



Interpretable Machine Learning

Do you know what your model is doing?

Marcel Spitzer



Mannheim, 15.05.2019



Marcel Spitzer

Big Data Scientist @ inovex

- Applied Mathematics, Data Science
- SW Engineering, Machine Learning
- Big Data, Hadoop, Spark

 mspitzer@inovex.de

 @mspitzer243





Interpretation
is the process of
giving explanations
to humans.

~ B. Kim, Google Brain, Interpretable
Machine Learning (ICML 2017)

“Interpretability is the degree to which an observer can understand the cause of a decision.”

~ Miller T., 2017, Explanation in AI: Insights from the Social Sciences

- humans **create** decision systems
- humans are **affected** by decisions
- humans **demand** for explanations



Bias towards Accuracy

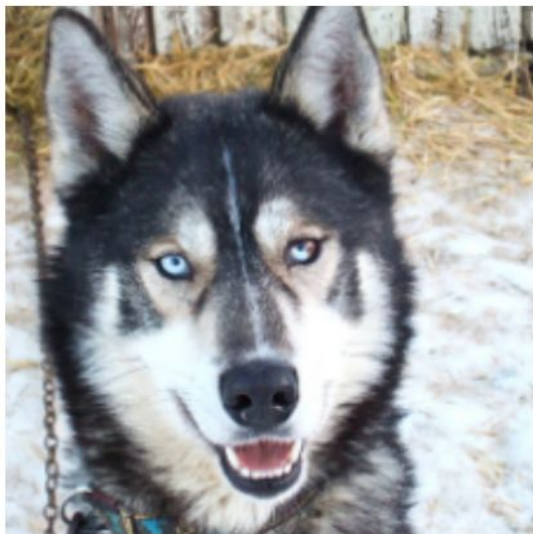
“The machine learning community focuses too much on predictive performance. But machine learning models are always a small part of a complex system.”

~ C. Molnar, 2019, One Model to Rule Them All

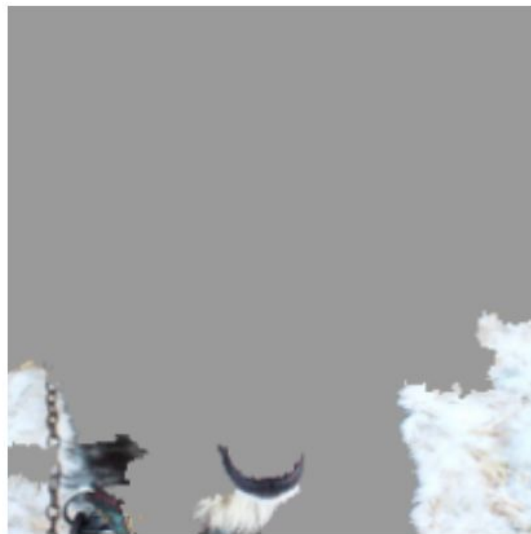
Do also consider asking yourself:

- Am I solving the **right problem**?
- How to make people **trust** my algorithm?
- Is there any **bias**? Are the training data representative?
- What is the **impact** in a real-world setting?





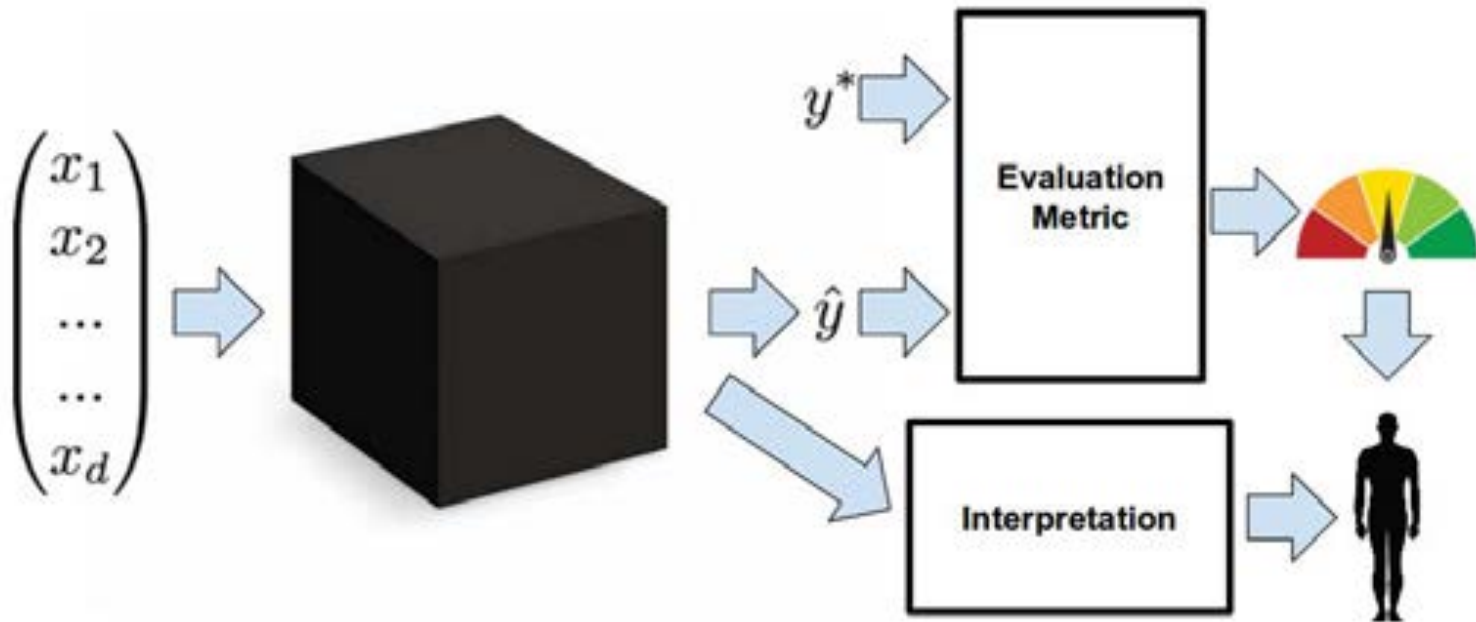
(a) Husky classified as wolf



(b) Explanation

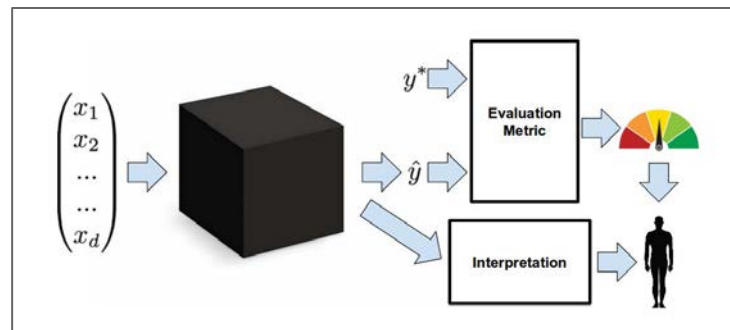
Figure 11: Raw data and explanation of a bad model’s prediction in the “Husky vs Wolf” task.

The additional need for interpretability



The additional need for interpretability

The decision process of a model should be **consistent to the domain knowledge** of an expert.



In particular, it ...

- should not encode bias
- should not pick up random correlation
- should not use leaked information

hard to capture in a single quantity

Use models that are **intrinsically interpretable** and known to be easy for humans to understand.

1

2 Train a black box model and apply **post-hoc interpretability techniques** to provide explanations.

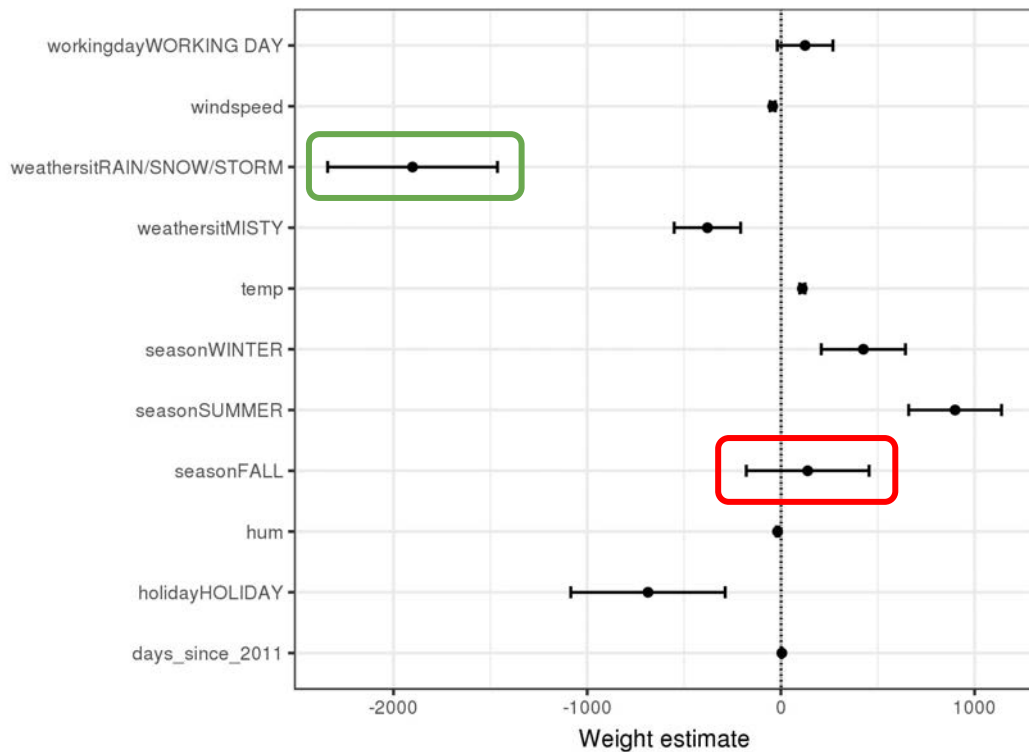
Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \epsilon$$

Linearity makes model easy to interpret

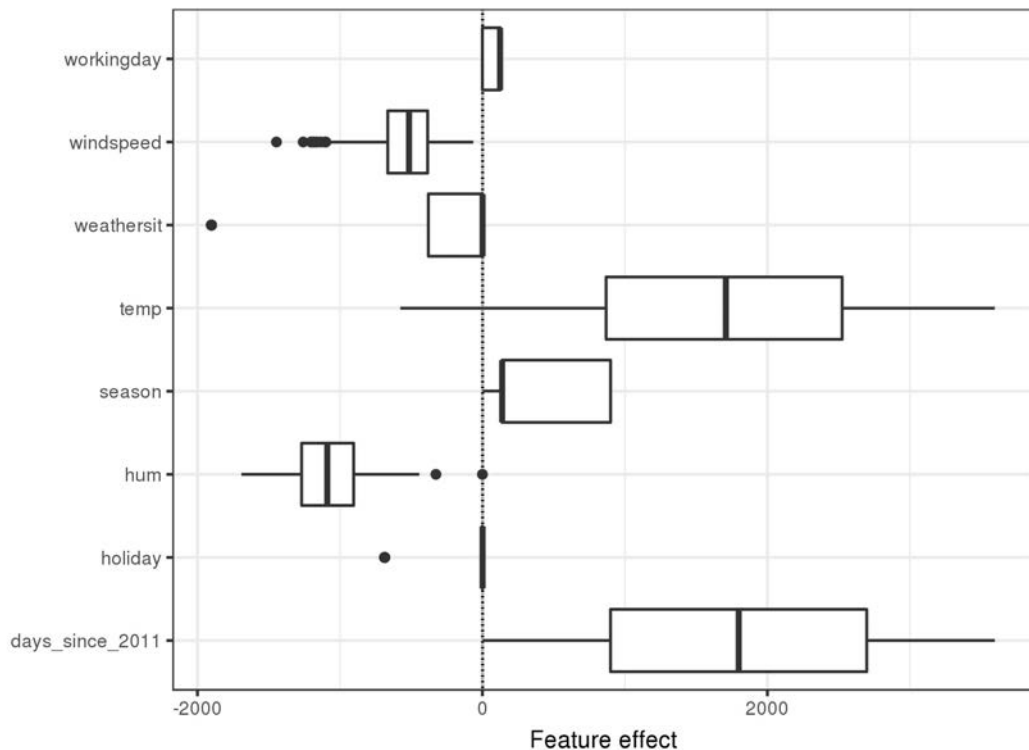
- **learned weights** can be used to explain feature effects
- predictions can be **decomposed** into individual attributions
- **confidence intervals** express uncertainty

Linear Regression: Model Internals



- Coefficients indicate direction of influence
- Confidence intervals express significance and uncertainty
- Not comparable across features unless standardized

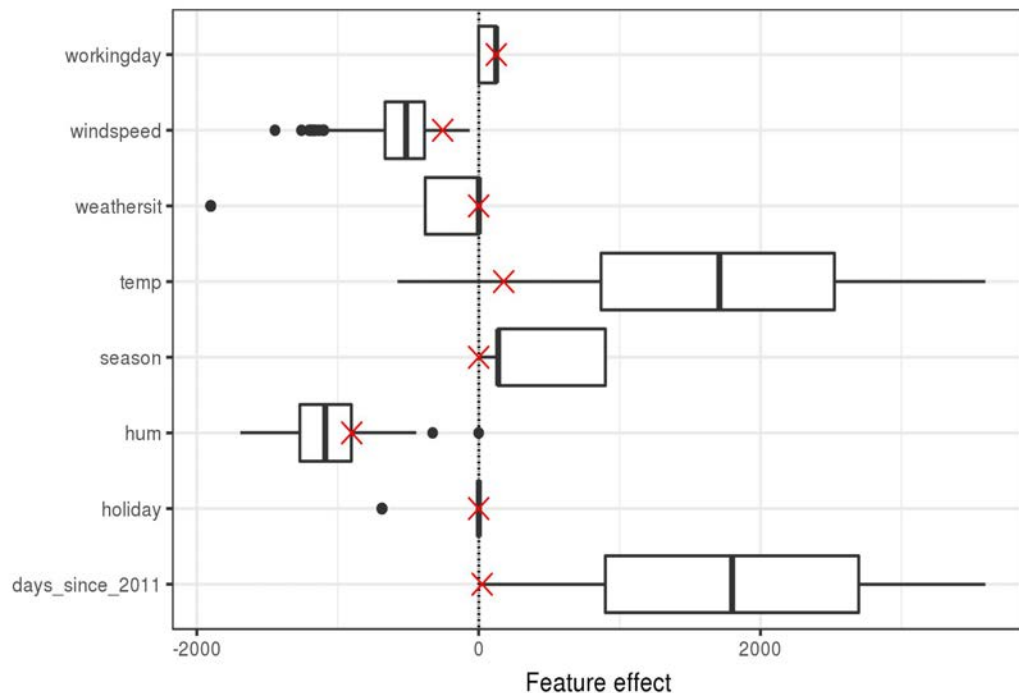
Linear Regression: Feature Effects



- Decompose predictions into individual feature attributions
- Scale-independent, hence comparable across features
- Enables us to draw conclusions about feature importance

Linear Regression: Feature Effects

Predicted value for instance: 1571
Average predicted value: 4504
Actual value: 1606



- Decompose predictions into individual feature attributions
- Scale-independent, hence comparable across features
- Enables us to draw conclusions about feature importance
- Individual effects comparable to overall distribution

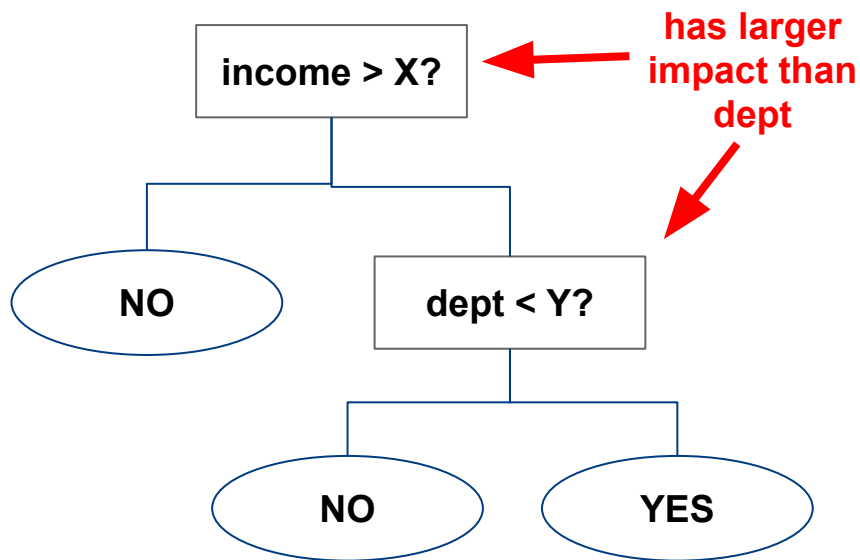
Decision Trees

$$\hat{y} = \hat{f}(x) = \sum_{m=1}^M c_m I\{x \in R_m\}$$

Intuitive decision process makes model easy to interpret

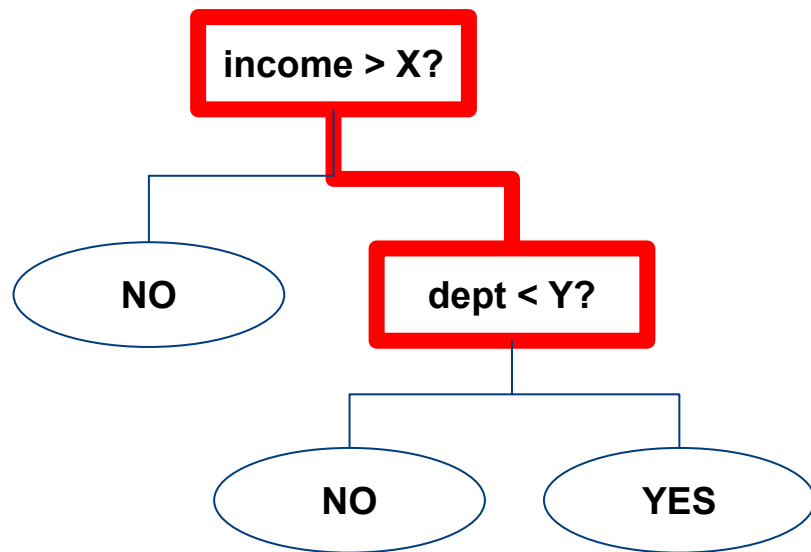
- captures **nonlinear** dependencies and interactions
- model is **simple** and **self-describing**
- **rule-based** prediction feels natural for humans

Decision Trees: Model Internals



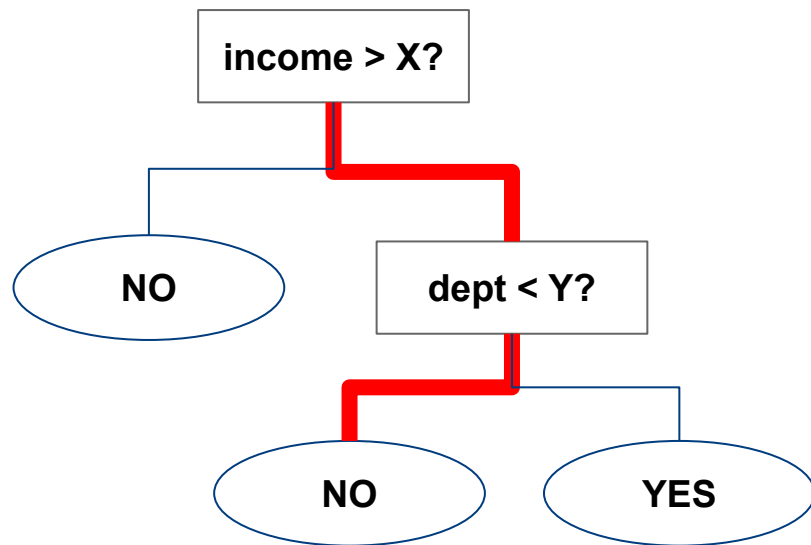
- Tree structure lets us assess feature importance

Decision Trees: Model Internals



- Tree structure lets us assess feature importance
- Feature interactions can be determined by following paths from root to leaf

Decision Trees: Individual Explanations



- Tree structure lets us assess feature importance
- Feature interactions can be determined by following paths from root to leaf
- To explain a particular prediction, just take a look at the path from root to leaf

Wrap Up: Desirable Properties

Intrinsically interpretable models *simplify* answering:

- Which features are **relevant**?
- How do they **influence** predictions?
- How do features **interact**?
- How **certain** is a prediction?

Both for the entire model as well as individual predictions

Wrap Up: Desirable Properties

Intrinsically interpretable

*How to answer these questions
if models were black boxes?*

- Which features are **relevant**?
- How do they **influence** predictions?
- How do features **interact**?
- How **certain** is a prediction?

Both for the entire model as well as individual predictions

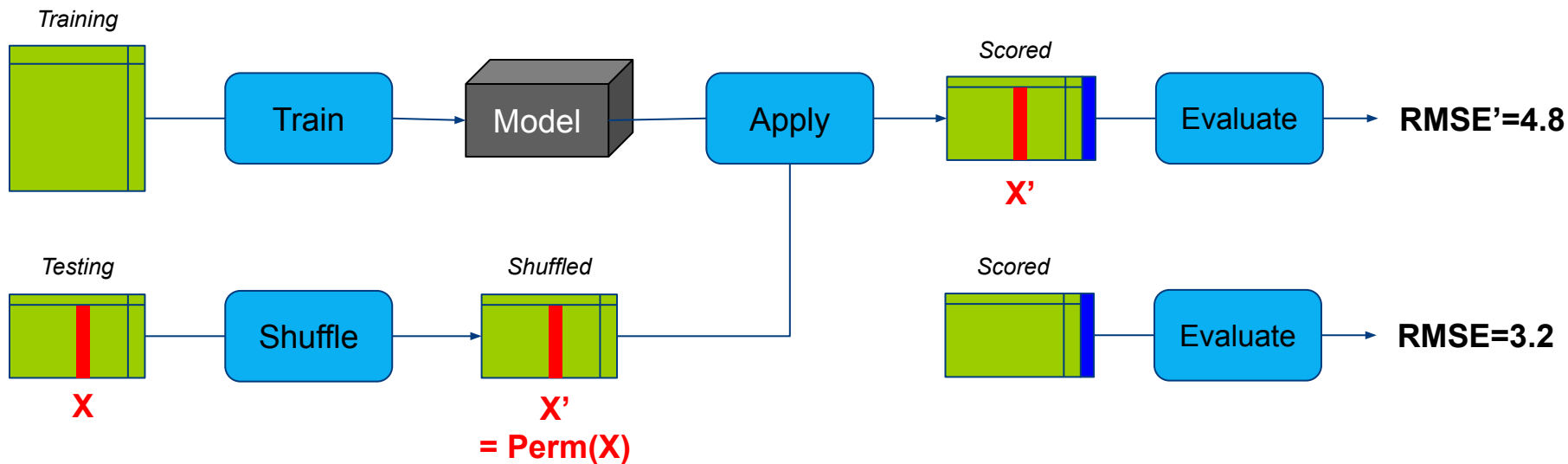


Use models that are **intrinsically interpretable** and known to be easy for humans to understand.

1

2 Train a black box model and apply **post-hoc interpretability techniques** to provide explanations.

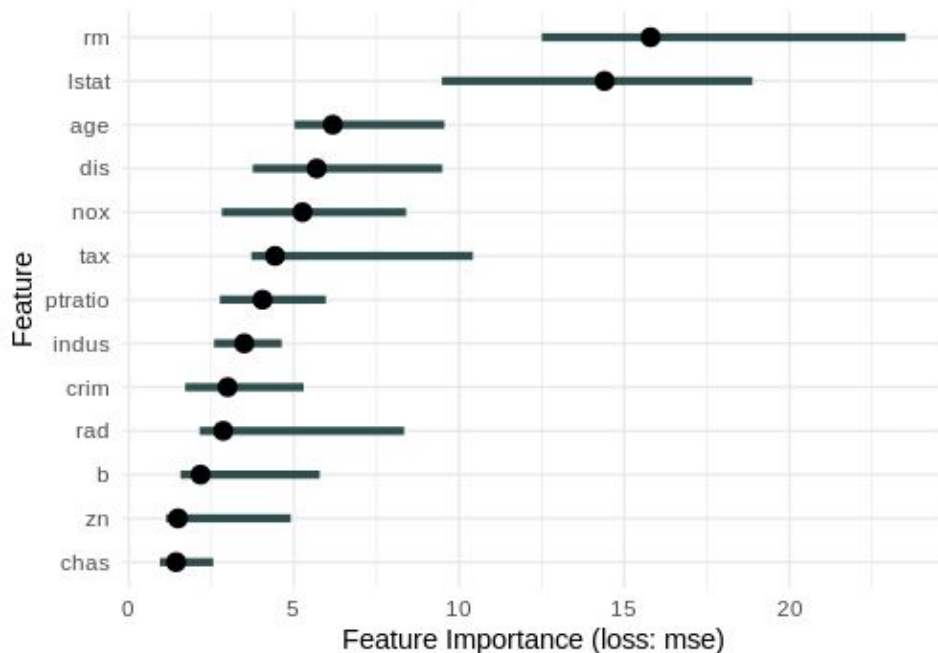
Feature Shuffling



**For every Feature X:
Repeat this process N times**

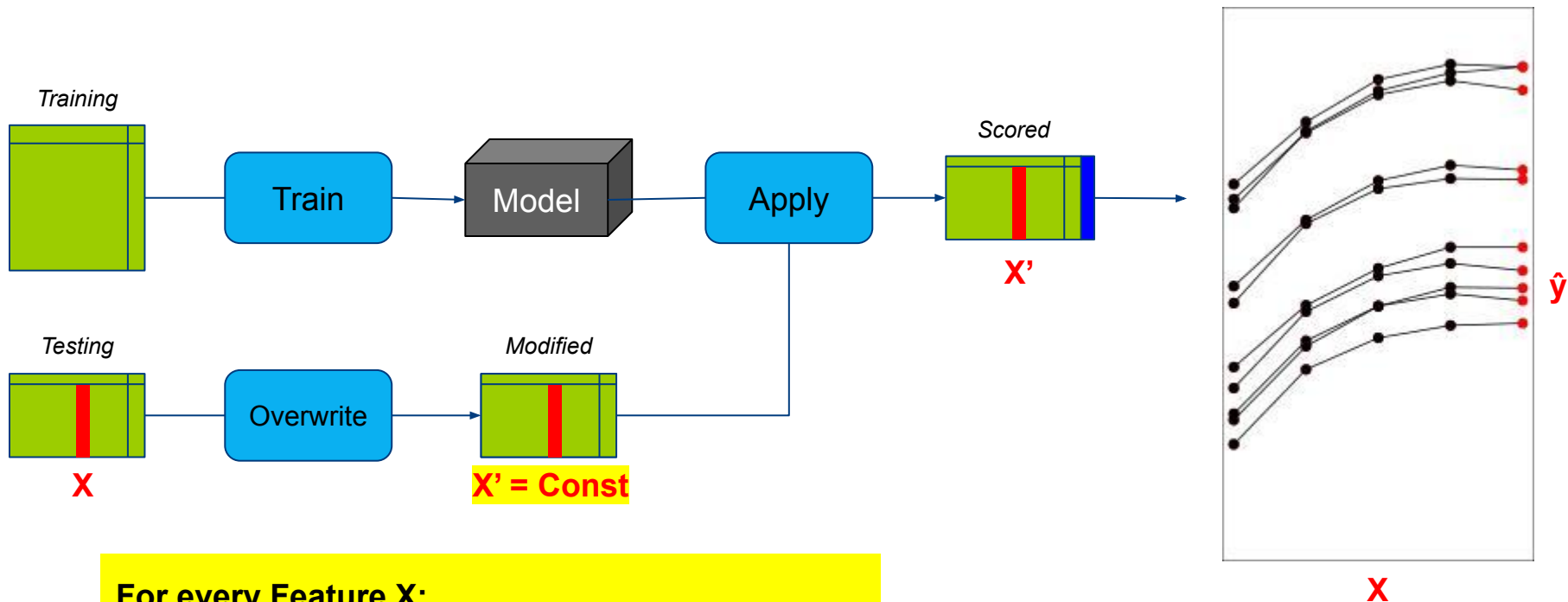
$$\text{Degradation}(\text{RMSE}, X) = 4.8 - 3.2 = 1.6$$

Feature Shuffling



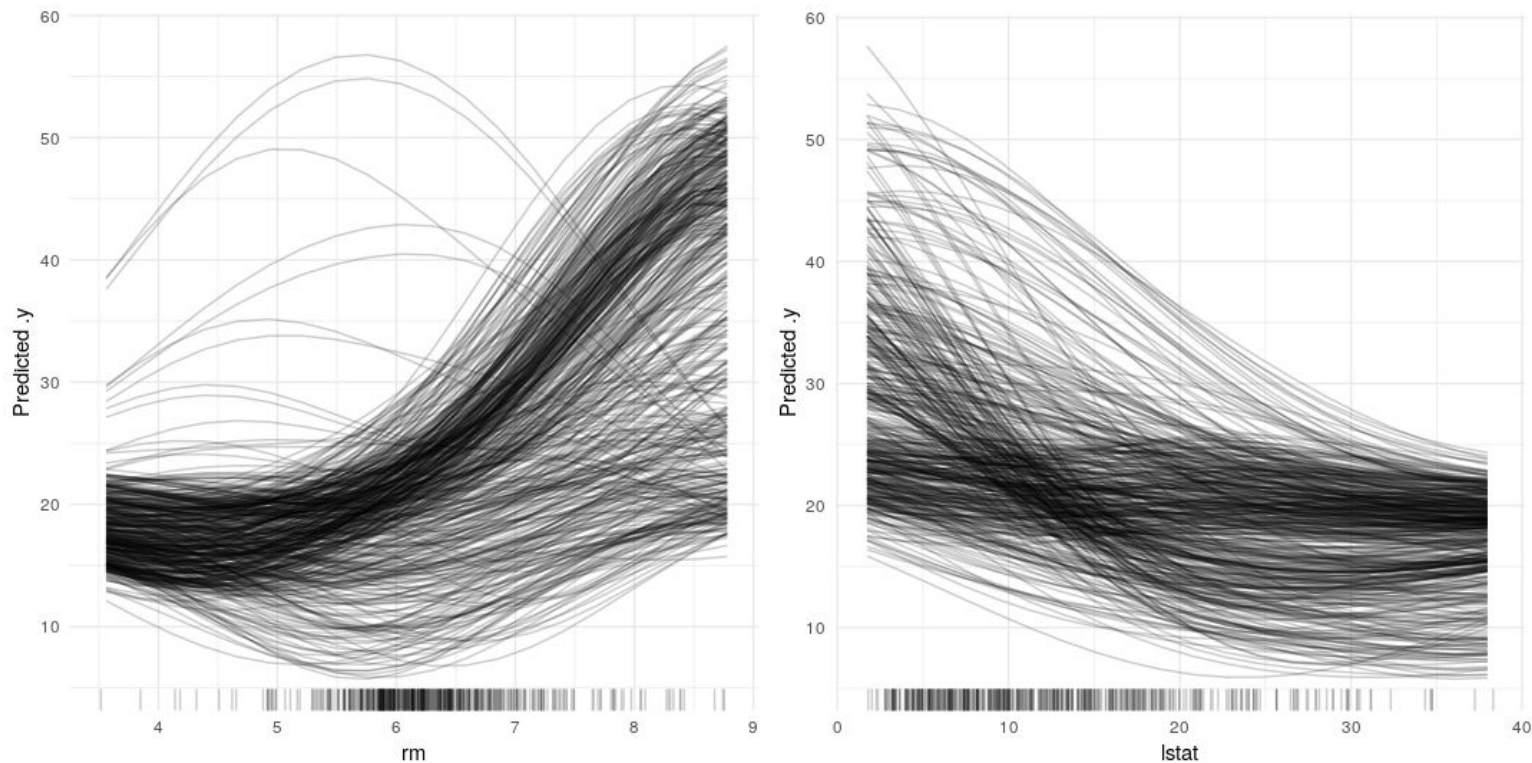
- Estimates Feature importance by averaging degradation
- Tied to certain loss function
- Not applicable in high dimensional domains

Individual Conditional Expectation (ICE)

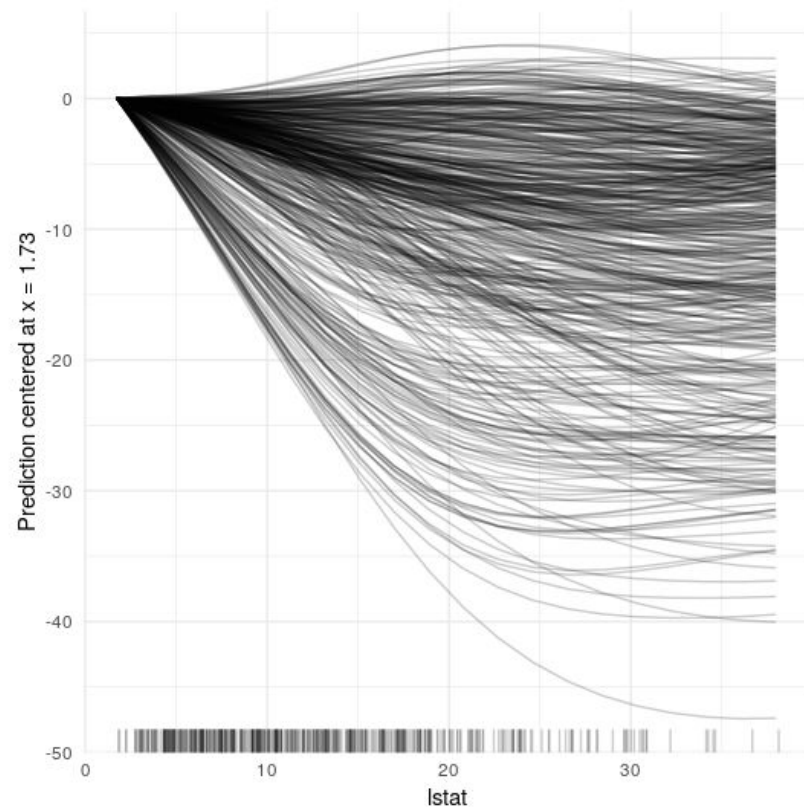
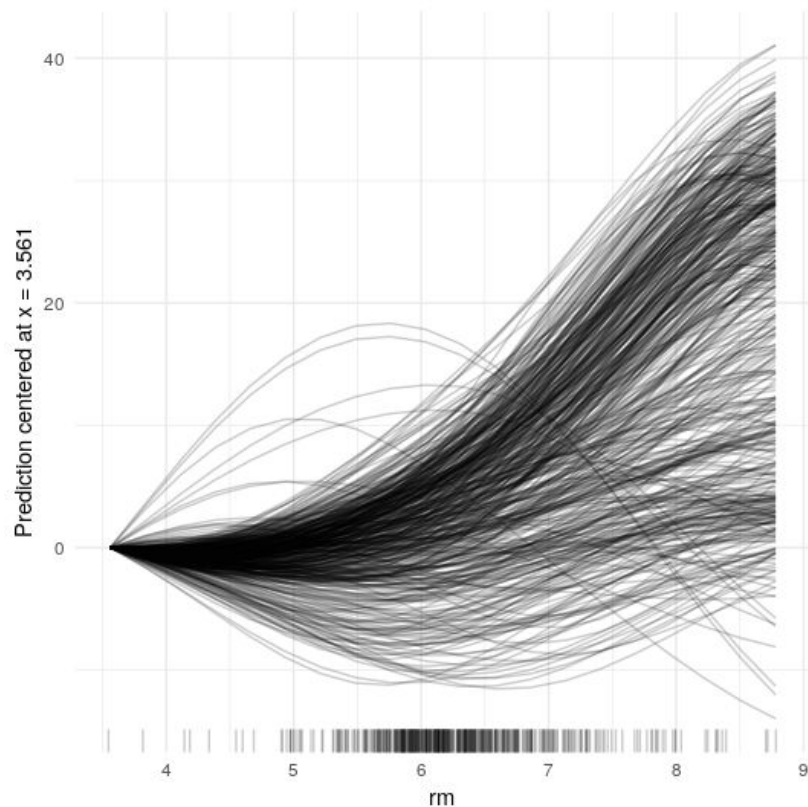


**For every Feature X:
Repeat this process using some constants C_x**

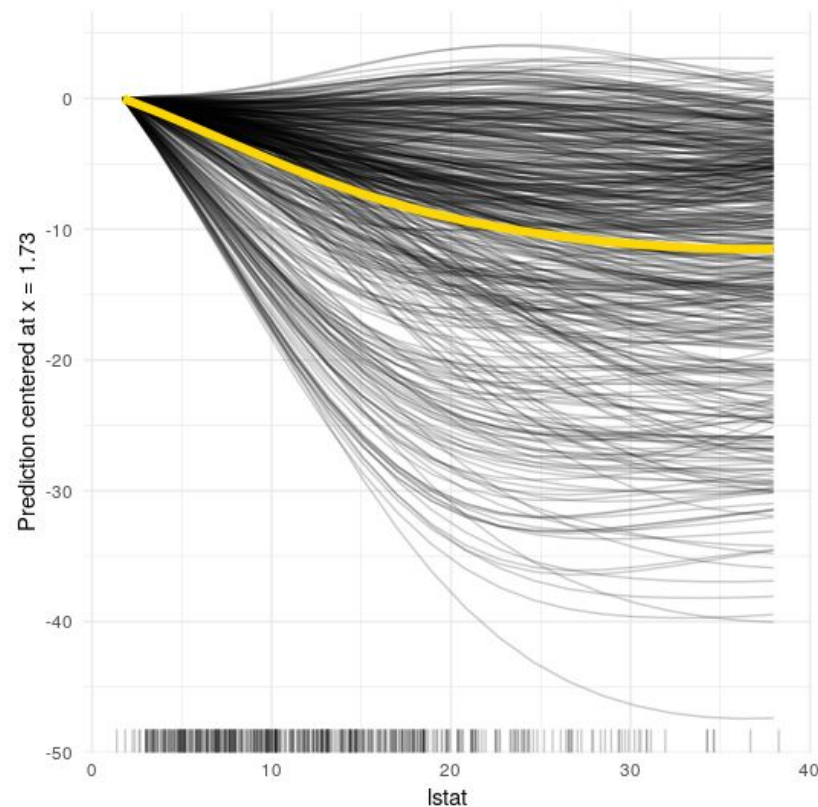
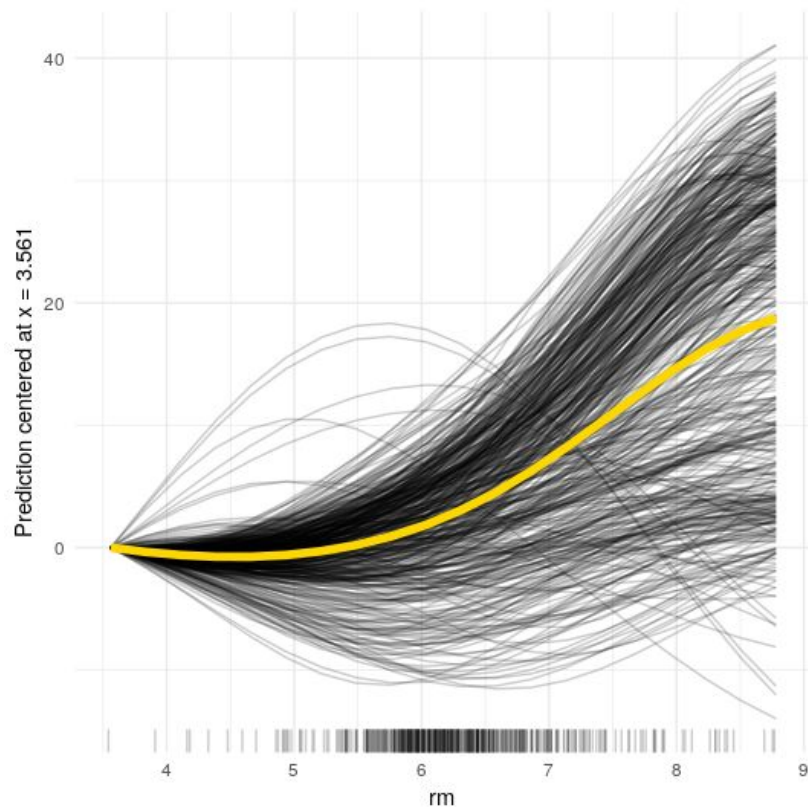
Individual Conditional Expectation (ICE)



Individual Conditional Expectation (ICE)



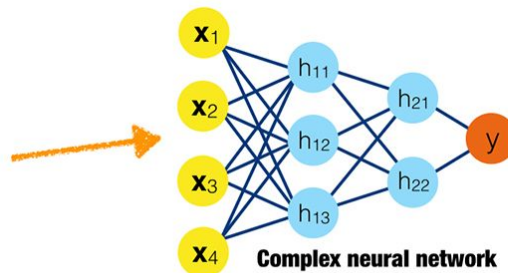
Individual Conditional Expectation (ICE)



Global Surrogate Models

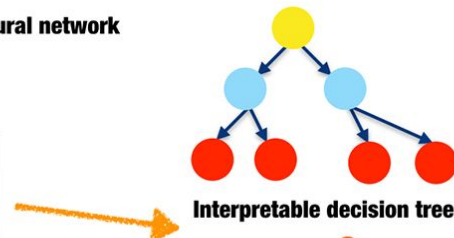
BAD	CUSTOMER_DTI	LOAN_PURPOSE	CHANNEL
0	0.18	MORT	7
1	0.42	HELOC	10
0	0.11	MORT	10
0	0.21	MORT	1

1. Train a complex machine learning model

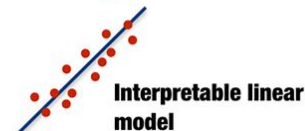


BAD	PREDICTED_BAD	CUSTOMER_DTI	LOAN_PURPOSE	CHANNEL
0	0.47	0.18	MORT	7
1	0.82	0.42	HELOC	10
0	0.18	0.11	MORT	10
0	0.12	0.21	MORT	1

2. Train an interpretable model on the original inputs and the predicted target values of the complex model

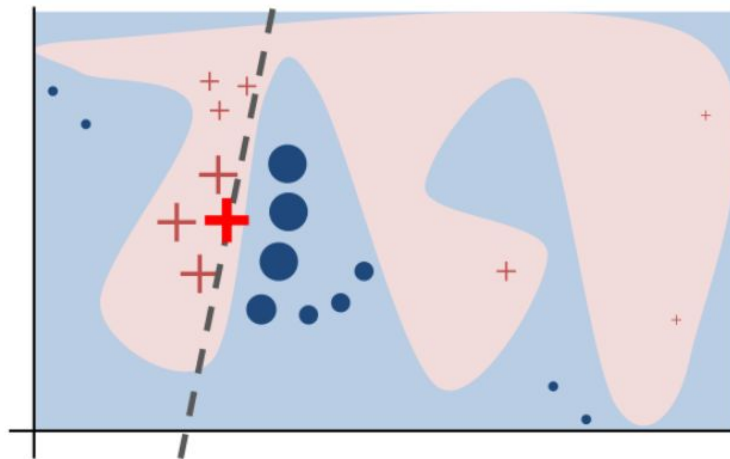


Or



Local Surrogate Models (LIME)

- Feeds original model with small variations of instance to be explained
- Sampled instances are weighted by proximity to the instance of interest
- Interpretable models are fit locally on observed outcome



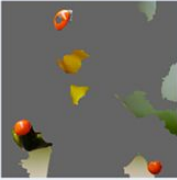





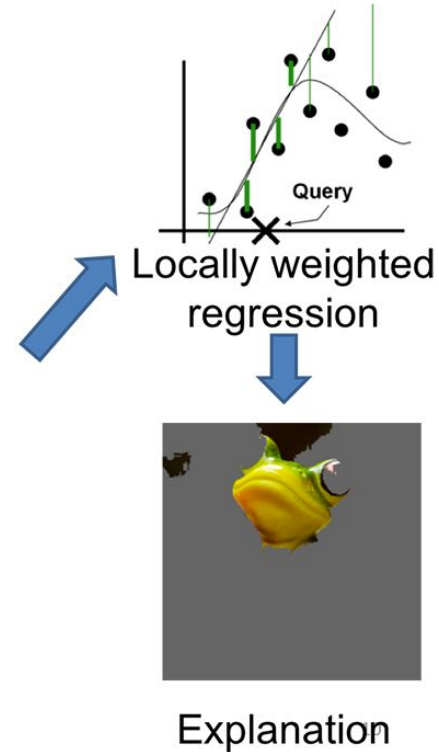
Local Surrogate Models (LIME)



Original Image
 $P(\text{tree frog}) = 0.54$

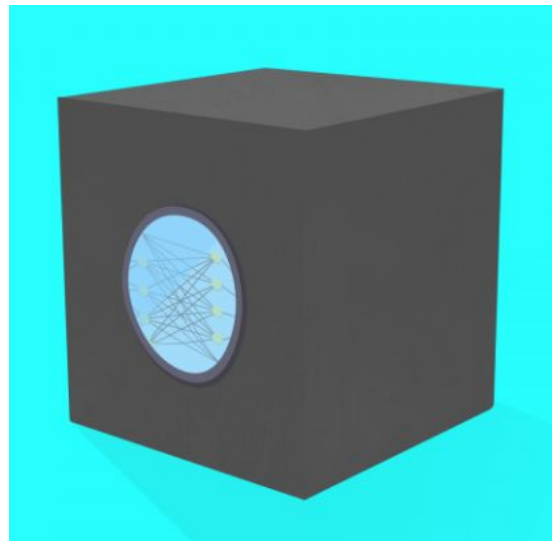


Perturbed Instances	$P(\text{tree frog})$
	 0.85
	 0.00001
	 0.52



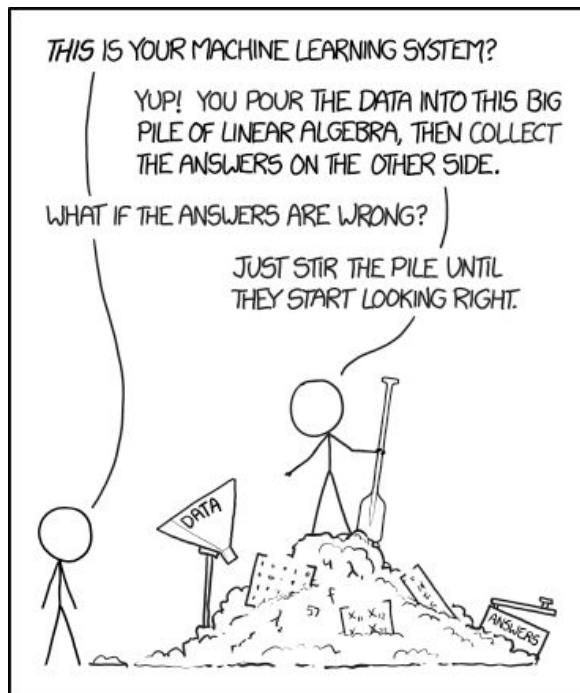
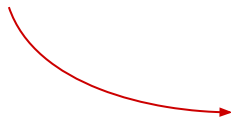
Conclusion

- performance metrics are crucial for evaluation, but they **lack explanations**
- criteria like fairness and consistency are much harder if not **impossible to quantify**
- the problem with blackboxes is the **lack of trust** caused by their opaque nature
- **transparency is key** to achieving trust and acceptance in the mainstream



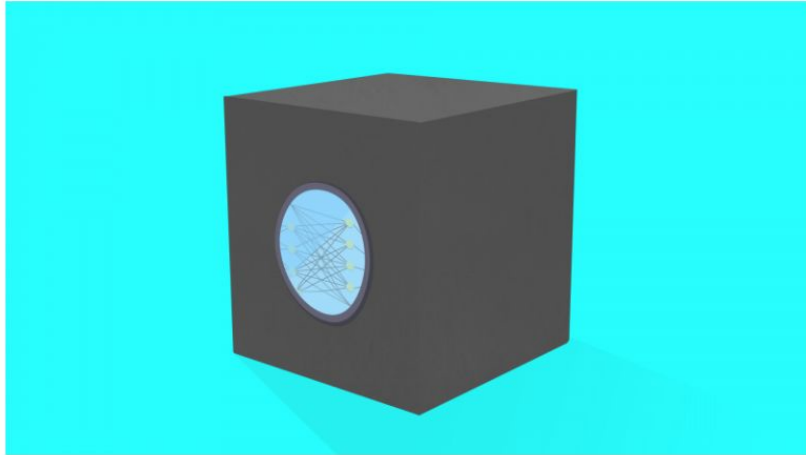
Conclusion

don't end up like this!



Resources

- Molnar C., 2018, Interpretable Machine Learning - A Guide for Making Black Box Models Explainable
- Gill N., Hall P., 2018, An Introduction to Machine Learning Interpretability
- Zhao Q., Hastie T., 2017, Causal Interpretations of Black-Box Models
- Kim B., Doshi-Velez F., 2017, Interpretable Machine Learning: The fuss, the concrete and the questions
- Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. Why should i trust you? Explaining the predictions of any classifier



Machine Learning Interpretability: Do You Know What Your Model Is Doing?

Gepostet am: 13. Februar 2019

Marcel Spitzer



Vielen Dank

Marcel Spitzer

Big Data Scientist

mspitzer@inovex.de

inovex GmbH

Schanzenstraße 6-20

Kupferhütte 1.13

51063 Köln

