

# SHAP: Interpreting ML Models with IML

...

March 7, 2020  
SatRday



# Who am I?



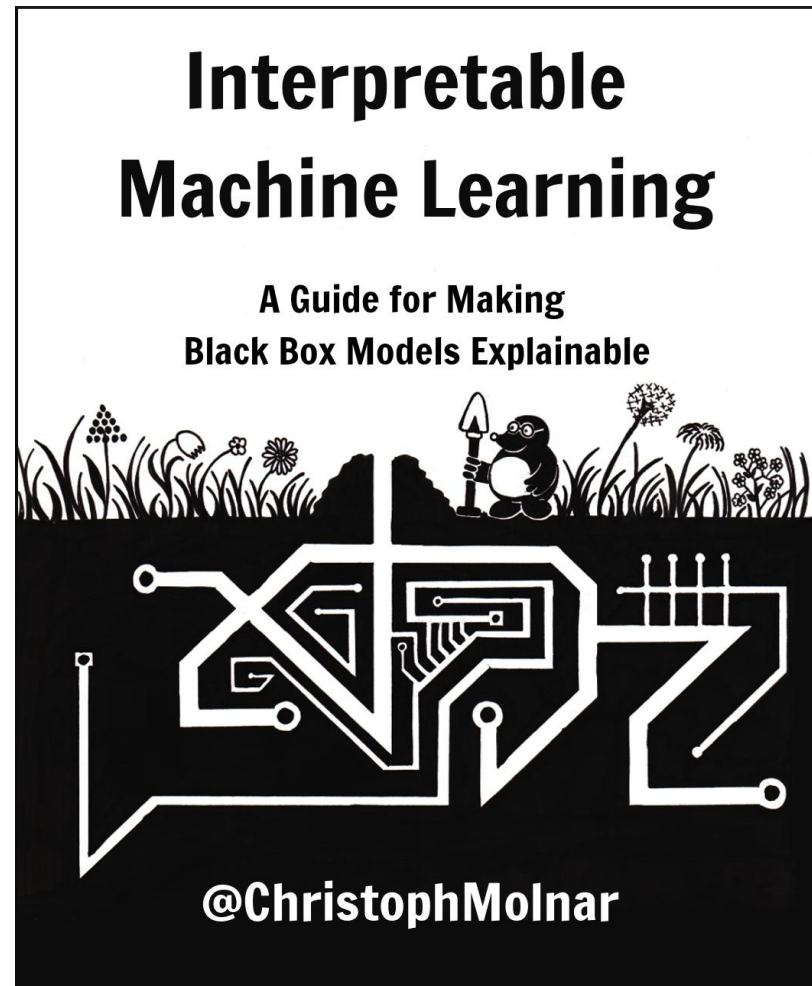
## Drikus du Toit

Data Scientist  
Decision Science  
Capitec Bank



# Resource

Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019.  
<https://christophm.github.io/interpretable-ml-book/>.



**What is ML interpretability and why is it important?**

## The Client

Applies for loan, gets rejected.

### Client Questions:

- Why did I not get the loan?
- What should I do to improve my credit score?

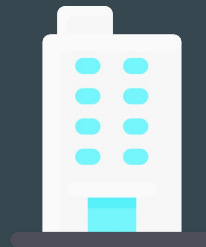


## The Business

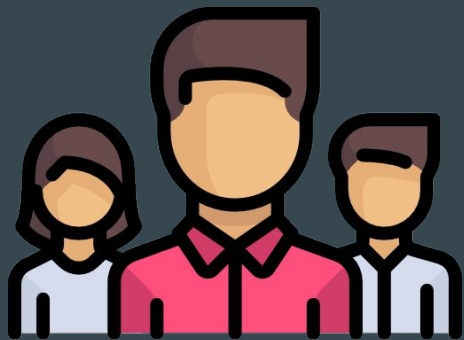
Build a credit default model. High accuracy usually goes with high complexity.

### Business Questions:

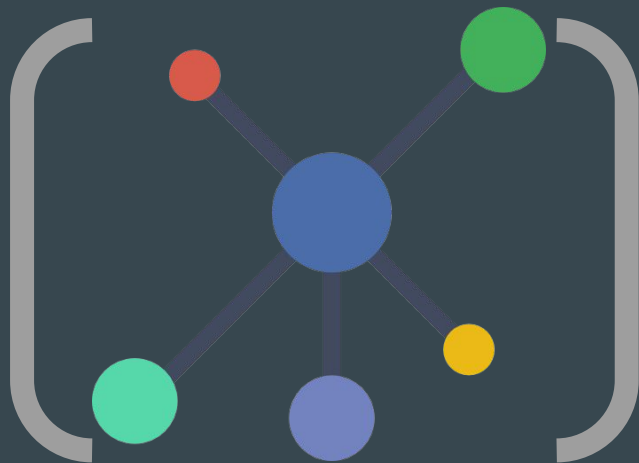
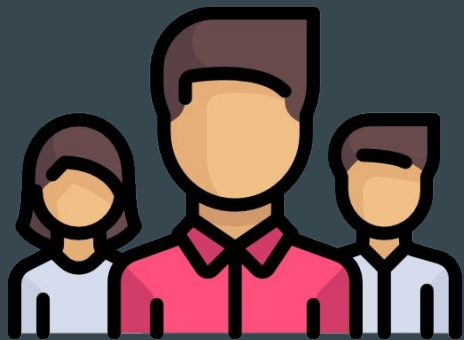
- Is there bias in our model?
- Do we understand our underlying data?
- What will cause the model to break?
- Are we within regulatory framework?



# Context..



# Context..



# Context..

Interpretable

Accurate

Complex Model

Simple Model



# Context..

Interpretable

Accurate

Complex Model

Simple Model



# Context..

Interpretable




Accurate

Complex Model






Simple Model





# Context..

	Interpretable	Accurate
Complex Model		
Simple Model		

# Context..

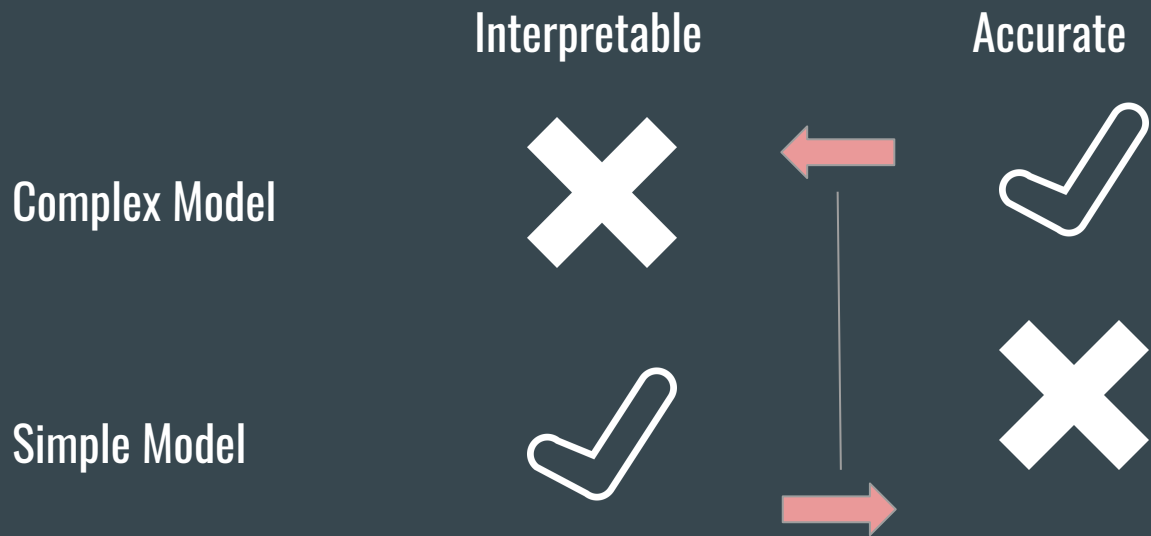
	Interpretable	Accurate
Complex Model		
Simple Model		

# Context..

	Interpretable	Accurate
Complex Model		
Simple Model		

A vertical line and a red arrow point from the Simple Model row towards the Accurate column.

# Context..

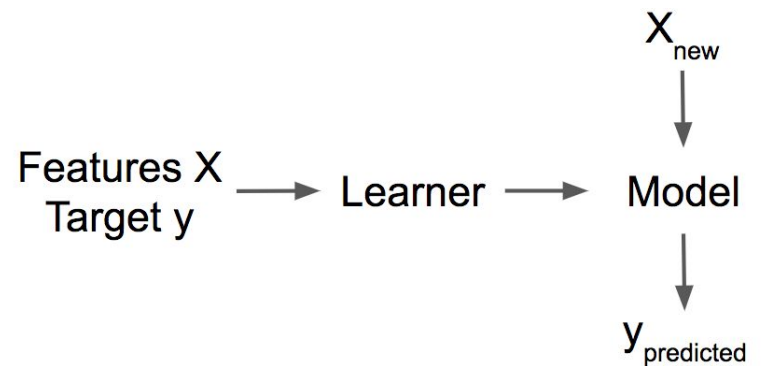


# Machine Learning Model

“Machine Learning is a set of methods that allow computers to learn from data to make and improve predictions (for example cancer, weekly sales, credit default) .”

- Christopher Molnar

# Machine Learning Model





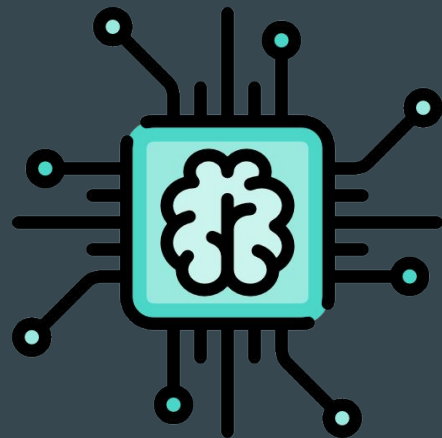
# Interpretability

“Interpretability is the degree to which a human can understand the cause of a decision.” - [2]

“Interpretability is the degree to which a human can consistently predict the model’s result.” - [3]

# Importance of Interpretability (what or why)

1. **Human curiosity and learning**
2. **Goal of science**
3. **Safety measures**
4. **Detecting bias**
5. **Manage social interactions**
6. **Debugged and audited**



# Interpretability Techniques



- **Reveal its internal mechanisms**
- Fully understood by looking at their parameters
- Also called interpretable models



- **Does not reveal its internal mechanisms**
- Cannot be understood by looking at their parameters (e.g. a neural network)

# Black Box Models (interpretability techniques)



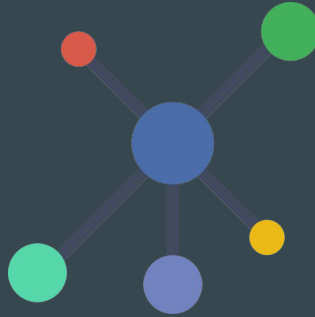
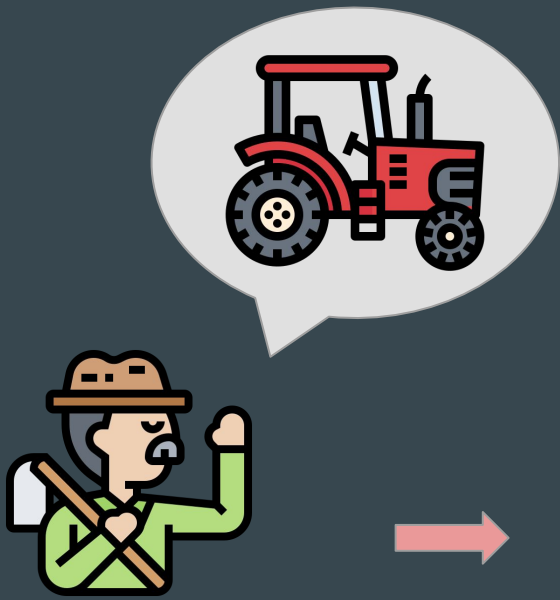
Model Agnostic Methods	Example Based Explanations
Partial Dependence Plot (PDP)	Counterfactual Explanations
Individual Conditional Expectation (ICE)	Adversarial Examples
Accumulated Local Effects (ALE) Plot	Prototypes and Criticisms
Feature Interaction	Influential Instances
Permutation Feature Importance	
Global Surrogate	
Local Surrogate (LIME)	
Scoped Rules (Anchors)	
Shapley Values	
SHAP (SHapley Additive exPlanations)	

# Black Box Models (interpretability techniques)



Model Agnostic Methods	Example Based Explanations
Partial Dependence Plot (PDP)	Counterfactual Explanations
Individual Conditional Expectation (ICE)	Adversarial Examples
Accumulated Local Effects (ALE) Plot	Prototypes and Criticisms
Feature Interaction	Influential Instances
Permutation Feature Importance	
Global Surrogate	
Local Surrogate (LIME)	
Scoped Rules (Anchors)	
Shapley Values	
SHAP (SHapley Additive exPlanations)	

# Shapley Values



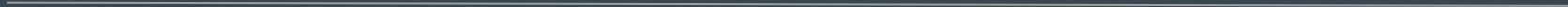
22 %







0





Average Pred  
15%

0





Average Pred  
15%  
 $E[f(x)]$

Joe  
23%  
 $f(x)$

0

?



Average Pred  
15%  
 $E[f(x)]$

0



Base  
 $\Phi(0)$



Average Pred  
15%  
 $E[f(x)]$

0

19%

Base  
 $\Phi(0)$

Income not verified  
 $\Phi(1)$



Average Pred  
15%  
 $E[f(x)]$

0

19%

21%

Base  
 $\Phi(0)$

Income not verified  
 $\Phi(1)$

DTI = 30  
 $\Phi(2)$



Average Pred  
15%  
 $E[f(x)]$

0

19%

21%

23%

Base  
 $\Phi(0)$

Income not verified  
 $\Phi(1)$

DTI = 30  
 $\Phi(2)$

Delinquent 10 months  
ago  
 $\Phi(3)$



Average Pred  
15%  
 $E[f(x)]$

0

19%

21%

23%

Base  
 $\Phi(0)$

Income not verified  
 $\Phi(1)$

DTI = 30  
 $\Phi(2)$

Delinquent 10 months ago  
 $\Phi(3)$

No recent account openings  
 $\Phi(4)$





Average Pred  
15%  
 $E[f(x)]$

0

19%

21%

23%

Base  
 $\Phi(0)$

Income not verified  
 $\Phi(1)$

DTI = 30  
 $\Phi(2)$

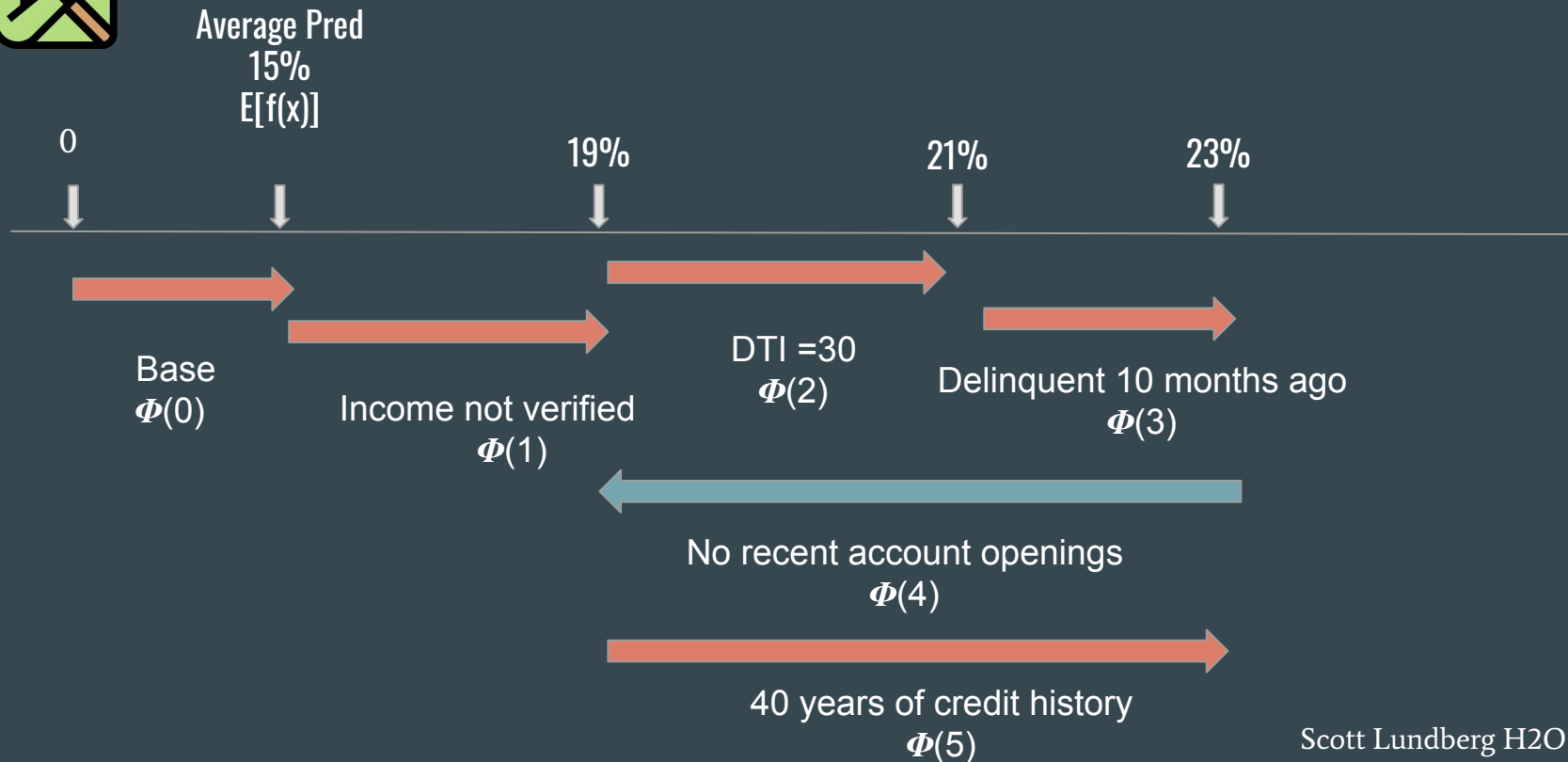
Delinquent 10 months ago  
 $\Phi(3)$

No recent account openings  
 $\Phi(4)$

40 years of credit history  
 $\Phi(5)$

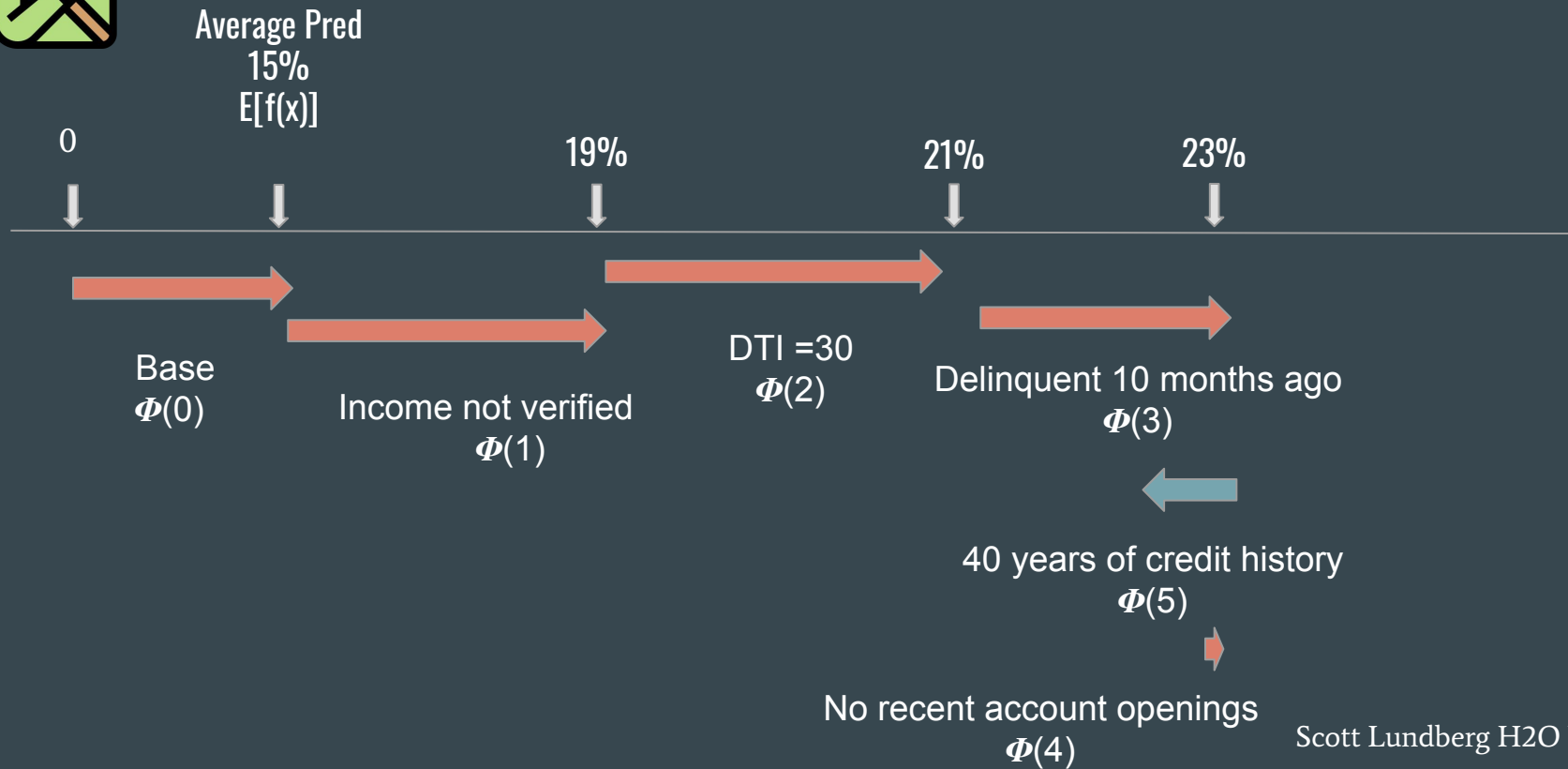


## What about the order?



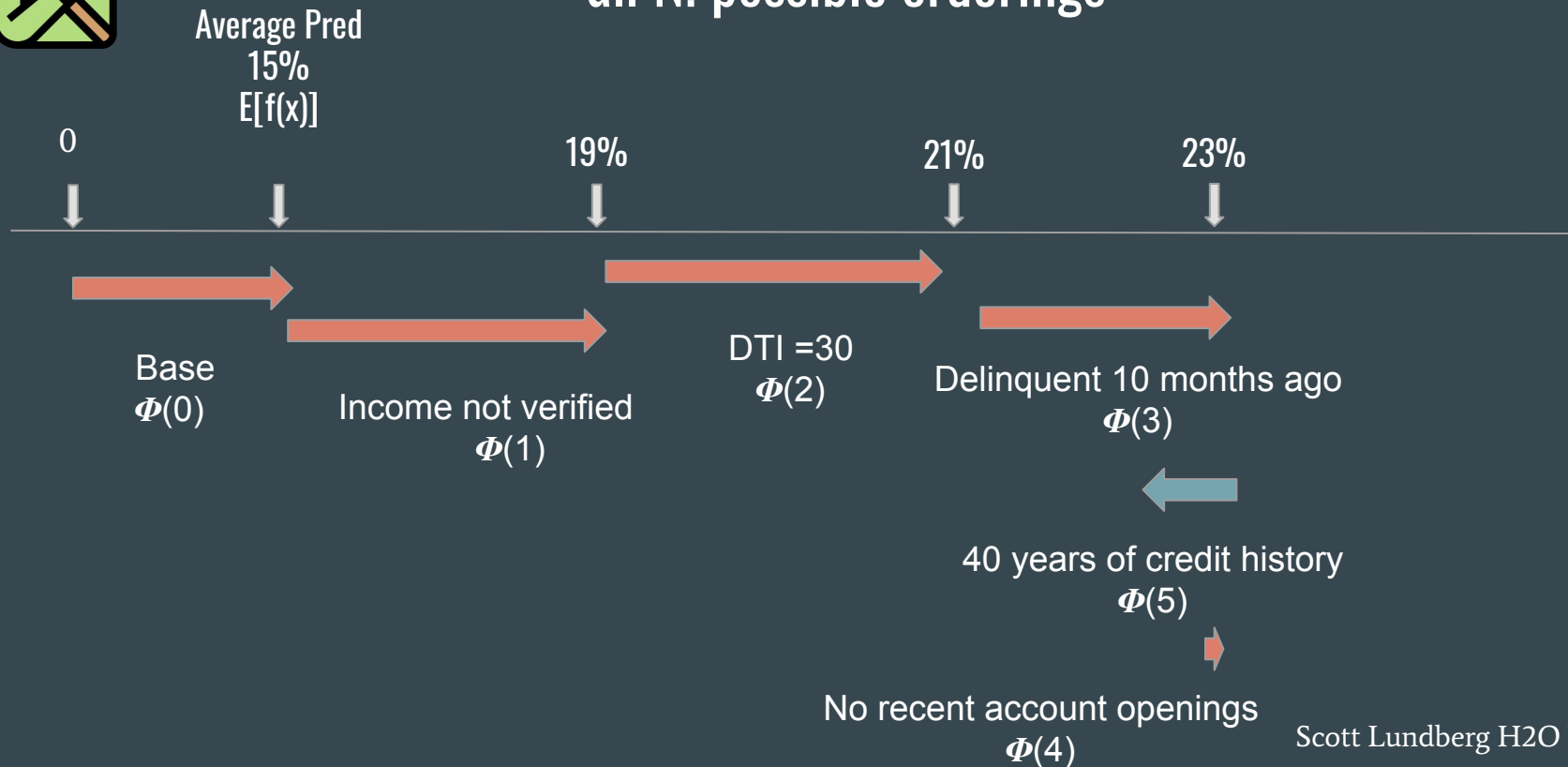


## What about the order?





# Shapley values results from averaging over all $N!$ possible orderings

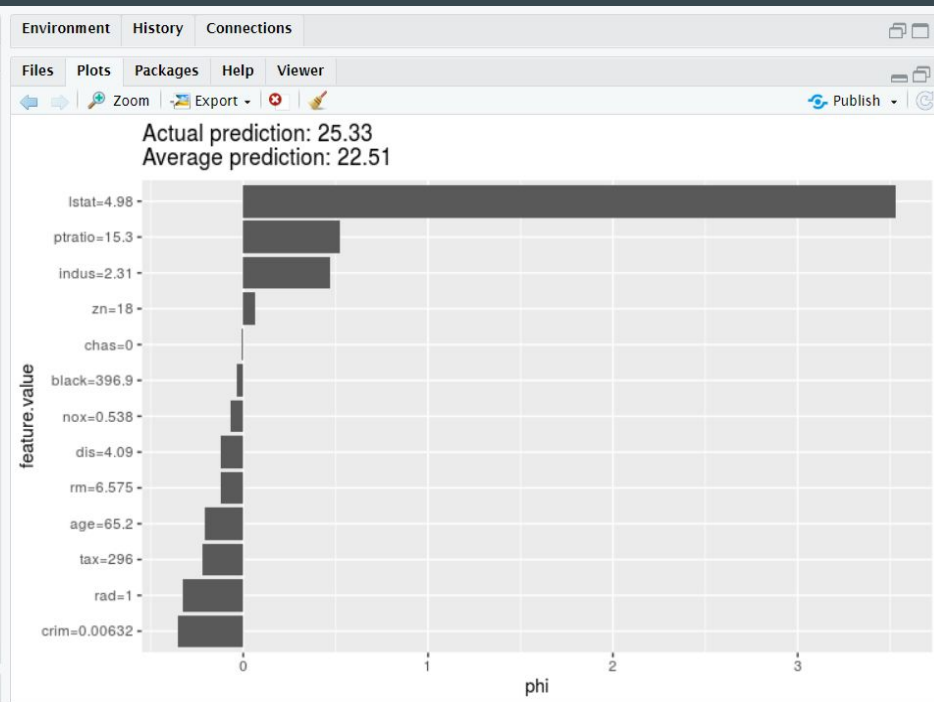


# Examples & Interpretation

```
1 # First we fit a machine learning model on the Boston housing data
2
3
4 data("Boston", package = "MASS")
5 rf = randomForest(medv ~ ., data = Boston, ntree = 50)
6 X = Boston[-which(names(Boston) == "medv")]
7 mod = Predictor$new(rf, data = X)
8
9 # Then we explain the first instance of the dataset with the Shapley method:
10
11 x.interest = X[1,]
12 shapley = Shapley$new(mod, x.interest = x.interest)
13
14 # plot
15 |
16 plot(shapley)
17
```

15:1 (Top Level) R Script

Console



# SHAP (Shapley Additive Explanations)

## KernelSHAP

An alternative, kernel-based estimation approach for Shapley values inspired by local surrogate models

## TreeSHAP

An efficient estimation approach for tree-based models

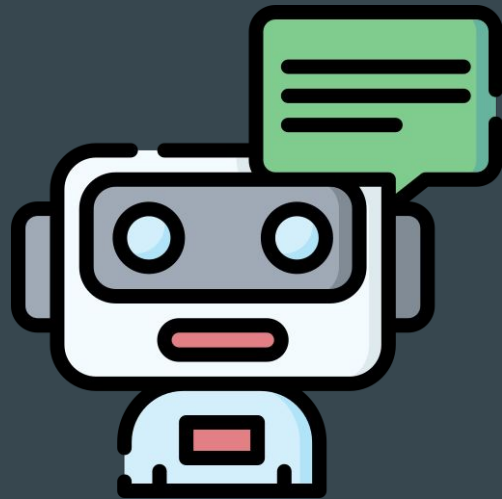
SHAP comes with many global interpretation methods based on aggregations of Shapley values

# The future of interpretability

The focus will be on model-agnostic interpretability tools.

Machine learning will be automated and, with it, interpretability.

Robots and programs will explain themselves



# Resource

Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv Preprint arXiv:1706.07269. (2017). [2]

Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretability." Advances in Neural Information Processing Systems (2016). [3]

Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning," no. ML: 1–13. <http://arxiv.org/abs/1702.08608> ( 2017)

## Package iml

<https://www.youtube.com/watch?v=ngOBhhINWb8>

Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. <https://christophm.github.io/interpretable-ml-book/>.

Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. 2017.

---



# Resource

Icons:

Icons made by <a href="https://www.flaticon.com/authors/smashicons" title="Smashicons">Smashicons</a> from <a href="https://www.flaticon.com/" title="Flaticon"> www.flaticon.com</a>  
Icons made by <a href="https://www.flaticon.com/authors/freepik" title="Freepik">Freepik</a> from <a href="https://www.flaticon.com/" title="Flaticon"> www.flaticon.com</a>  
Icons made by <a href="https://www.flaticon.com/authors/monkik" title="monkik">monkik</a> from <a href="https://www.flaticon.com/" title="Flaticon"> www.flaticon.com</a>  
Icons made by <a href="https://www.flaticon.com/authors/dinosoftlabs" title="DinosoftLabs">DinosoftLabs</a> from <a href="https://www.flaticon.com/" title="Flaticon"> www.flaticon.com</a>  
Icons made by <a href="https://www.flaticon.com/authors/vitaly-gorbachev" title="Vitaly Gorbachev">Vitaly Gorbachev</a> from <a href="https://www.flaticon.com/" title="Flaticon"> www.flaticon.com</a>  
Icons made by <a href="https://www.flaticon.com/authors/flat-icons" title="Flat Icons">Flat Icons</a> from <a href="https://www.flaticon.com/" title="Flaticon"> www.flaticon.com</a>  
Icons made by <a href="https://www.flaticon.com/authors/photo3idea-studio" title="photo3idea\_studio">photo3idea\_studio</a> from <a href="https://www.flaticon.com/" title="Flaticon"> www.flaticon.com</a>  
Icons made by <a href="https://www.flaticon.com/authors/srip" title="srip">srip</a> from <a href="https://www.flaticon.com/" title="Flaticon"> www.flaticon.com</a>  
Icons made by <a href="https://www.flaticon.com/authors/eucalyp" title="Eucalyp">Eucalyp</a> from <a href="https://www.flaticon.com/" title="Flaticon"> www.flaticon.com</a>  
Icons made by <a href="https://www.flaticon.com/authors/ultimatearm" title="ultimatearm">ultimatearm</a> from <a href="https://www.flaticon.com/" title="Flaticon"> www.flaticon.com</a>

---



Questions?