

Machine Learning 520

Advanced Machine Learning

Lesson 09: Model Interpretability

Today's Agenda

- MACHINE BIAS
- Model Interpretability
- LIME
- Shapley



Machine Bias & Model Interpretability



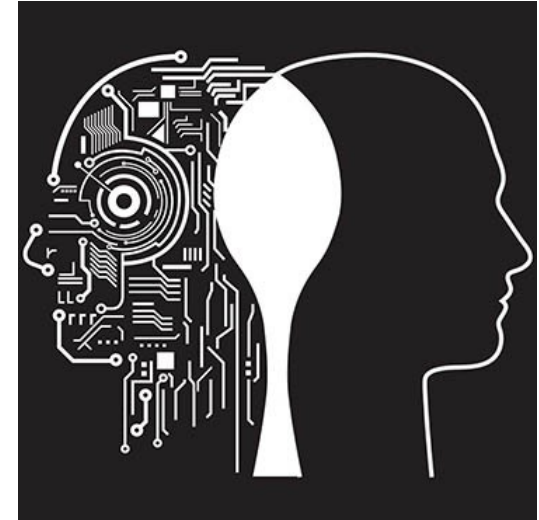
Criminal Risk Scoring or How not to create Risk Scoring Systems


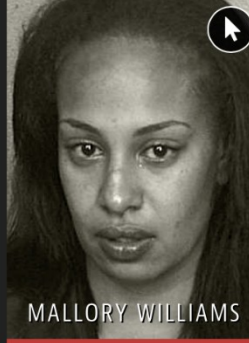
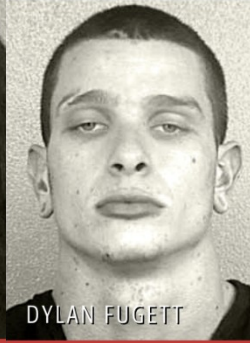


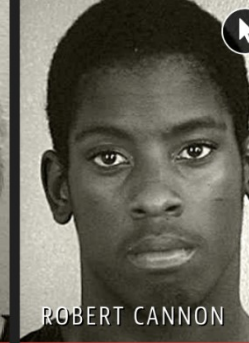


Why do we need Transparency, Accountability and Fairness in Machine Learning?

One way to achieve these goals is through embedding explanations in Machine Learning

Problems with the Criminal Risk Scoring System should be used as guides for other domains to not make the same mistakes

Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin use Criminal Risk Scoring systems in the Judicial System in the US



							
GREGORY LUGO	MALLORY WILLIAMS	DYLAN FUGETT	BERNARD PARKER	JAMES RIVELLI	ROBERT CANNON	VERNON PRATER	BRISHA BORDEN
LOW RISK 1	MEDIUM RISK 6	LOW RISK 3	HIGH RISK 10	LOW RISK 3	MEDIUM RISK 6	LOW RISK 3	HIGH RISK 8
GREGORY LUGO	MALLORY WILLIAMS	DYLAN FUGETT	BERNARD PARKER	JAMES RIVELLI	ROBERT CANNON	VERNON PRATER	BRISHA BORDEN
Prior Offenses 3 DULs, 1 battery	Prior Offenses 2 misdemeanors	Prior Offense 1 attempted burglary	Prior Offense 1 resisting arrest without violence	Prior Offenses 1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking	Prior Offense 1 petty theft	Prior Offenses 2 armed robberies, 1 attempted armed robbery	Prior Offenses 4 juvenile misdemeanors
Subsequent Offenses 1 domestic violence battery	Subsequent Offenses None	Subsequent Offenses 3 drug possessions	Subsequent Offenses None	Subsequent Offenses 1 grand theft	Subsequent Offenses None	Subsequent Offenses 1 grand theft	Subsequent Offenses None

Bias in Criminal Risk Scoring Systems

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

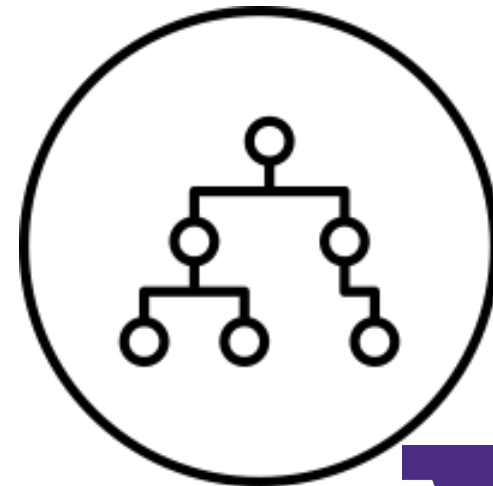
MACHINE BIAS

“Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice, they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society.”

- United States Attorney General, Eric Holder 2014

Criminal Risk Scoring Systems

- Northpointe (for-profit company)
- Most widely used assessment tools in the country
- Risk score calculation is not publically disclosed (proprietary)
- Impossible for defendants to argue about bias that may be present in these systems.



ProPublica Study



- Obtained risk scores of 7,000 people arrested in Broward County, Florida, in 2013 and 2014
- Checked to see how many were charged with new crimes over the next two years (benchmark used by the creators of the algorithm)
- The algorithm is better than random
- 61% of high risk offenders actually re-offended
- Misclassification Problems
 - Falsely labeled black defendants as future criminals twice as many times as white defendants
 - White defendants were mislabeled as low risk more often than black defendants

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>



LESSONS TO BE LEARNED

Don't trust Arbitrary Scoring Systems unless they make domain sense

“A guy who has molested a small child every day for a year could still come out as a low risk because he probably has a job, Boessenecker said. Meanwhile, a drunk guy will look high risk because he's homeless. These risk factors don't tell you whether the guy ought to go to prison or not; the risk factors tell you more about what the probation conditions ought to be.” ProPublica

Even well-intentioned engineering can lead to adverse social effects

The effects of what we are seeing for Criminal Scoring Systems is not what its creators had intended

Need for open and fair systems in machine learning

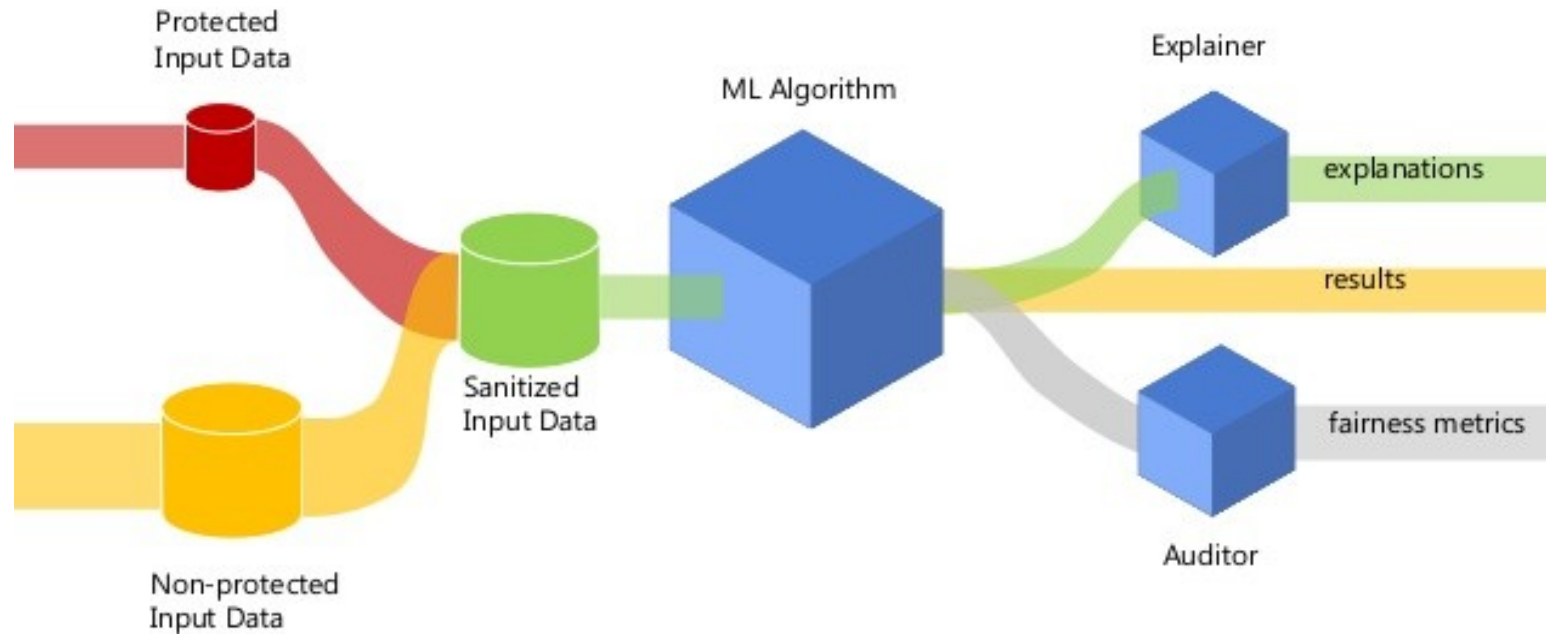
Open and fair machine learning is needed in order to give additional information the users of machine learning so that such systems can be audited in case of a wrong assessment or prediction

Need to educate users of ML systems

Educate users of machine learning systems with respect to what the models are predicting vs. what they might think the system is doing (predictive vs. prescriptive vs. descriptive vs. causal models)



Outlook: What Future ML Systems Could Look Like



Need for Explanations in Machine Learning

- **Decision Making:**
 - Why is the machine learning algorithm making a prediction?
 - What if the reasons for making the prediction are wrong?
- **Fairness:**
 - Are different populations in the dataset being scored fairly?
- **Regulation:**
 - GDPR's right to explanation



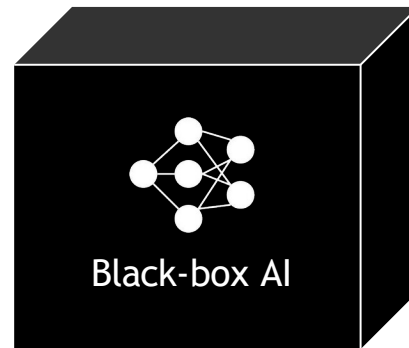


Model Interpretability



Problem: Machine Learning is a Black box

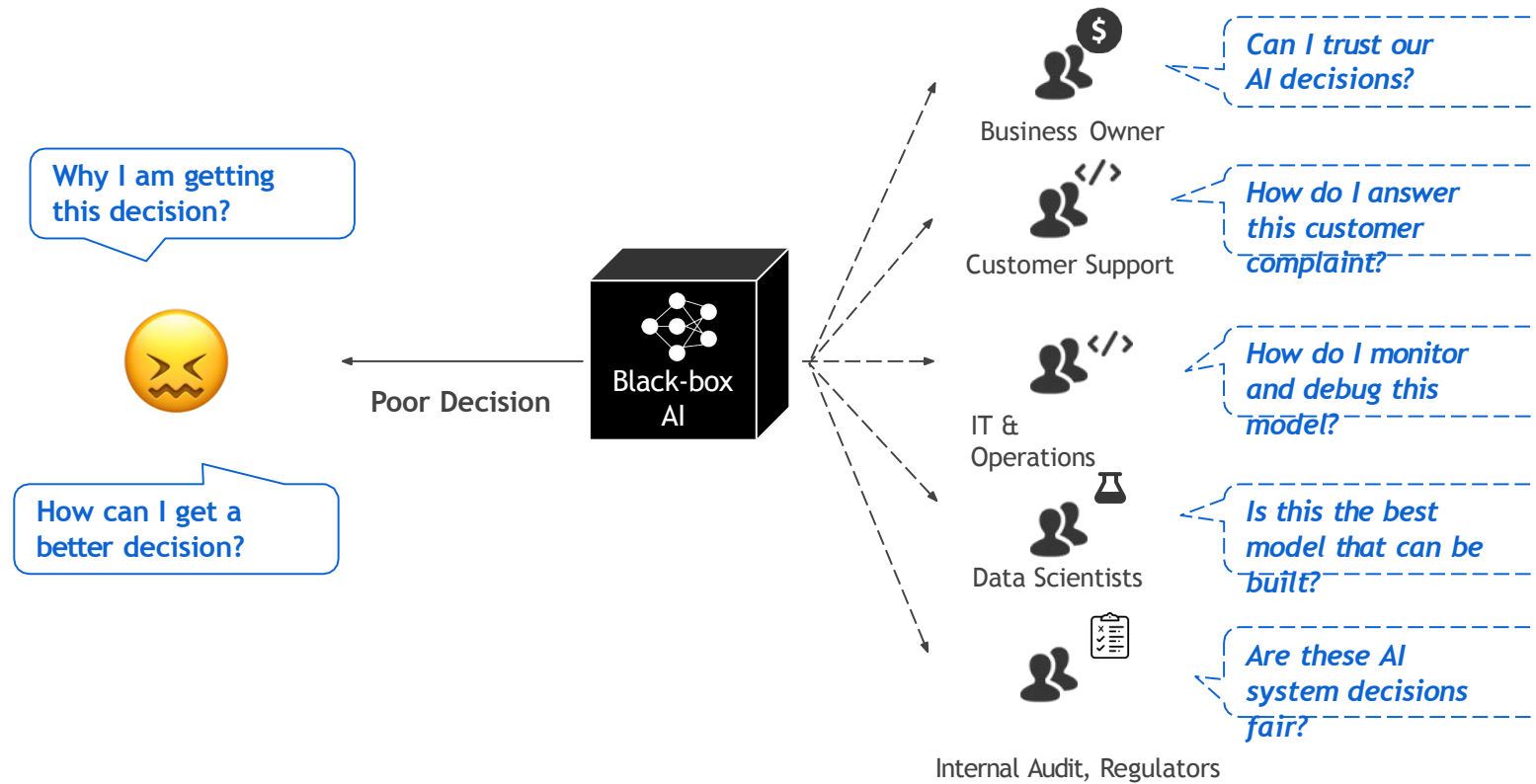
Output
(Label, sentence, next word, game position)



Input
(Data, image, sentence,
etc.)



Black-box AI creates confusion and doubt



Early History

- Explanations and interpretability are currently at the cutting edge of ML
- Many people assume that this a new topic
- The earliest work in AI was rooted in logic to make mechanical reasoning transparent and easily understood
- Logic Theorist (1955-56) could prove mathematical theorems given a problem



World's first programmable Automata 12th century

The Al-Jazari orchestra, one of the earliest examples of known automata. Credit: Freer Gallery of Art



Why do we need explanations now?

- Machine Learning models have shown tremendous success in various prediction tasks
- Availability of large amounts of data
- Availability of cheap computing power
- Ease of data collection
- Deployment of ML systems in multiple domains





Need for ML Explanations

Case Study of Model with Erroneous Recommendations

- Algorithms to predict which pneumonia patients should be admitted to hospitals and which treated as outpatients
- Neural nets were far more accurate than classical statistical methods
- However, it turned out that both the regression and the neural net had inferred that pneumonia patients with asthma have a lower risk of dying, and shouldn't be admitted
- But the opposite is true. Patients with asthma are high-risk. This was not captured in the data because asthma patients with pneumonia usually were admitted not just to the hospital but directly to the ICU, treated aggressively, and survived.

Patient ID		Has Asthma	Risk of Death	
84	...	Yes	...	5%
85	...	Yes	...	6%
86	...	No	...	12%
87	...	No	...	15%
...

Feature Importance (Higher risk of death): Low  High

Feature Importance (Lower risk of death): Low  High

With Context:

Patients with asthma have a lower risk of death from pneumonia because they receive more intensive care.



When is Model Interpretability Needed?

- Any context where humans are required to provide explanations so that people cannot hide behind machine learning models
- Where the decision based on the model predictions can have far reaching consequences, e.g., recommend an operation

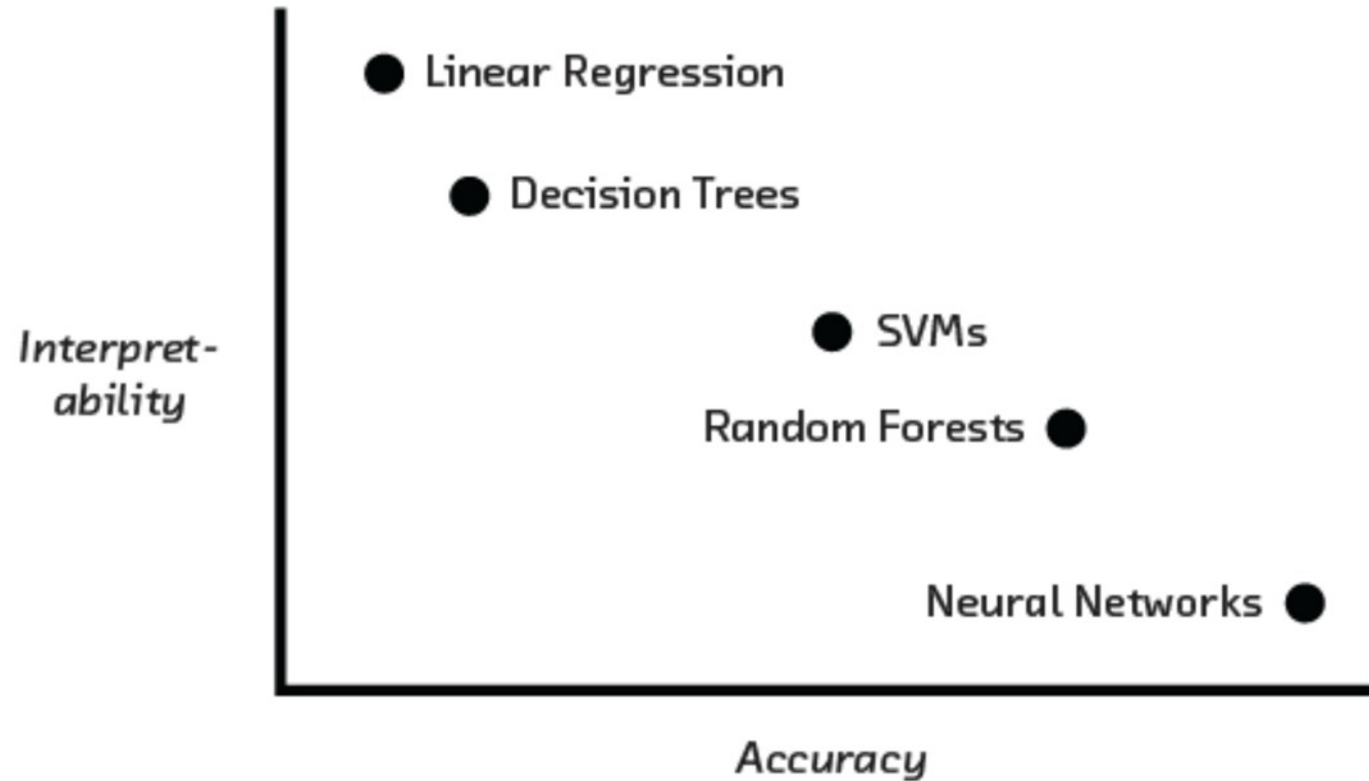


When are explanations not needed?

- When are Contexts where Model Interpretability is not needed
 - Explanation Agnostic contexts, e.g., ED Staffing Prediction
 - Models which have shown to be highly effective empirically i.e., e.g., 98% accuracy for a diagnosis for the AI system
 - Systems that have theoretical guarantees on their performance

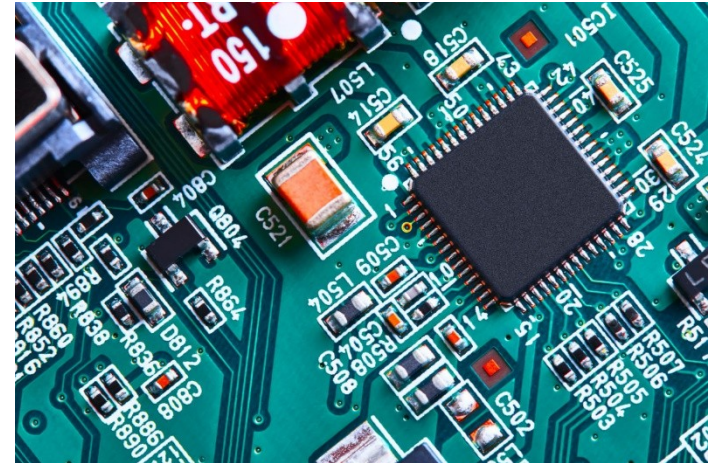


Accuracy vs. Interpretability



The Problem of Generalizability

- Adrian Thompson (Sussex University) used neural networks to design circuits that could distinguish between 2 audio tunes
- Resulted in a system with near perfect performance and fewer inputs than what humans use
- Patterns picked up by the system were not generalizable
- The algorithms can't know which of those patterns exist nowhere else
- The ML system was overfitting without realizing it



Defining Interpretable Machine Learning

- Interpretable machine learning refers to giving **explanations** of machine learning models to **humans** with **domain knowledge**
- Explanation: Why is the prediction being made?
- Explanation to Human: The explanation should be comprehensible to humans in (i) natural language (ii) easy to understand
- Domain Knowledge: The explanation should make sense to a domain expert

$$\begin{aligned} \mathcal{L}_{SM} = & -\frac{1}{2}\partial_\nu g_\mu^a \partial_\nu g_\mu^a - g_s f^{abc} \partial_\mu g_\nu^a g_\mu^b g_\nu^c - \frac{1}{4}g_s^2 f^{abc} f^{ade} g_\mu^b g_\nu^c g_\mu^d g_\nu^e - \partial_\nu W_\mu^+ \partial_\nu W_\mu^- - \\ & M^2 W_\mu^+ W_\mu^- - \frac{1}{2}\partial_\nu Z_\mu^0 \partial_\nu Z_\mu^0 - \frac{1}{2c_w^2} M^2 Z_\mu^0 Z_\mu^0 - \frac{1}{2}\partial_\mu A_\nu \partial_\mu A_\nu - igc_w (\partial_\nu Z_\mu^0 (W_\mu^+ W_\nu^- - W_\nu^+ W_\mu^-) - \\ & Z_\nu^0 (W_\mu^+ \partial_\nu W_\mu^- - W_\mu^- \partial_\nu W_\mu^+) + Z_\nu^0 (W_\nu^+ \partial_\mu W_\mu^- - W_\mu^- \partial_\nu W_\nu^+)) - ig s_w (\partial_\nu A_\mu (W_\mu^+ W_\nu^- - \\ & W_\nu^+ W_\mu^-) - A_\nu (W_\mu^+ \partial_\nu W_\mu^- - W_\mu^- \partial_\nu W_\mu^+) + A_\mu (W_\nu^+ \partial_\nu W_\mu^- - W_\mu^- \partial_\nu W_\nu^+)) - \\ & \frac{1}{2}g^2 W_\mu^+ W_\mu^- W_\nu^+ W_\nu^- + \frac{1}{2}g^2 W_\mu^+ W_\nu^- W_\mu^- W_\nu^+ + g^2 c_w^2 (Z_\mu^0 W_\nu^+ Z_\nu^0 W_\mu^- - Z_\mu^0 Z_\nu^0 W_\nu^+ W_\mu^-) + \\ & g^2 s_w^2 (A_\mu W_\nu^+ A_\nu W_\mu^- - A_\mu A_\nu W_\nu^+ W_\mu^-) + g^2 s_w c_w (A_\mu Z_\nu^0 (W_\mu^+ W_\nu^- - W_\nu^+ W_\mu^-) - \\ & 2A_\mu Z_\mu^0 W_\nu^+ W_\nu^-) - \frac{1}{2}\partial_\mu H \partial_\mu H - 2M^2 \alpha_h H^2 - \partial_\mu \phi^+ \partial_\mu \phi^- - \frac{1}{2}\partial_\mu \phi^0 \partial_\mu \phi^0 - \\ & \beta_h \left(\frac{2M^2}{g^2} + \frac{2M}{g} H + \frac{1}{2}(H^2 + \phi^0 \phi^0 + 2\phi^+ \phi^-) \right) + \frac{2M^4}{g^2} \alpha_h - g \alpha_h M (H^3 + H \phi^0 \phi^0 + 2H \phi^+ \phi^-) - \\ & \frac{1}{8}g^2 \alpha_h (H^4 + (\phi^0)^4 + 4(\phi^+ \phi^-)^2 + 4(\phi^0)^2 \phi^+ \phi^- + 4H^2 \phi^+ \phi^- + 2(\phi^0)^2 H^2) - g M W_\mu^+ W_\mu^- H - \\ & \frac{1}{2}g \frac{M}{c_w} Z_\mu^0 Z_\mu^0 H - \frac{1}{2}ig (W_\mu^+ (\phi^0 \partial_\mu \phi^- - \phi^- \partial_\mu \phi^0) - W_\mu^- (\phi^0 \partial_\mu \phi^+ - \phi^+ \partial_\mu \phi^0)) + \\ & \frac{1}{2}g (W_\mu^+ (H \partial_\mu \phi^- - \phi^- \partial_\mu H) + W_\mu^- (H \partial_\mu \phi^+ - \phi^+ \partial_\mu H)) + \frac{1}{2}g \frac{1}{c_w} (Z_\mu^0 (H \partial_\mu \phi^0 - \phi^0 \partial_\mu H) + \\ & M (\frac{1}{c_w} Z_\mu^0 \partial_\mu \phi^0 + W_\mu^+ \partial_\mu \phi^- + W_\mu^- \partial_\mu \phi^+) - ig \frac{2M}{c_w} M Z_\mu^0 (W_\mu^+ \phi^- - W_\mu^- \phi^+) + ig s_w M A_\mu (W_\mu^+ \phi^- - \\ & W_\mu^- \phi^+) - ig \frac{1-2c_w^2}{2c_w} Z_\mu^0 (\phi^+ \partial_\mu \phi^- - \phi^- \partial_\mu \phi^+) + ig s_w A_\mu (\phi^+ \partial_\mu \phi^- - \phi^- \partial_\mu \phi^+) - \\ & \frac{1}{4}g^2 W_\mu^+ W_\mu^- (H^2 + (\phi^0)^2 + 2\phi^+ \phi^-) - \frac{1}{8}g^2 \frac{1}{c_w} Z_\mu^0 Z_\mu^0 (H^2 + (\phi^0)^2 + 2(2s_w^2 - 1)\phi^+ \phi^-) - \\ & \frac{1}{2}g^2 \frac{2s_w^2}{c_w} Z_\mu^0 \phi^0 (W_\mu^+ \phi^- + W_\mu^- \phi^+) - \frac{1}{2}ig^2 \frac{s_w^2}{c_w} Z_\mu^0 H (W_\mu^+ \phi^- - W_\mu^- \phi^+) + \frac{1}{2}g^2 s_w A_\mu \phi^0 (W_\mu^+ \phi^- + \\ & W_\mu^- \phi^+) + \frac{1}{2}ig^2 s_w A_\mu H (W_\mu^+ \phi^- - W_\mu^- \phi^+) - g^2 \frac{s_w^2}{c_w} (2c_w^2 - 1) Z_\mu^0 A_\mu \phi^+ \phi^- - g^2 s_w^2 A_\mu A_\mu \phi^+ \phi^- + \\ & \frac{1}{2}ig s \lambda_{ij}^a (\bar{q}_i^\sigma \gamma^\mu q_j^\sigma) g_\mu^a - \bar{e}^\lambda (\gamma^\mu \partial + m_e^\lambda) e^\lambda - \bar{\nu}^\lambda (\gamma^\mu \partial + m_\nu^\lambda) \nu^\lambda - \bar{u}_j^\lambda (\gamma^\mu \partial + m_u^\lambda) u_j^\lambda - \bar{d}_j^\lambda (\gamma^\mu \partial + m_d^\lambda) d_j^\lambda + \\ & ig s_w A_\mu \left(-(\bar{e}^\lambda \gamma^\mu e^\lambda) + \frac{2}{3}(\bar{u}_j^\lambda \gamma^\mu u_j^\lambda) - \frac{1}{3}(\bar{d}_j^\lambda \gamma^\mu d_j^\lambda) \right) + \frac{ig}{4c_w} Z_\mu^0 \{ (\bar{\nu}^\lambda \gamma^\mu (1 + \gamma^5) \nu^\lambda) + (\bar{e}^\lambda \gamma^\mu (4s_w^2 - \\ & 1 - \gamma^5) e^\lambda) + (\bar{d}_j^\lambda \gamma^\mu (\frac{2}{3}s_w^2 - 1 - \gamma^5) d_j^\lambda) + (\bar{u}_j^\lambda \gamma^\mu (1 - \frac{8}{3}s_w^2 + \gamma^5) u_j^\lambda) \} + \\ & \frac{ig}{2\sqrt{2}} W_\mu^+ \left((\bar{\nu}^\lambda \gamma^\mu (1 + \gamma^5) U^{lep}{}_{\lambda\kappa} e^\kappa) + (\bar{u}_j^\lambda \gamma^\mu (1 + \gamma^5) C_{\lambda\kappa} d_j^\kappa) \right) + \\ & \frac{ig}{2\sqrt{2}} W_\mu^- \left((\bar{e}^\kappa U^{lep}{}_{\kappa\lambda} \gamma^\mu (1 + \gamma^5) \nu^\lambda) + (\bar{d}_j^\kappa C_{\kappa\lambda}^\dagger \gamma^\mu (1 + \gamma^5) u_j^\lambda) \right) + \\ & \frac{ig}{2M\sqrt{2}} \phi^- \left(-m_e^\lambda (\bar{\nu}^\lambda U^{lep}{}_{\lambda\kappa} (1 - \gamma^5) e^\kappa) + m_\nu^\lambda (\bar{\nu}^\lambda U^{lep}{}_{\lambda\kappa} (1 + \gamma^5) e^\kappa) + \right. \\ & \left. \frac{ig}{2M\sqrt{2}} \phi^- \left(m_e^\lambda (\bar{e}^\lambda U^{lep}{}_{\lambda\kappa}^\dagger (1 + \gamma^5) \nu^\kappa) - m_\nu^\lambda (\bar{e}^\lambda U^{lep}{}_{\lambda\kappa}^\dagger (1 - \gamma^5) \nu^\kappa) - \frac{g}{2} \frac{m_\nu^\lambda}{M} H (\bar{\nu}^\lambda \nu^\lambda) - \right. \right. \\ & \left. \left. \frac{g}{2} \frac{m_\lambda^\lambda}{M} H (\bar{e}^\lambda e^\lambda) + \frac{ig}{2} \frac{m_\lambda^\lambda}{M} \phi^0 (\bar{\nu}^\lambda \gamma^5 \nu^\lambda) - \frac{ig}{2} \frac{m_\lambda^\lambda}{M} \phi^0 (\bar{e}^\lambda \gamma^5 e^\lambda) - \frac{1}{4} \bar{\nu}_\lambda M_{\lambda\kappa}^R (1 - \gamma_5) \bar{\nu}_\kappa - \right. \right. \\ & \left. \left. \frac{1}{4} \bar{\nu}_\lambda M_{\lambda\kappa}^R (1 - \gamma_5) \bar{\nu}_\kappa + \frac{ig}{2M\sqrt{2}} \phi^+ \left(-m_d^\lambda (\bar{u}_j^\lambda C_{\lambda\kappa} (1 - \gamma^5) d_j^\kappa) + m_u^\lambda (\bar{u}_j^\lambda C_{\lambda\kappa} (1 + \gamma^5) d_j^\kappa) + \right. \right. \\ & \left. \left. \frac{ig}{2M\sqrt{2}} \phi^- \left(m_d^\lambda (\bar{d}_j^\lambda C_{\lambda\kappa}^\dagger (1 + \gamma^5) u_j^\kappa) - m_u^\lambda (\bar{d}_j^\lambda C_{\lambda\kappa}^\dagger (1 - \gamma^5) u_j^\kappa) - \frac{g}{2} \frac{m_\lambda^\lambda}{M} H (\bar{u}_j^\lambda u_j^\lambda) - \frac{g}{2} \frac{m_\lambda^\lambda}{M} H (\bar{d}_j^\lambda d_j^\lambda) + \right. \right. \\ & \left. \left. \frac{ig}{2} \frac{m_\lambda^\lambda}{M} \phi^0 (\bar{u}_j^\lambda \gamma^5 u_j^\lambda) - \frac{ig}{2} \frac{m_\lambda^\lambda}{M} \phi^0 (\bar{d}_j^\lambda \gamma^5 d_j^\lambda) \right) \right) \end{aligned}$$

Standard Model Lagrangian



Limits of Explanations

“You can ask a human, but, you know, what cognitive psychologists have discovered is that when you ask a human you’re not really getting at the decision process. They make a decision first, and then you ask, and then they generate an explanation and that may not be the true explanation.”

- Peter Norvig



Desiderata for Interpretability in ML

- Causality
 - The model can find causal links between variables and the outcome e.g., smoking and lung cancer. *Note:* Supervised Learning Algorithms learn associations which may not be causal.
- Transferability
 - The results should generalize from the model building setting to the deployment setting even beyond issues related to overfitting e.g., be resistant to adversarial attacks.
- Informativeness
 - Provide useful information about the model variables and the target
- Fair Decision Making
 - Free from bias



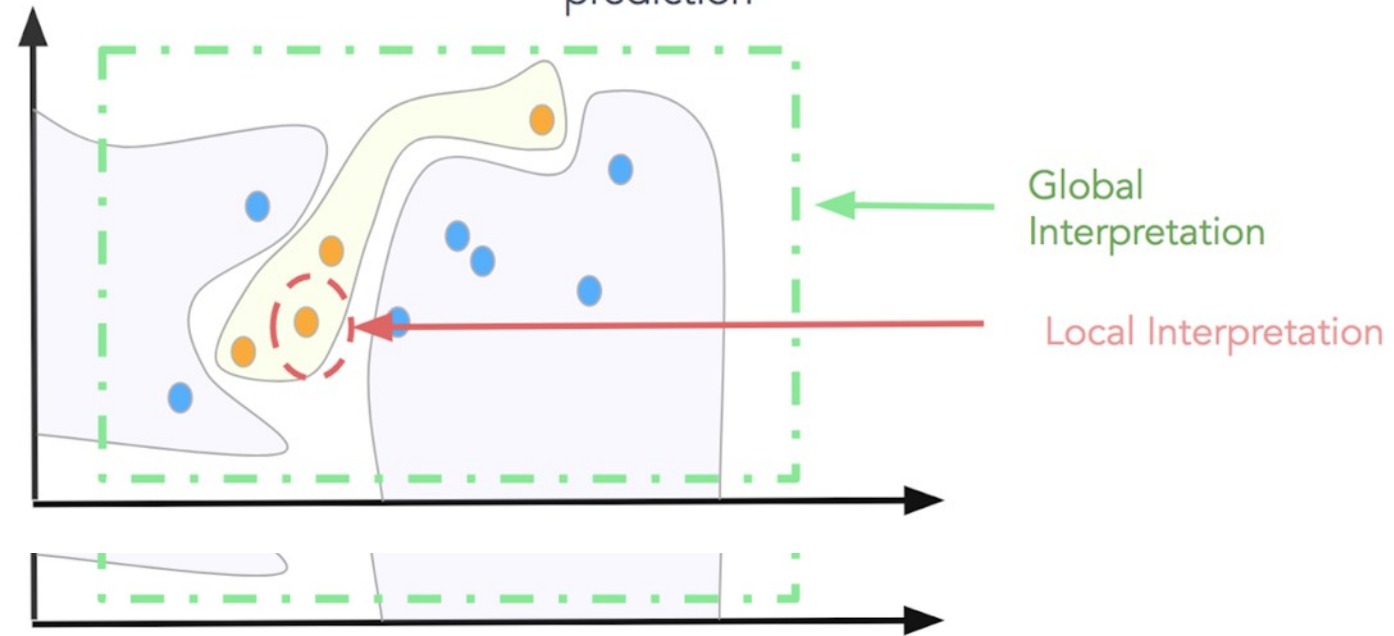
Global vs. Local

Global Interpretation

Being able to explain the conditional interaction between dependent(*response*) variables and independent(*predictor, or explanatory*) variables based on the complete dataset

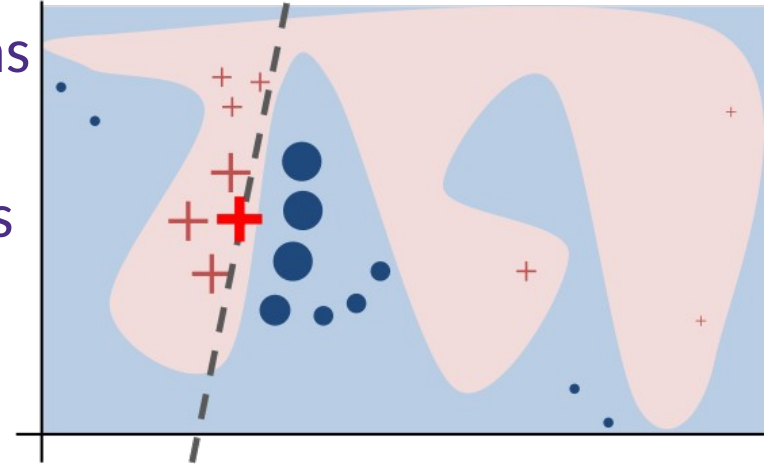
Local Interpretation

Being able to explain the conditional interaction between dependent(*response*) variables and independent(*predictor, or explanatory*) variables wrt to a single prediction

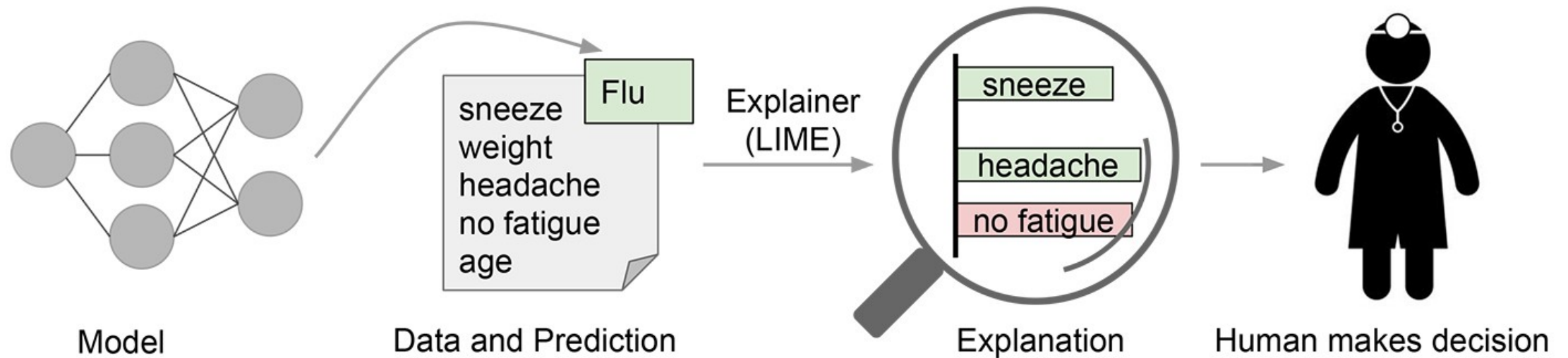


LIME

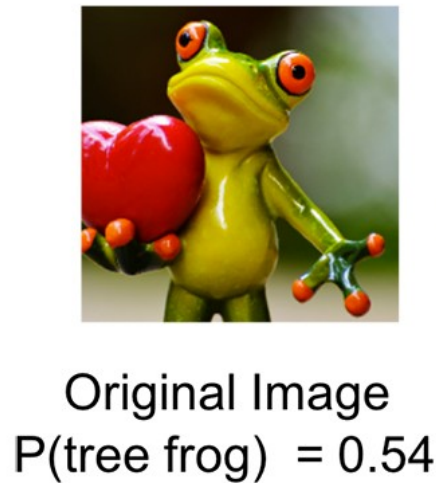
- Locally Interpretable Model-agnostic Explanations
- Given a non-interpretable model create local interpretable models that are good approximates of the data at the local level
- The local model can be linear models like regression models
- LIME simulates the distribution of the data locally
- Can be problematic if the decision boundary is not linear locally





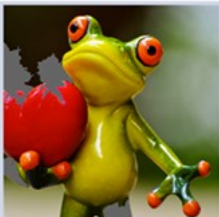



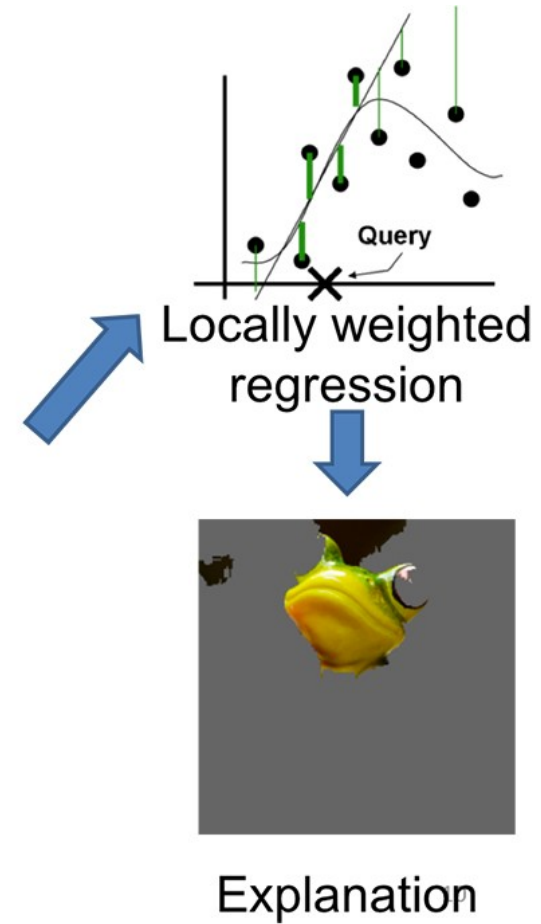
LIME: Model Explanations



LIME: Explanations



Perturbed Instances	$P(\text{tree frog})$
	 0.85
	 0.00001
	 0.52

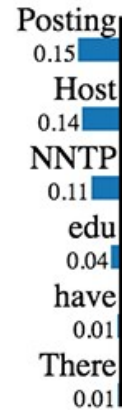


LIME: Explanations

Prediction probabilities



atheism



christian

Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

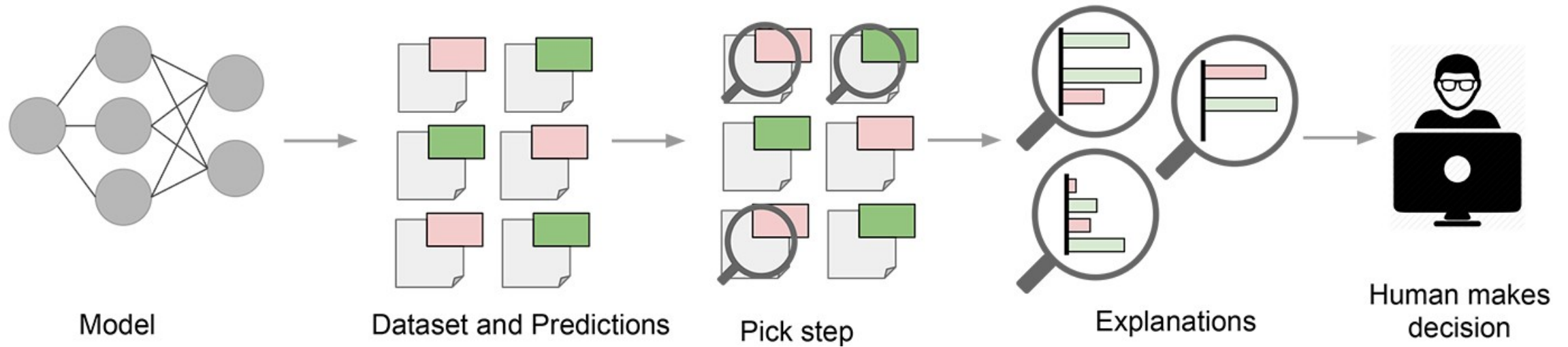
Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the net. If anyone has a contact please post on the net or email me.

Example of predicting the right label because of an artifact of the data



LIME: Explanation Actionability



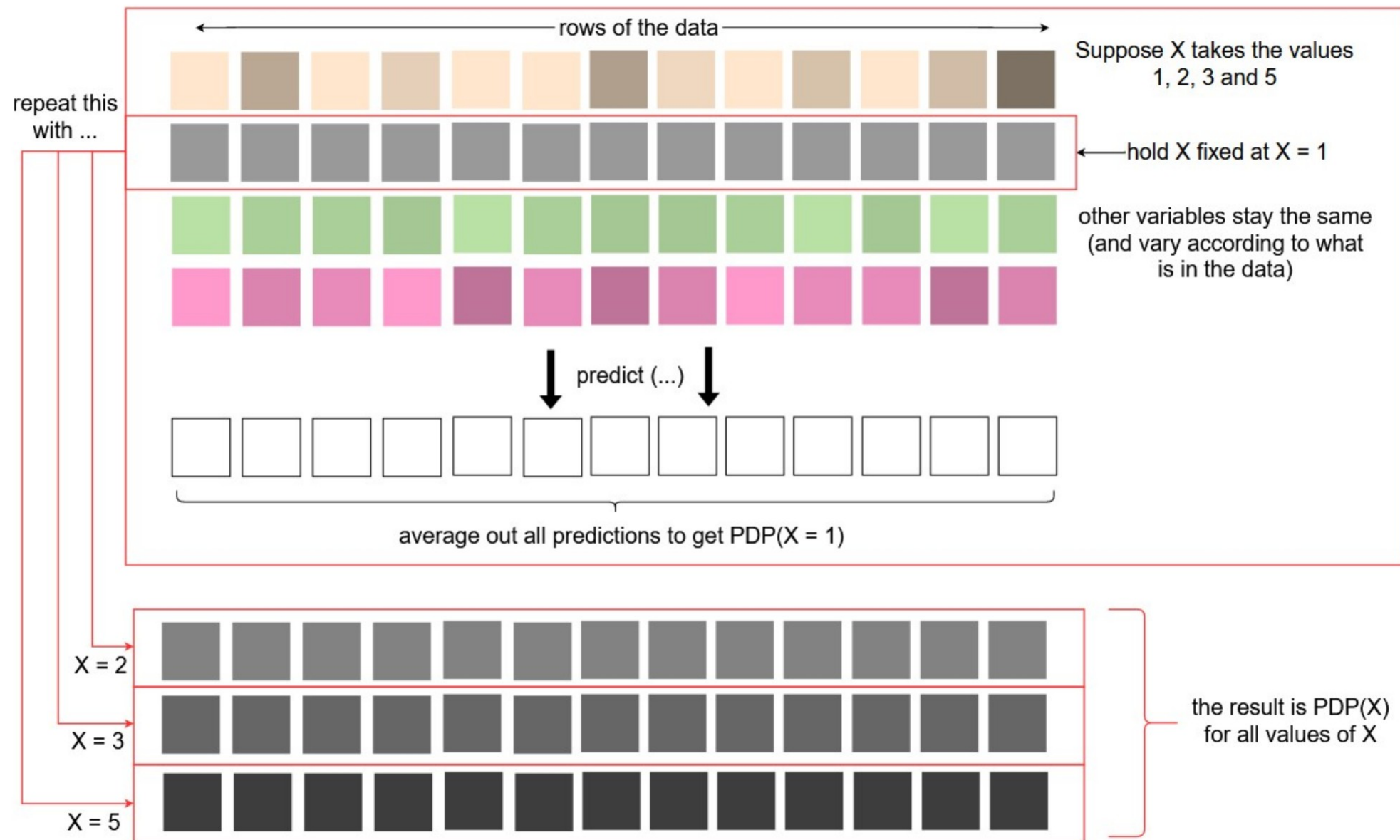
partial dependence plots

Let S be a subset of the feature set, i.e. $S \subset \{X_1, \dots, X_p\}$. The PDP of S is

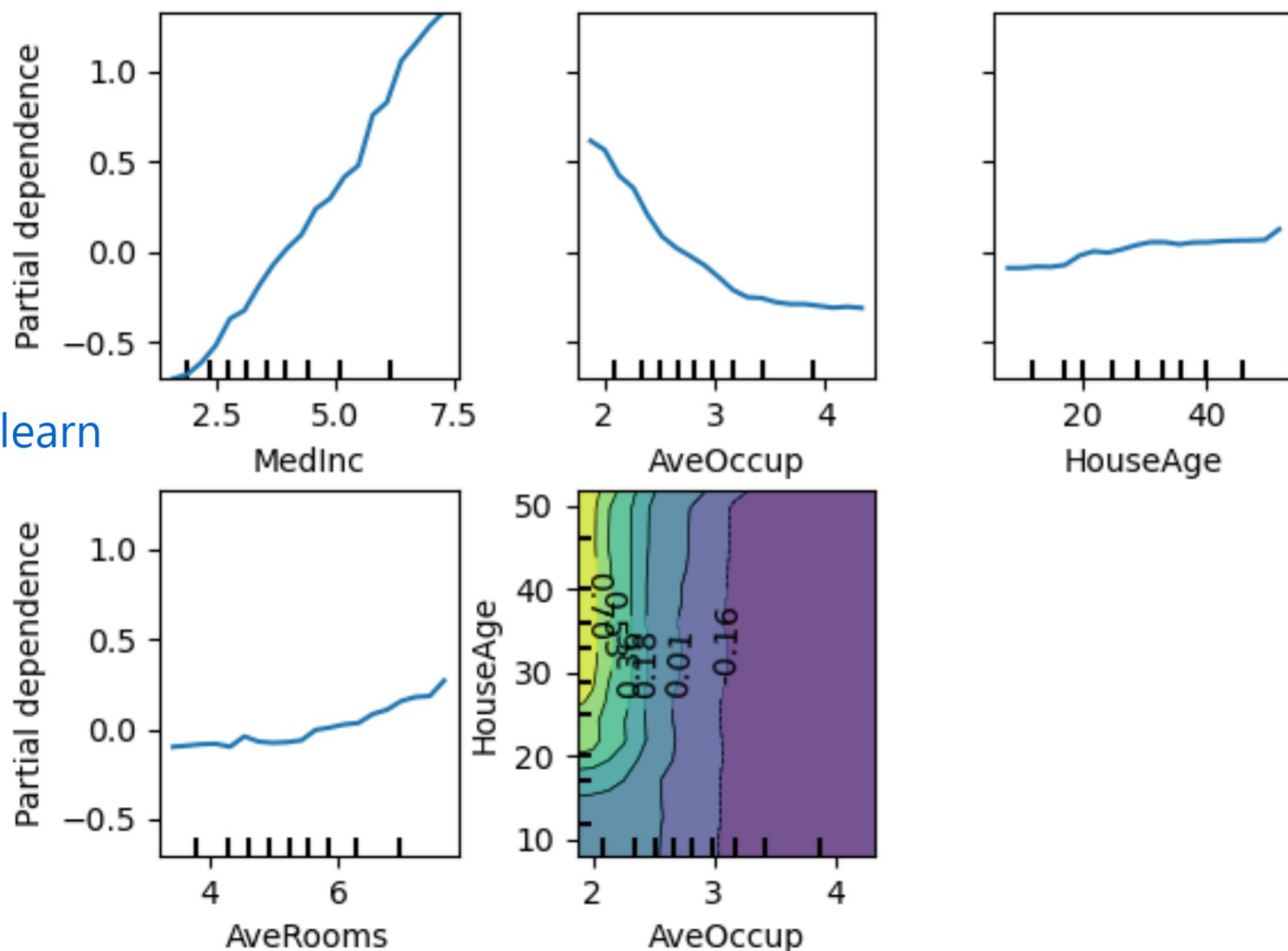
$$\hat{f}_S(x \in S) = \frac{1}{n} \sum_{i=1}^n \hat{f}(x, S')$$

where S' is the complement of S (i.e. S' is the set of all *other* features), and \hat{f} is the prediction function of some **black box** model we want to explain.

The PDP is a function of S , that fixes all features in S to some value x and the remaining features are marginalized. So we reduce the joint probability over all features to the probability over S . We then average the predictions over the resulting probability distribution.



Partial dependence of house value on non-location features for the California housing dataset, with Gradient Boosting



source: [sklearn](#)

discussion

Here are some of the properties of PDPs. Read each and try to find a good reason for **why** PDPs have these properties.

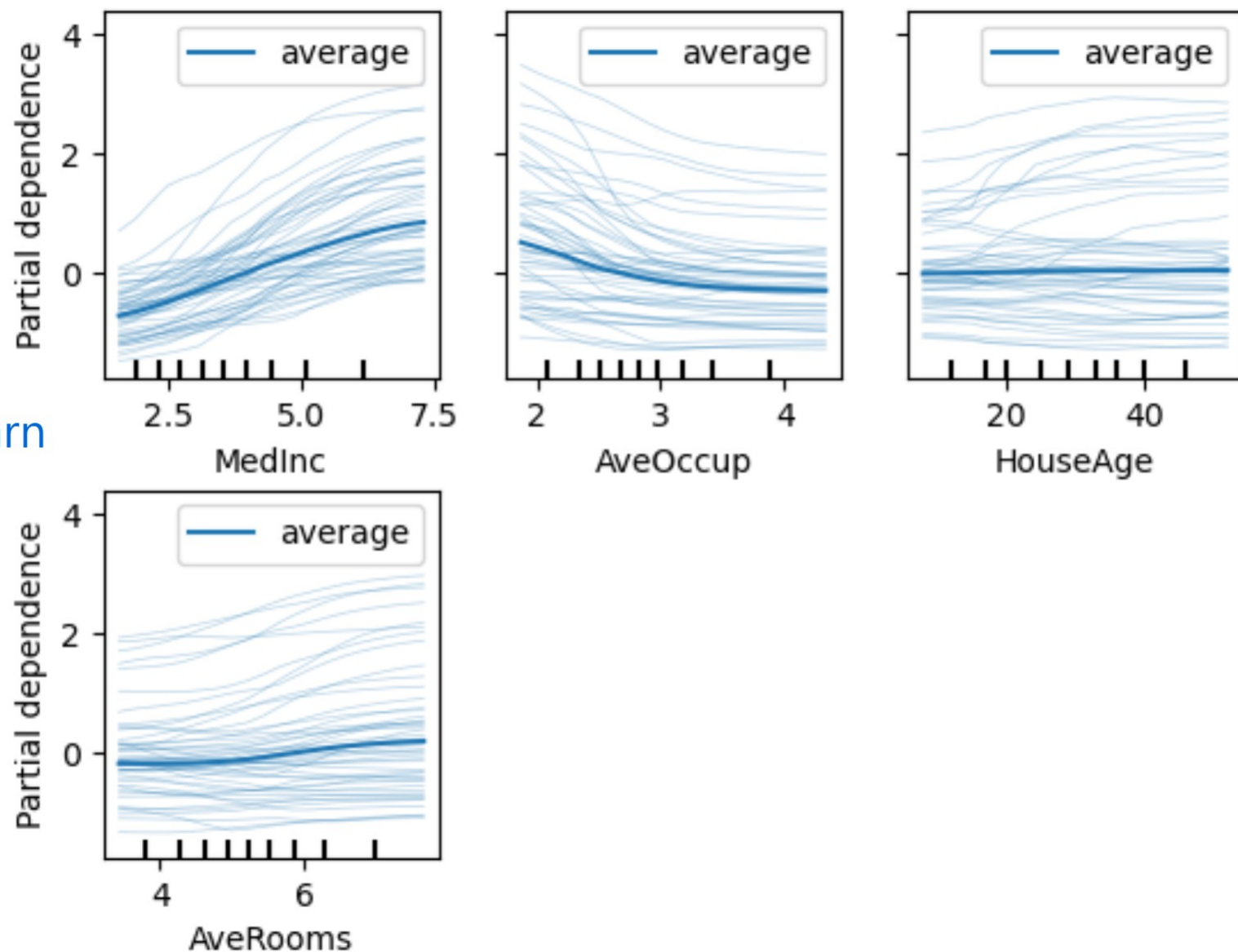
- in most cases, S usually contains **only one feature** $S = \{X_j\}$, **or two at most** $S = \{X_i, X_j\}$. Otherwise visualizing results is a challenge.
- PDPs work great when the feature size is small, and the features are mostly independent, but strong interactions can throw off PDPs
- it's important to take the distribution of S into account when looking at a PDP, because low density ranges can be unreliable

individual conditional expectation

If we skip the averaging in the PDP calculation, we end up with one PDP plot **per instance** instead of one over the whole data: we call this the individual conditional expectation (ICE), unfortunately a bit of a misnomer.

- ICE plots can provide a hint as to how the plotted feature might interact with other features in the data (can you guess how?)
- ICE plots are sometimes centered from the starting point to make comparisons easier
- it's easy to overplot with ICE plots, so sampling is recommended. If we expect interactions with a known categorical feature, we can try color coding to see if it shows

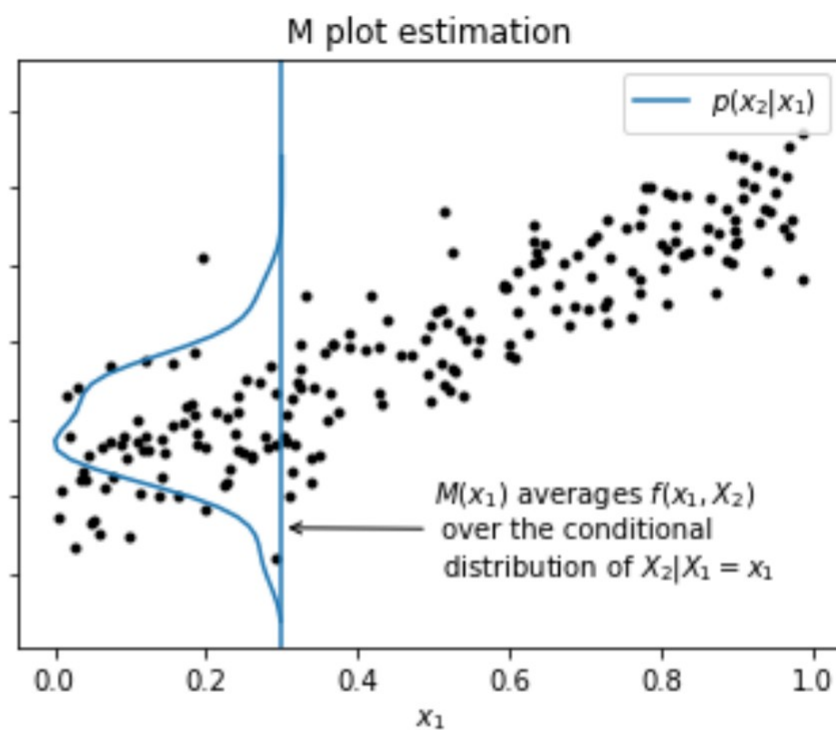
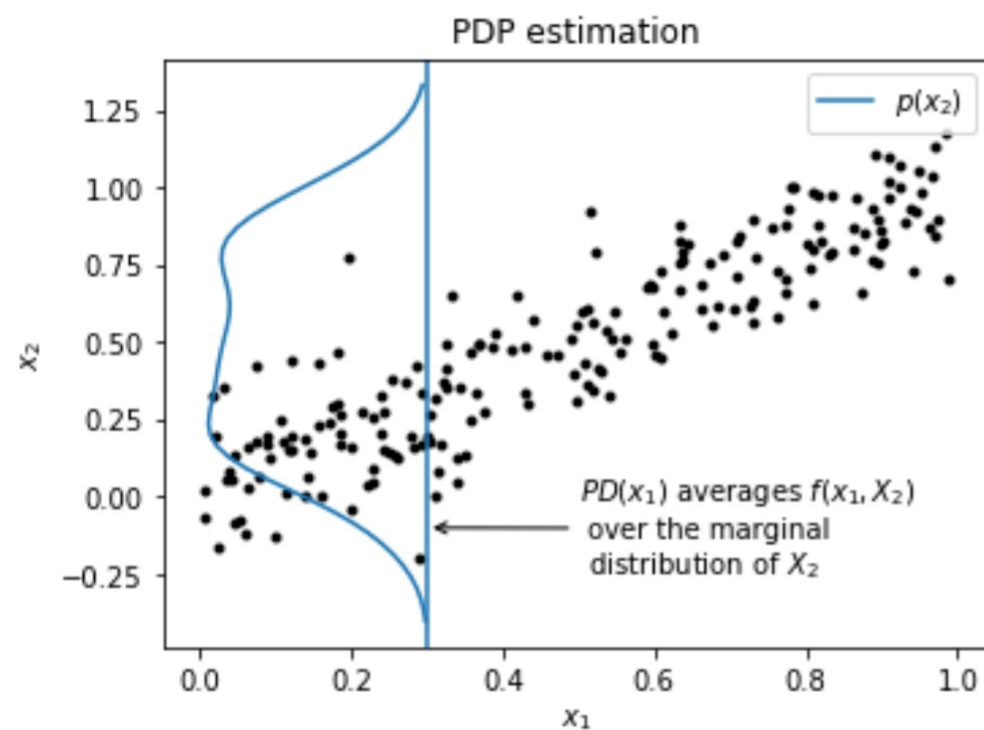
Partial dependence of house value on non-location features
for the California housing dataset, with MLPRegressor



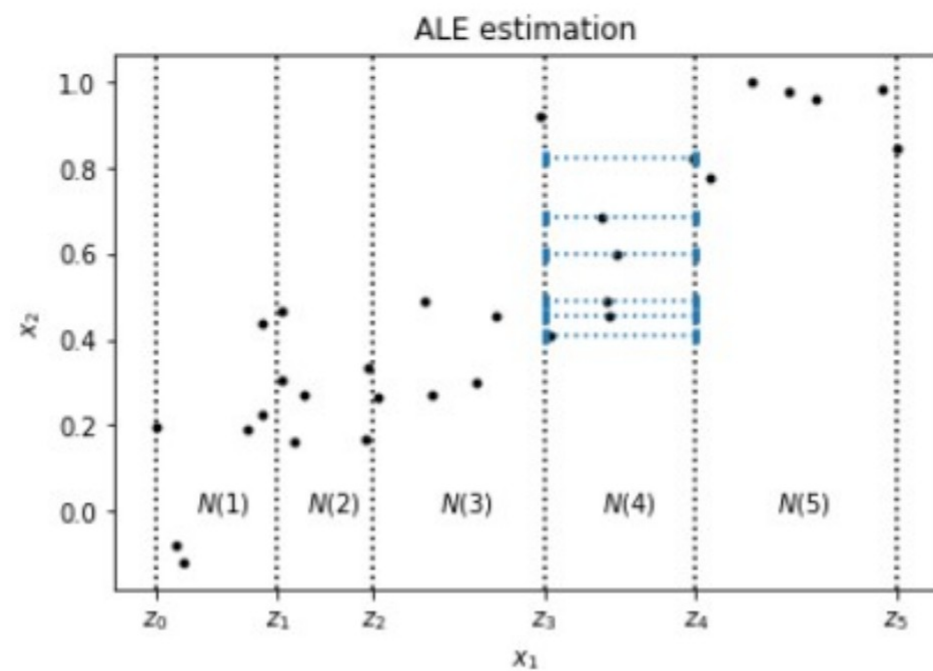
source: [sklearn](#)

improving on PDP

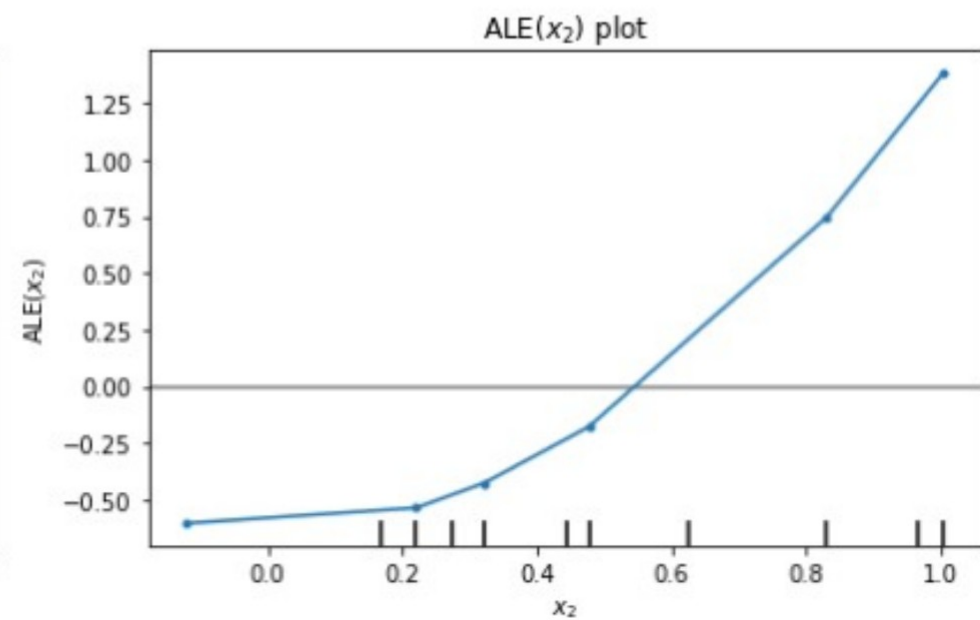
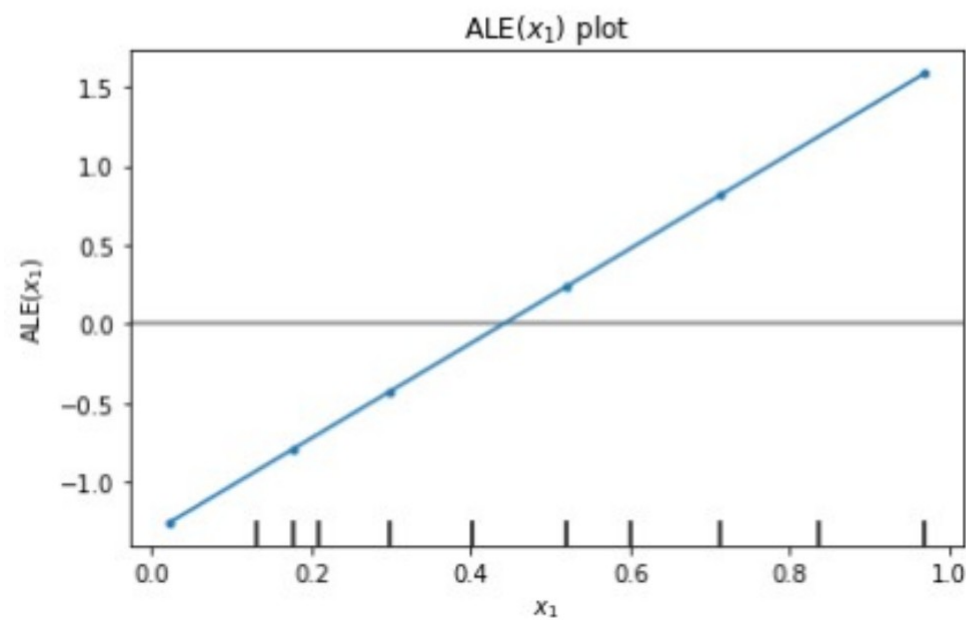
- if we use the **conditional distribution** $\hat{f}(x|S')$ instead of $\hat{f}(x, S')$ in the calculation, the result is called an **M-plot**
 - using the conditional distribution makes it more likely to find "reasonable" values for S' given the fixed features in S (see the next slide)
 - the result is that instead of averaging over S' over the whole data like with PDPs, we average only over a subset of the data in the neighborhood of x
- if instead of averaging the predictions, we average the *difference* in prediction in a neighborhood around x , the result is called an **accumulated local effects** (ALE) plot: unlike PDP and M-plots, ALE plots isolate the effect of main effects from that of their interactions
 - we can plot a separate ALE plot for the interactive term just to see effect of the interaction, but looking at too many plots is not good for one's sanity



source: [alibi](#)



source: [alibi](#)



To summarize how each type of plot (PDP, M, ALE) calculates the effect of a feature at a certain grid value v :

Partial Dependence Plots: “Let me show you what the model predicts on average when each data instance has the value v for that feature. I ignore whether the value v makes sense for all data instances.”

M-Plots: “Let me show you what the model predicts on average for data instances that have values close to v for that feature. The effect could be due to that feature, but also due to correlated features.”

ALE plots: “Let me show you how the model predictions change in a small “window” of the feature around v for data instances in that window.”

Shapley Value

- Classic result in game theory on distributing the total gain from a **cooperative game**
- Introduced by **Lloyd Shapley** in **1953**¹, who later won the **Nobel Prize in Economics** in the 2012
- Popular tool in studying cost-sharing, market analytics, voting power, and most recently **explaining ML models**



Lloyd Shapley in 1980

¹ "A Value for n-person Games". Contributions to the Theory of Games 2.28 (1953): 307-317



Cooperative Game

- Players $\{1, \dots, M\}$ collaborating to generate some **gain**
 - Think: Employees in a company creating some profit
 - Described by a **set function** $v(S)$ specifying the gain for any subset $S \subseteq \{1, \dots, M\}$
- **Shapley values** are a fair way to attribute the total gain to the players
 - Think: Bonus allocation to the employees
 - Shapley values are commensurate with the player's contribution



Shapley Value Algorithm [Conceptual]

$$\phi_i(v) = \mathbb{E}_{\mathbf{O} \sim \pi(M)} [v(\text{pre}_i(\mathbf{O}) \cup \{i\}) - v(\text{pre}_i(\mathbf{O}))]$$

- Consider all possible permutations $\pi(M)$ of players ($M!$ possibilities)
- In each permutation $\mathbf{O} \sim \pi(M)$
 - Add players to the coalition in that order
 - Note the marginal contribution of each player i to set of players before it in the permutation, i.e. $v(\text{pre}_i(\mathbf{O}) \cup \{i\}) - v(\text{pre}_i(\mathbf{O}))$
- The average marginal contribution across all permutations is the Shapley Value



Quiz

A company with two employees **Alice** and

Bob

- No employees, no profit $[v(\{\}) = 0]$
- Alice alone makes 20 units of profit $[v(\{\text{Alice}\}) = 20]$
- Bob alone makes 10 units of profit $[v(\{\text{Bob}\}) = 10]$
- Alice and Bob make 50 units of profit $[v(\{\text{Alice}, \text{Bob}\}) = 50]$

What should the bonuses be?



Quiz

A company with two employees **Alice** and **Bob**

- No employees, no profit $[v(\{\}) = 0]$
- Alice alone makes 20 units of profit $[v(\{\text{Alice}\}) = 20]$
- Bob alone makes 10 units of profit $[v(\{\text{Bob}\}) = 10]$
- Alice and Bob make 50 units of profit $[v(\{\text{Alice}, \text{Bob}\}) = 50]$

What should the bonuses be?

Permutation	Marginal for Alice	Marginal for Bob
Alice, Bob	20	30
Bob, Alice	40	10
Shapley Value	30	20



Axiomatic Justification

Shapley values are **unique** under four simple axioms

- **Dummy:** A player that doesn't contribute to any subset of players must receive zero attribution
- **Efficiency:** Attributions must add to the total gain
- **Symmetry:** Symmetric players must receive equal attribution
- **Linearity:** Attribution for the (weighted) sum of two games must be the same as the (weighted) sum of the attributions for each of the games



Computing Shapley Values

Exact computation

- **Permutations-based approach** (Complexity: $O(M!)$)

$$\phi_i(v) = \mathbb{E}_{\mathbf{O} \sim \pi(M)} [v(\text{pre}_i(\mathbf{O}) \cup \{i\}) - v(\text{pre}_i(\mathbf{O}))]$$

- **Subsets-based approach** (Complexity: $O(2^M)$)

$$\phi_i(v) = \mathbb{E}_S \left[\frac{2^{M-1}}{M} \binom{M-1}{|S|}^{-1} (v(S \cup \{i\}) - v(S)) \right]$$

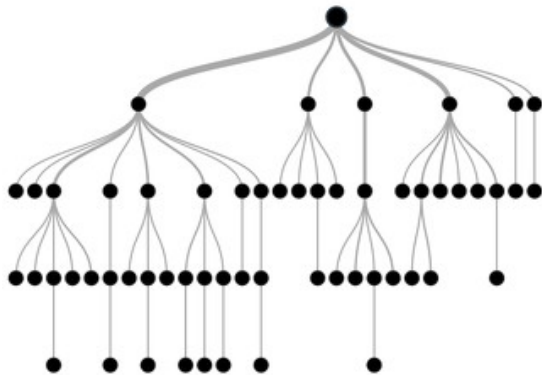
- KernelSHAP: Solve a weighted least squares problem (Complexity: $O(2^M)$)

$$\phi = \arg \min_{\phi} \sum_{S \subseteq \mathcal{M}} \frac{M-1}{\binom{M}{|S|} |S| (M-|S|)} \left(v(S) - \sum_{i=1}^M \phi_i \right)^2$$



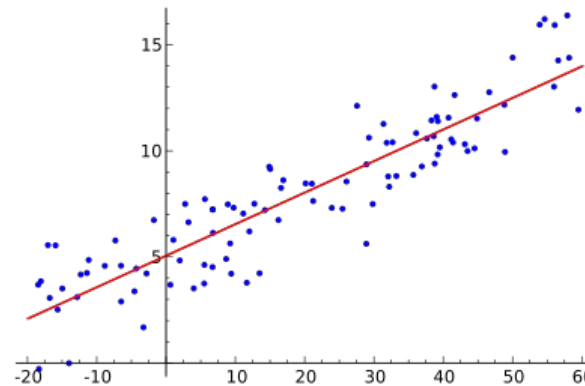
Taxonomy of Interpretable Models By Model Output

Rule Based Models



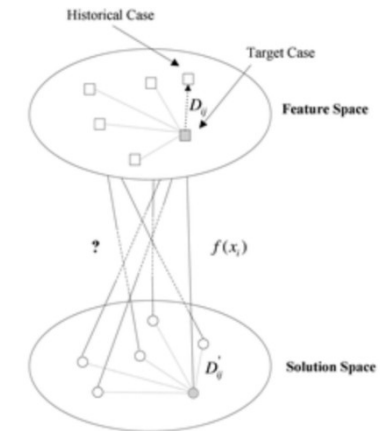
A set of rules that describe how the prediction was made e.g., propositional rule learners, Bayesian rule lists, decision trees etc.

Relative Variable Importance



Linear models and related families of predictors that show relative importance and ordering of variables

Case Based Models



Similarity function based models that use similar cases for prediction i.e., a prediction was made because this case looked like these k other cases



Taxonomy of Interpretable Models

By Model Composition

First Order Models

Models which have explanations built in as part of the model building process

Examples:

Decision Trees, GAMs, CENs etc.

Second Order Models

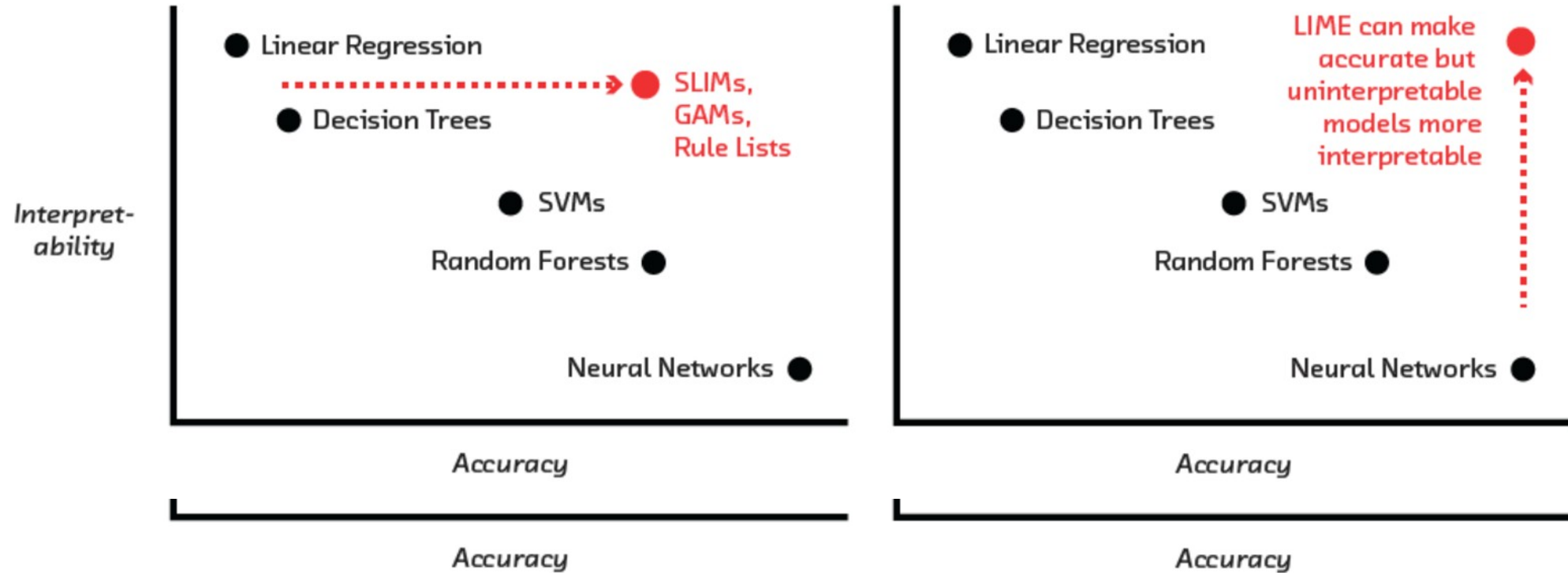
Models that are built on top of black box models to provide explanations.

Local Models: LIME family of models, ASTRID, Random Forest Explainers,

Global/Student Models: Student models that train on model outputs of other models



Accuracy vs. Interpretability



Feature Importance

- Global feature importance values give an indication of the magnitude of a feature's contribution to model predictions for all observations
- Unlike regression parameters, they are often unsigned and typically not directly related to the numerical predictions of the model
- Local feature importance describes how the combination of the learned model rules or parameters and an individual observation's attributes affect a model's prediction for that observation while taking nonlinearity and interactions into effect.



The Future: All AI Systems Explain Themselves

- Given the proliferation of ML systems in healthcare and the potential impact on the lives of patients, the need for explanations will only increase
- Only a matter of time before regulations comes to effect
- Automation will bring greater scrutiny from practitioners
- Lack of vigilance may lead to a 'driverless car crash' moment in ML in healthcare

