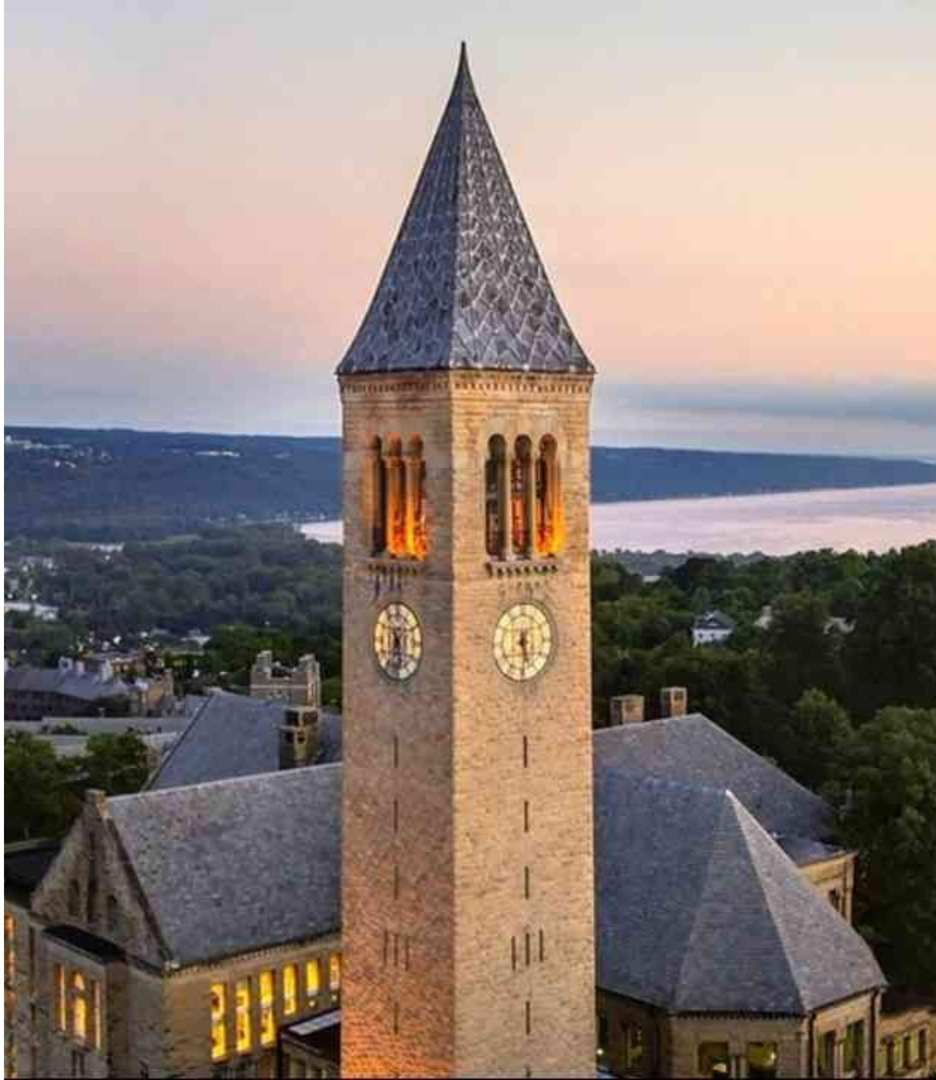
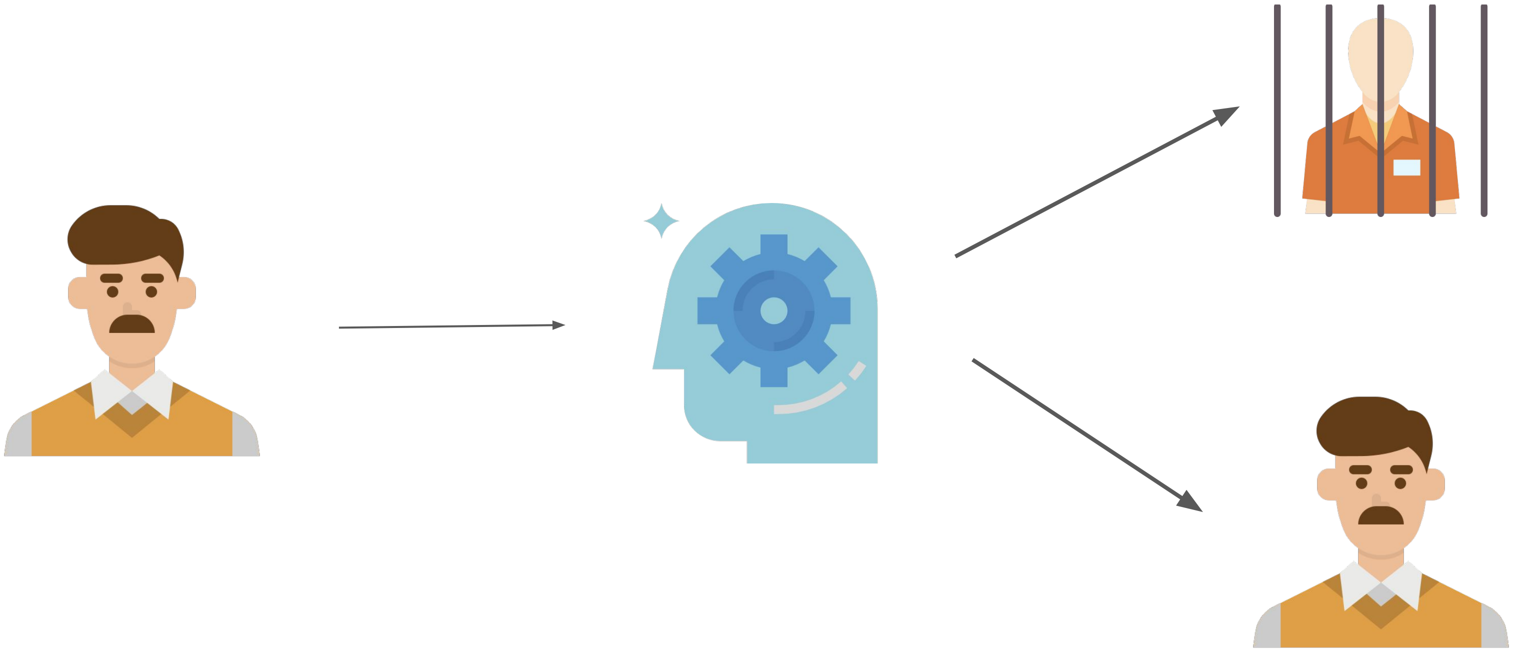
# Explainable Machine Learning

ORIE 4371
November 2, 2021



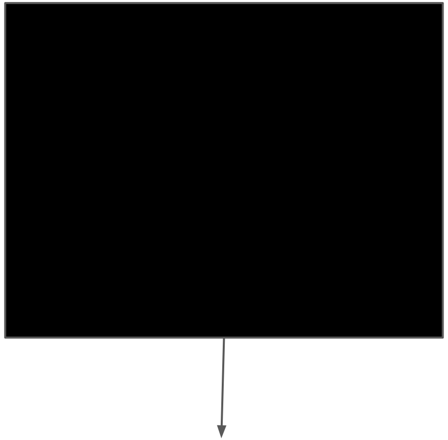
# Why should we care about interpreting machine learning models?

# Socially Sensitive Machine Learning



# Two Approaches

*Un-interpretable*



*No Parole*

*Interpretable*

$$\left[(\text{Priors}\geq 3) \text{ and } (\text{Age}\leq 45) \text{ and } (\text{Score Factor} = \text{TRUE})\right]$$
$$\text{OR}$$
$$\left[(\text{Priors}\geq 20) \text{ and } (\text{Age}\geq 45)\right]$$

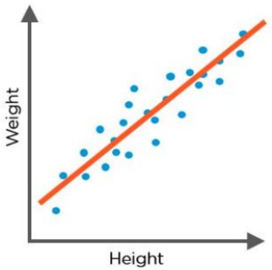
# COMPAS

# Legal Issues + Procedural Fairness

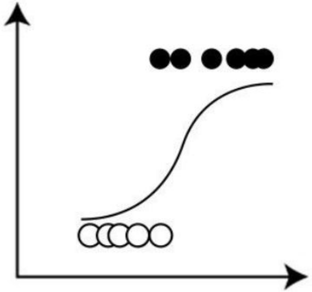
*GDPR establishes a right for all individuals to obtain “meaningful explanations of the logic involved” when “automated (algorithmic) individual decision-making”, including profiling, takes place.*

# What are some strategies we’ve seen to interpret a model before?

# Simple Models




Linear Regression



Logistic Regression


Decision Tree

$$\left[(\text{Priors}\geq 3) \text{ and } (\text{Age}\leq 45) \text{ and } (\text{Score Factor} = \text{TRUE})\right]$$
$$\text{OR}$$
$$\left[(\text{Priors}\geq 20) \text{ and } (\text{Age}\geq 45)\right]$$

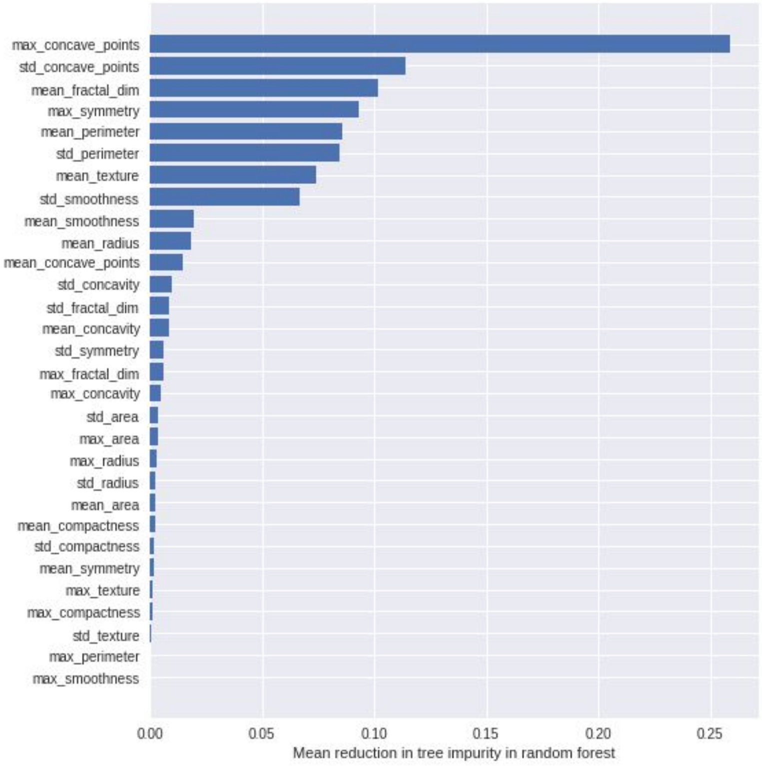
Rule Sets and Scorecards

PREDICT PATIENT HAS OBSTRUCTIVE SLEEP APNEA IF SCORE $> 1$

| | | | |
|---|---|---|---|
| 1. | *Age* $\geq$ *60* | 4 points | $\cdots$ |
| 2. | *Hypertension* | 4 points | $+ \cdots$ |
| 3. | *BMI* $\geq$ *30* | 2 points | $+ \cdots$ |
| 4. | *BMI* $\geq$ *40* | 2 points | $+ \cdots$ |
| 5. | *Female* | -6 points | $+ \cdots$ |
| | ADD POINTS FROM ROWS 1 – 5 | SCORE | $= \cdots$ |

Sparse Linear Models

# Feature Importance



# Hierarchy of Interpretability

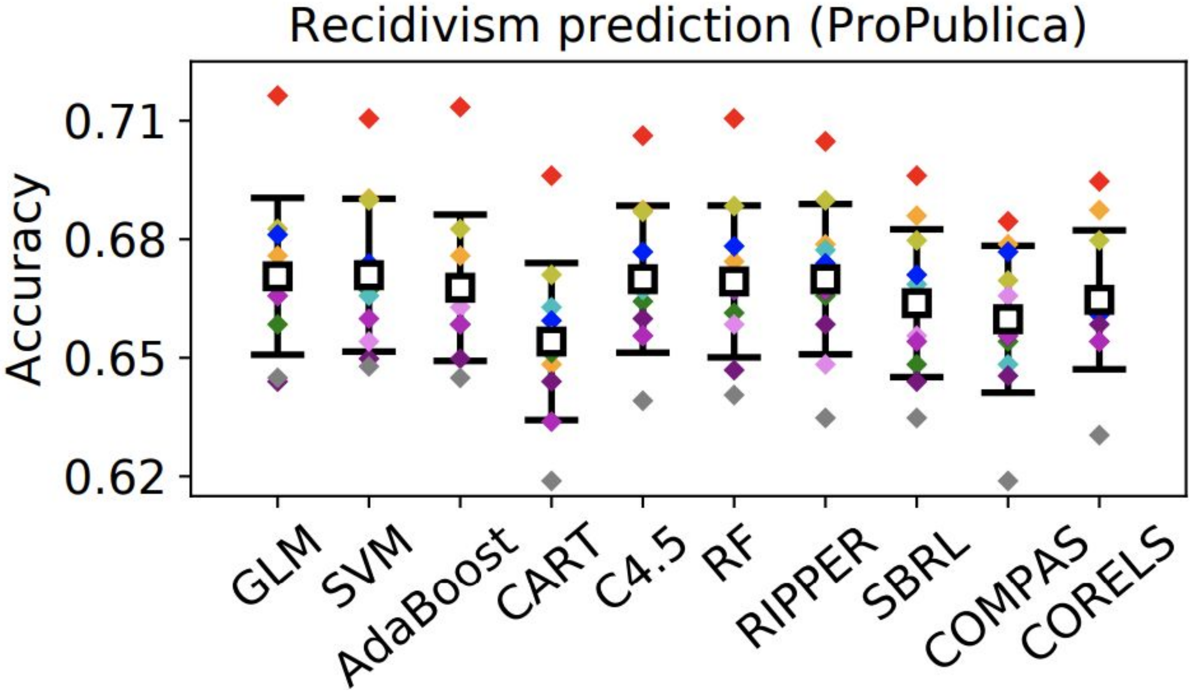
**Best** Simple Interpretable Model

Complex Interpretable Model

Explain Black Box

**Worst** Complete Black Box

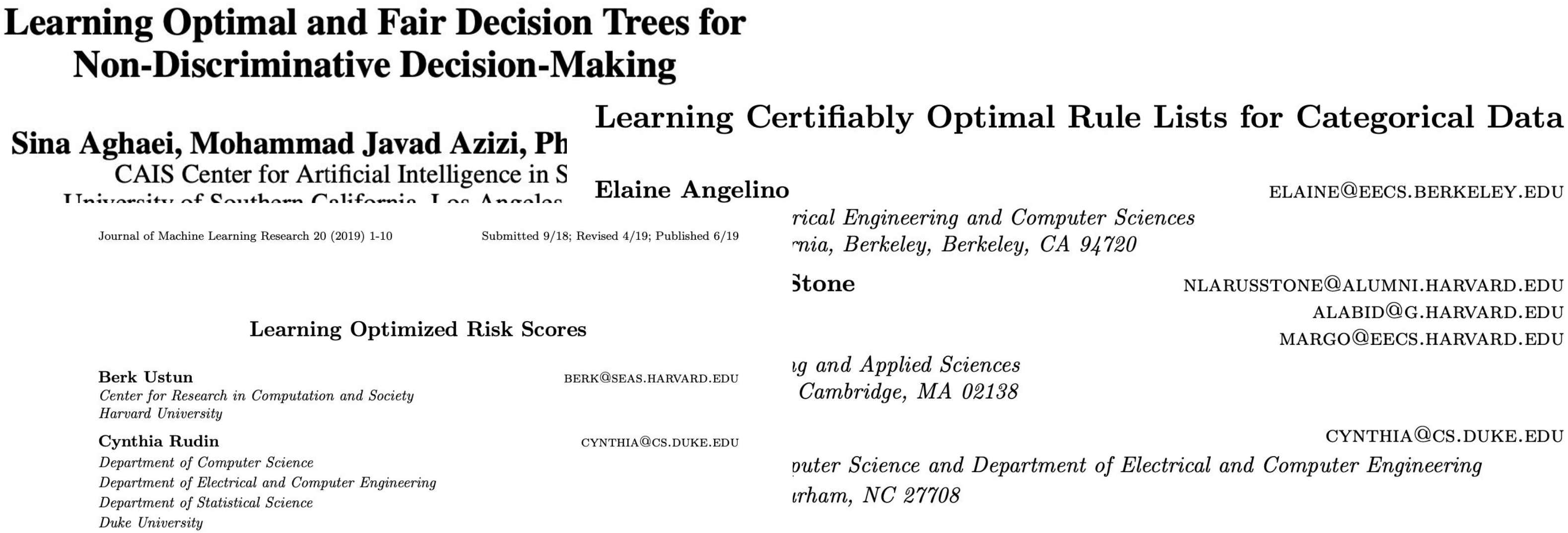
# Why not always use simple models?




*Learning Certifiably Optimal Rule Sets. Angelino et al. (2016)

# Revisiting Simple Models

Researchers are working on using modern optimization approaches to find better simple models for practical datasets!



**Learning Optimal and Fair Decision Trees for Non-Discriminative Decision-Making**

Sina Aghaei, Mohammad Javad Azizi, Ph

CAIS Center for Artificial Intelligence in S

**Learning Certifiably Optimal Rule Lists for Categorical Data**

Elaine Angelino ELAINE@EECS.BERKELEY.EDU

rical Engineering and Computer Sciences

rnia, Berkeley, Berkeley, CA 94720

Stone NLARUSSTONE@ALUMNI.HARVARD.EDU

ALABID@G.HARVARD.EDU

MARGO@EECS.HARVARD.EDU

ng and Applied Sciences

Cambridge, MA 02138

CYNTHIA@CS.DUKE.EDU

puter Science and Department of Electrical and Computer Engineering

urham, NC 27708

Journal of Machine Learning Research 20 (2019) 1-10 Submitted 9/18; Revised 4/19; Published 6/19

**Learning Optimized Risk Scores**

Berk Ustun BERK@SEAS.HARVARD.EDU

*Center for Research in Computation and Society*

*Harvard University*

Cynthia Rudin CYNTHIA@CS.DUKE.EDU

*Department of Computer Science*

*Department of Electrical and Computer Engineering*

*Department of Statistical Science*

*Duke University*

***Come chat with Connor if you're interested about this!**

# Hierarchy of Interpretability

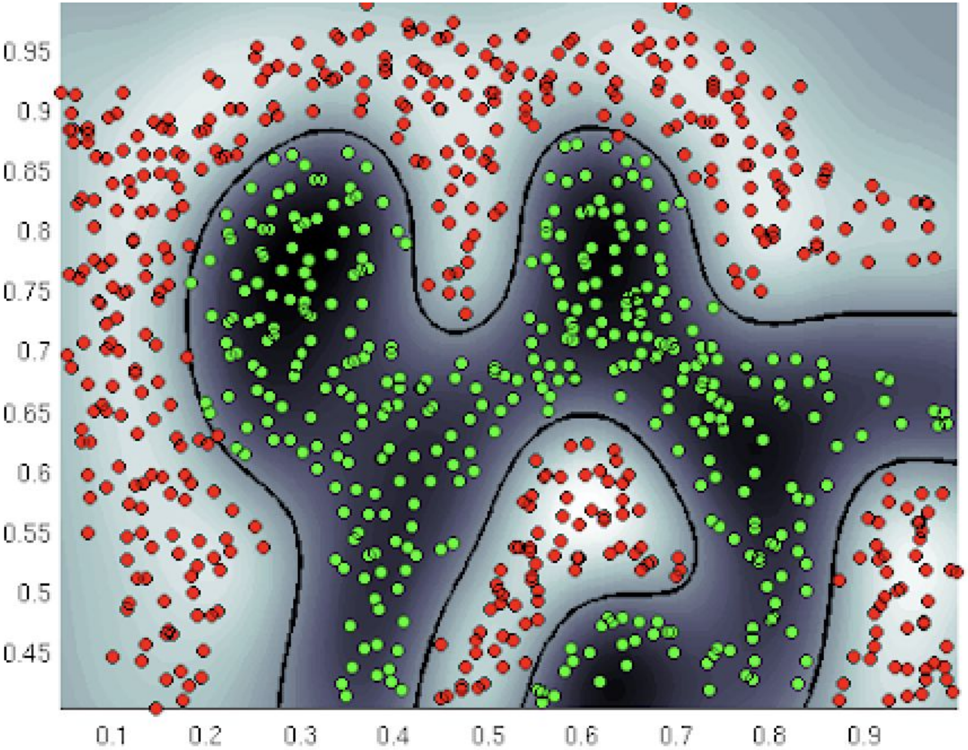
**Best** Simple Interpretable Model

Complex Interpretable Model

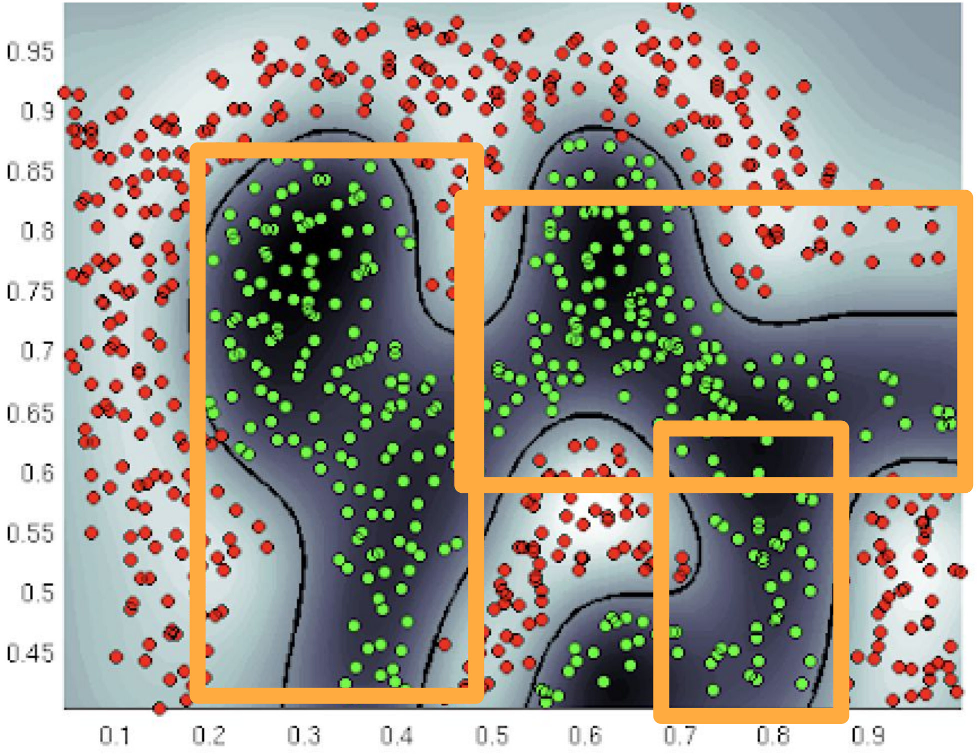
Explain Black Box

**Worst** Complete Black Box

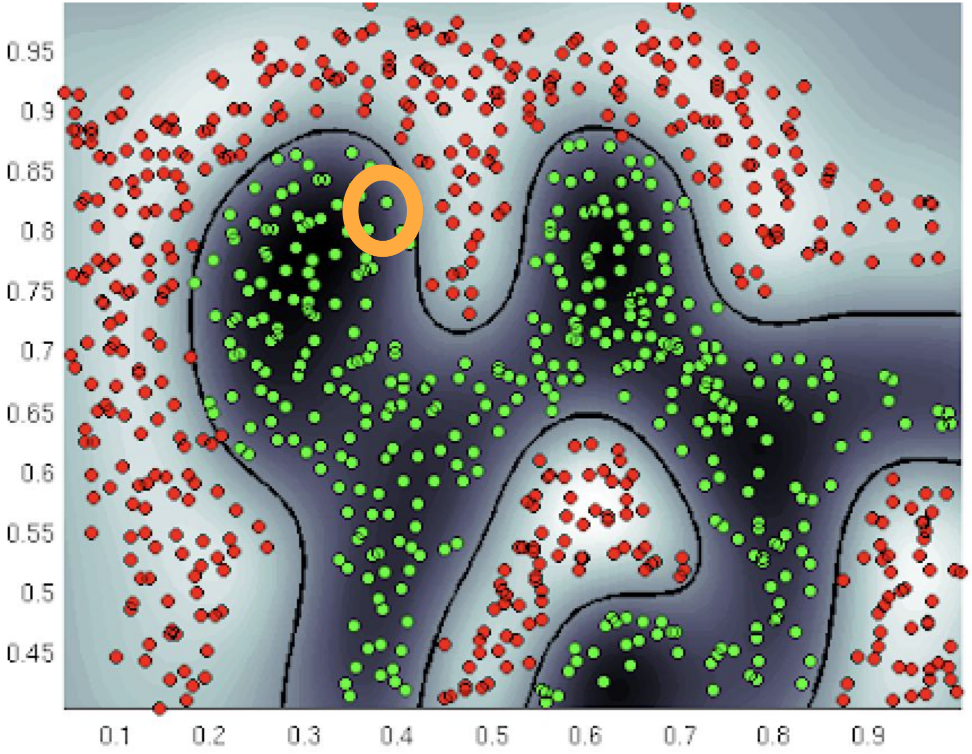
# Global vs. Local Explanations



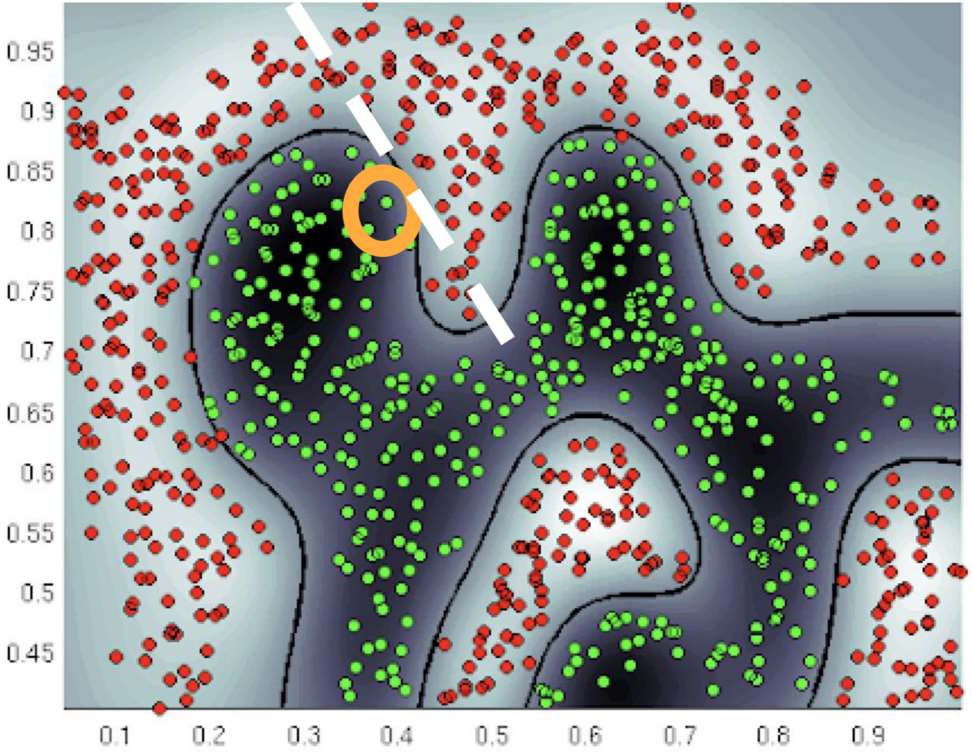
# **Global** vs. Local Explanations



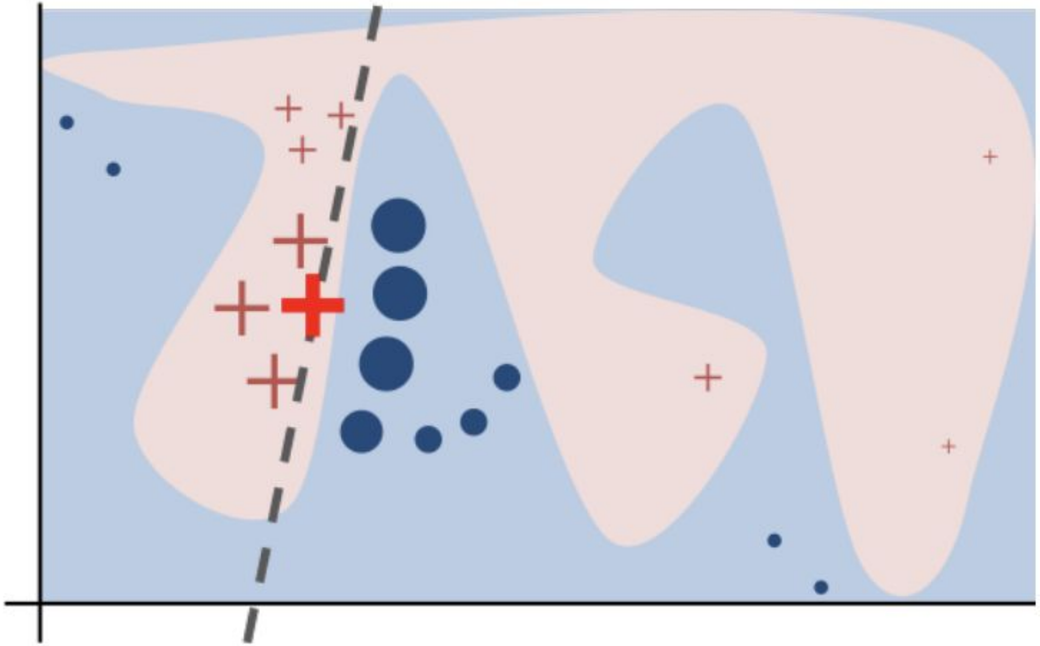
# Global vs. **Local** Explanations



# Global vs. **Local** Explanations



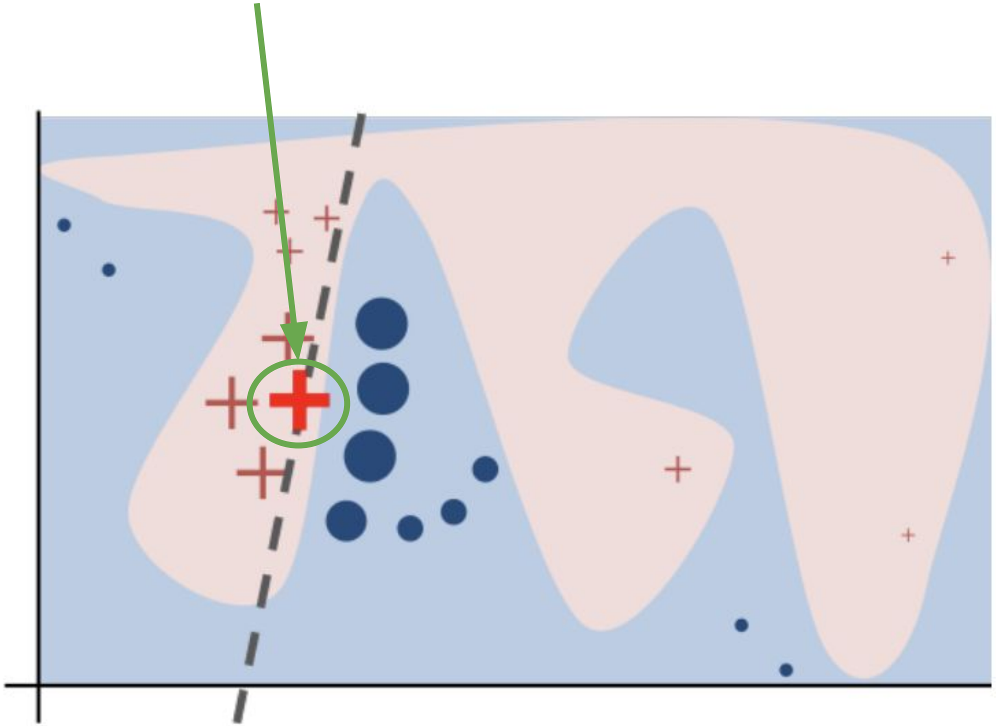
# **L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanations



***"Why Should I Trust You?" Explaining output of any classifier. Ribeiro et al. 2016**

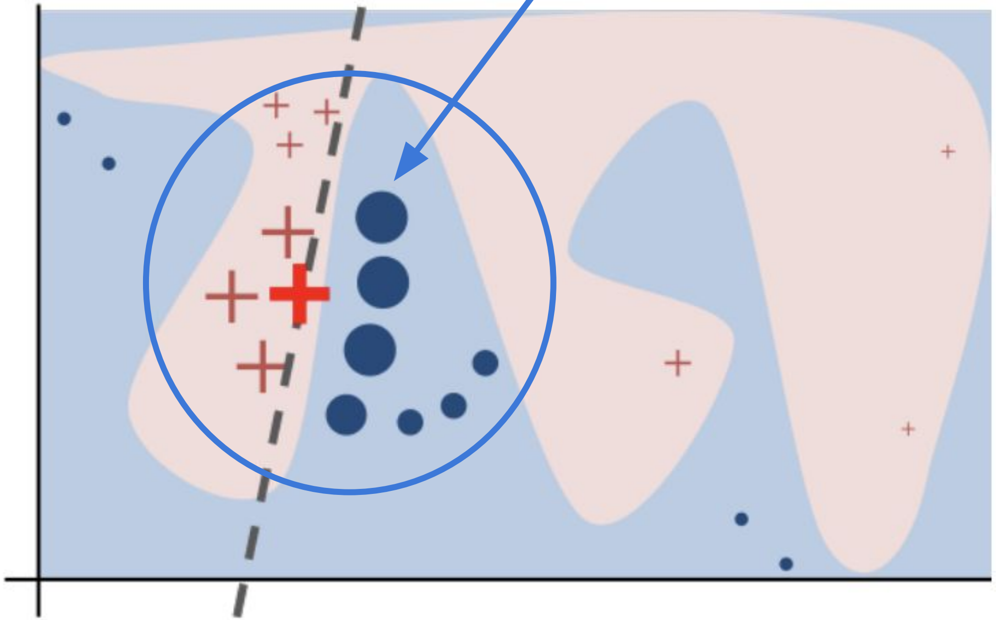
# LIME

High-Level Idea: ***Pick point to explain***, sample points around it, and construct a linear classifier.



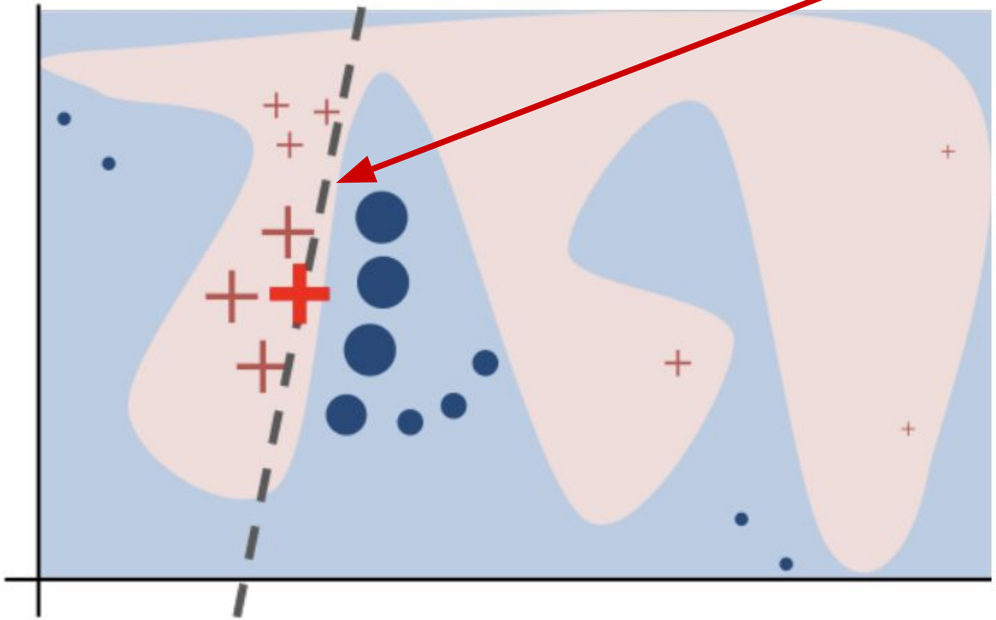
# LIME

High-Level Idea: Pick point to explain, ***sample points around it,*** and construct a linear classifier.



# LIME

High-Level Idea: Pick point to explain, sample points around it, and ***construct a linear classifier.***



# LIME ‘Formally’

*‘Black Box’*

*Locality Kernel*

$$\underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

*‘Explanation’ Loss*

*Complexity Loss*

*Interpretable Model Class*

# What could be a pick for our ‘complexity’ loss if G is linear models?

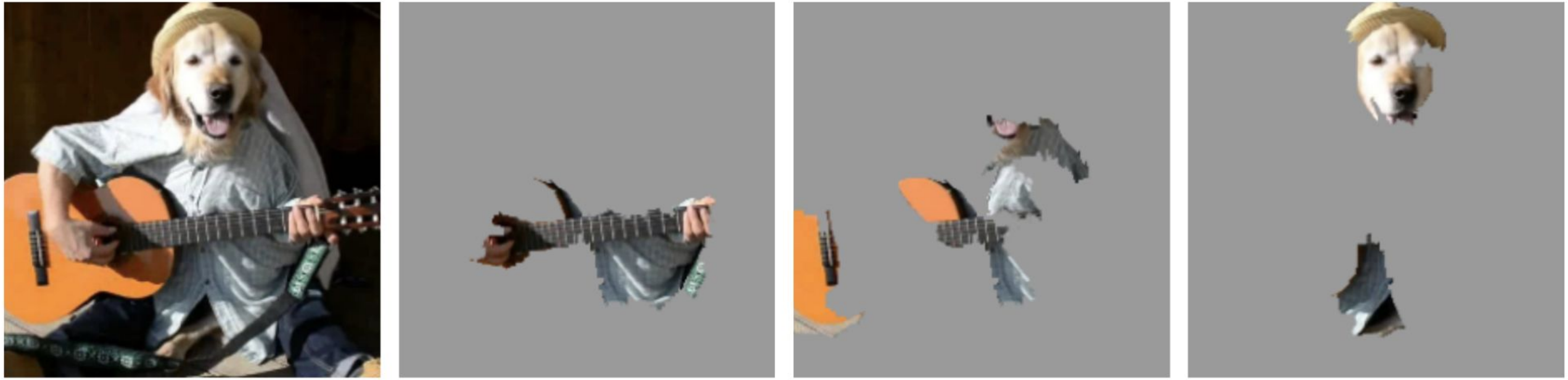
# LIME Algorithm

**Algorithm 1** Sparse Linear Explanations using LIME

**Require:** Classifier $f$, Number of samples $N$
**Require:** Instance $x$, and its interpretable version $x'$
**Require:** Similarity kernel $\pi_x$, Length of explanation $K$

$\mathcal{Z} \leftarrow \{\}$
**for** $i \in \{1, 2, 3, ..., N\}$ **do**
  $z'_i \leftarrow sample\_around(x')$
  $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
**end for**
$w \leftarrow \text{K-Lasso}(\mathcal{Z}, K)$ ▷ with $z'_i$ as features, $f(z)$ as target
**return** $w$

# Image Classification Example



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

**Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)**

# **SH**apley **A**dditive ex**P**lanation

Shapely values are a tool from game theory to **distribute payout from a collaborative game 'fairly'.** In our context, the **'payout' is going to be our final prediction.**

*Payout for player i*

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[ f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

*All possible subsets of players*

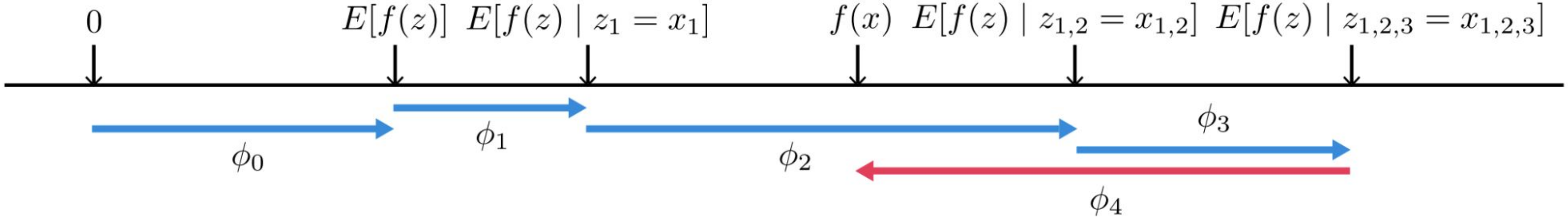
*Value for including 'player' i*

*Value without 'player' i*

# What’s a good pick for our function f?

# SHapley Additive exPlanation

We'll use the conditional expectation! Note that the order of features we consider matters when our classifier is non-linear, so we 'average' the phi for each ordering.



# Permutation Importance

A final simple tool, we can use is just shuffle a column and see how it impacts performance.

| Height at age 20 (cm) | Height at age 10 (cm) | ... | Socks owned at age 10 |
|---|---|---|---|
| 182 | 155 | ... | 20 |
| 175 | 147 | ... | 10 |
| ... | ... | ... | ... |
| 156 | 142 | ... | 8 |
| 153 | 130 | ... | 24 |

# Permutation Importance

Positive numbers mean shuffling a column decreased performance (i.e. is important), small or negative numbers mean less.

| Weight | Feature |
|---|---|
| 0.1750 ± 0.0848 | Goal Scored |
| 0.0500 ± 0.0637 | Distance Covered (Kms) |
| 0.0437 ± 0.0637 | Yellow Card |
| 0.0187 ± 0.0500 | Off-Target |
| 0.0187 ± 0.0637 | Free Kicks |
| 0.0187 ± 0.0637 | Fouls Committed |
| 0.0125 ± 0.0637 | Pass Accuracy % |
| 0.0125 ± 0.0306 | Blocked |
| 0.0063 ± 0.0612 | Saves |
| 0.0063 ± 0.0250 | Ball Possession % |
| 0 ± 0.0000 | Red |
| 0 ± 0.0000 | Yellow & Red |
| 0.0000 ± 0.0559 | On-Target |
| -0.0063 ± 0.0729 | Offsides |
| -0.0063 ± 0.0919 | Corners |
| -0.0063 ± 0.0250 | Goals in PSO |
| -0.0187 ± 0.0306 | Attempts |
| -0.0500 ± 0.0637 | Passes |