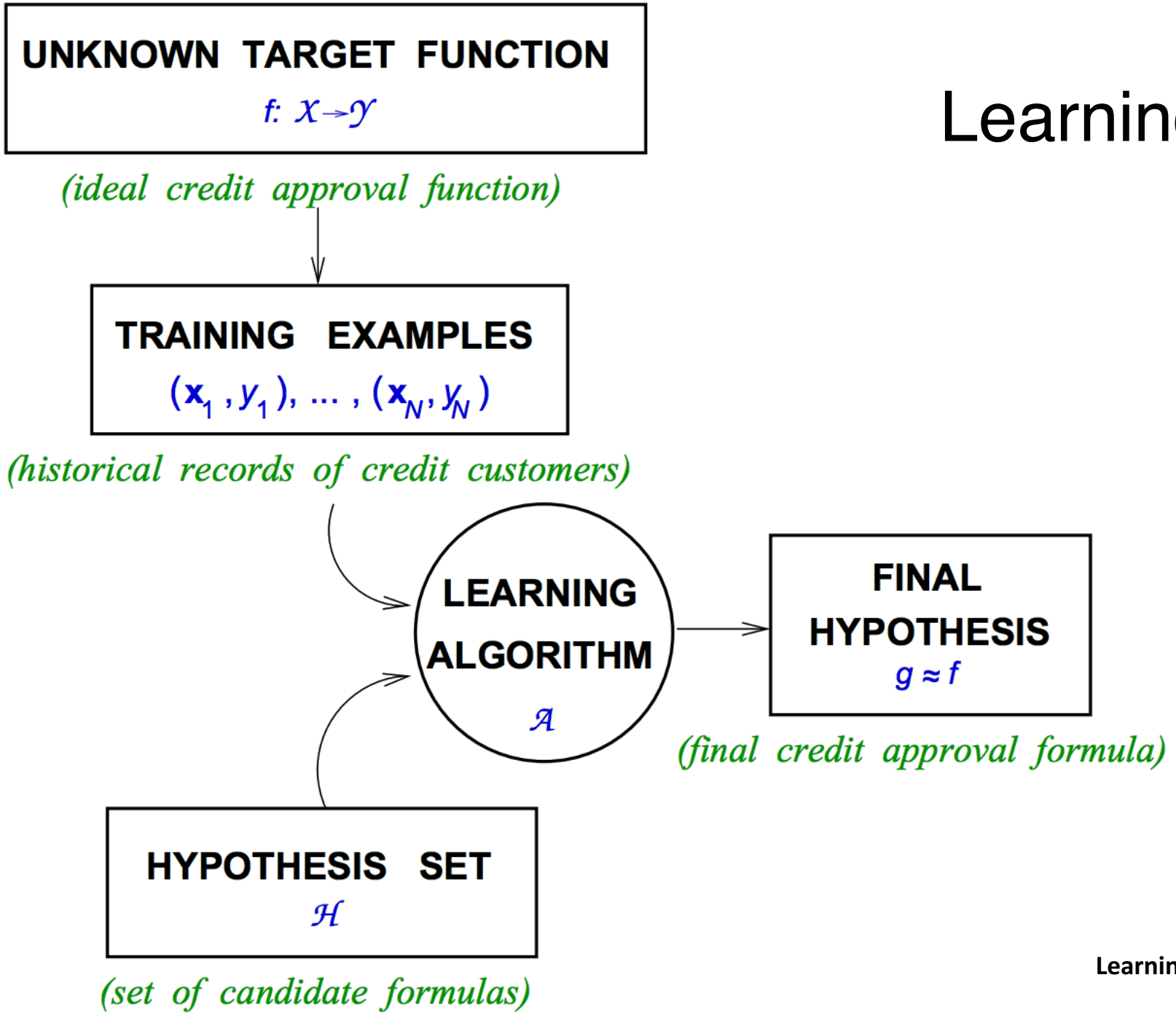
Ideas on Machine Learning Interpretability

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

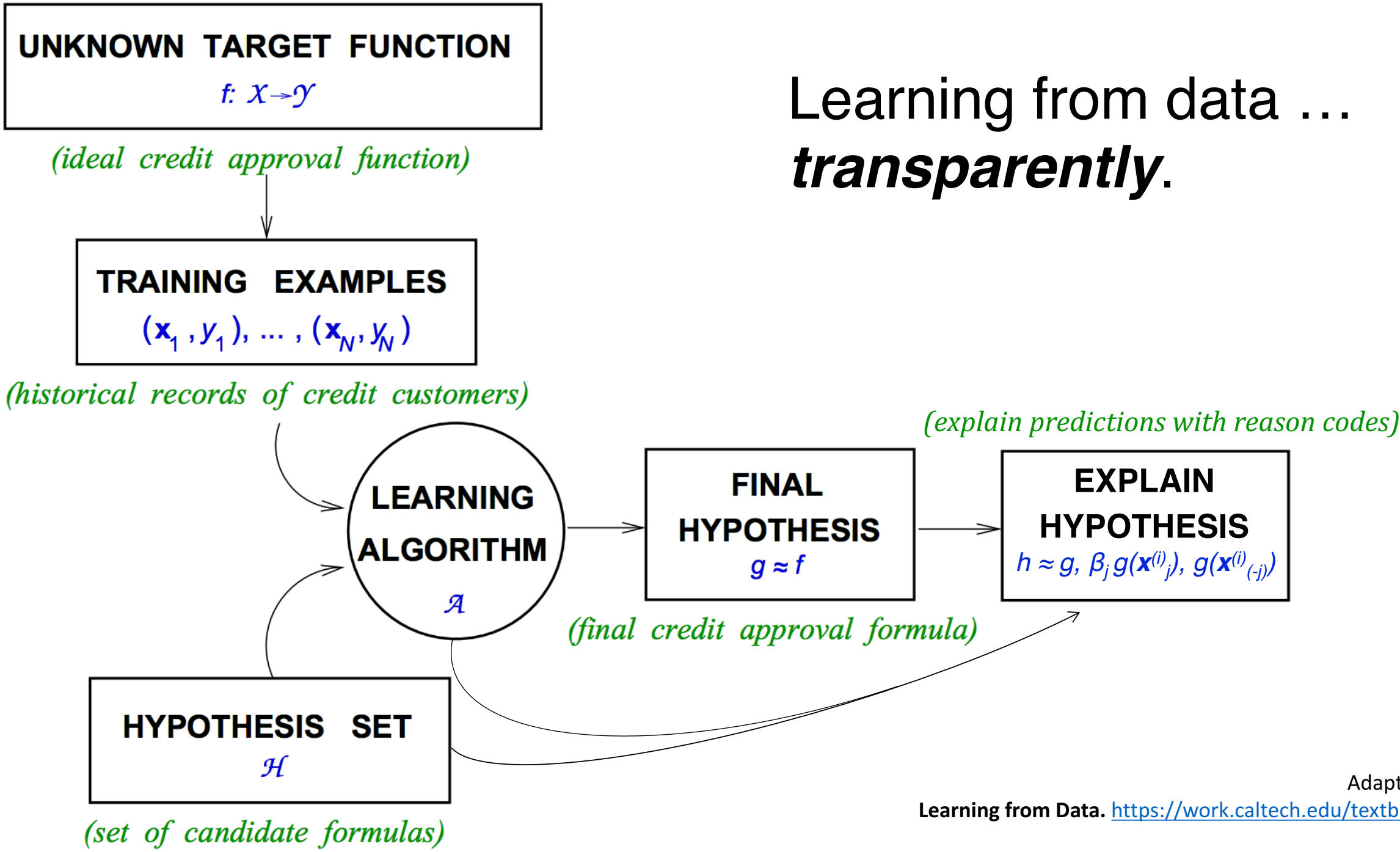
H2O.ai

Big Ideas

Learning from data …



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



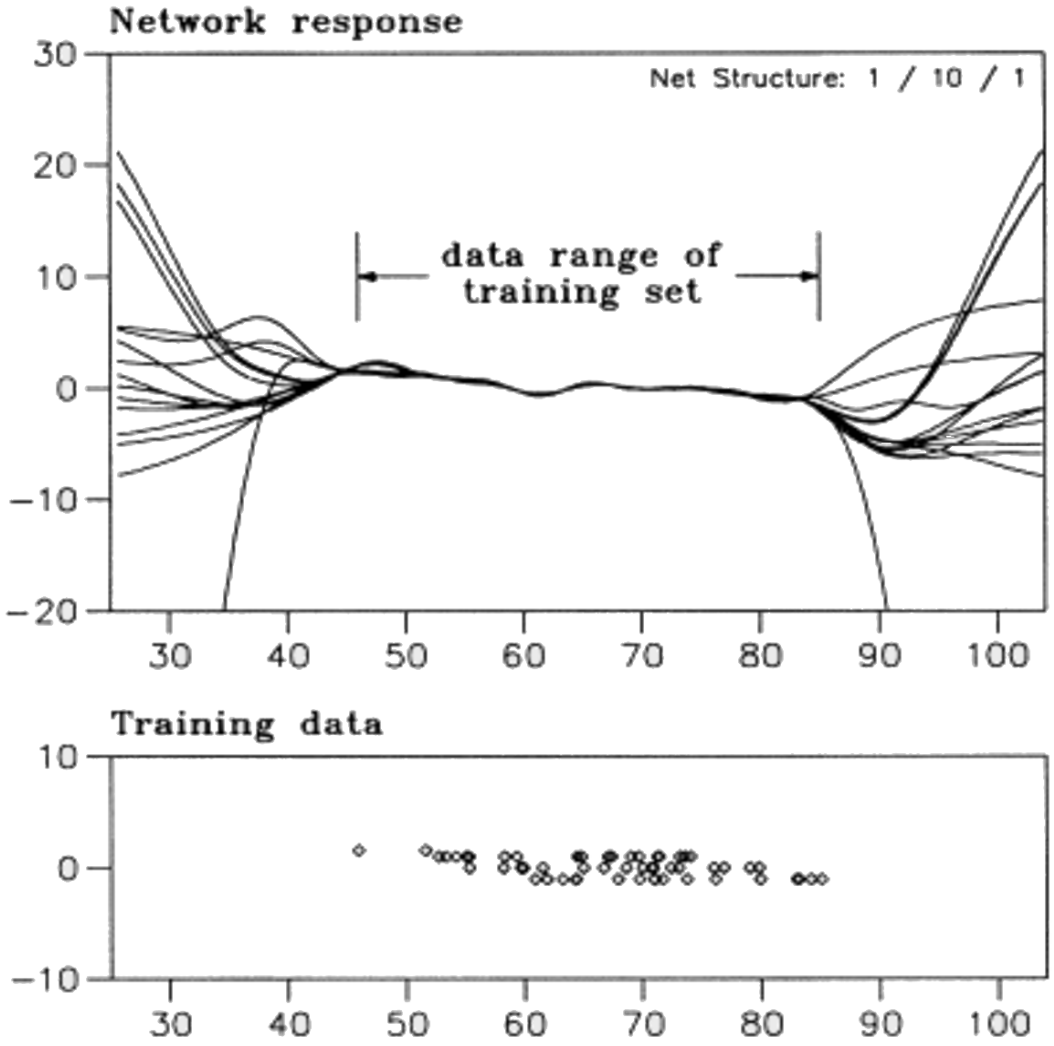
Learning from data … ***transparently***.

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

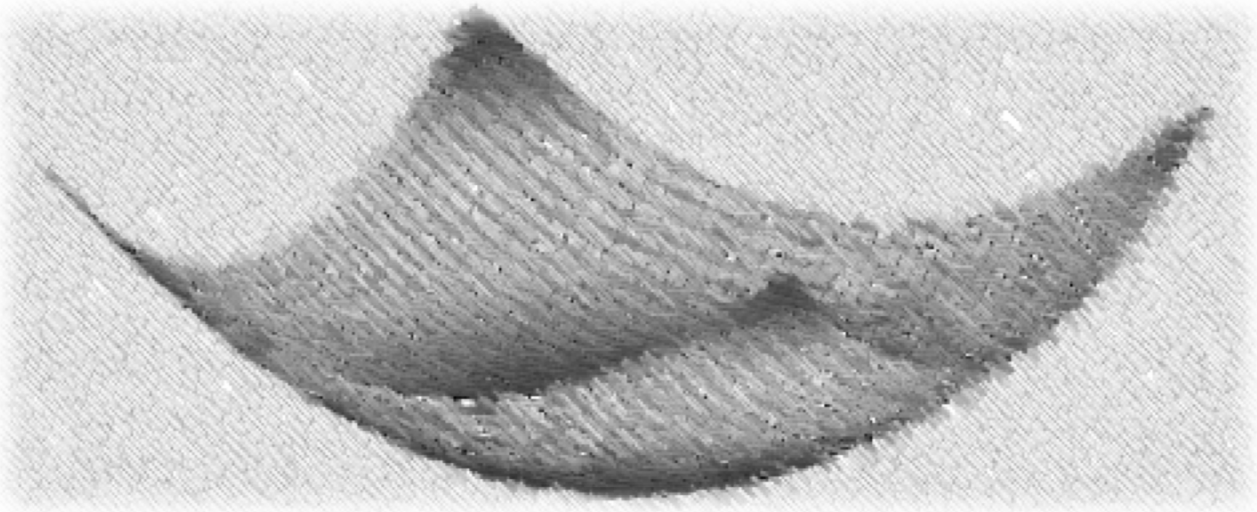
Big Challenges

H2O.ai

Linear Models

Strong model locality

Usually stable models and explanations

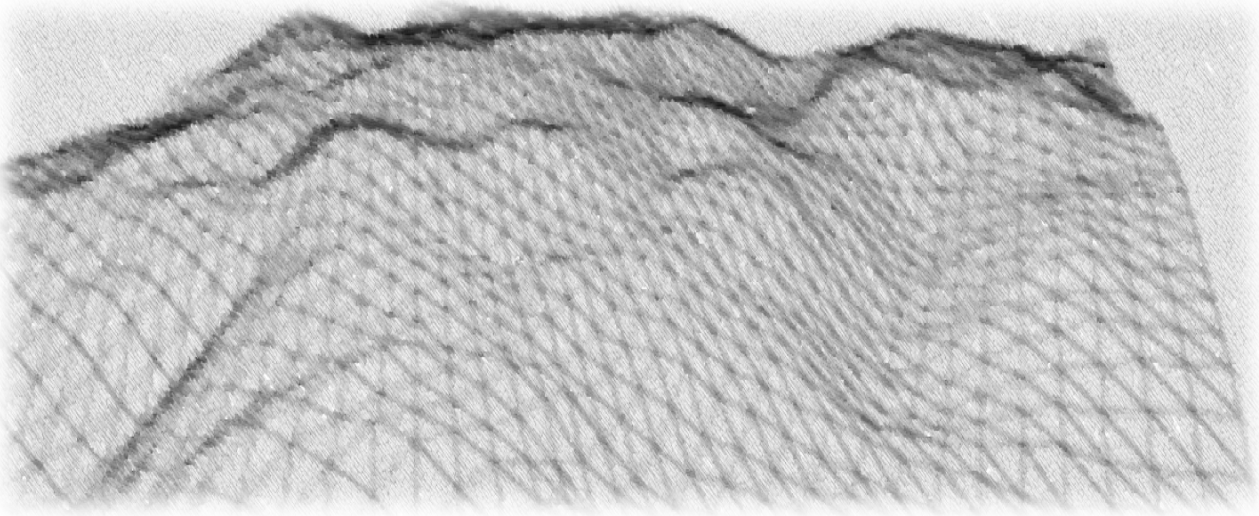


Machine Learning

Weak model locality

Sometimes unstable models and explanations

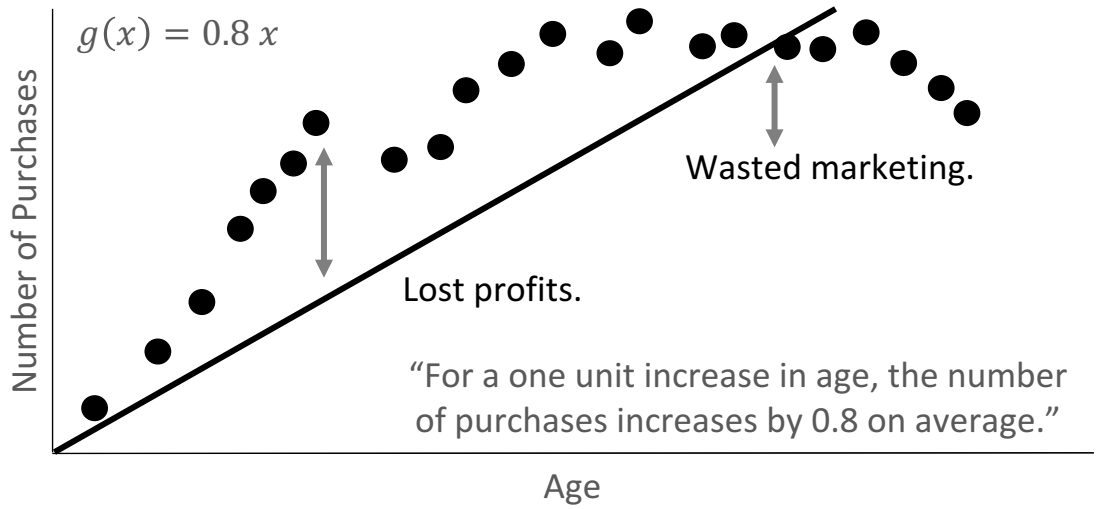
(a.k.a. The Multiplicity of Good Models)



H2O.ai

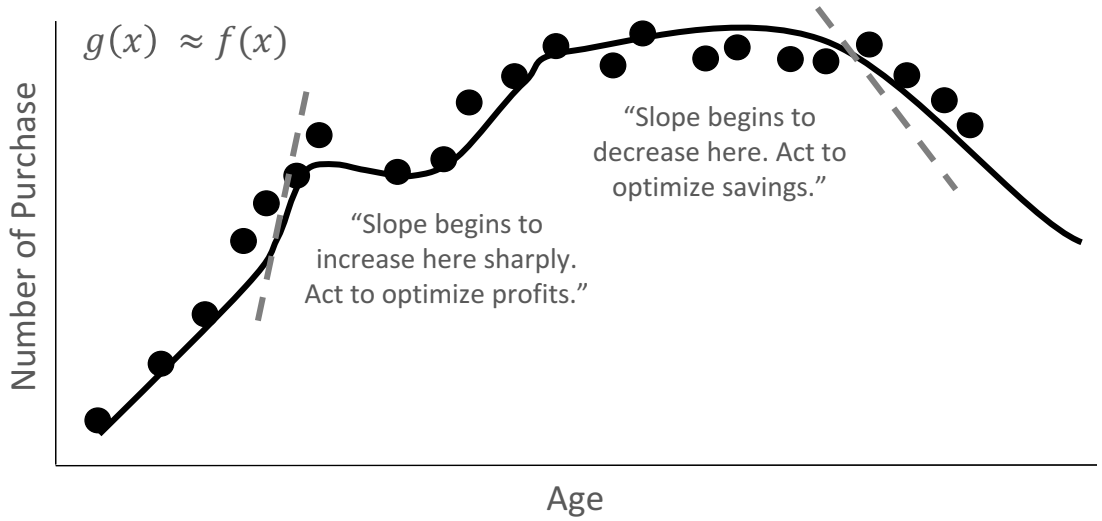
Linear Models

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



Machine Learning

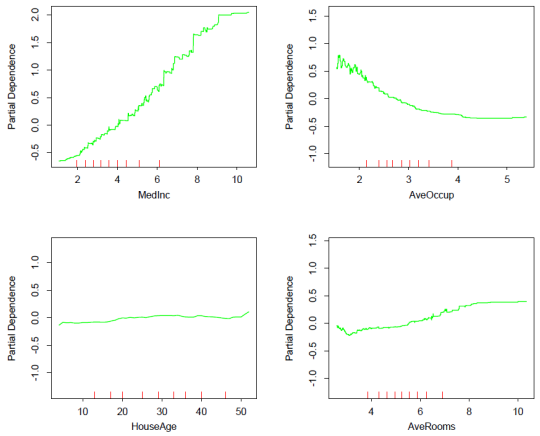
Approximate explanations for ***exact*** models.

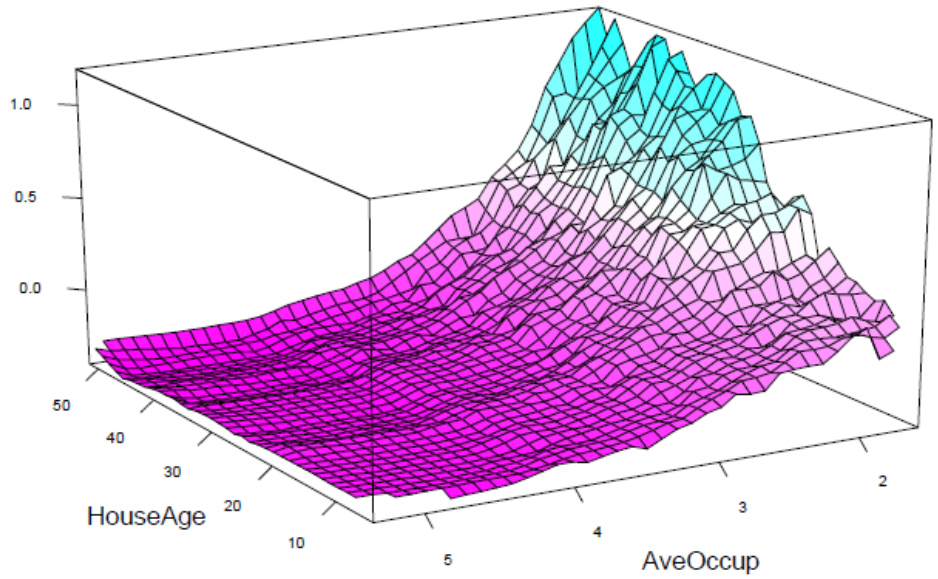


H2O.ai

A Few of Our Favorite Things

Partial dependence plots





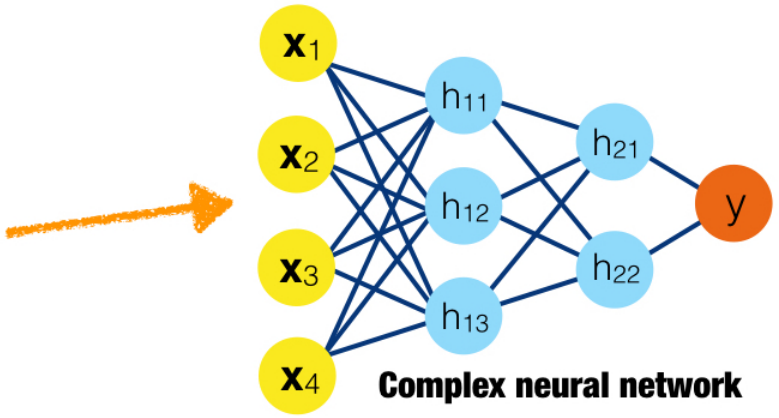
HomeValue ~ MedInc + AveOccup + HouseAge + AveRooms

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

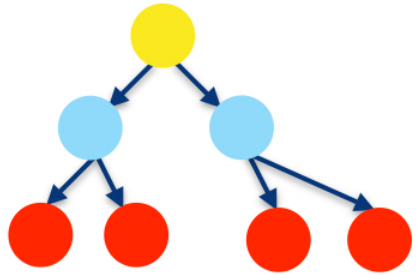
1. Train a complex machine learning model



Complex neural network

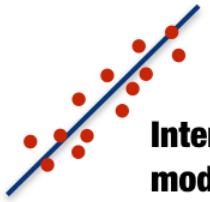
BAD	PREDICTED_BAD	CUSTOMER_DTI	LOAN_PURPOSE	CHANNEL
0	0.47	0.18	MORT	7
1	0.82	0.42	HELOC	10
0	0.18	0.11	MORT	10
0	0.12	0.21	MORT	1

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



Interpretable decision tree

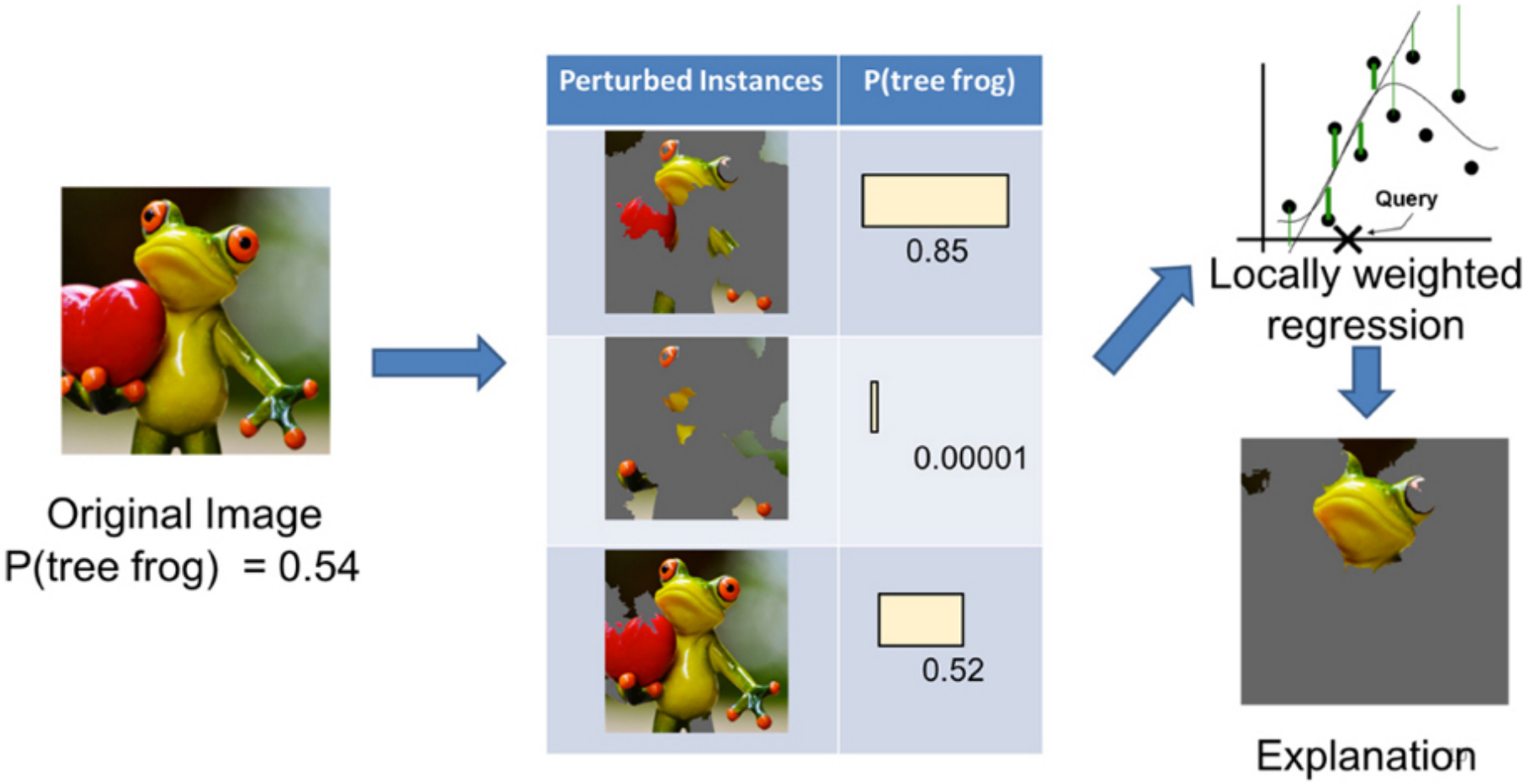
Or



Interpretable linear model

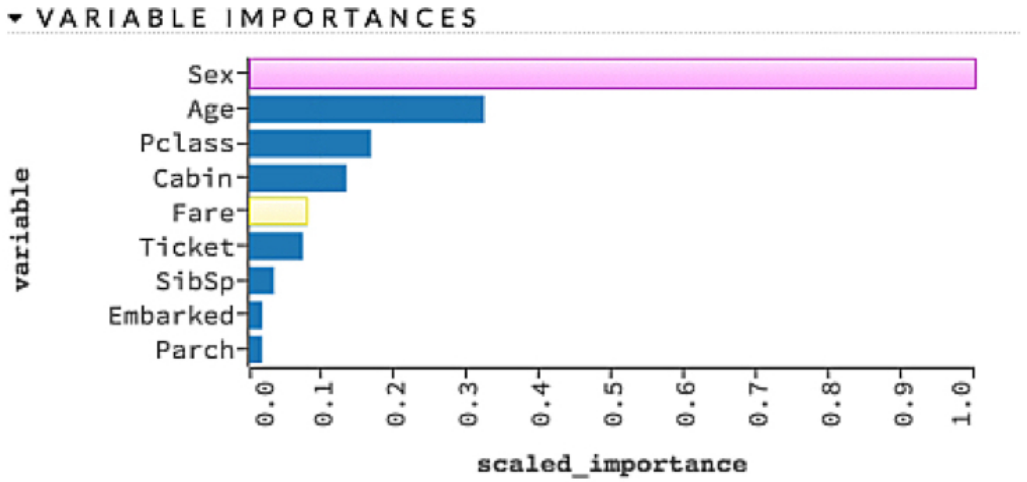
H2O.ai

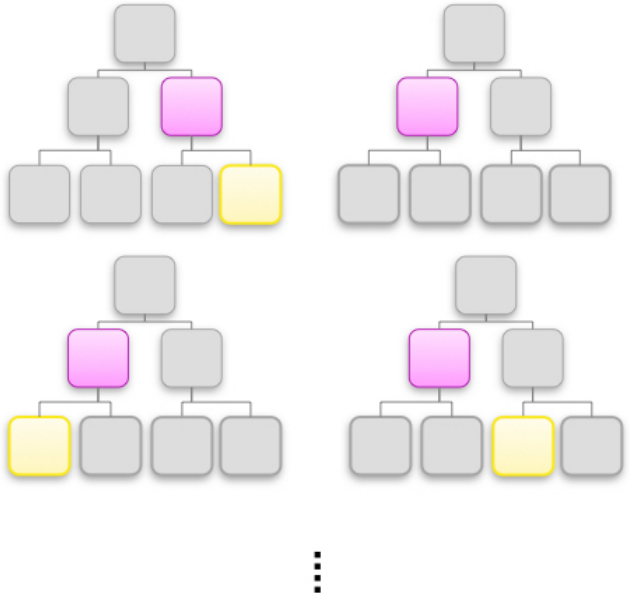
Local interpretable model-agnostic explanations



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

Resources

Machine Learning Interpretability with H2O Driverless AI

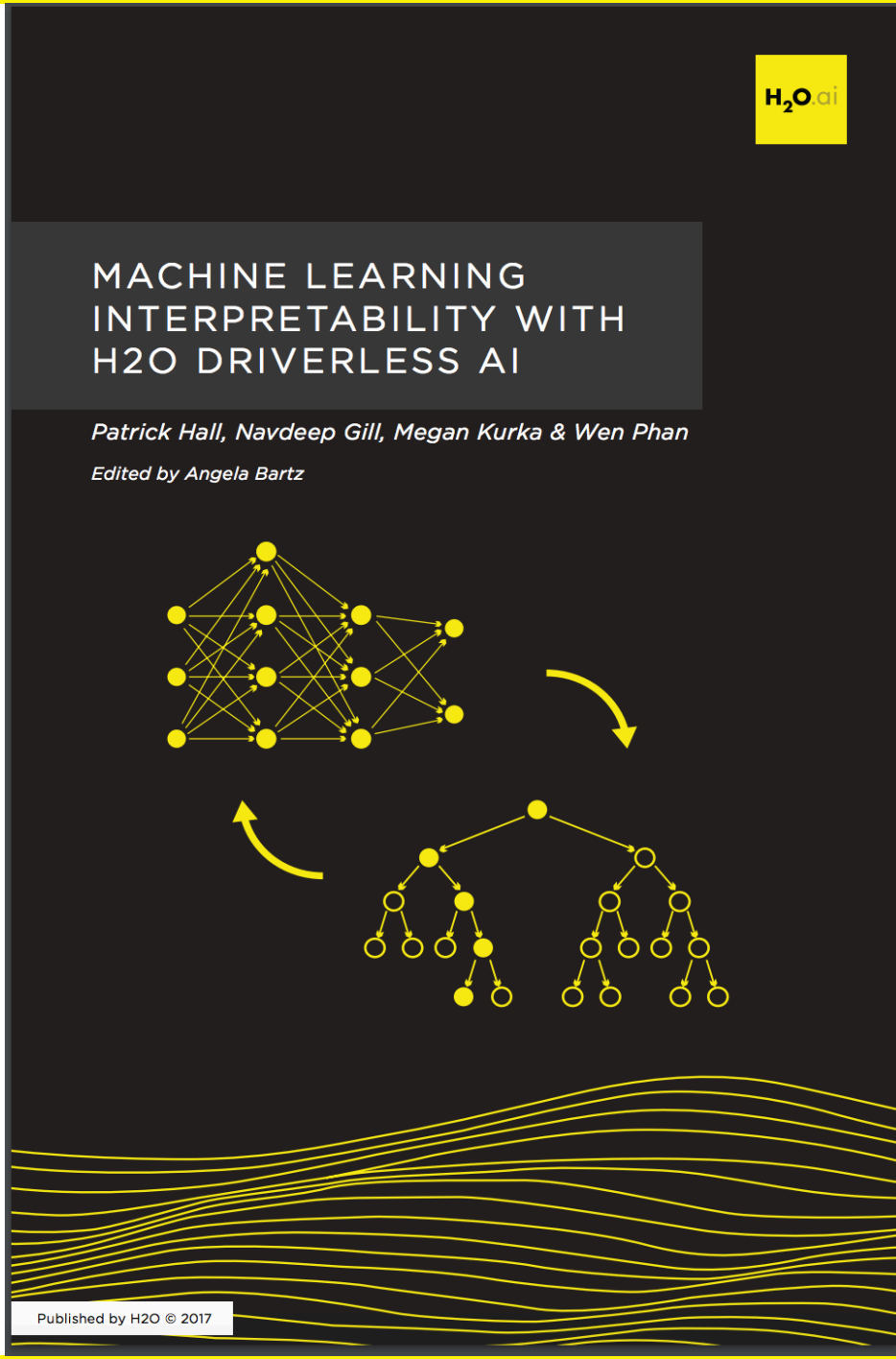
https://www.h2o.ai/wp-content/uploads/2017/09/MLI.pdf
(OR come by the booth!!)

Ideas on Interpreting Machine Learning

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

FAT/ML

http://www.fatml.org/



Questions?