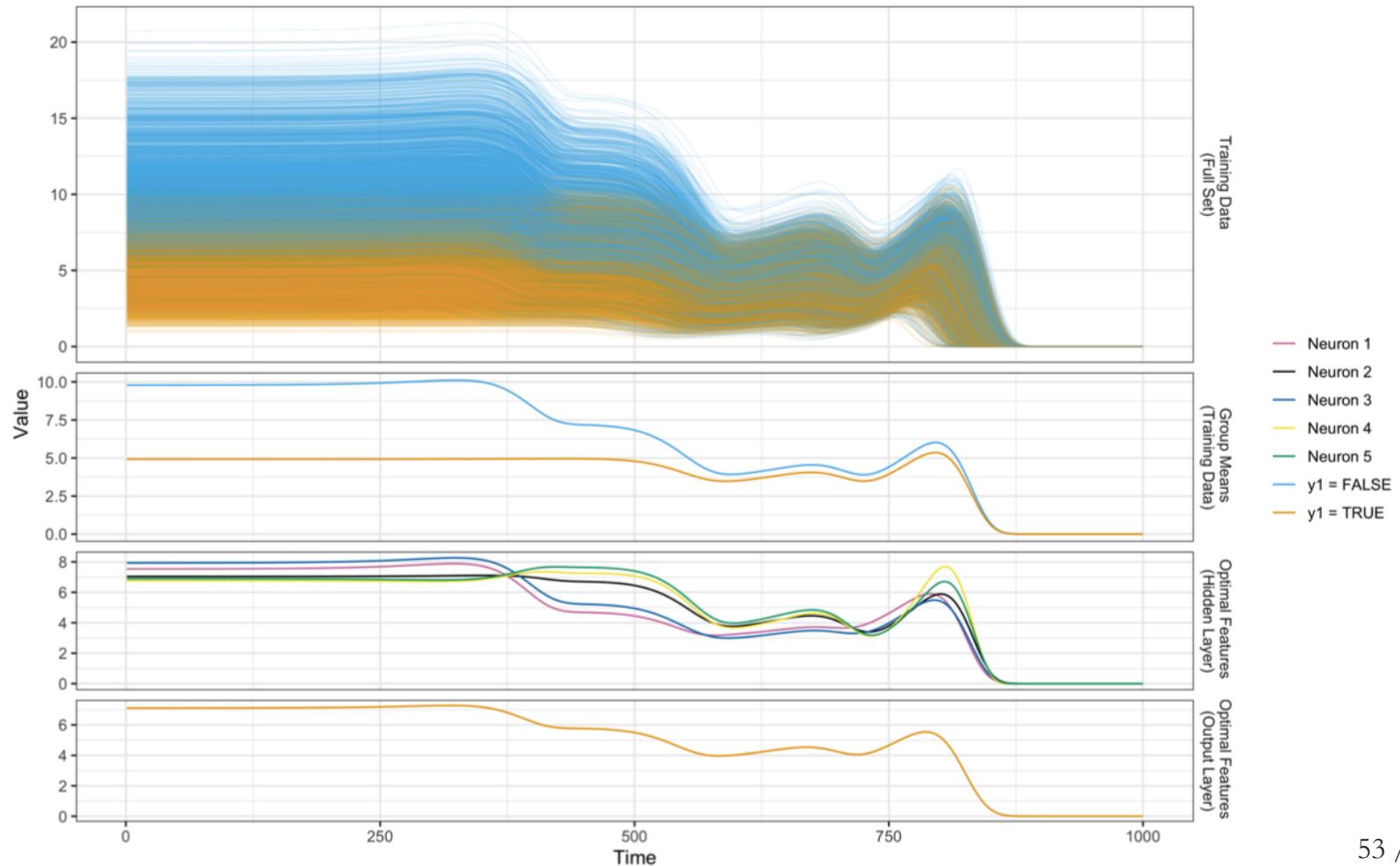
$$\sigma(v) = \frac{1}{1 + e^{-v}}$$

- Feature visualization optimization:

$$\arg \max_{PC} \sigma \left( \hat{\beta}_{0,i} + \hat{\beta}_{1,i} PC_1 + \dots + \hat{\beta}_{5,i} PC_5 \right)$$

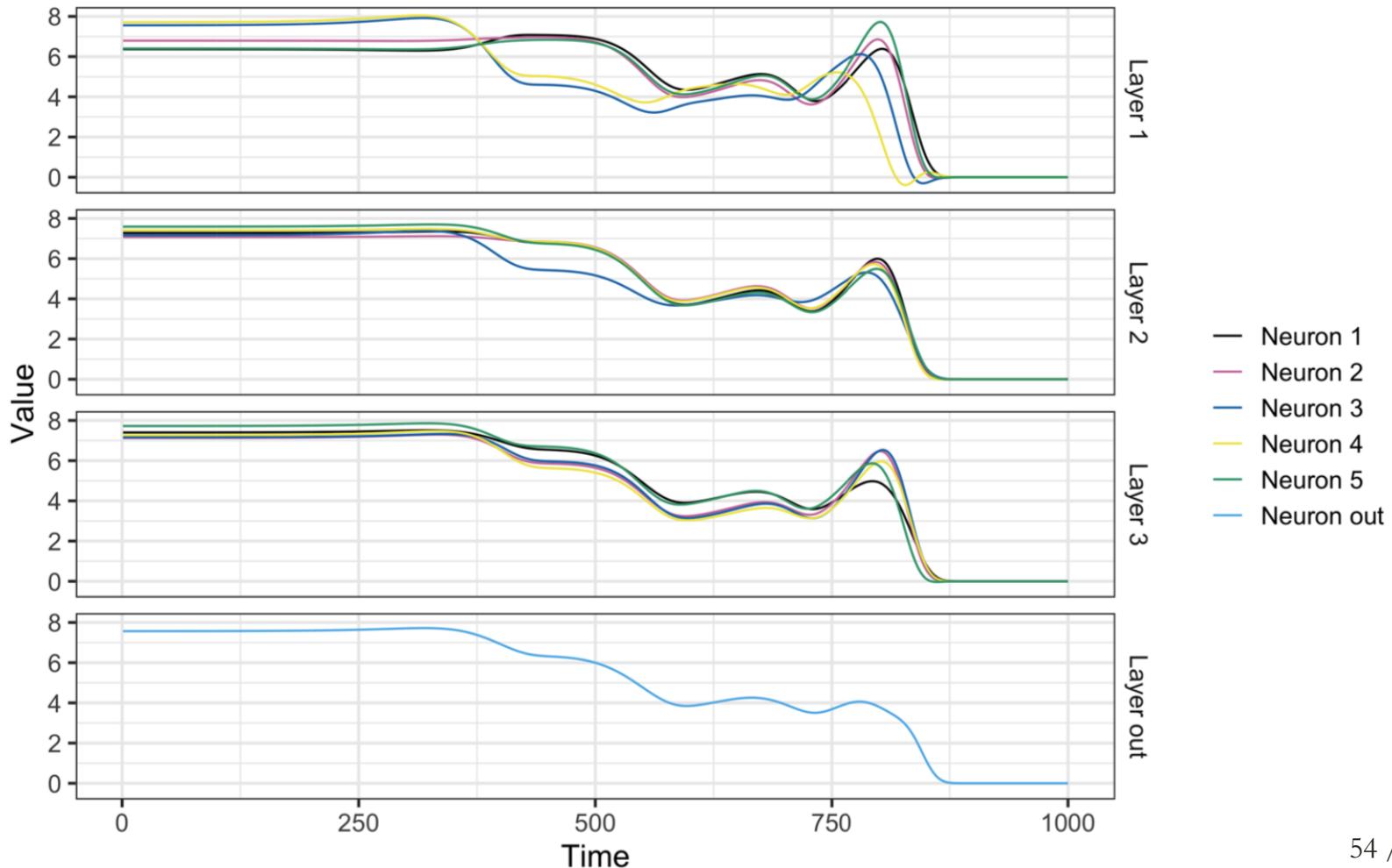
# Feature Visualization

Comparing back-transformed optimal features to the observed data



# Feature Visualization

Plots of the back transformed PCs that optimized the activation functions



# Ideas for Future Work

Feature visualization adjustments

- Visualize the functional principal components

Applications/extensions of other methods

- Permutation feature importance, saliency maps, and partial dependence plots

Visualizations of the paths of an observation through the network

- Example: flow (Halnaut, Giot, Bourqui, et al., 2020)

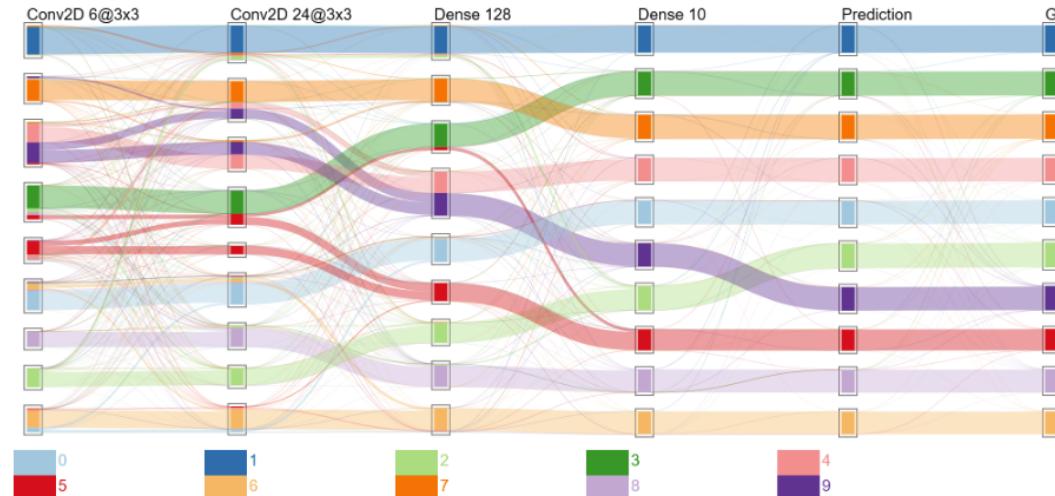


Figure 2: Visualization result on a LeNet5-inspired model evaluated on MNIST.

# Timeline for Completion

# Timeline for Completion

## Summer 2020

- Chapter 1: Finish writing (possibly submit paper)
- Chapter 3: Work on content (Sandia internship)

## Fall 2020

- Chapter 1: Submit paper (if not already done)
- Chapter 2: Work on content
- Chapter 3: Write up work

## Spring 2021

- Chapter 2: Write up work
- Defend

# Discussion Points

# Publication of Chapter 1

How to divide up the material from chapter 1 for publication?

Current ideas:

## Paper 1: Survey paper on LIME

- Explain LIME in a statistical context
- Use visualizations to help explain the procedure
- Use diagnostic visualizations to assess LIME
- Highlight issues with LIME
- Use iris and sine data

## Paper 2: Diagnostic plots for LIME

- Motivate assessment of LIME using the bullet matching data (example of high stakes decision using machine learning)
- Demonstrate issues found with LIME explanations using diagnostic plots
- LIME should not be trusted to explain machine learning models when making high stakes decisions

# References

Need to format this eventually using **start** and **end** options

Altmann, A, L. Toloşı, O. Sander, et al. (2010). "Permutation importance: a corrected feature importance measure". In: *Bioinformatics* 26.10, pp. 1340-1347. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq134.

Apley, D. W. and J. Zhu (2016). "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models".

Beckett, C. (2018). "Rfviz: An Interactive Visualization Package for Random Forests in R". In: *DigitalCommons@USU All Graduate Plan B and otherReports*.

Biggio, B. and F. Roli (2018). "Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning". In: *Pattern Recognition* 84, pp. 317-331. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2018.07.023.

Breiman, L. (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5-32. ISSN: 0885-6125. DOI: 10.1023/a:1010933404324.

Casalicchio, G, C. Molnar, and B. Bischl (2019). "Visualizing the Feature Importance for Black Box Models" , pp. 655-670. ISSN: 2190-5053. DOI: 10.1007/978-3-030-10925-7\_40.

Craven, M. W. and J. W. Shavlik (1996). "Extracting Tree-Structured Representations of Trained Networks". In: *Advances in Neural Information Processing Systems* 8.

Fisher, A, C. Rudin, and F. Dominici (2018). "Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the “Rashomon” Perspective".

Friedman, J. H. (2001). "Greedy function approximation: A gradient boosting machine." In: *The Annals of Statistics* 29.5. ISSN: 0090-5364. DOI: 10.1214/aos/1013203451.

# Appendix

Add application of diagnostics to bullet examples