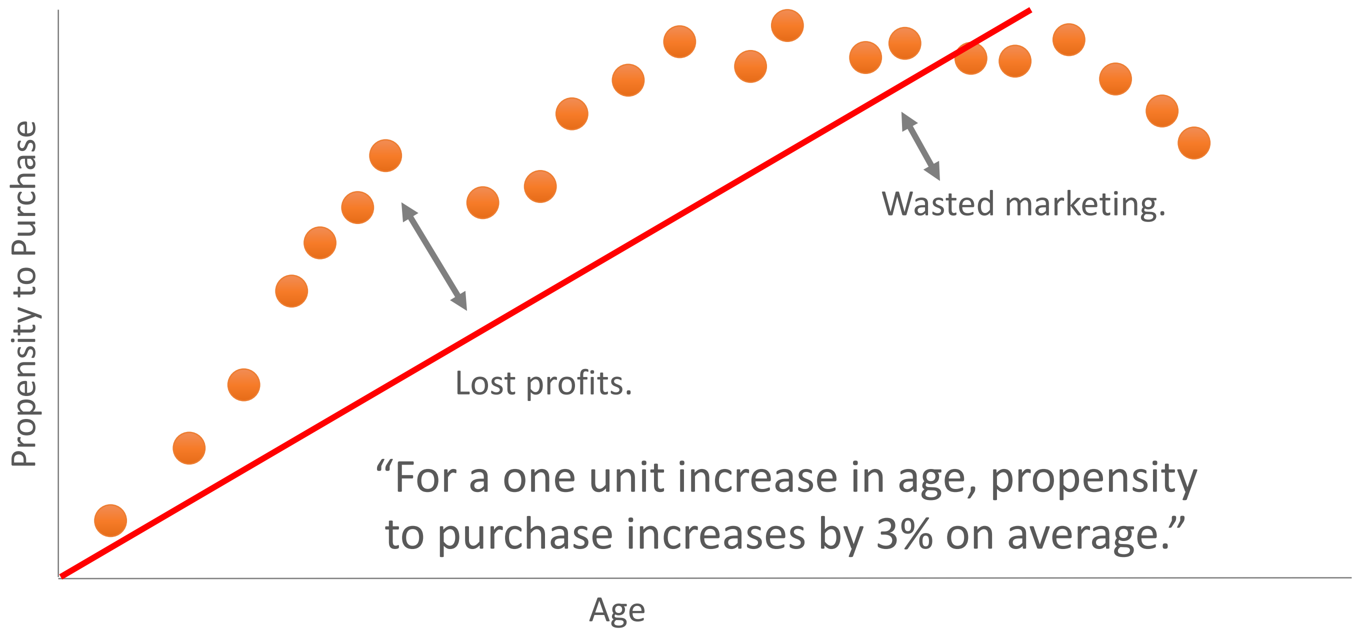
Ideas on Machine Learning Interpretability

Patrick Hall, Wen Phan, SriSatish Ambati and the H2O.ai team

Big Ideas

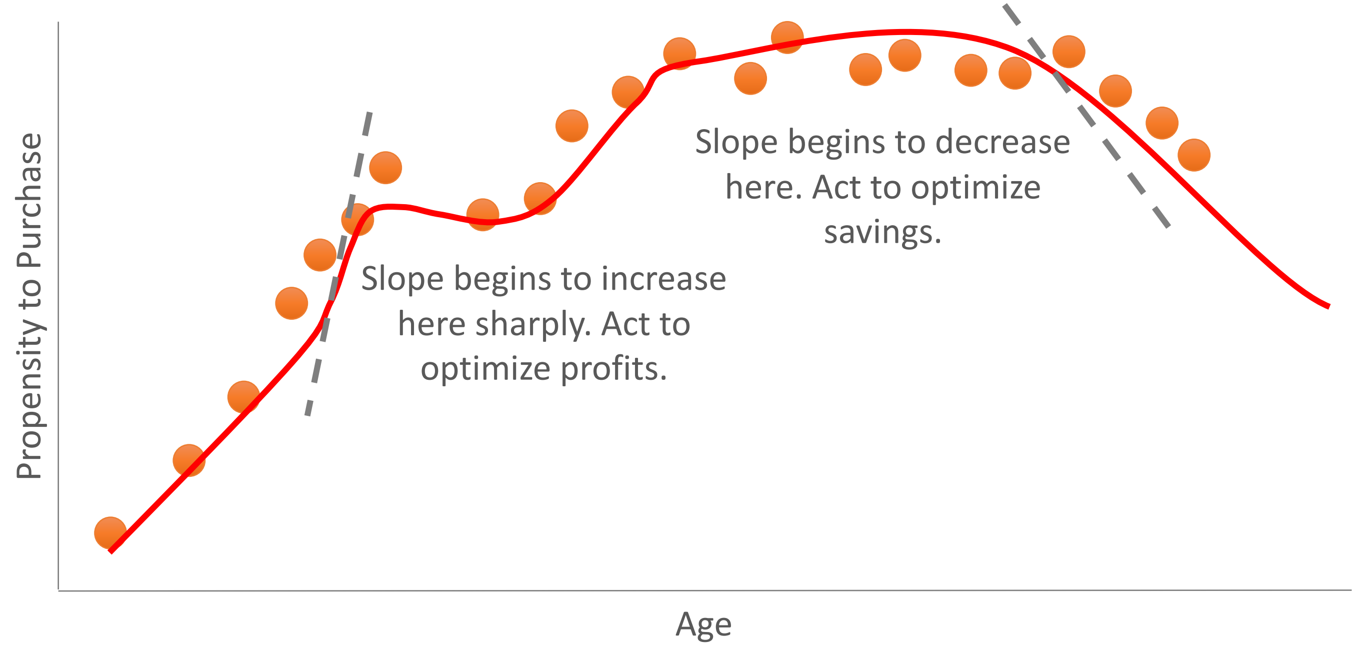
Linear Models

Exact explanations for ***approximate*** models.



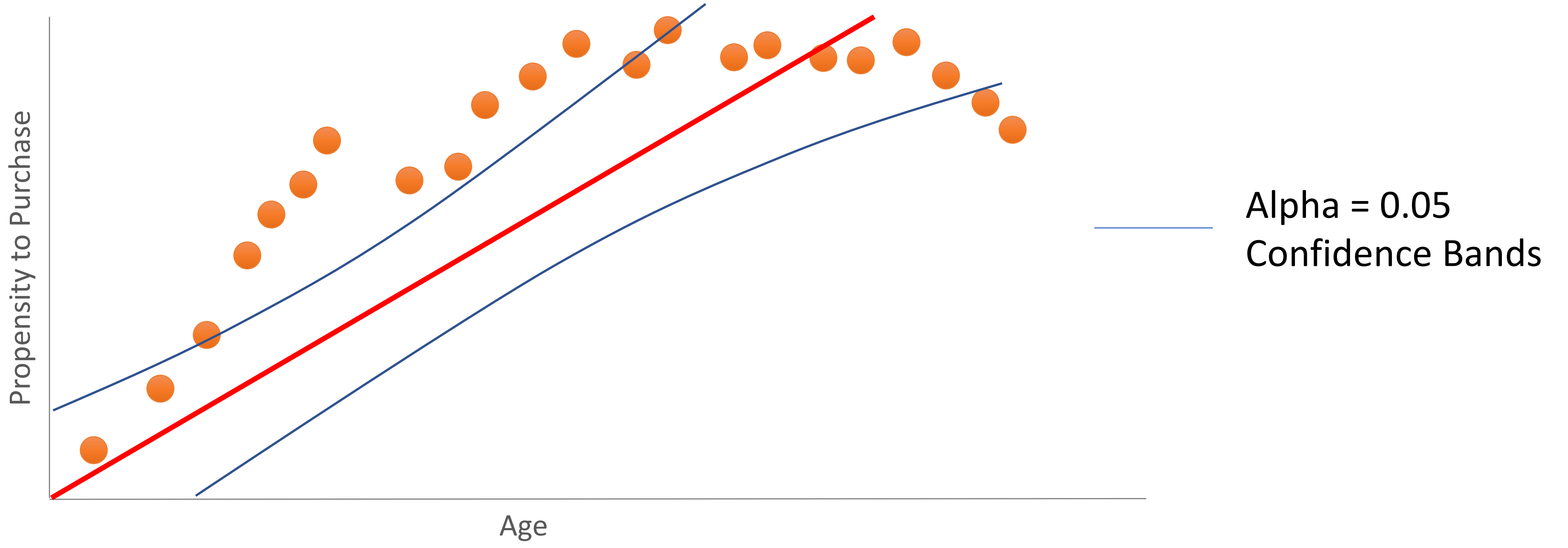
Machine Learning

Approximate explanations for ***exact*** models.



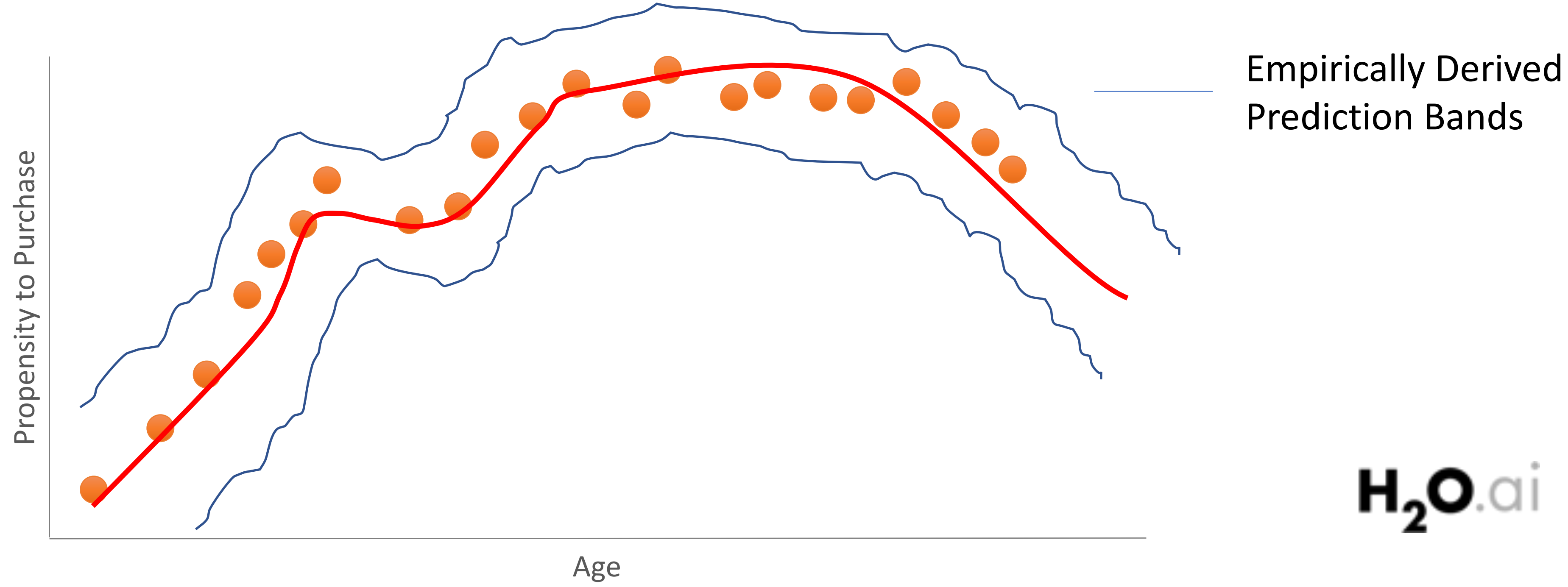
Linear Models

Risk is well defined ...
Theoretically ...
Based on **strong assumptions**.



Machine Learning

Risk is **empirically quantifiable** ...
But it's **hard work**.



H2O.ai

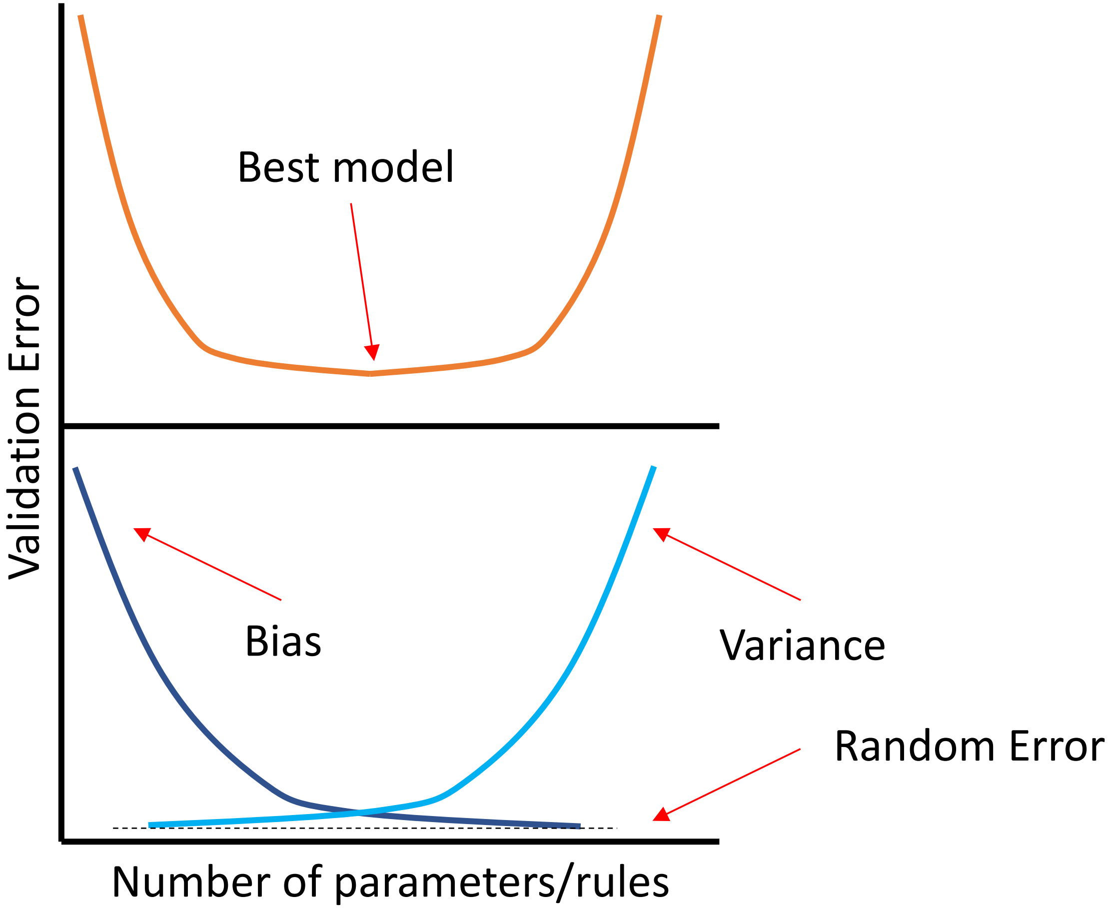
Nobody really believes that multivariate data is multivariate normal, but that data model occupies a large number of pages in every graduate textbook on multivariate statistical analysis.

-- Leo Breiman

Risk from Unwanted Bias and Prediction Variance

Total Error = Bias + Variance + Random Error =

$$(\hat{f}(x) - f(x))^2$$



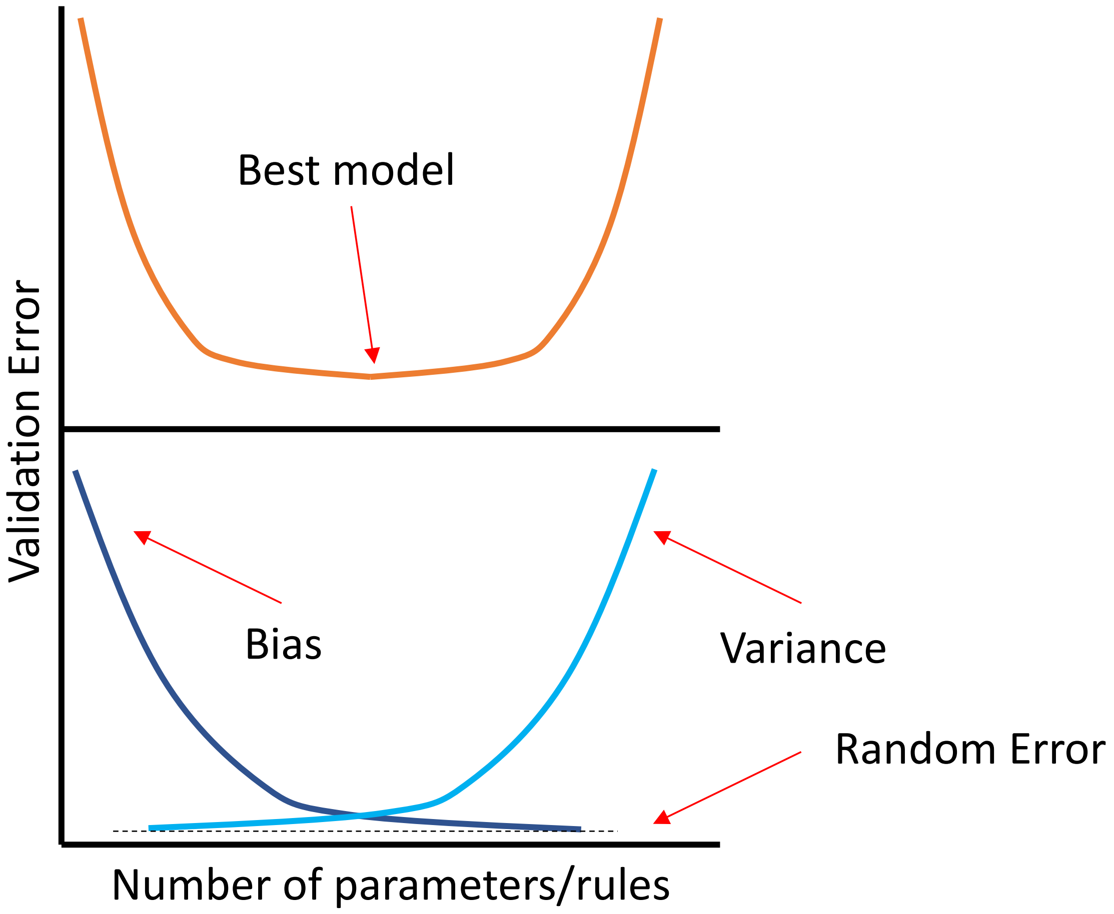
Bias = $E[\hat{f}(x)] - f(x)$ or the error that arises from a model's inability to replicate the fundamental phenomena represented by a data set.

Variance = $(\hat{f}(x) - E[\hat{f}(x)])^2$ or the error that arises from a model's ability to produce differing predictions from the values in a new data set.

Risk from Unwanted Bias and Prediction Variance

Total Error = Bias + Variance + Random Error =

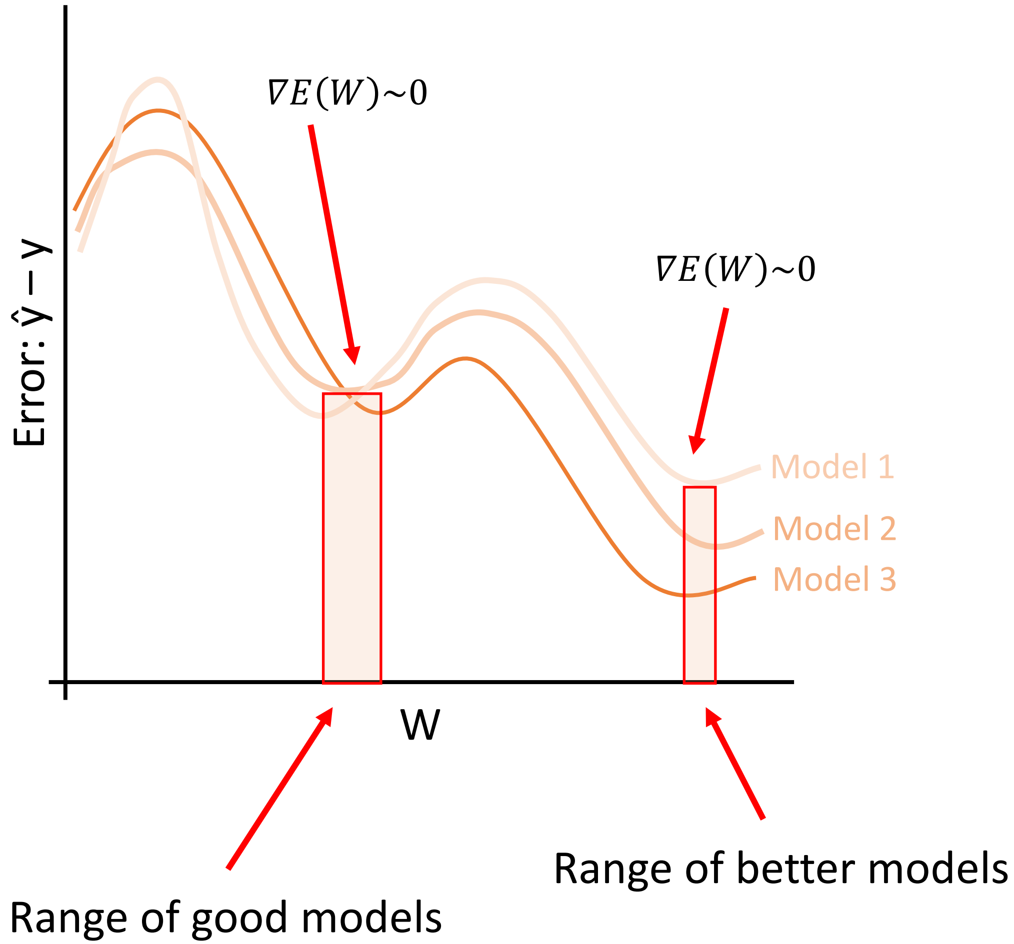
$$(\hat{f}(x) - f(x))^2$$



Risk from Unwanted Bias: Your model includes contributions from race, gender, disability status, marital status, or other unwanted latent features.

Risk from Prediction Variance: Your model is unpredictable outside of the training domain.

The Multiplicity of Good Models



Training ML models often involves solving non-convex optimization problems with multiple local minima.

Different solutions for a good ML model produce the same distribution of model outputs, *but* the actual numeric predictions produced can be slightly different.

These small differences will result in slightly different explanations. Mathematically this is ok … it's a different model after all, *but* it poses serious philosophical and regulatory problems.

H2O.ai

A framework for interpretability

Complexity of learned functions:
- Linear, monotonic
- Nonlinear, monotonic
- Nonlinear, non-monotonic

Scope of interpretability:
Global vs. local

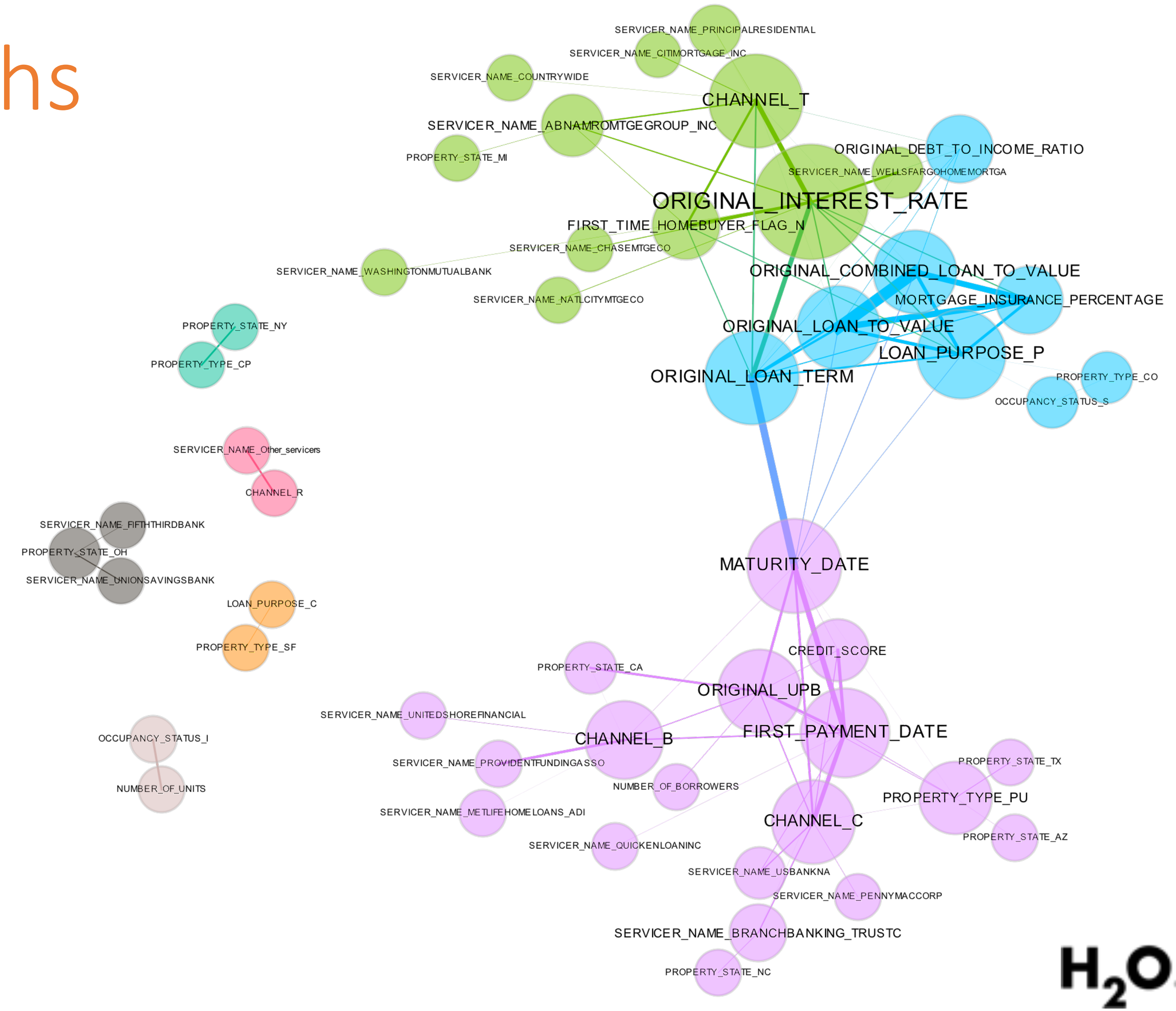
Enhancing trust and understanding:
the mechanisms and results of an interpretable model should be both transparent AND dependable.

Application domain:
Model-agnostic vs. model-specific

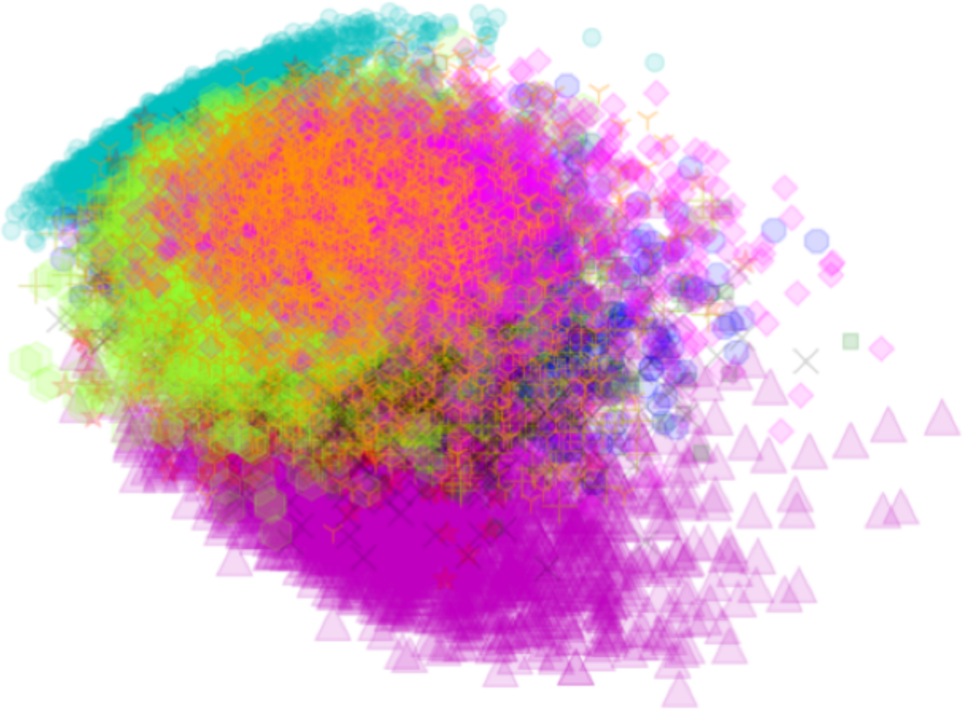
A Few of Our Favorite Things

Correlation graphs

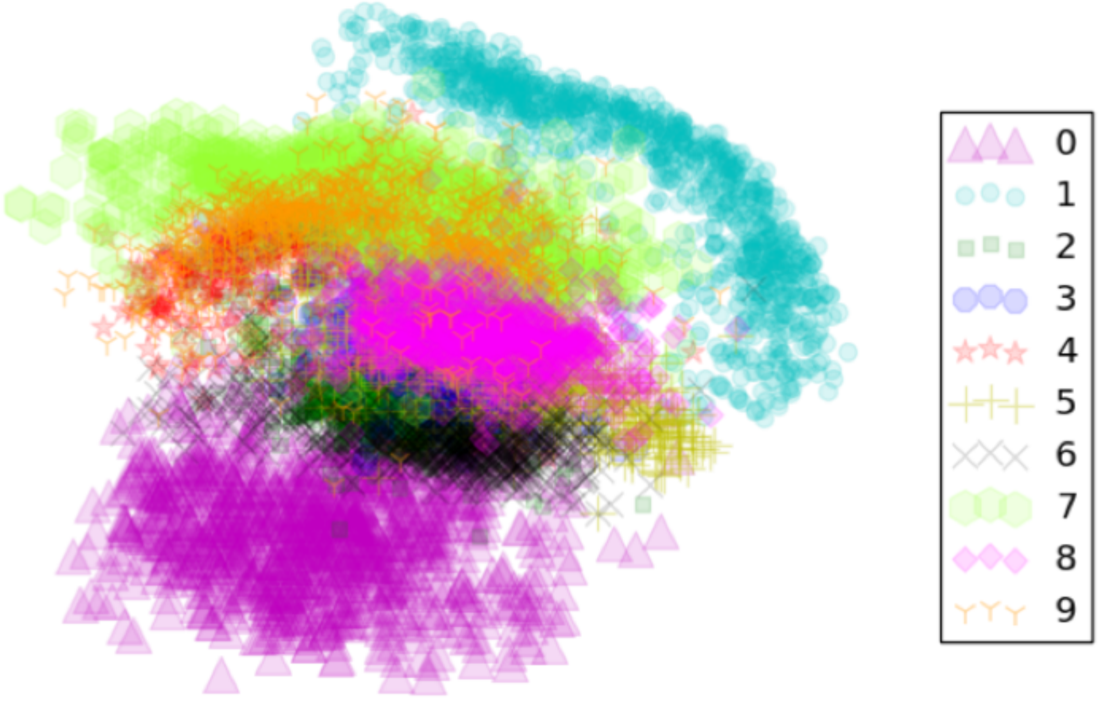
The nodes of this graph are the variables in a data set. The weights between the nodes are defined by the absolute value of their pairwise Pearson correlation.



2-D projections

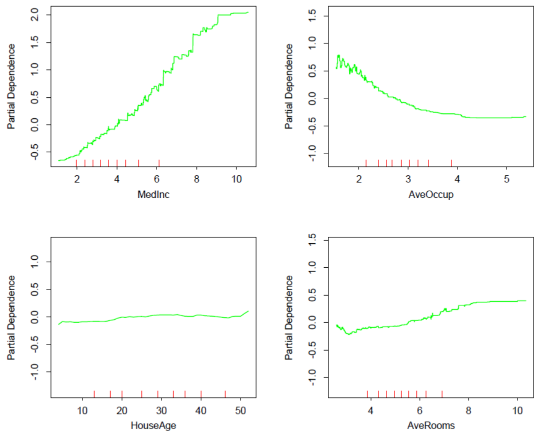


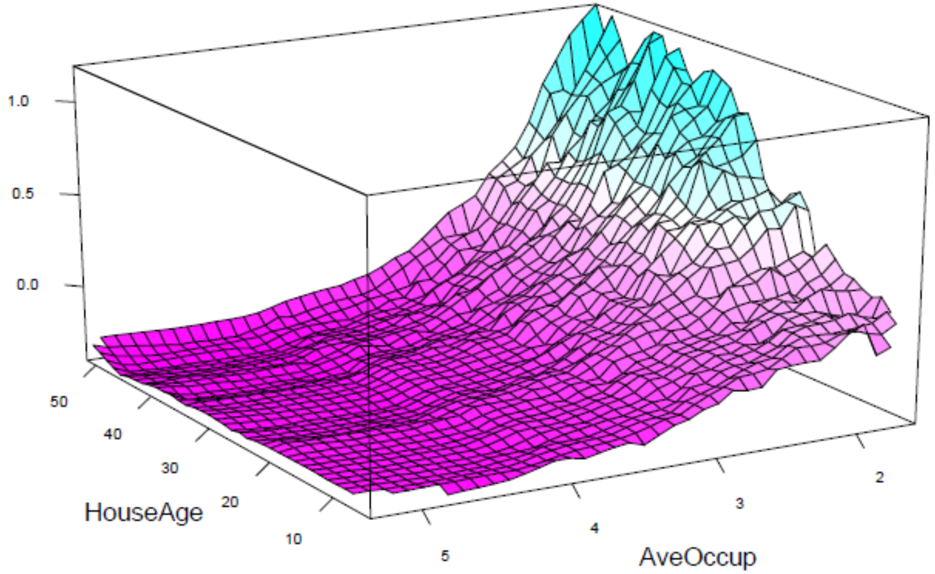
784 dimensions to 2 dimensions with PCA



784 dimensions to 2 dimensions with autoencoder network

Partial dependence plots

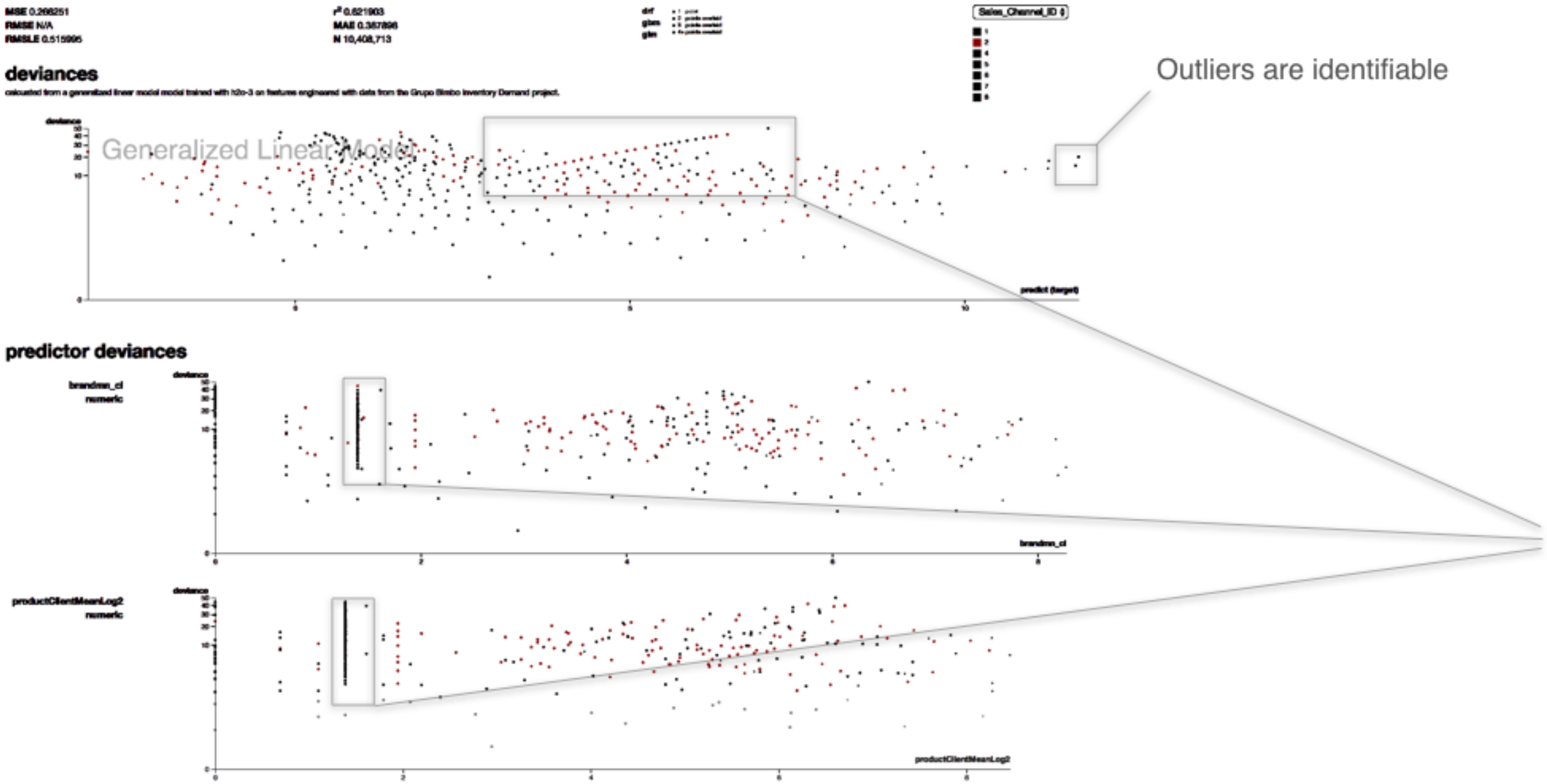




HomeValue ~ MedInc + AveOccup + HouseAge + AveRooms

Source: http://statweb.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf

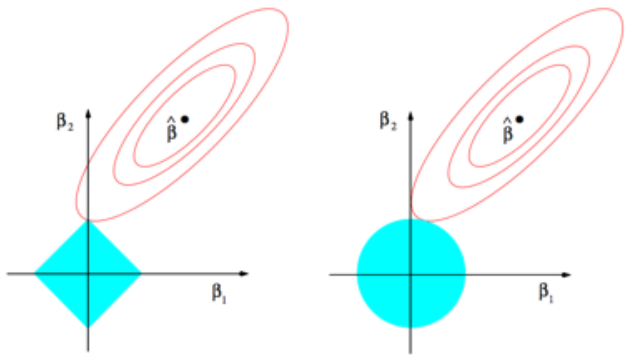
Residual analysis



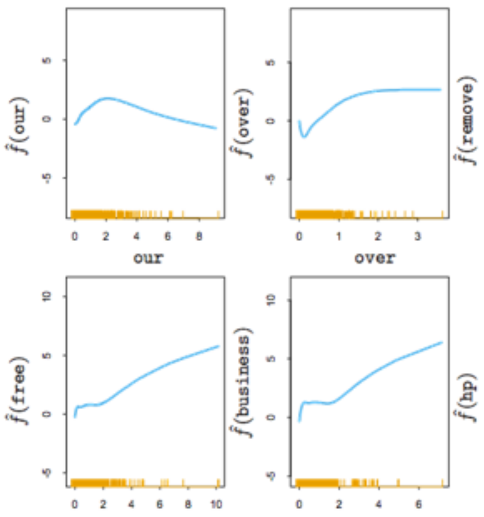
Residuals from a machine learning model should be randomly distributed

obvious patterns in residuals can indicate problems with data preparation or model specification

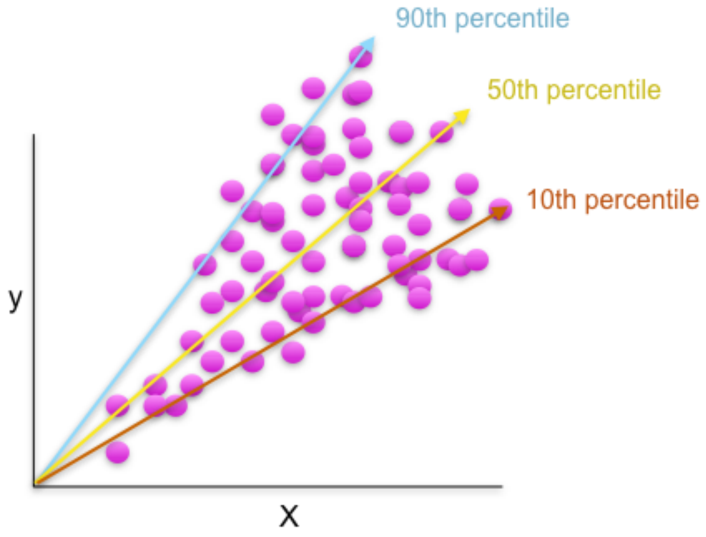
OLS regression alternatives



Penalized Regression



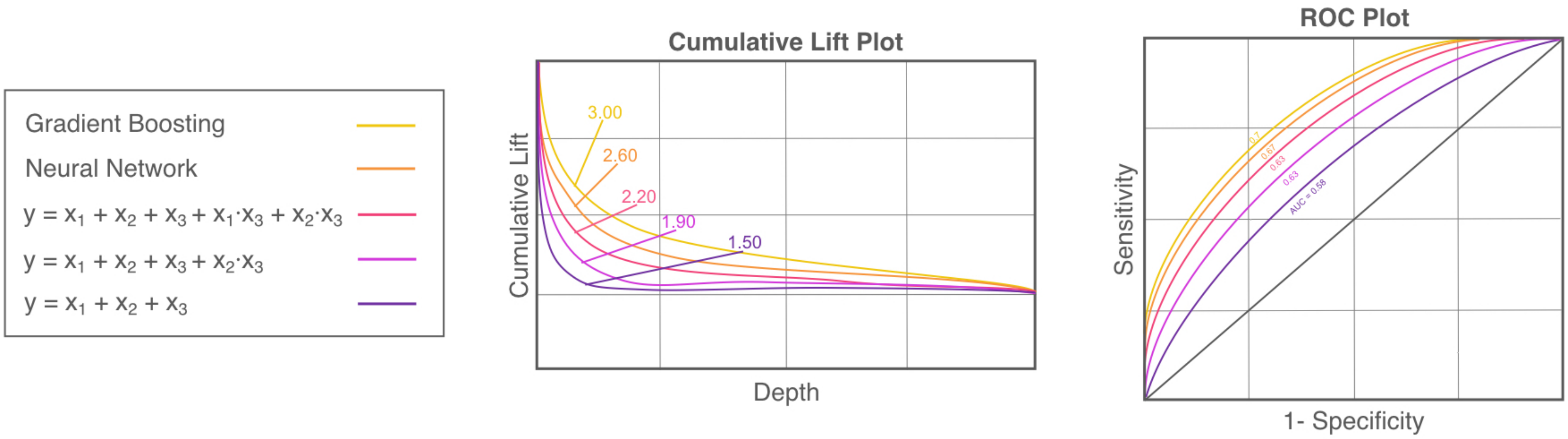
Generalized Additive Models



Quantile Regression

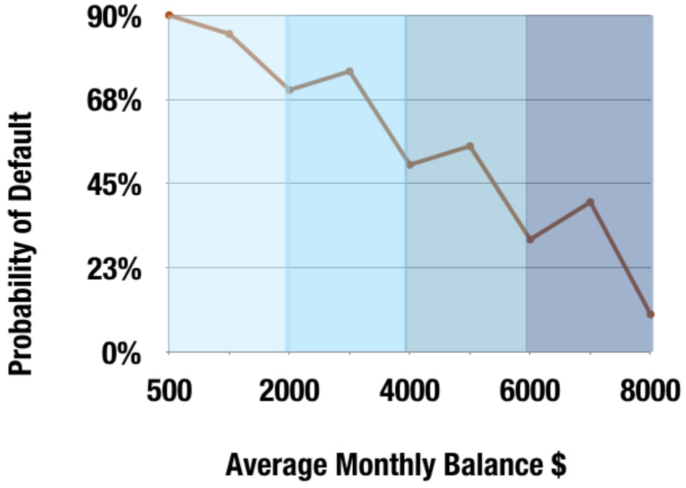
Source: http://statweb.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf

Build toward ML model benchmarks

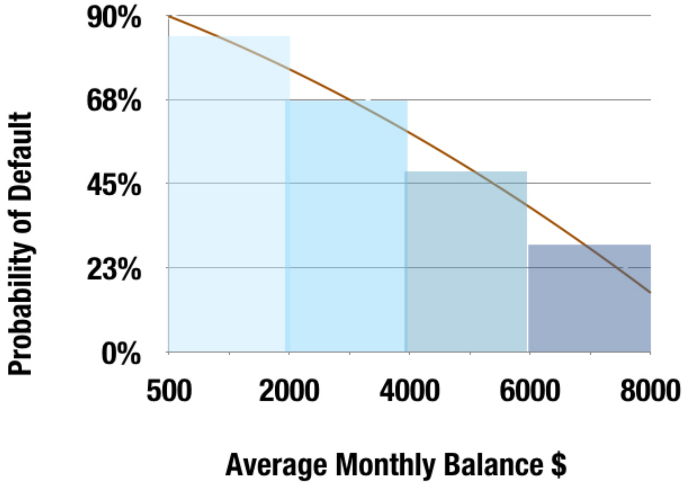


Incorporate interactions and piecewise linear components to increase the accuracy of linear models relative to machine learning models.

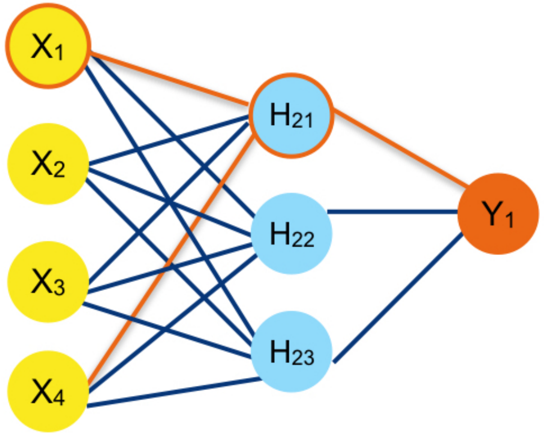
Monotonicity constraints



Average Monthly Balance is a nonnegative quantity, but is not monotonic with respect to Probability of Default.



By discretization, the Average Monthly Balance can be transformed to be monotonic with respect to the target.



When all inputs are nonnegative and monotonic with respect to the target, and model weights are constrained to be nonnegative, it's easier understand the impact of individual features and to find interactions.

Rule-based models

IF Age: 30 <= 40 …

AND Income: 70K <= 80K …

AND Zip Code == 20009 …

AND Marital Status == SINGLE …

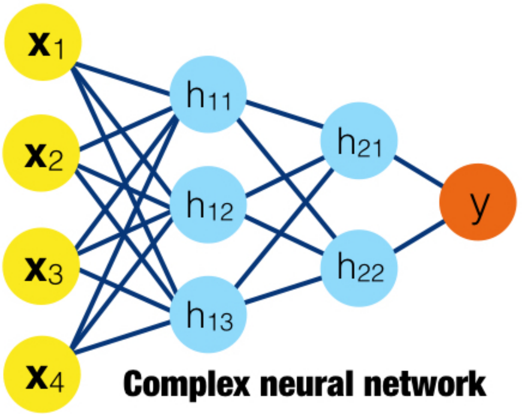
AND Loyalty Status == SILVER …

THEN Probability = 0.7463

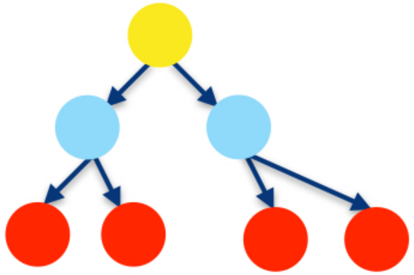
Surrogate models

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

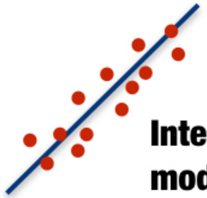


Complex neural network



Interpretable decision tree

Or



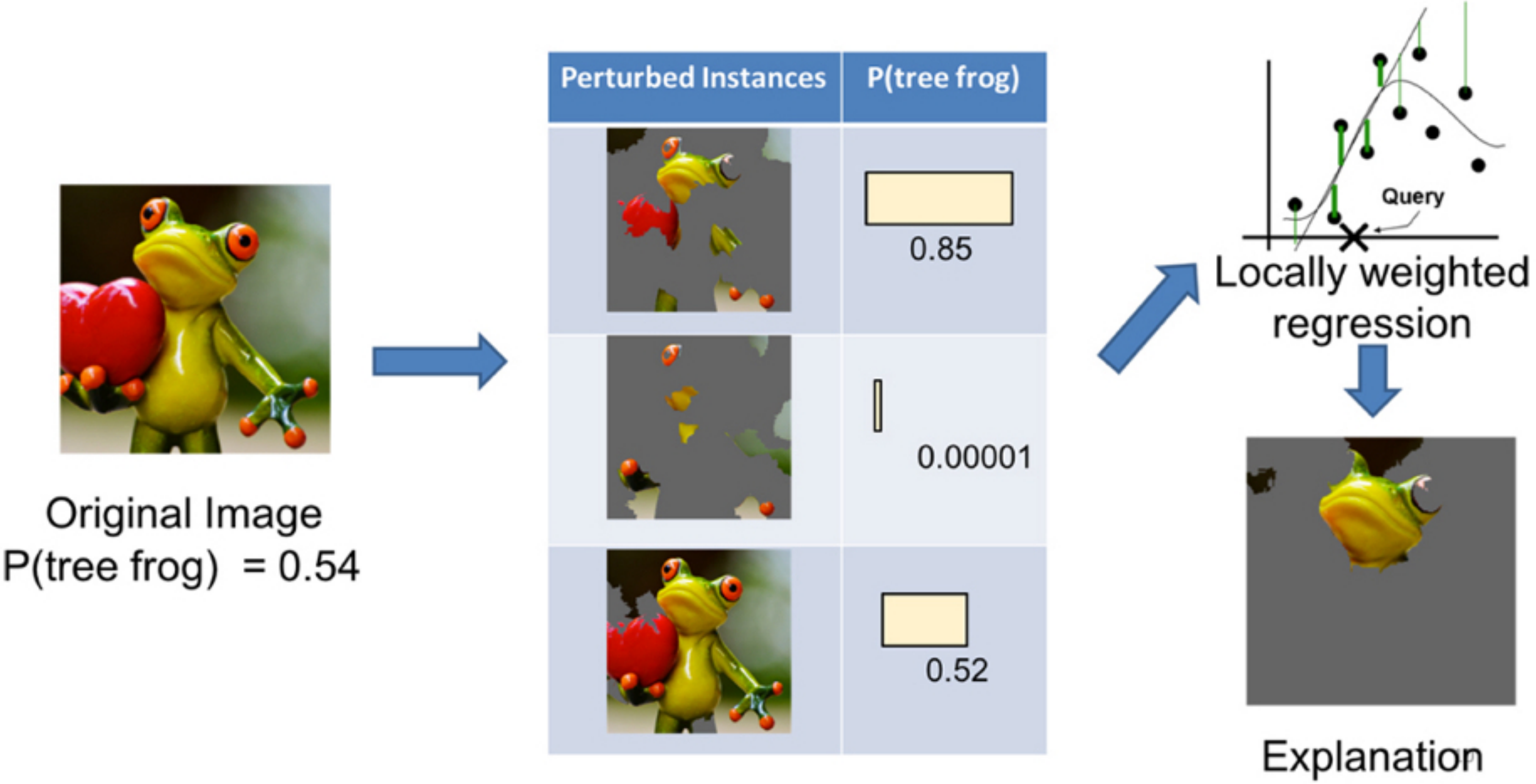
Interpretable linear model

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

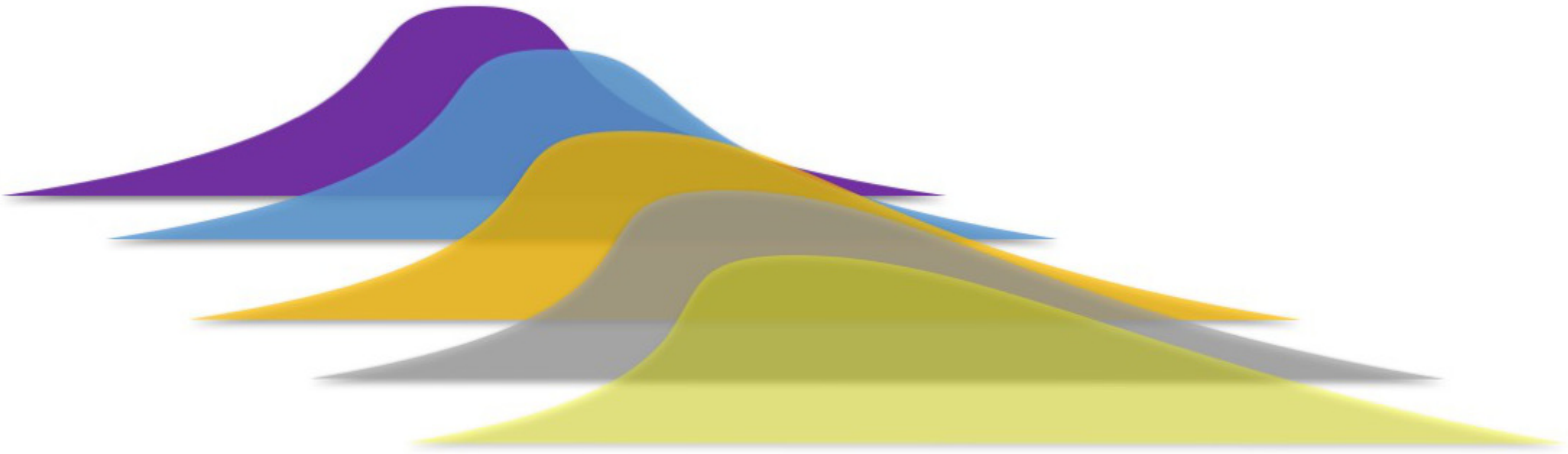
H2O.ai

Local interpretable model-agnostic explanations



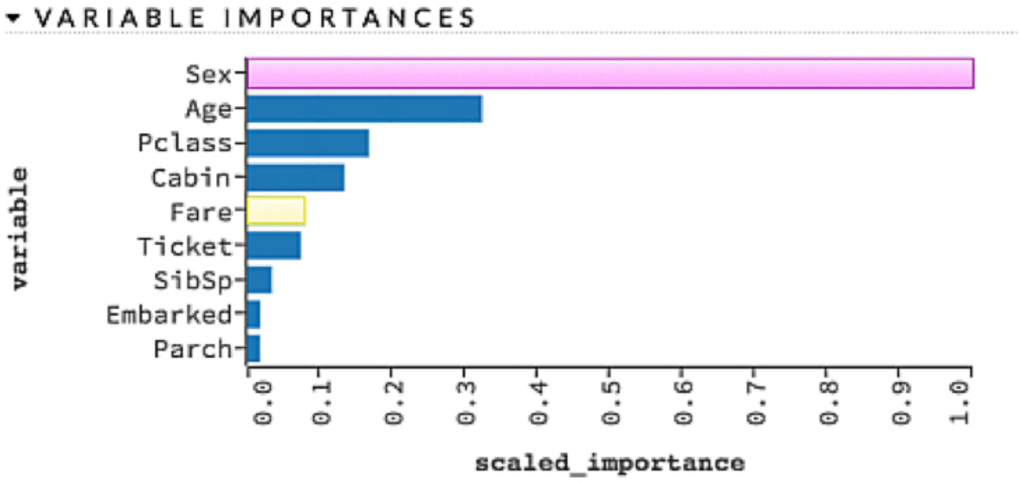
Source: https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

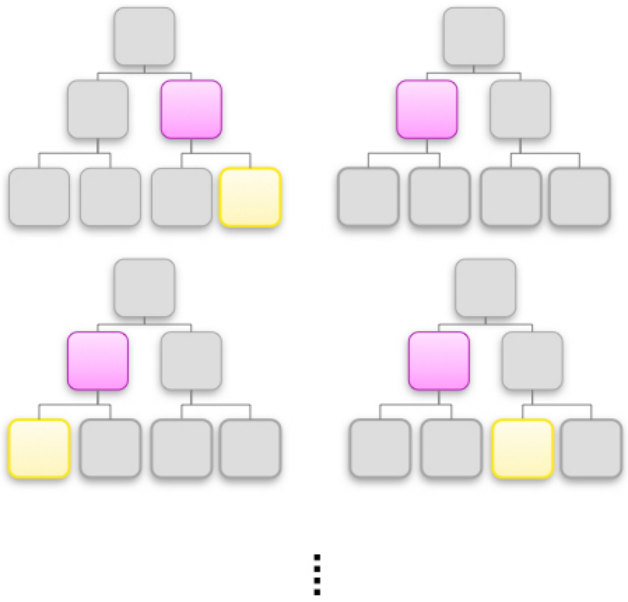
Sensitivity analysis



Data distributions shift over time. How will your model handle these shifts?

Variable importance measures





Global variable importance indicates the impact of a variable on the model for the entire training data set.

Sex	Age	…	Fare	$\hat{y}$	$\hat{y}_{(-Sex)}$	$\hat{y}_{(-Age)}$	…	$\hat{y}_{(-Fare)}$
M	11	…	8.45	0.2	0.01	0.1	…	0.21
F	34	…	51.86	0.8	0.6	0.65	…	0.78
M	26	…	21.08	0.5	0.2	0.3	…	0.53
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Local variable importance can indicate the impact of a variable for each decision a model makes – similar to reason codes.

H2O.ai

Questions?