

# Machine Learning Interpretability



The good, the bad, and the ugly

# Acknowledgments

**Makers:** Navdeep Gill, Mark Chan, Doug Deloy, Megan Kurka, Michal Kurka, Wen Phan, Sri Satish Ambati, Lingyao Meng, Mathias Müller



**Advisors:** Leland Wilkinson, Trevor Hastie, Rob Tibshirani



**Community:** O'Reilly Strata and AI (accepted talks and tutorials), FAT/ML Conference (accepted tutorial), ASA Symposium on Data Science and Statistics (invited talk), Joint Statistical Meetings (accepted talk)

# What is Machine Learning Interpretability?

*“The ability to explain or to present in understandable terms to a human.”*

-- Finale Doshi-Velez and Been Kim. “Towards a rigorous science of interpretable machine learning.” arXiv preprint. 2017. <https://arxiv.org/pdf/1702.08608.pdf>

FAT\*: <https://www.fatml.org/resources/principles-for-accountable-algorithms>

XAI: <https://www.darpa.mil/program/explainable-artificial-intelligence>

# Why Should You Care About Machine Learning Interpretability?

“The now-contemplated field of data science amounts to a superset of the fields of statistics and machine learning, which adds some technology for “scaling up” to “big data.” This chosen superset is motivated by commercial rather than intellectual developments. **Choosing in this way is likely to miss out on the really important intellectual event of the next 50 years.**”

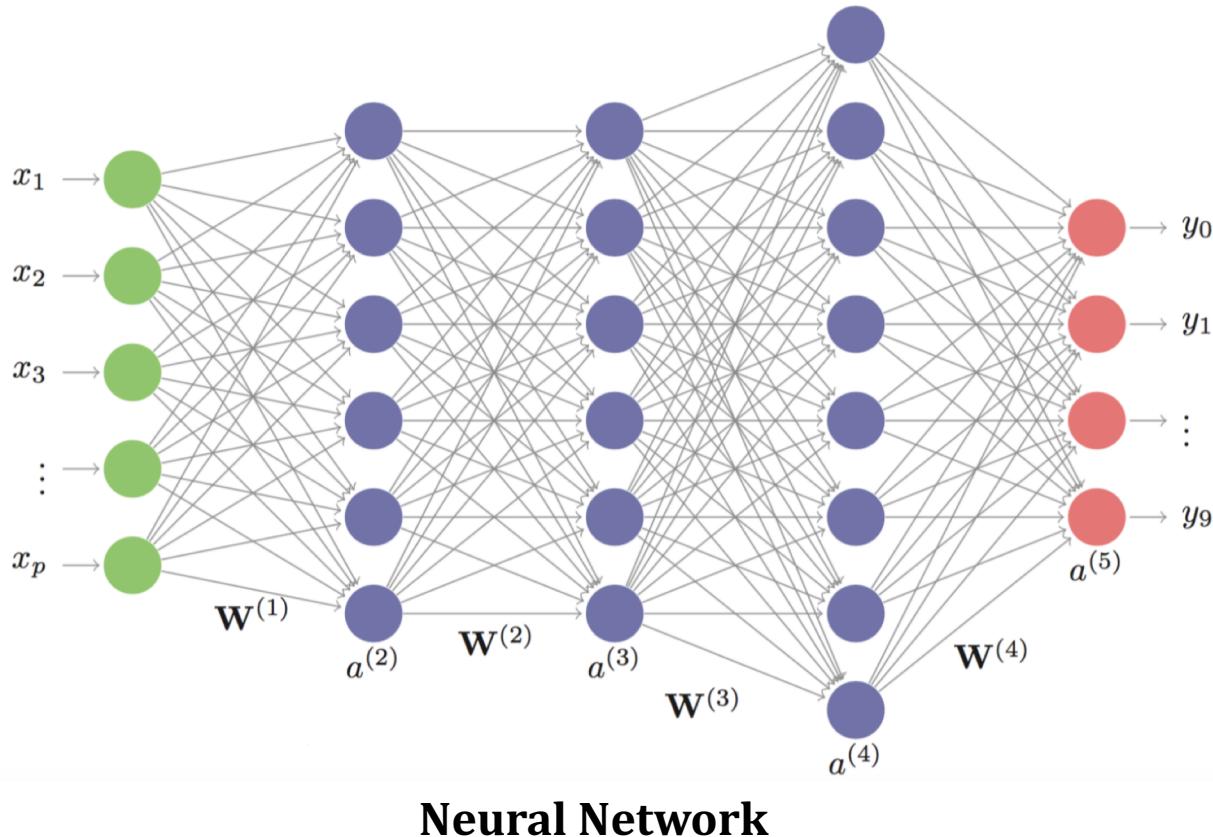
-- David Donoho. “50 years of Data Science.” Tukey Centennial Workshop, 2015. <http://bit.ly/2GQOh1J>

**Social Motivation:** Interpretability plays a critical role in the increased convenience, automation, and organization in our day-to-day lives promised by AI.

**Commercial Motivation:** Interpretability is required for regulated industry to adopt machine learning.

- Check and balance against accidental or intentional discrimination.
  - “Right to explanation.”
- Hacking and adversarial attacks.
- Improved revenue, i.e. Equifax NeuroDecision: [https://www.youtube.com/watch?v=9Z\\_GW9WDS2c](https://www.youtube.com/watch?v=9Z_GW9WDS2c)

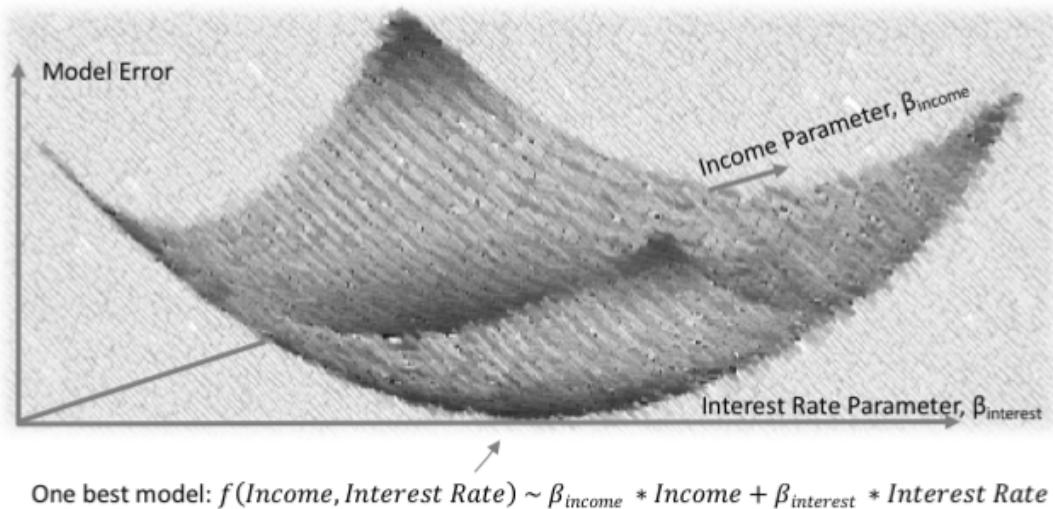
# Why is Machine Learning Interpretability Difficult?



Machine learning algorithms intrinsically consider high-degree interactions between input features.

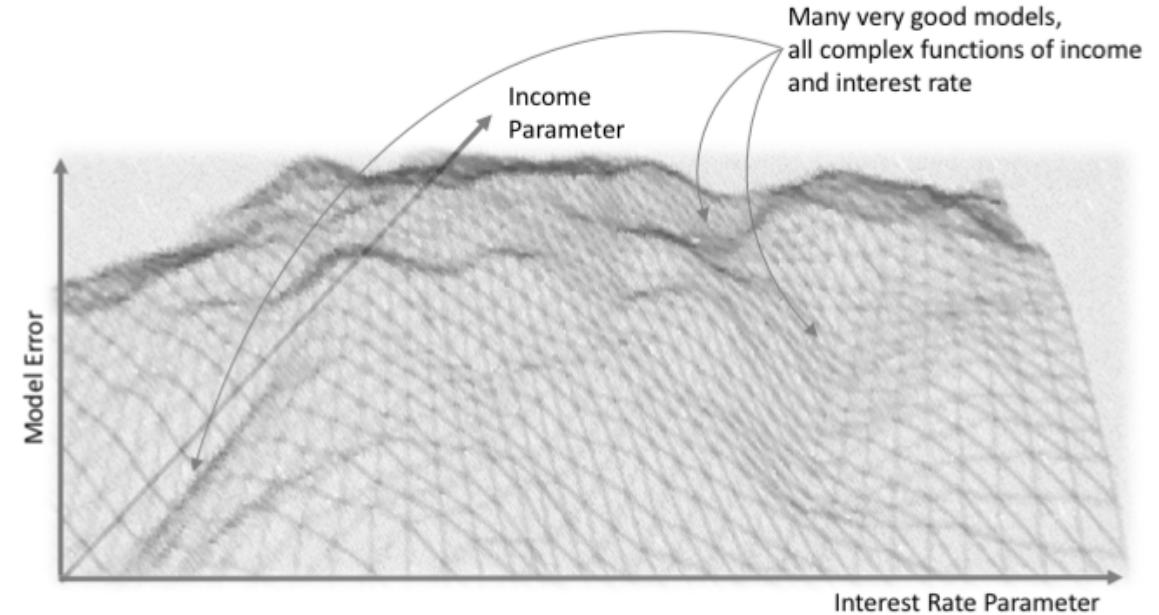
Disaggregating such functions into *reason codes* based on single input features is difficult.

# Why is Machine Learning Interpretability Difficult?



## Linear Models

For a given well-understood dataset there is usually **one best model**.



## Machine Learning

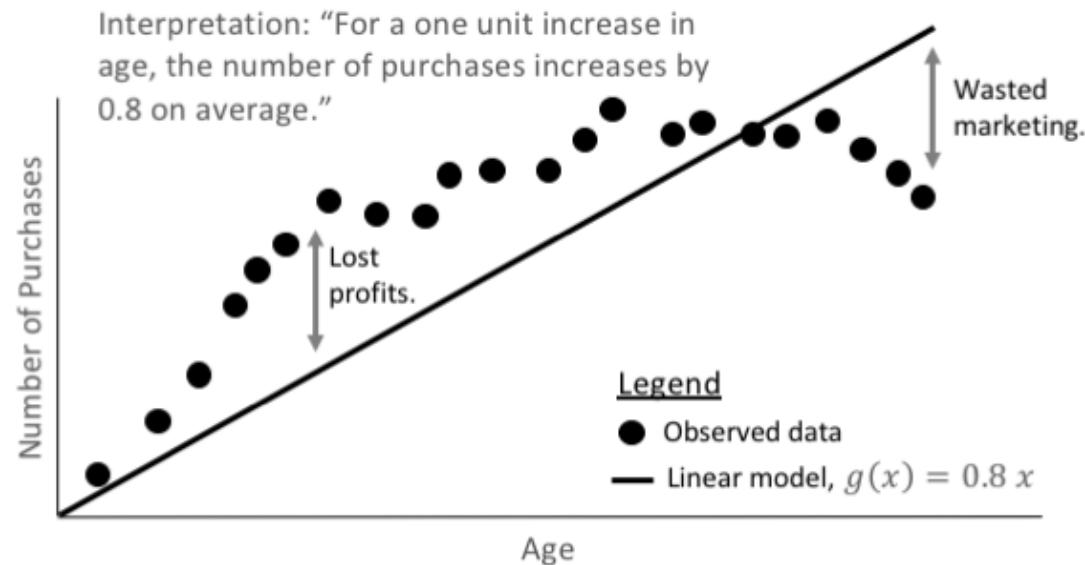
For a given well-understood dataset there are usually **many good models**. This is often referred to as “the multiplicity of good models.”

-- Leo Breiman. “Statistical modeling: The two cultures (with comments and a rejoinder by the author).” Statistical Science. 2001.  
<http://bit.ly/2pwz6m5>

# What is the Value Proposition of Machine Learning Interpretability?

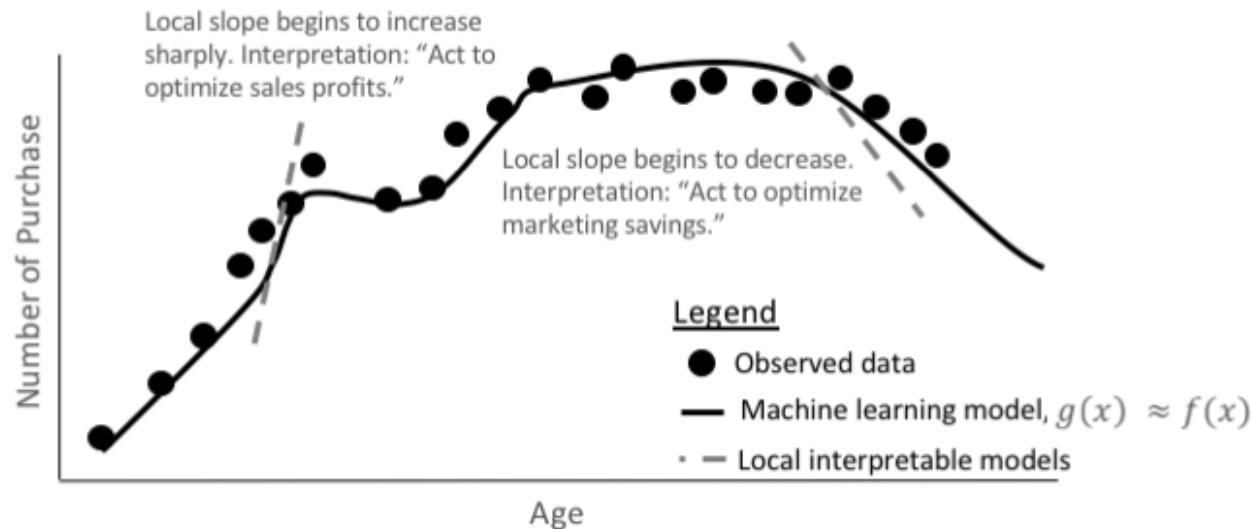
Linear Models

*Exact* explanations for  
*approximate* models.



Machine Learning

*Approximate* explanations  
for *exact* models.

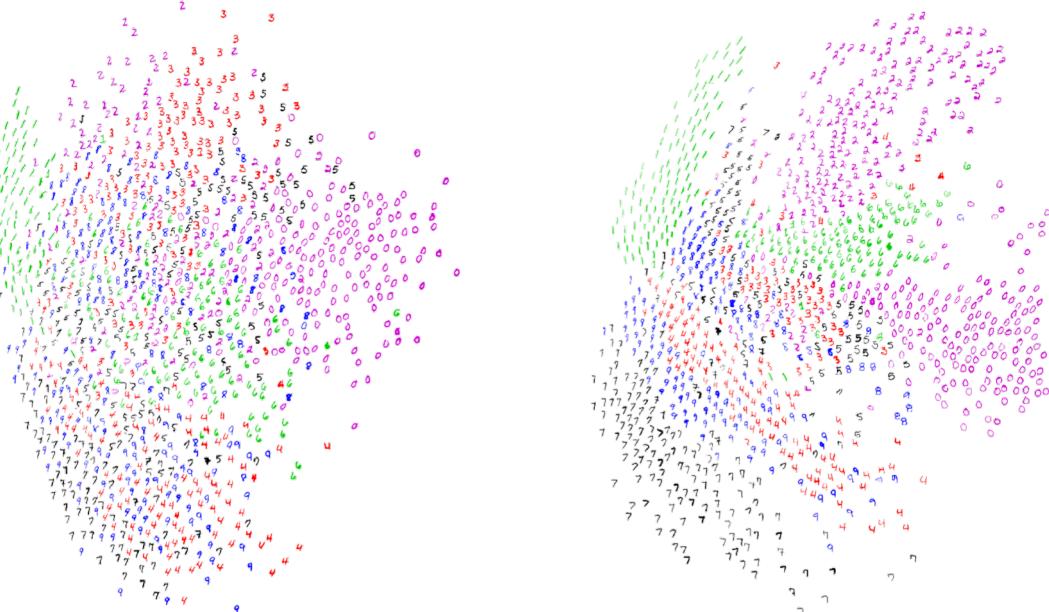


# How Can Machine Learning Interpretability Be Practiced?

By seeing and understanding relationships and structures in training, test, and new data.

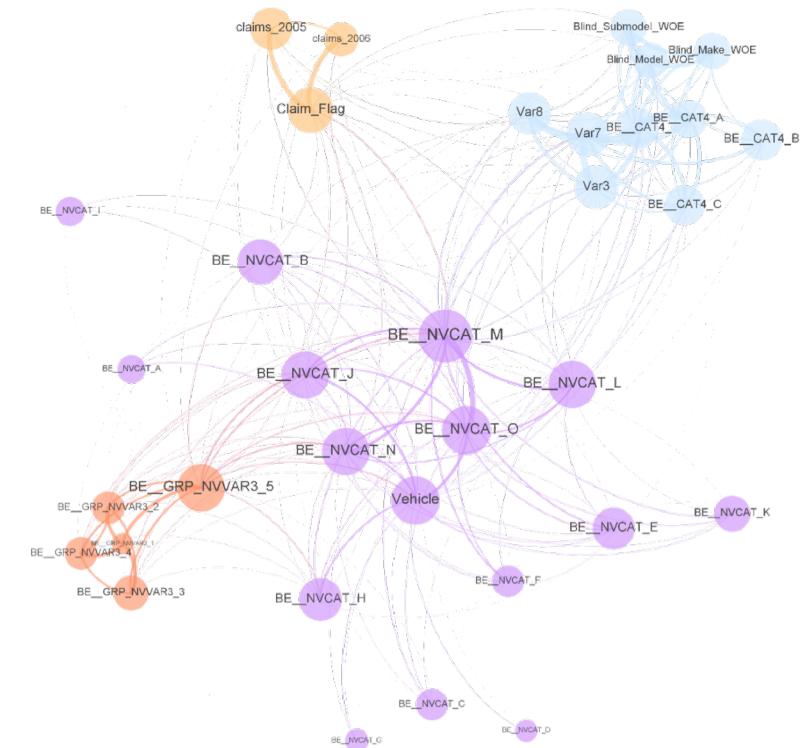
## 2-D Projections

<https://www.cs.toronto.edu/~hinton/science.pdf>



## Correlation Graphs

[https://github.com/jphall663/corr\\_graph](https://github.com/jphall663/corr_graph)



# How Can Machine Learning Interpretability Be Practiced?

By training interpretable (“white-box”) models.

## Decision Trees

- References:
  - Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen. Classification and regression trees. CRC press, 1984.
  - The Elements of Statistical Learning (ESL):  
[https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII\\_print12.pdf](https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf)
- OSS:
  - rpart: <https://cran.r-project.org/web/packages/rpart/index.html>
  - scikit-learn (various functions): <https://github.com/scikit-learn/scikit-learn>

## Monotonic Gradient Boosted Machines (GBMs)

- Reference: XGBoost Documentation:  
<http://xgboost.readthedocs.io/en/latest/tutorials/monotonic.html>
- OSS: XGBoost: <https://github.com/dmlc/xgboost>

## Logistic, elastic net, GAM, and quantile regression

- References:
  - ESL
  - Koenker, R. *Quantile regression* (No. 38). Cambridge University Press, 2005.
- OSS:
  - gam: <https://cran.r-project.org/web/packages/gam/index.html>
  - glmnet: <https://cran.r-project.org/web/packages/glmnet/index.html>
  - h2o: <https://github.com/h2oai/h2o-3>
  - quantreg: <https://cran.r-project.org/web/packages/quantreg/index.html>
  - scikit-learn (various functions): <https://github.com/scikit-learn/scikit-learn>

## Rule-based models

- Reference: *An Introduction to Data Mining*, Chapter 6:  
<https://www-users.cs.umn.edu/~kumar001/dmbook/ch6.pdf>
- OSS:
  - RuleFit:  
[http://statweb.stanford.edu/~jhf/R\\_RuleFit.html](http://statweb.stanford.edu/~jhf/R_RuleFit.html)
  - arules:  
<https://cran.r-project.org/web/packages/arules/index.html>
  - FP-Growth:  
<http://spark.apache.org/docs/2.2.0/mllib-frequent-pattern-mining.html>

## Supersparse Linear Integer Models (SLIMs)

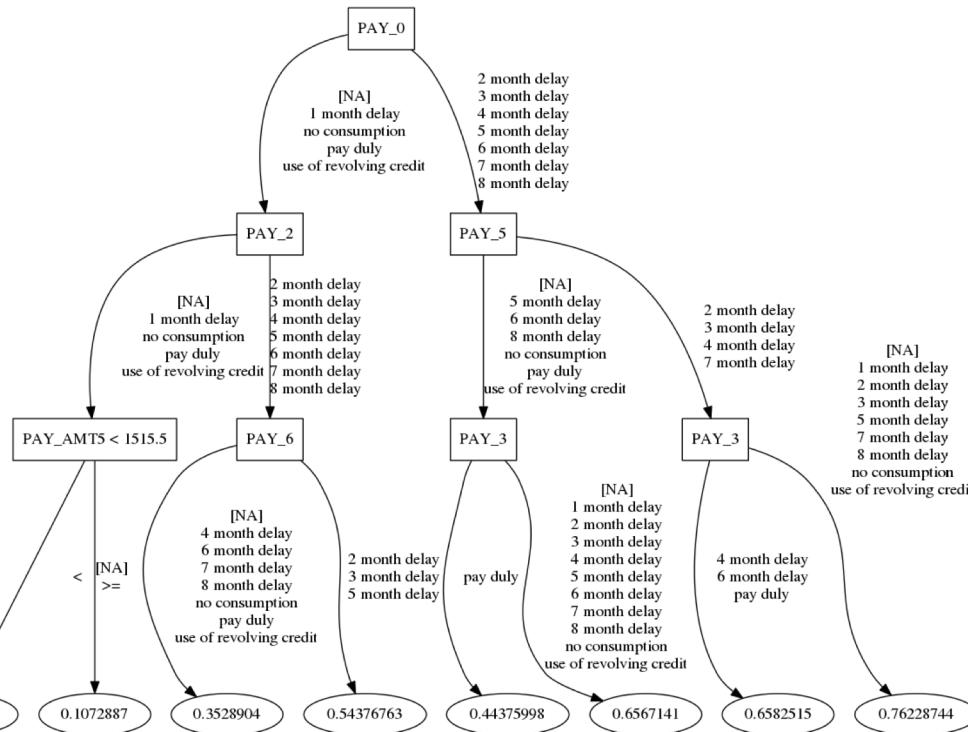
- Reference: Supersparse Linear Integer Models for Optimized Medical Scoring Systems:  
<https://link.springer.com/content/pdf/10.1007%2Fs10994-015-5528-6.pdf>

# How Can Machine Learning Interpretability Be Practiced?

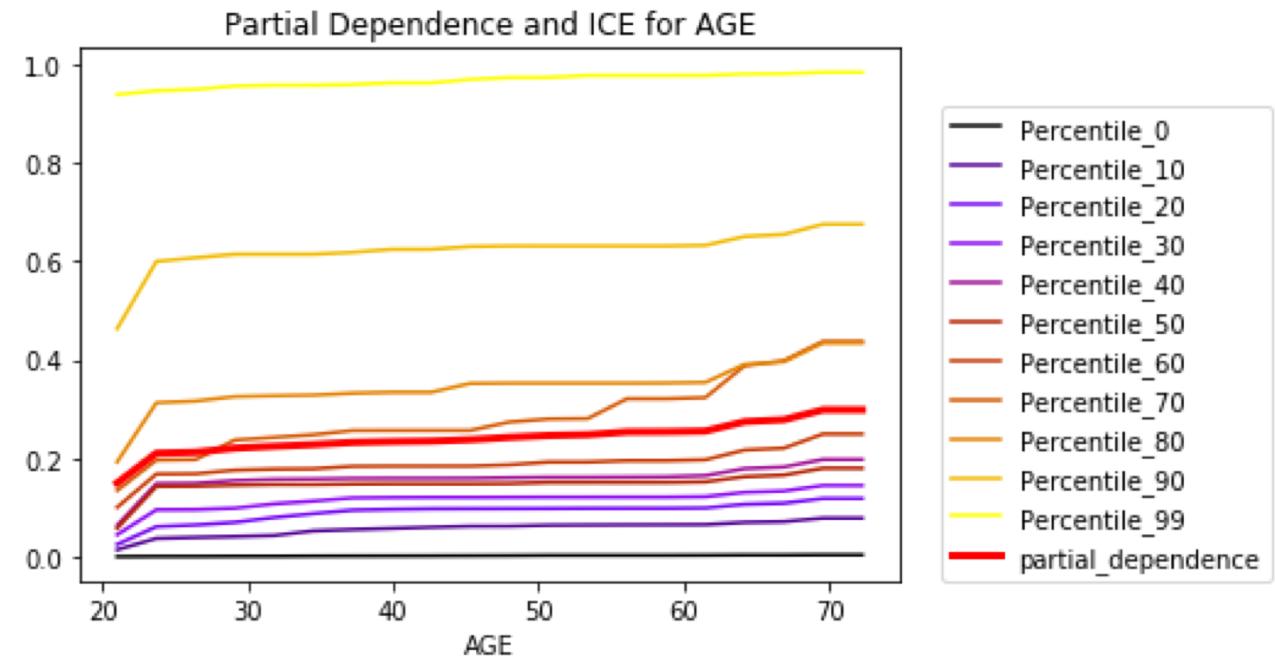
With complimentary 2-D visualizations of trained models that enable understanding of learned high-degree interactions.

## Decision Tree Surrogate Models

Credit Card Default Decision Tree Surrogate



Partial Dependence and Individual Conditional Expectation

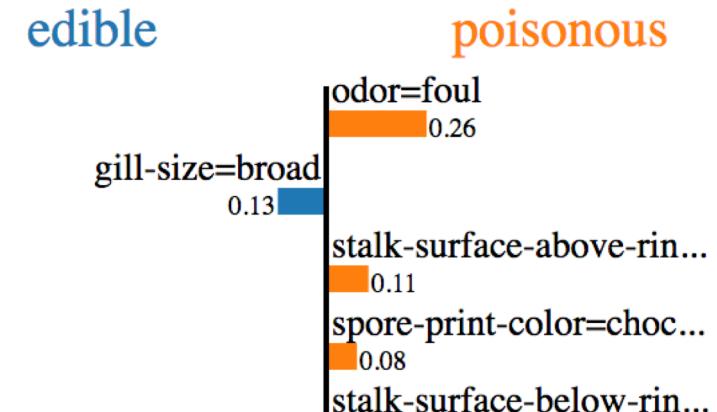


# How Can Machine Learning Interpretability Be Practiced?

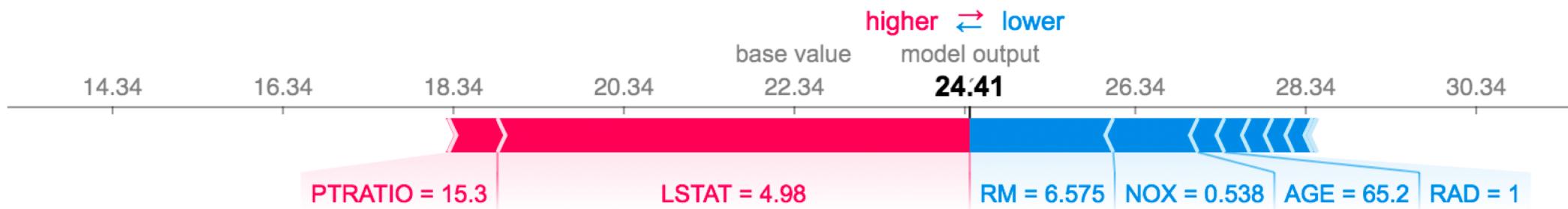
By calculating approximate local feature importance values and ranking them to create *reason codes* for every prediction.

y (score -1.178) top features	
Contribution?	Feature
+0.017	num12
+0.008	num5
+0.006	<BIAS>
-0.009	num11
-0.015	num4
-0.018	num3
-0.018	num10
-0.021	num6
-0.036	num2
-0.041	num7
-0.136	num8
-0.274	num1
-0.642	num9

Treeinterpreter:  
<https://github.com/andosa/treeinterpreter>



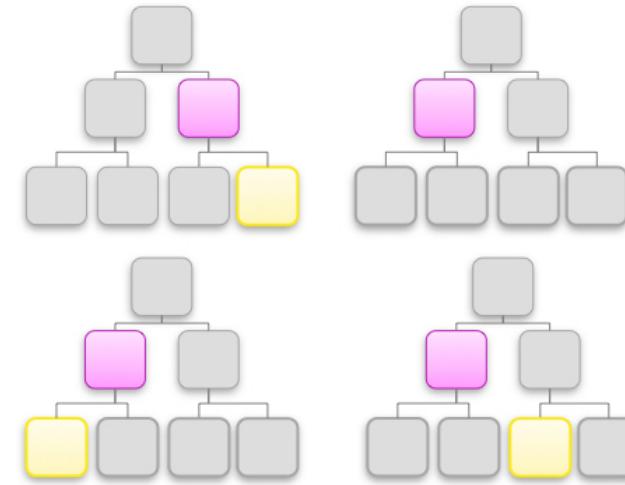
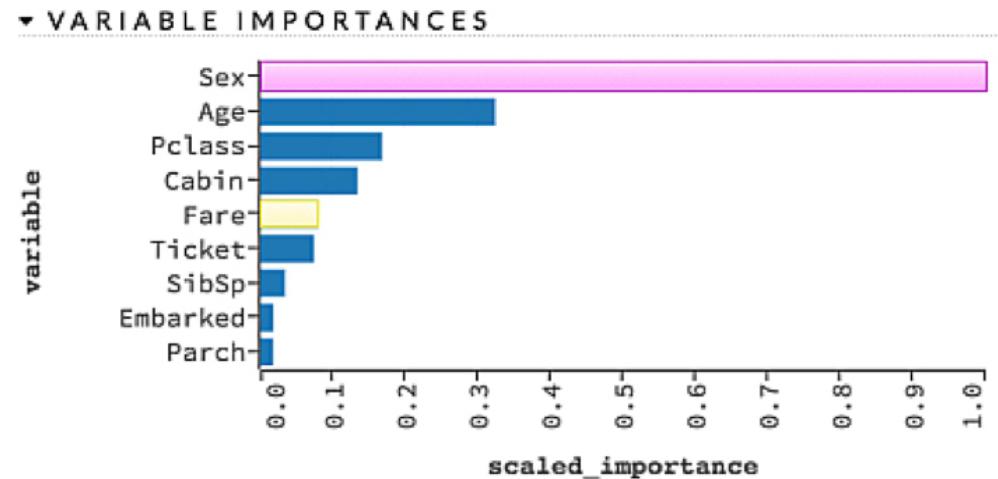
LIME:  
<https://github.com/marcotcr/lime>



shap:  
<https://github.com/slundberg/shap>

# How Can Machine Learning Interpretability Be Practiced?

By calculating approximate global feature importance to understand how each feature impacts the model in general.

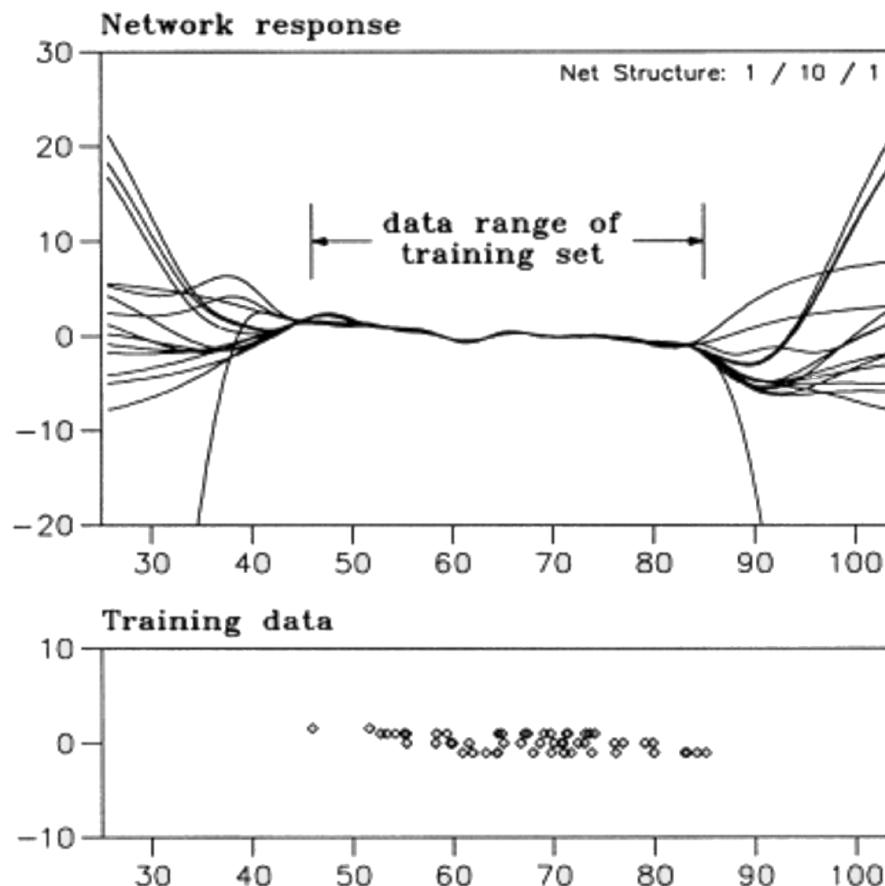


⋮

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

# How Can Machine Learning Interpretability Be Practiced?

By using sensitivity analysis to test machine learning model predictions for accuracy and stability. *If you are using a machine learning model, you should probably be conducting sensitivity analysis.*



# Can Machine Learning Interpretability Be Tested?

Yes.

## ***Simulated data***

You can use simulated data with known characteristics to test explanations. For instance, models trained on totally random data with no relationship between a number of input variables and a prediction target should not give strong weight to any input variable nor generate compelling local explanations or reason codes. Conversely, you can use simulated data with a known signal generating function to test that explanations accurately represent that known function.

[https://github.com/h2oai/mli-resources/tree/master/lime\\_shap\\_treeint\\_compare](https://github.com/h2oai/mli-resources/tree/master/lime_shap_treeint_compare)

## ***Explanation stability under data perturbation***

Trustworthy explanations likely should not change drastically for minor changes in input data. You can set and test thresholds for allowable explanation value changes automatically by perturbing input data. (Explanations or reason code values can also be averaged across a number of models to create more stable explanations.)

## ***Explanation stability with increased prediction accuracy***

If previously known, accurate explanations or reason codes from a simpler linear model are available, you can use them as a reference for the accuracy of explanations from a related, but more complex and hopefully more accurate, model. You can perform tests to see how accurate a model can become before its prediction's reason codes veer away from known standards.

# General Recommendations

- A very direct path to interpretable machine learning today is to train a monotonic GBM with XGBoost and to use Shapley explanations either in XGBoost or with the shap Python package.  
(Or just buy H2O Driverless AI <https://www.h2o.ai/driverless-ai/> ;)
- Use a combination of local and global explanatory techniques.
- Conduct sensitivity analysis and *random data attacks* on all machine learning models.
- Test your explanatory software.
- If possible, use model-specific explanatory techniques to generate reason codes.
- Open source explanation packages seem immature.
- Beware of uninterpretable features.

# LIME Recommendations

- LIME can give an indication of its own trustworthiness using fit statistics.
- LIME can fail, particularly in the presence of extreme nonlinearity or high-degree interactions.
- LIME is difficult to deploy, but there are highly deployable variants, e.g. H2O's K-LIME.
- Note that the intercept in LIME can account for the most important local phenomena.
- Generated LIME samples can contain large proportions of out-of-range data that can lead to unrealistically high or low intercept values.
- Try LIME on discretized input features and on manually constructed interactions.
- Use cross-validation to construct standard deviations or even confidence intervals for reason code values.

# Treeinterpreter and Shapley Recommendations

- Treeinterpreter and Shapley explanations do not give an indication of their own trustworthiness. We can only assume they are trustworthy ...
- Treeinterpreter appears to fail when used with regularized (L1/L2) XGBoost models.
- Due to theoretical support and robust implementation, Shapley explanations may be suitable for regulated applications.

# Questions?

References and resources:

<https://github.com/h2oai/mli-resources>