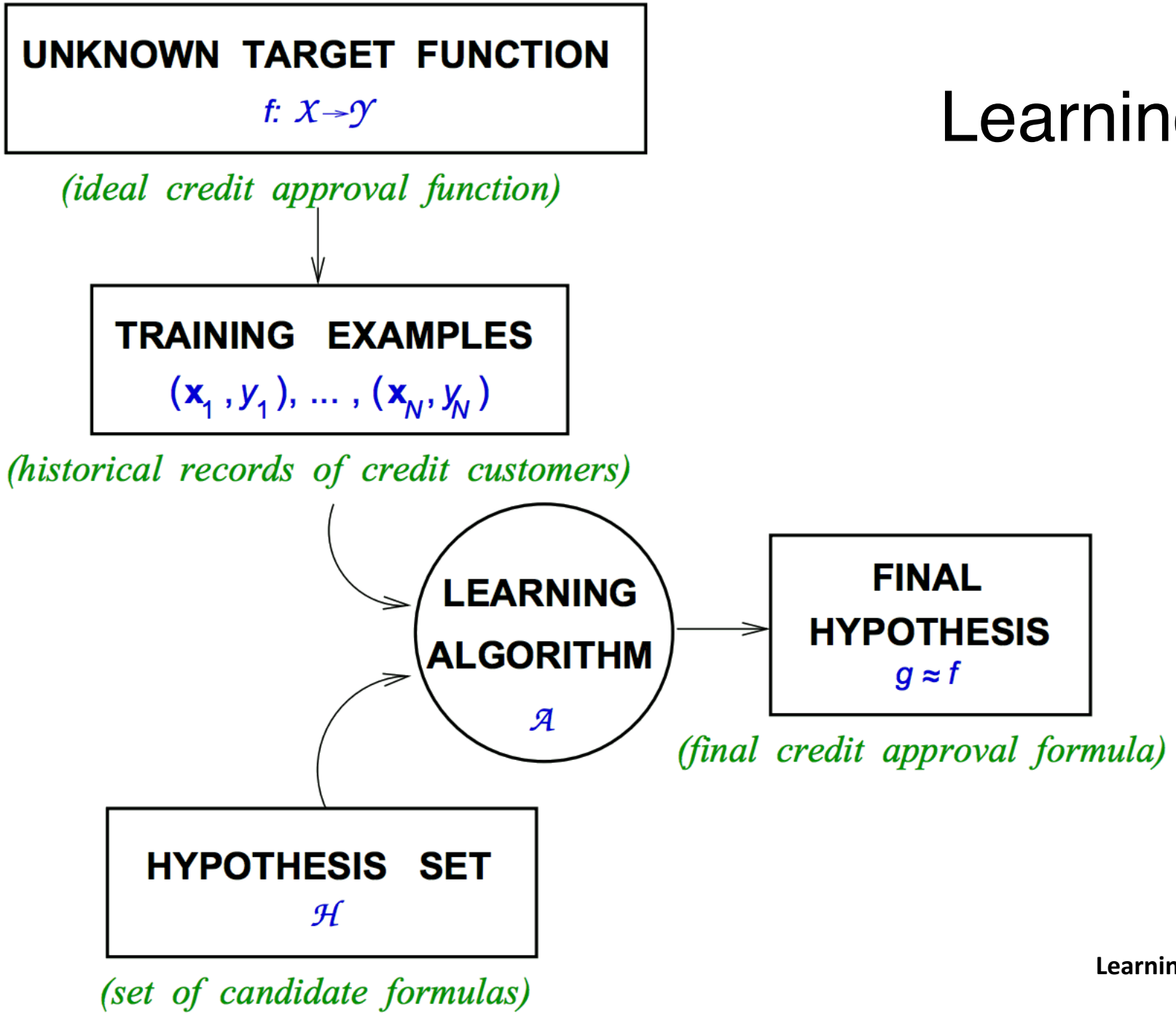
Ideas on Machine Learning Interpretability

Navdeep Gill, H2O.ai

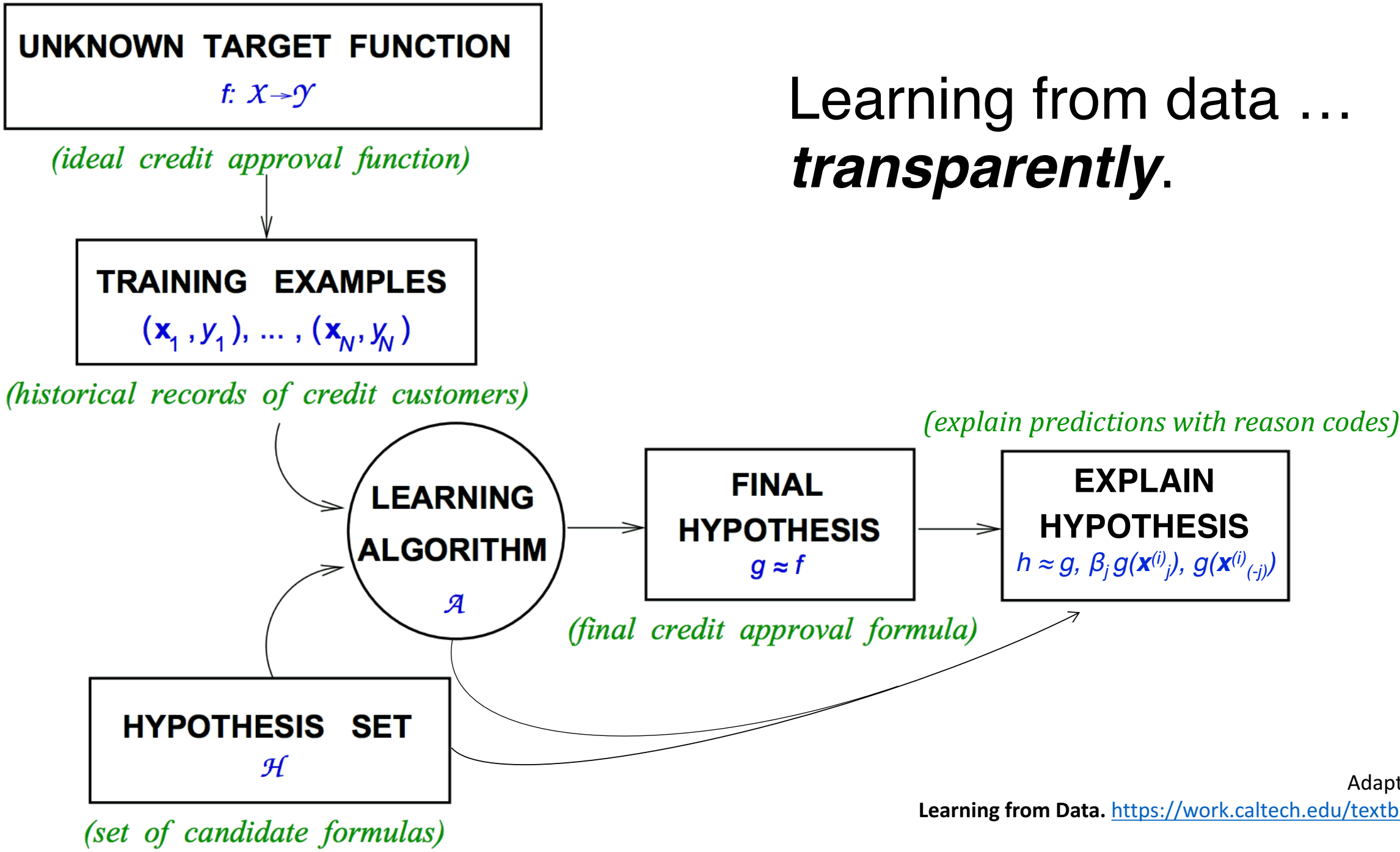
H2O.ai

Big Ideas

Learning from data …



Adapted from:
Learning from Data. https://work.caltech.edu/textbook.html



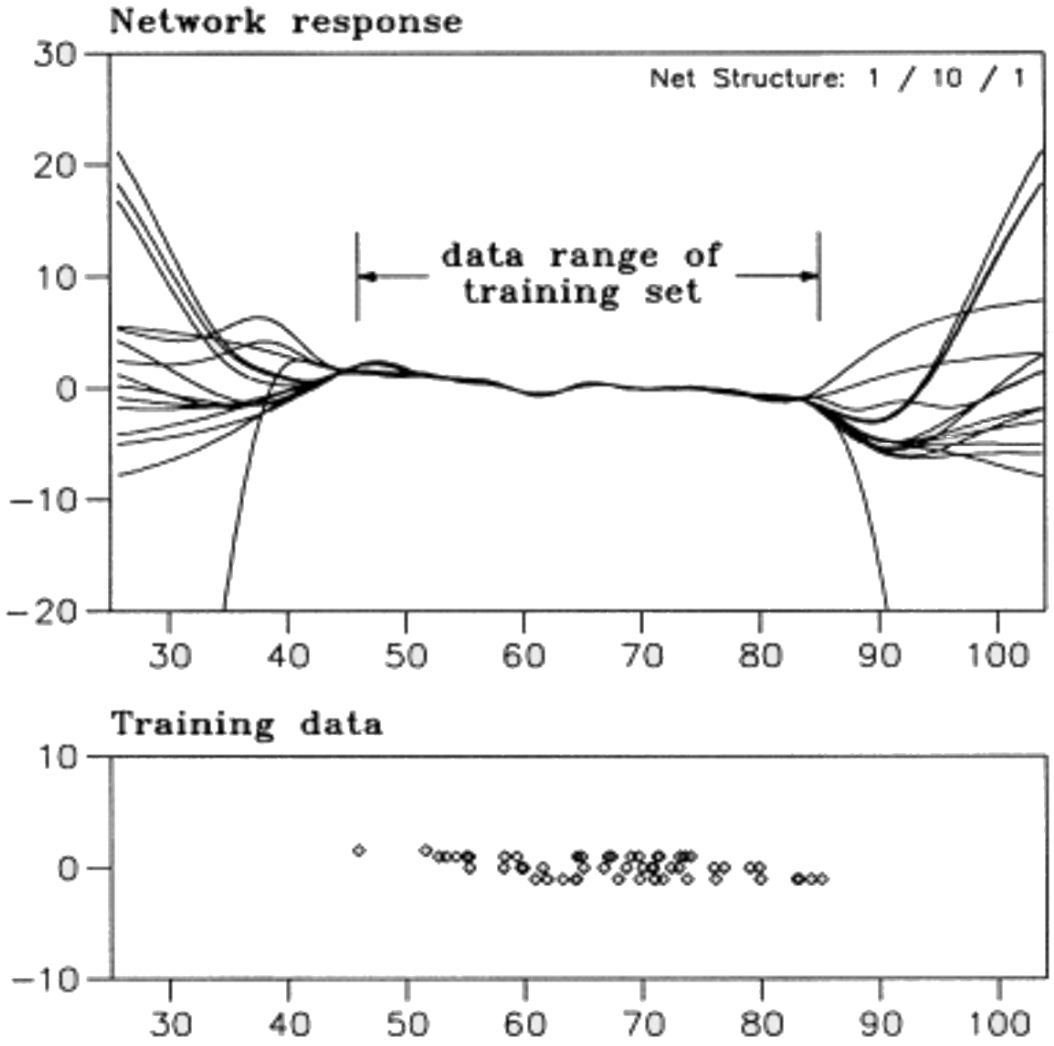
Learning from data … ***transparently***.

Adapted from:
Learning from Data. https://work.caltech.edu/textbook.html

Increasing fairness, accountability, and trust by decreasing unwanted sociological biases

Source: http://money.cnn.com/, Apple Computers

Increasing trust by quantifying prediction variance



Source: http://www.vias.org/tmdatanaleng/

A framework for interpretability

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

(~ Number of parameters/VC dimension)

Scope of interpretability:
Global vs. local

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

Understanding ~ transparency
Trust ~ fairness and accountability

Application domain:
Model-agnostic vs. model-specific

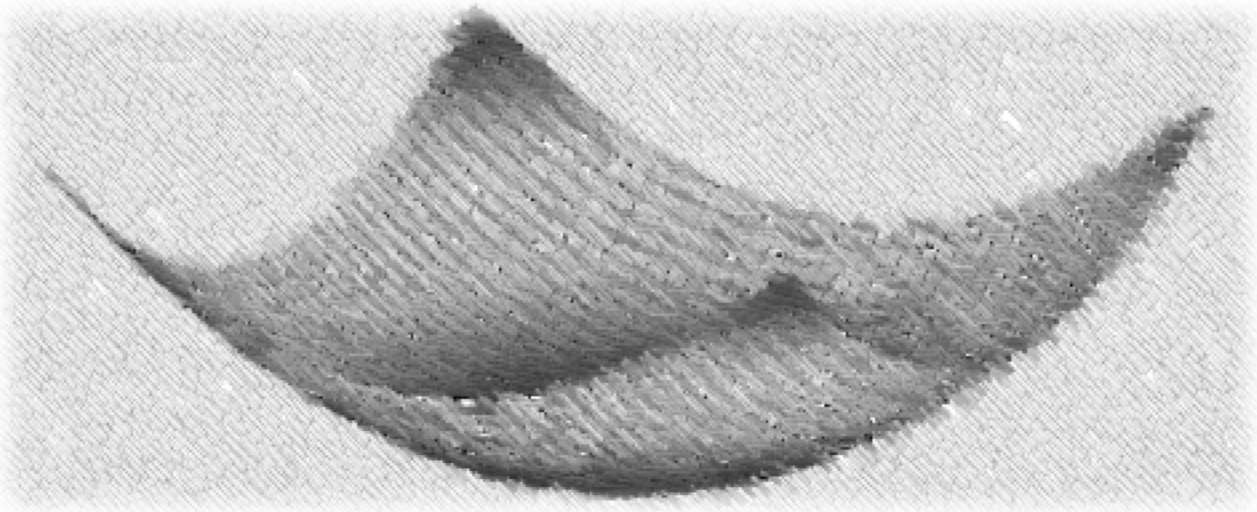
H2O.ai

Big Challenges

Linear Models

Strong model locality

Usually stable models and explanations

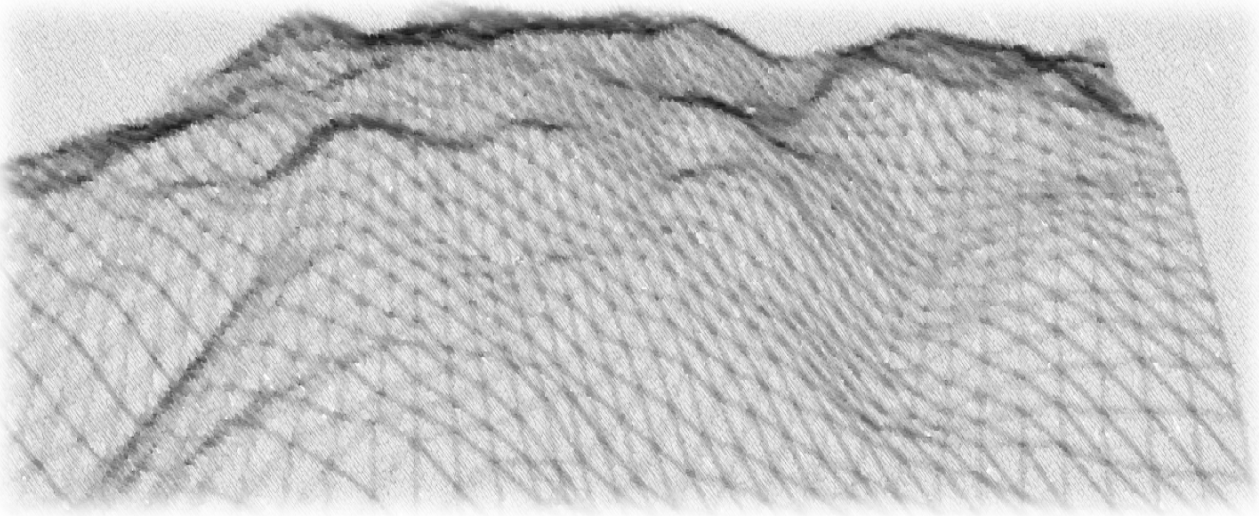


Machine Learning

Weak model locality

Sometimes unstable models and explanations

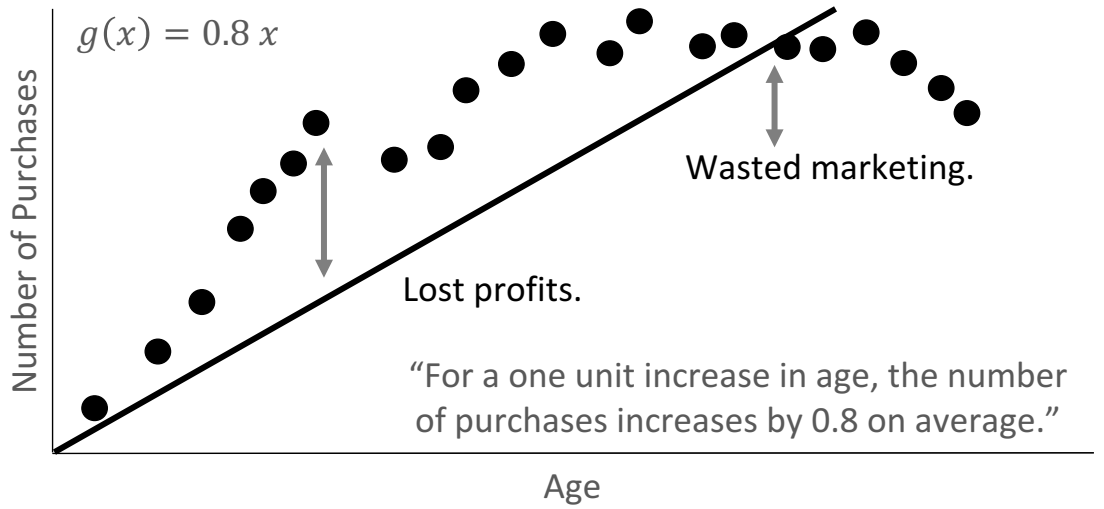
(a.k.a. The Multiplicity of Good Models)



H2O.ai

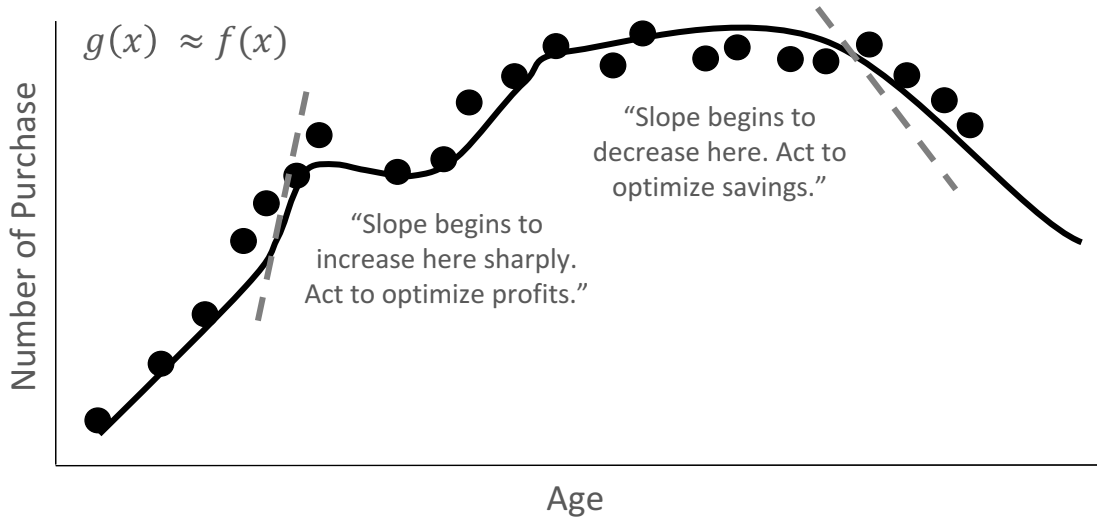
Linear Models

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



Machine Learning

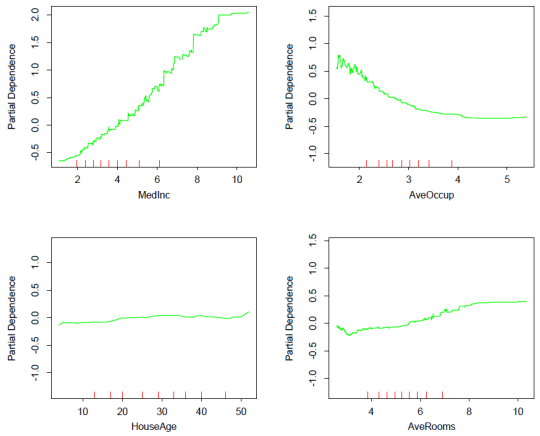
Approximate explanations for ***exact*** models.

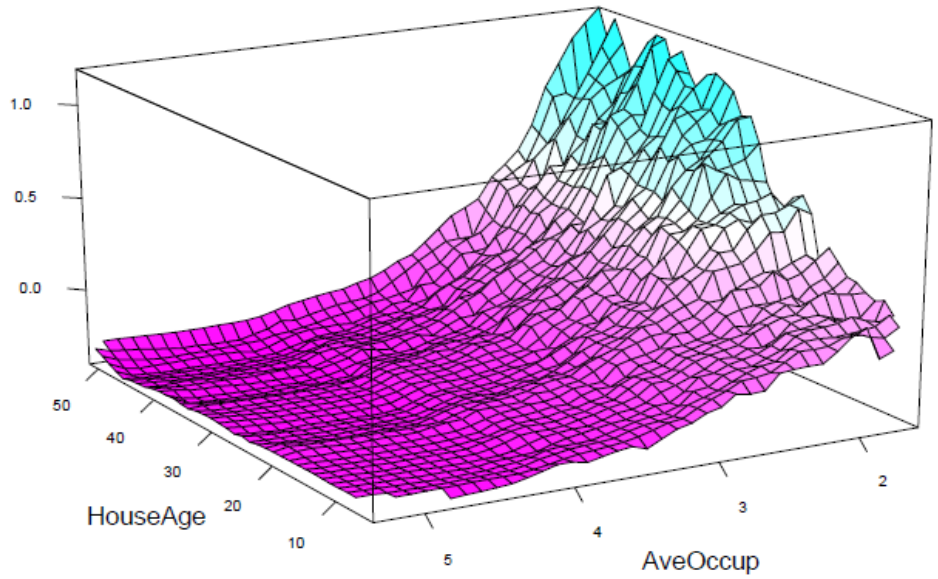


H2O.ai

A Few of Our Favorite Things

Partial dependence plots





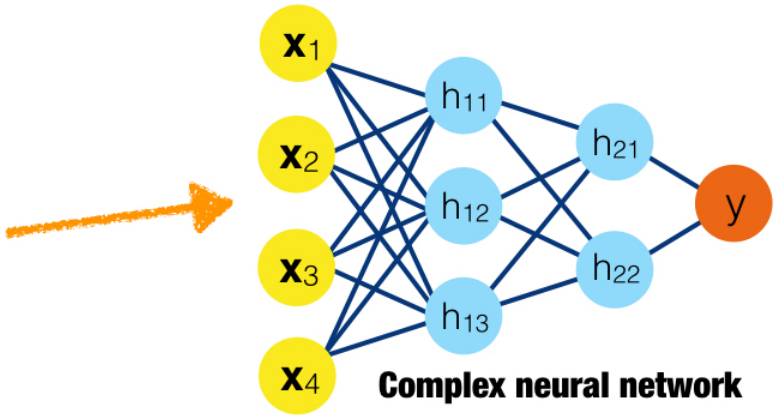
HomeValue ~ MedInc + AveOccup + HouseAge + AveRooms

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

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

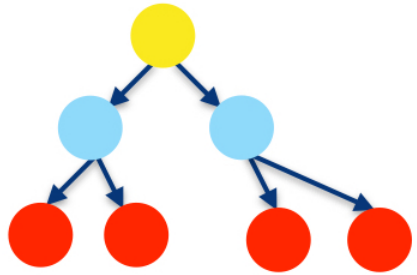
1. Train a complex machine learning model



Complex neural network

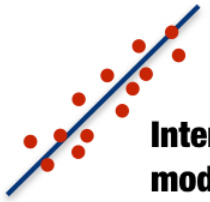
BAD	PREDICTED_BAD	CUSTOMER_DTI	LOAN_PURPOSE	CHANNEL
0	0.47	0.18	MORT	7
1	0.82	0.42	HELOC	10
0	0.18	0.11	MORT	10
0	0.12	0.21	MORT	1

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



Interpretable decision tree

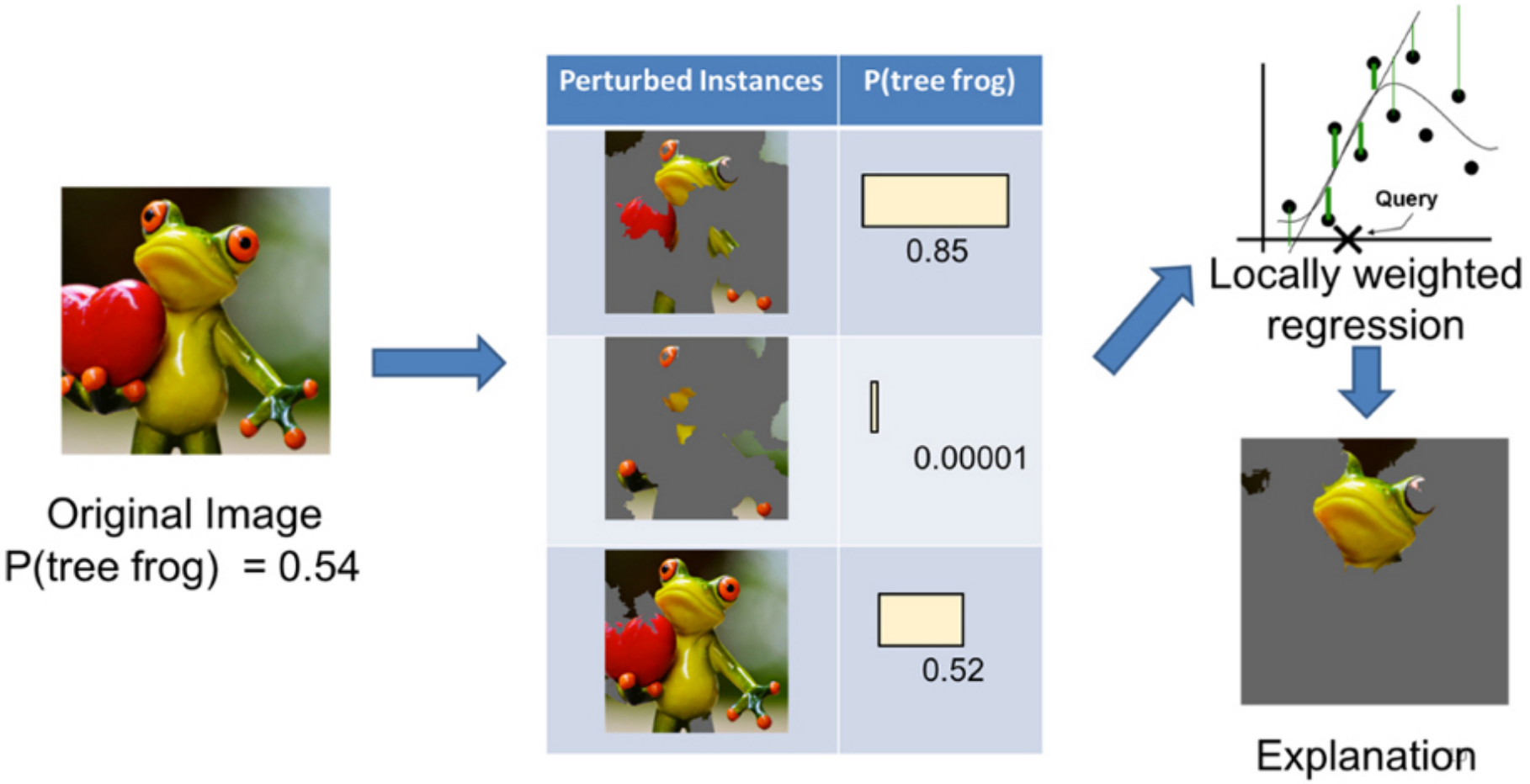
Or



Interpretable linear model

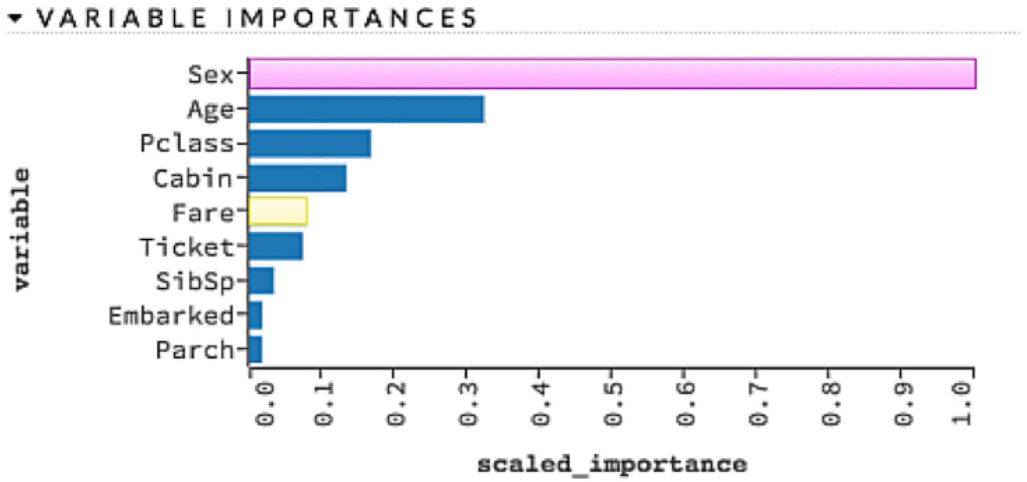
H2O.ai

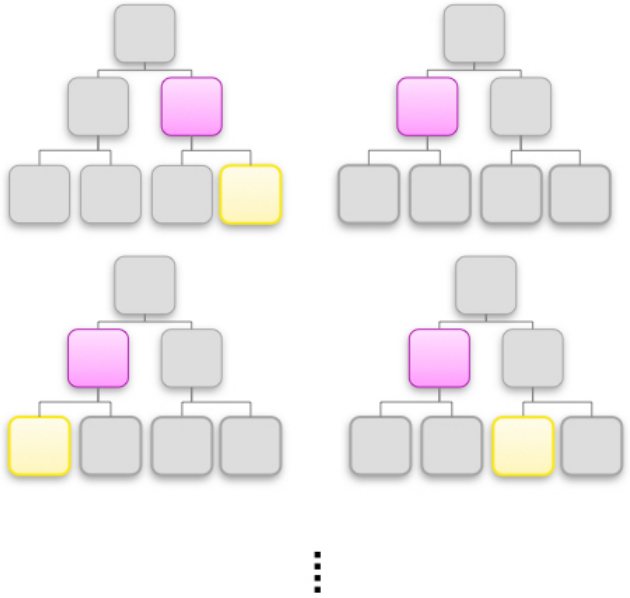
Local interpretable model-agnostic explanations



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

Variable importance measures





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

Sex	Age	…	Fare	ŷ	ŷ(-Sex)	ŷ(-Age)	…	ŷ(-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

H20.ai Driverless AI Demo

H2O.ai

H2O.ai Experiment c8b071

0.9.5+LOCAL_584D771

Show Experiments

TRAINING DATA

DATASET
creditcard_cat.csv

ROWS	COLUMNS	DROPPED COLS	TEST DATASET
24K	25	1	--

TARGET COLUMN

default.payment.next.month

TYPE	COUNT	UNIQUE	FREQ
int	23999	2	18630

STATUS: COMPLETE

INTERPRET THIS MODEL

SCORE ON ANOTHER DATASET

DOWNLOAD (HOLDOUT) TRAINING PREDICTIONS

DOWNLOAD TRANSFORMED TRAINING DATA

DOWNLOAD LOGS

DOWNLOAD SCORING PACKAGE

EXPERIMENT SETTINGS

ACCURACY	TIME	INTERPRETABILITY
5	5	5

CLASSIFICATION

SET RAND. SEED

ENABLE GPUS

SCORER

- R2
- MSE
- RMSE
- RMSLE
- MAE
- GINI
- AUC
- LOGLOSS

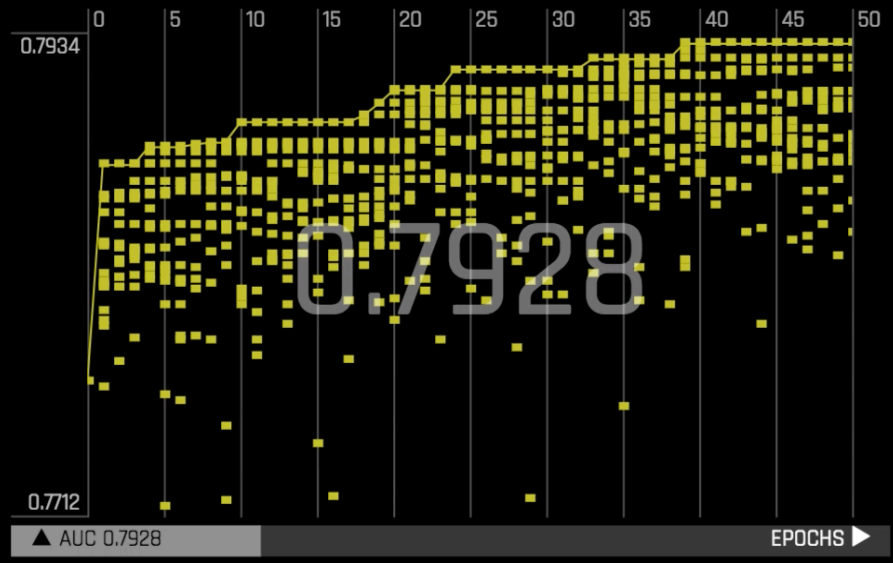
CPU / MEMORY

Trace

CPU

MEM

ITERATION SCORES - INTERNAL VALIDATION



VARIABLE IMPORTANCE

6_Freq__PAY_0	1.00
50_CV_TE_PAY_5_PAY_3_0	0.66
29_TruncSVD_PAY_AMT2_BILL_AMT1_BILL_AMT5_PAY_AMT1...	0.64
32_TruncSVD_BILL_AMT5_PAY_AMT2_LIMIT_BAL_0	0.55
36_TruncSVD_BILL_AMT1_BILL_AMT4_LIMIT_BAL_1	0.43
36_TruncSVD_BILL_AMT1_BILL_AMT4_LIMIT_BAL_0	0.41
46_TruncSVD_PAY_AMT6_BILL_AMT2_0	0.33
23_LIMIT_BAL	0.29
21_BILL_AMT1	0.29
37_TruncSVD_PAY_AMT1_PAY_AMT6_PAY_AMT3_0	0.29
17_PAY_AMT2	0.28
5_CV_TE_PAY_2_0	0.24
16_PAY_AMT1	0.24
44_CV_CatNumEnc_PAY_3__PAY_AMT5_mean	0.23

GPU USAGE

GPU1

GPU2

GPU3

GPU4

GPU5

GPU6

GPU7

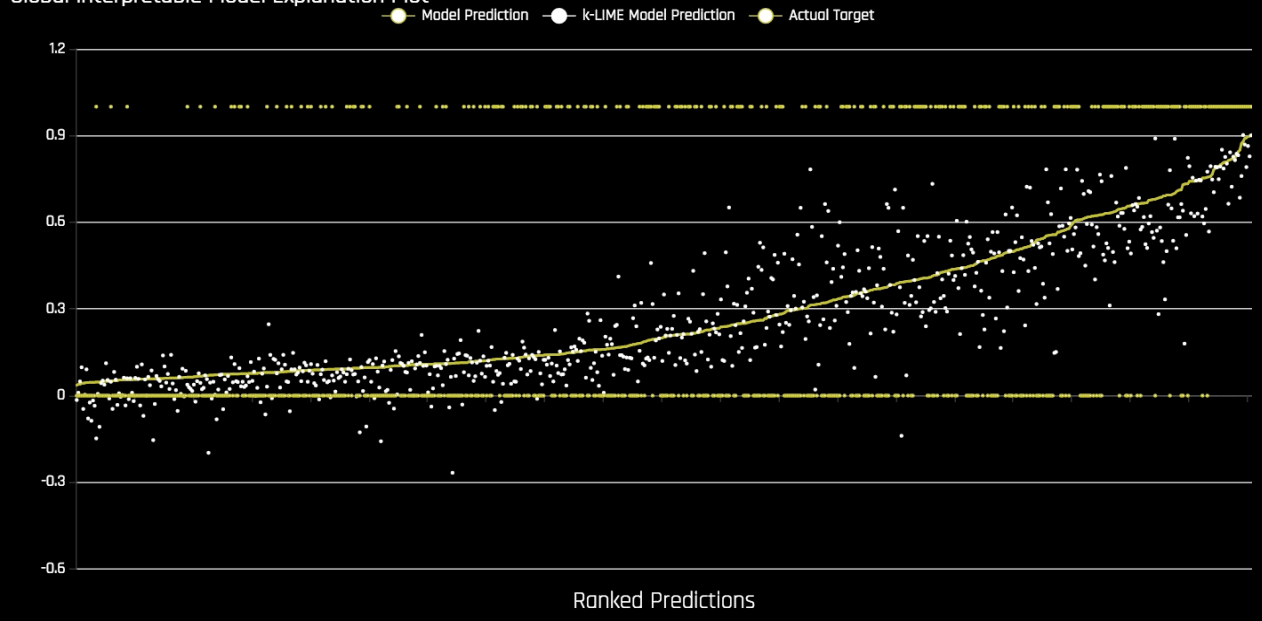
GPU8

Single Row Lookup

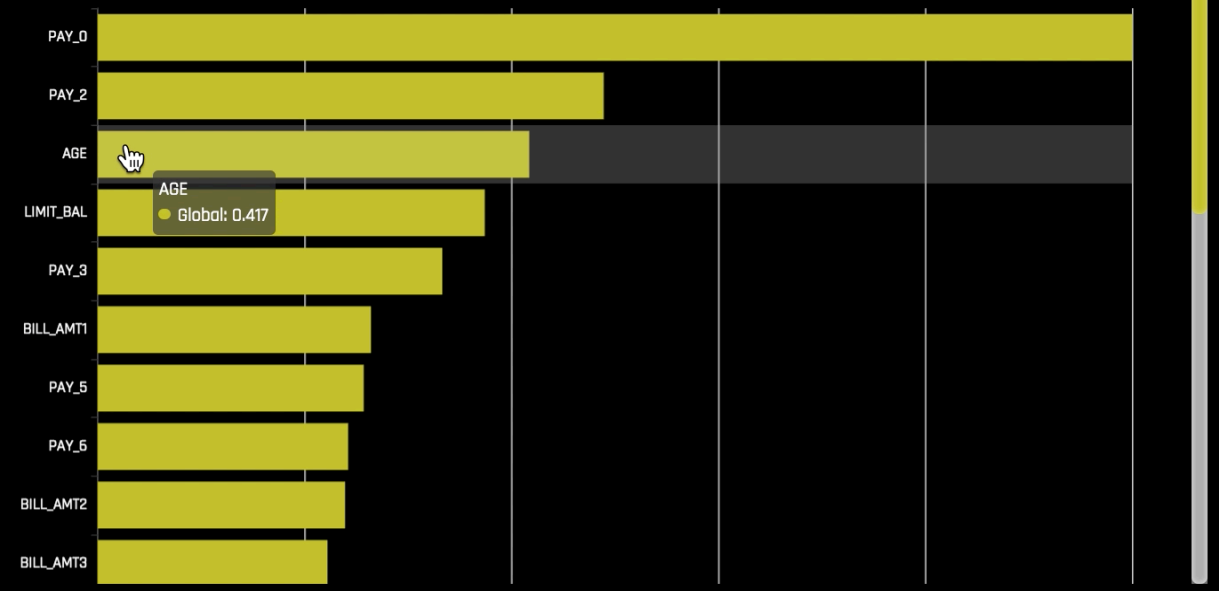
Column: H2O Frame Row # Value: SEARCH Plot: Global

EXPERIMENT EXPLANATIONS

Global Interpretable Model Explanation Plot

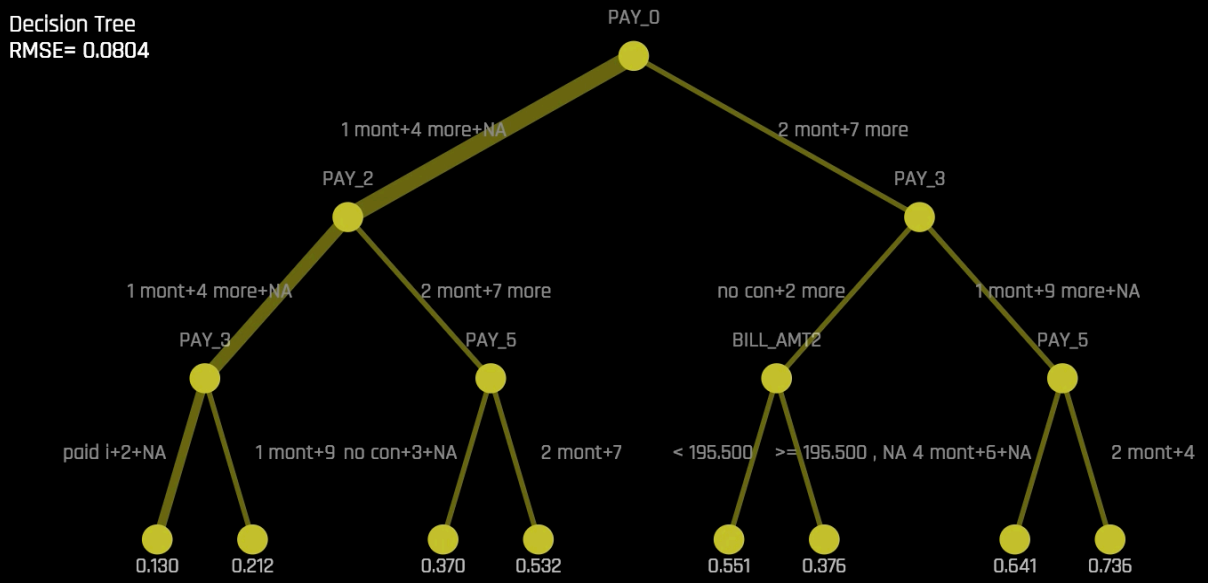


Variable Importance



Decision Tree

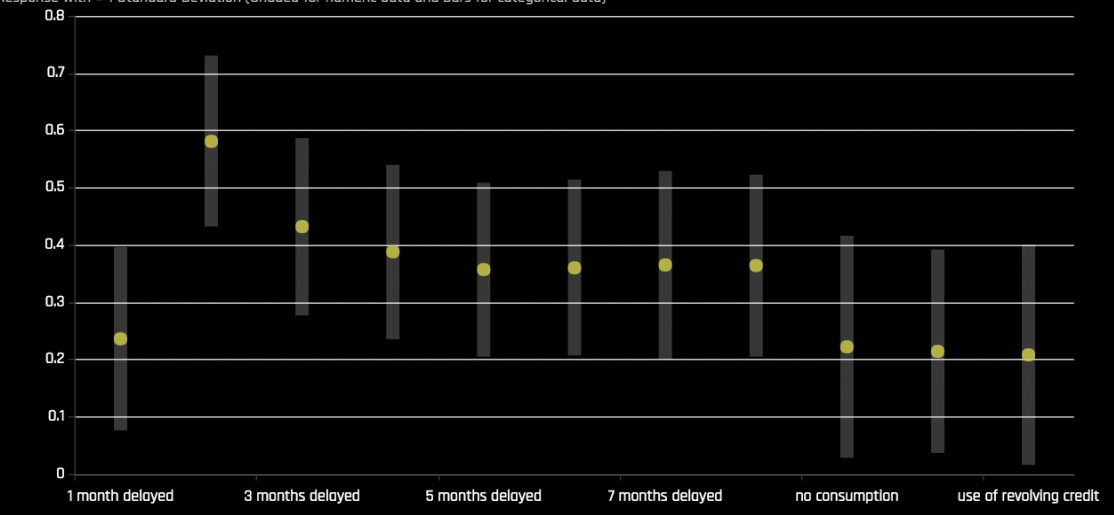
RMSE= 0.0804



Partial Dependence

PAY_0

Mean Response with ± 1 Standard Deviation (Shaded for numeric data and bars for categorical data)

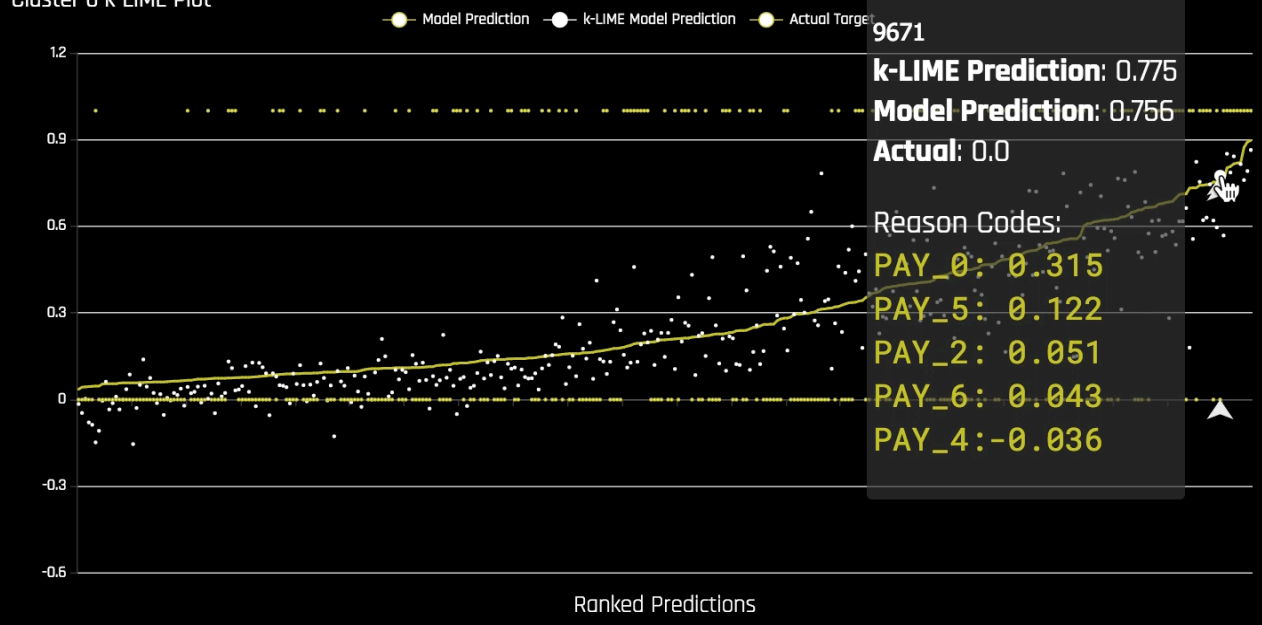


Single Row Lookup

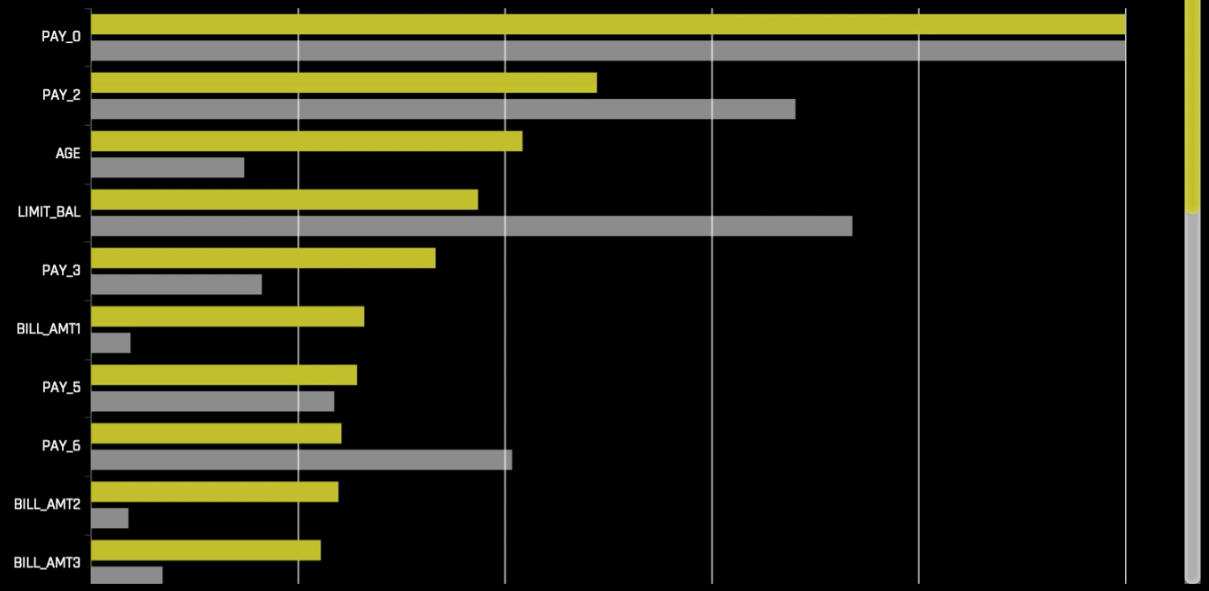
Column: H2O Frame Row # Value: 9671 SEARCH Plot: Cluster0

EXPERIMENT EXPLANATIONS

Cluster 0 k-LIME Plot

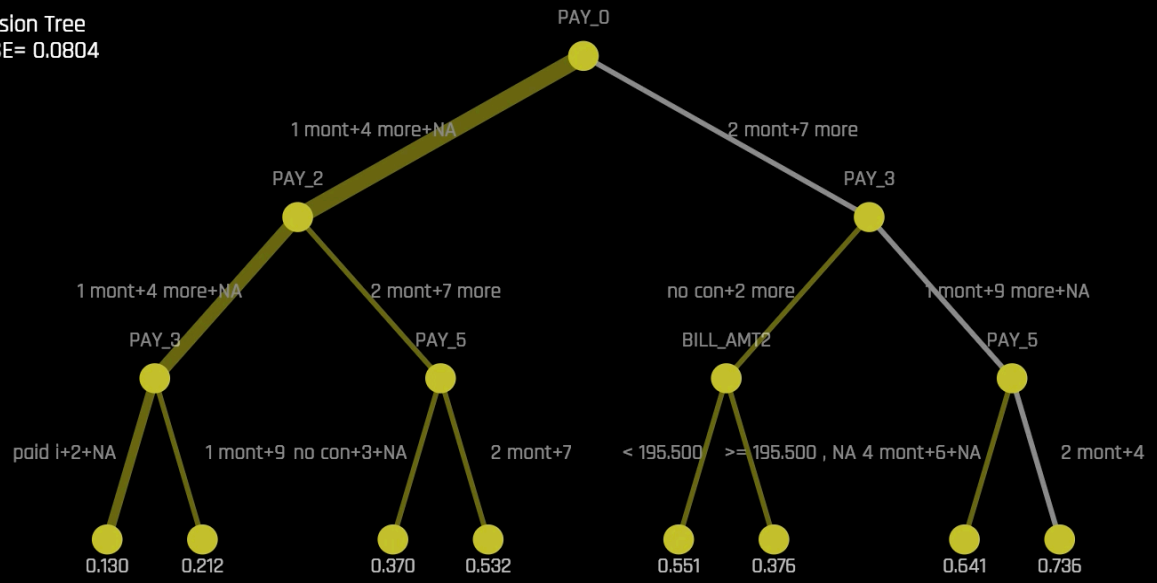


Variable Importance



Decision Tree

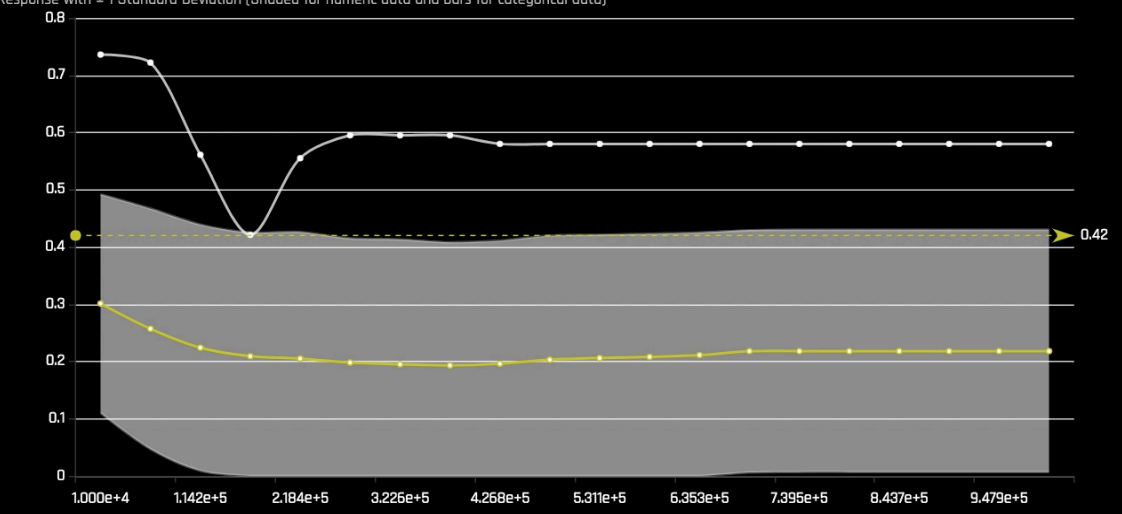
RMSE= 0.0804



Partial Dependence

LIMIT_BAL

Mean Response with ± 1 Standard Deviation (Shaded for numeric data and bars for categorical data)



CLOSE

Actual and Predicted Values

default.payment.next.month (Actual)	0
Model Prediction Value	0.7557
k-LIME Prediction Value	0.7746
k-LIME Prediction Accuracy (%)	97.5%

Local Reason Codes

k-LIME Local Attributions	Variable	with value	is associated with default.payment.next.month	
Top Positive Local Attributions				
	PAY_0	2	increase of	0.32
	PAY_5	7	increase of	0.12
	PAY_2	2	increase of	0.05
Skipped 7 additional attributions, click to view all ...				
Top Negative Local Attributions				
	PAY_4	7	decrease of	0.04
	LIMIT_BAL	170000	decrease of	0.03
	MARRIAGE	NaN	decrease of	0.01
Skipped 10 additional attributions, click to view all ...				
k-LIME Cluster Baseline Attribution				0.2801
k-LIME Prediction (k-LIME Local + Cluster Baseline Attribtions)				0.7746

Cluster 0 Reason Codes

k-LIME explains **89.14%** in **default.payment.next.month** for this cluster.

	Variable	with value	is associated with default.payment.next.month	
Top Positive Cluster Attributions				
	PAY_5	8	increase of	0.413
	PAY_2	6	increase of	0.407
	PAY_0	2	increase of	0.315

Resources

Machine Learning Interpretability with H2O Driverless AI

https://www.h2o.ai/wp-content/uploads/2017/09/MLI.pdf

Ideas on Interpreting Machine Learning

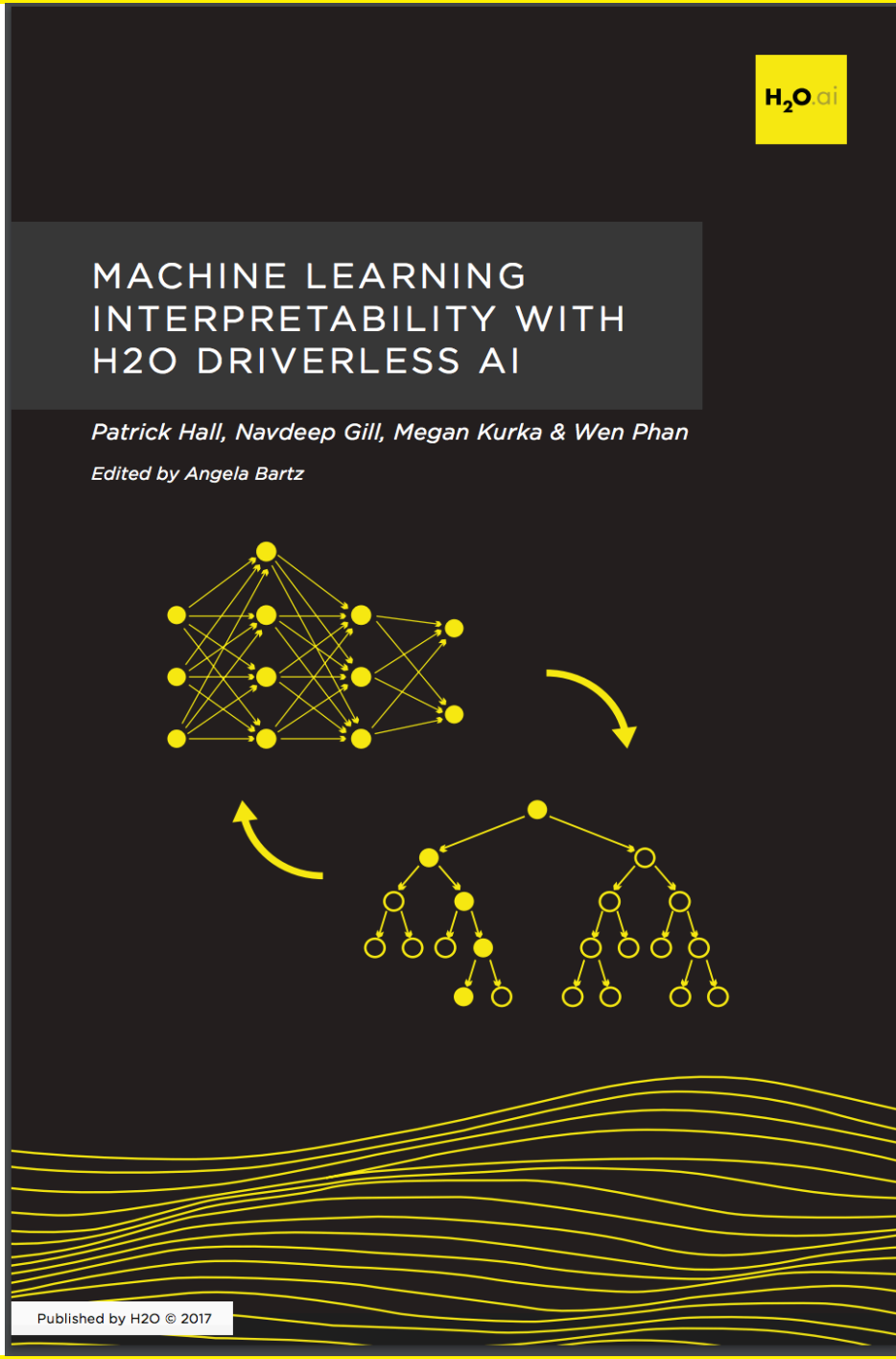
https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning

FAT/ML

http://www.fatml.org/

MLI Resources

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



Questions?