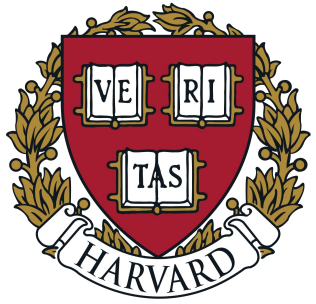
Bringing Order to Chaos: Probing the Disagreement Problem in XAI

Hima Lakkaraju

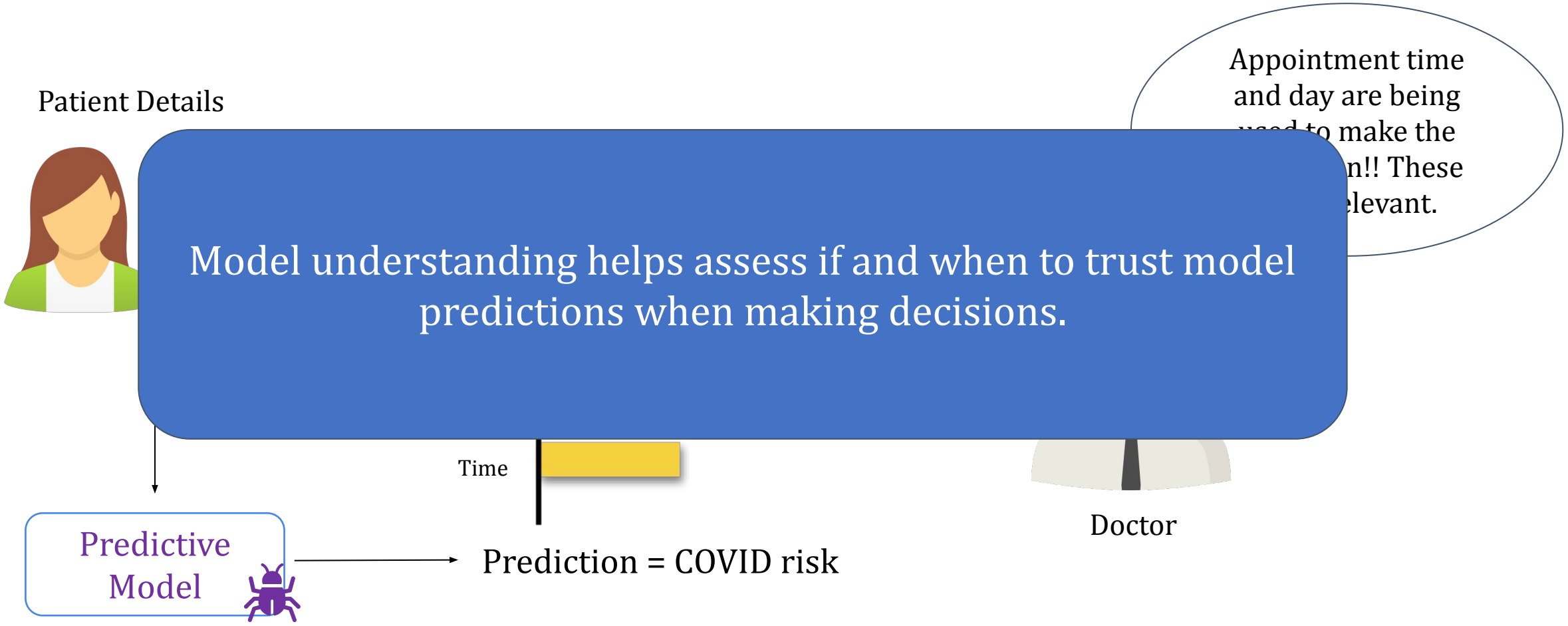


Motivation

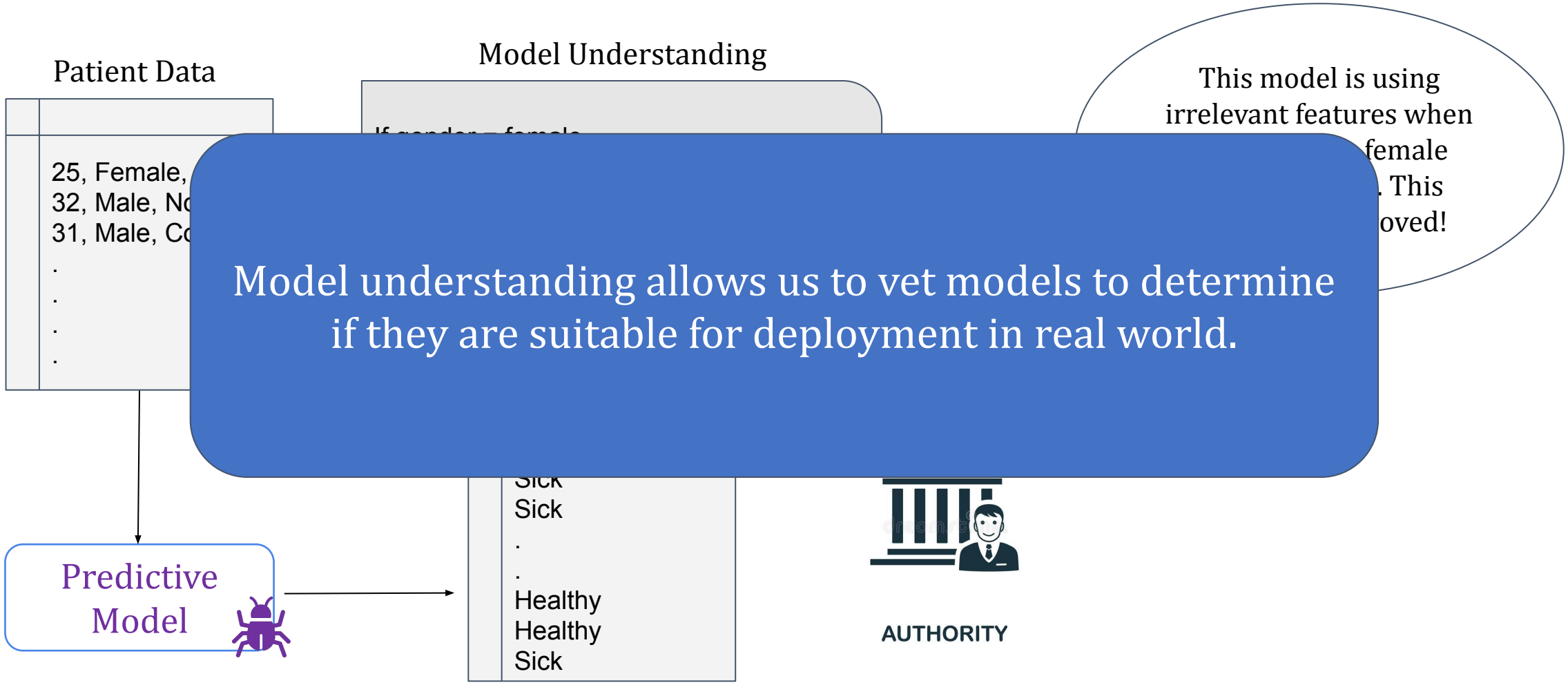
Model understanding is absolutely critical in several domains -- particularly those involving *high stakes decisions*!

[Larson et. al. 2016]

Motivation: Why Model Understanding?



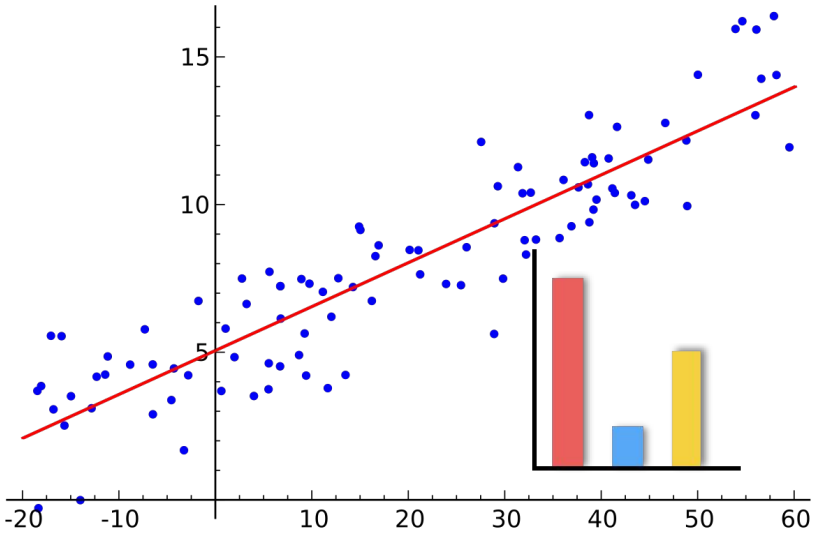
Motivation: Why Model Understanding?

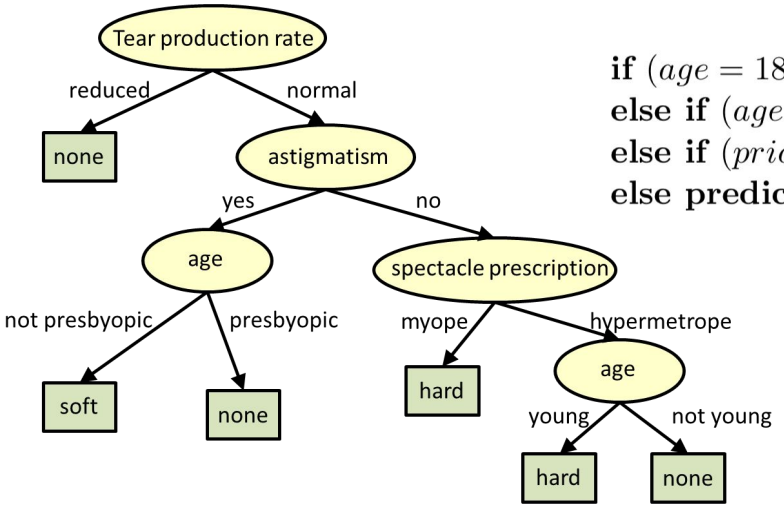


[Letham and Rudin 2015; Lakkaraju et. al. 2016]

Achieving Model Understanding

Take 1: Build *inherently interpretable* predictive models



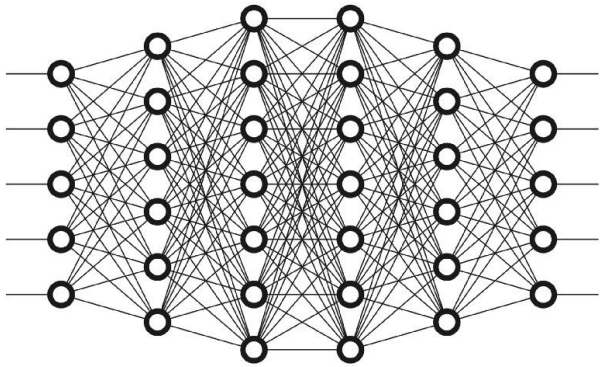


if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

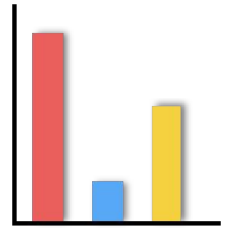
[Ribeiro et. al. 2016, 2018; Lakkaraju et. al. 2019]

Achieving Model Understanding

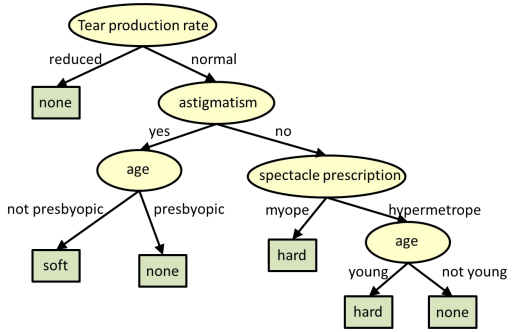
Take 2: *Explain* pre-built models *in a post-hoc manner*



Explainer

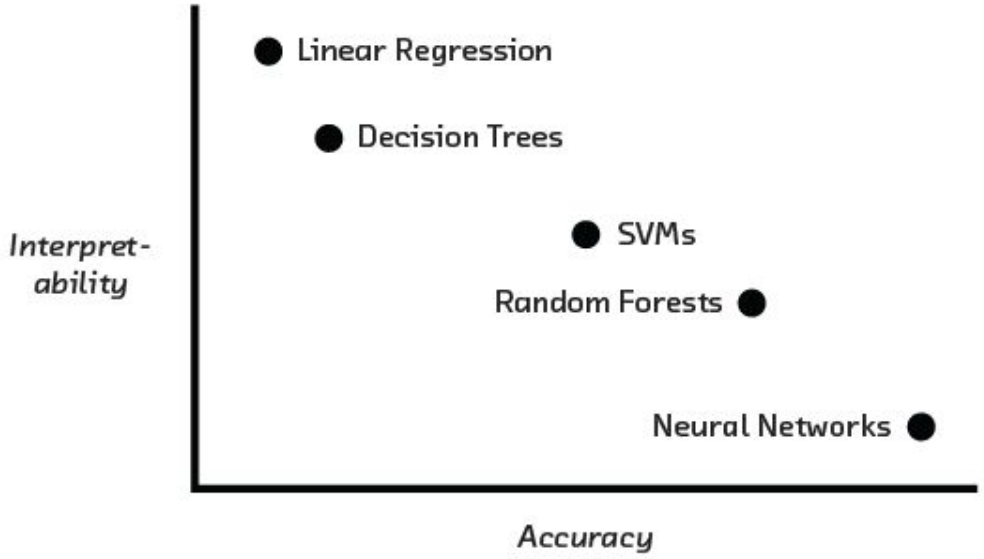


if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no



[Cireşan et. al. 2012, Caruana et. al. 2006, Frosst et. al. 2017, Stewart 2020]

Inherently Interpretable Models vs. Post hoc Explanations



In ***certain*** settings, *accuracy-interpretability trade offs* may exist.

In ***certain*** settings, you may just have access to a (proprietary) black box.

Feature Attribution Based Local Explanations

- Local explanations
  - explain individual predictions of any classifier
- Output feature attributions for individual instances, which capture the effect/contribution of each feature on the black box prediction
- Examples: LIME, SHAP, Gradient, Gradient times Input, SmoothGrad, Integrated Gradients

Disagreement Problem in XAI: Overview

- Study to understand:
  - if and how often feature attribution based explanation methods disagree with each other in practice
  - What constitutes disagreement between these explanations, and how to formalize the notion of explanation disagreement based on practitioner inputs?
  - How do practitioners resolve explanation disagreement?

Krishna and Han et. al., 2022

Practitioner Inputs on Explanation Disagreement

- 30 minute semi-structured interviews with 25 data scientists
- 84% of participants said they often encountered disagreement between explanation methods
- Characterizing disagreement:
  - Top features are different
  - Ordering among top features is different
  - Direction of top feature contributions is different
  - Relative ordering of features of interest is different

Practitioner Inputs on Explanation Disagreement

- Participants typically characterize explanation disagreement based on factors such as:
  - mismatch in top features,
  - feature ordering, and
  - directions of feature contributions,
  - But NOT on the feature importance values output by different explanation methods

- 24 out of 25 participants (96%) in our study opine that feature importance values output by different explanation methods are not directly comparable

Practitioner Inputs on Explanation Disagreement

- **Quote**: "*The values generated by different explanation methods are clearly different. So, I would not characterize disagreement based on that. But, I would at least want the explanations they output to give me consistent insights. The explanations should agree on what are the most important features, the ordering among them and so on for me to derive consistent insights. But, they don't!*"

Formalizing the Notion of Explanation Disagreement (Top K)

$$FeatureAgreement(E_a, E_b, k) = \frac{|top\_features(E_a, k) \cap top\_features(E_b, k)|}{k}$$

$$RankAgreement(E_a, E_b, k)$$

$$\frac{|\bigcup_{s \in S} \{s \mid s \in top\_features(E_a, k) \wedge s \in top\_features(E_b, k) \wedge rank(E_a, s) = rank(E_b, s)\}|}{k}$$

$$SignAgreement(E_a, E_b, k)$$

$$\frac{|\bigcup_{s \in S} \{s \mid s \in top\_features(E_a, k) \wedge s \in top\_features(E_b, k) \wedge sign(E_a, s) = sign(E_b, s)\}|}{k}$$

$$SignedRankAgreement(E_a, E_b, k)$$

$$\frac{|\bigcup_{s \in S} \{s \mid s \in top\_features(E_a, k) \wedge s \in top\_features(E_b, k) \wedge sign(E_a, s) = sign(E_b, s) \wedge rank(E_a, s) = rank(E_b, s)\}|}{k}$$

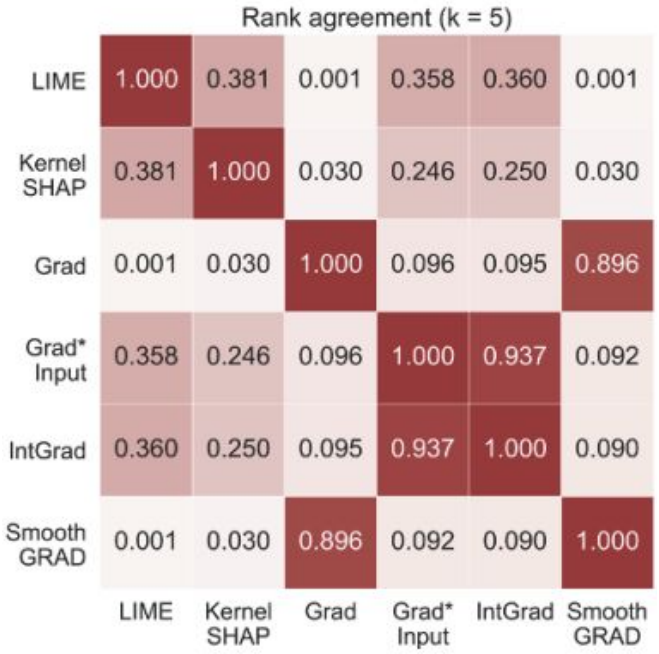
Formalizing the Notion of Explanation Disagreement (Features of Interest)

Spearman rank correlation coefficient computed over features of interest

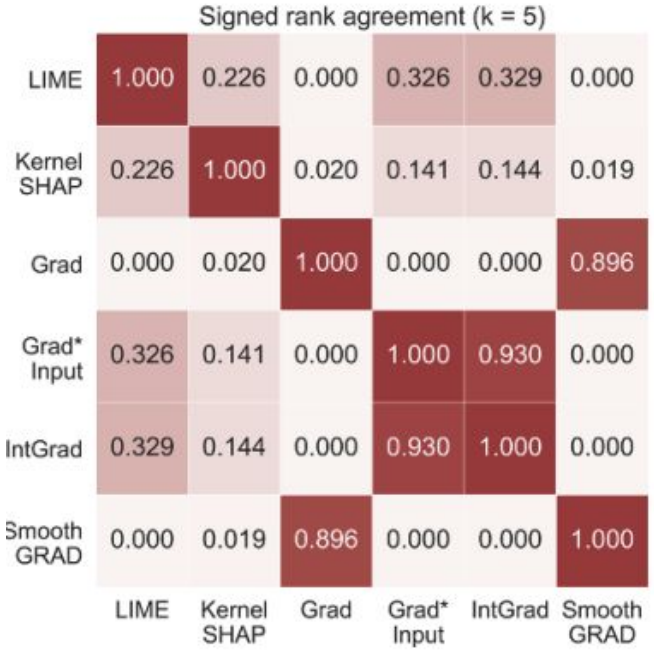
$$RankCorrelation(E_a, E_b, F) = r_s(Ranking(E_a, F), Ranking(E_b, F))$$

$$PairwiseRankAgreement(E_a, E_b, F) = \frac{\sum_{i,j \text{ for } i<j} \mathbb{1}[RelativeRanking(E_a, f_i, f_j) = RelativeRanking(E_b, f_i, f_j)]}{\binom{|F|}{2}}$$

Empirical Analysis of Explanation Disagreement



COMPAS data, NN

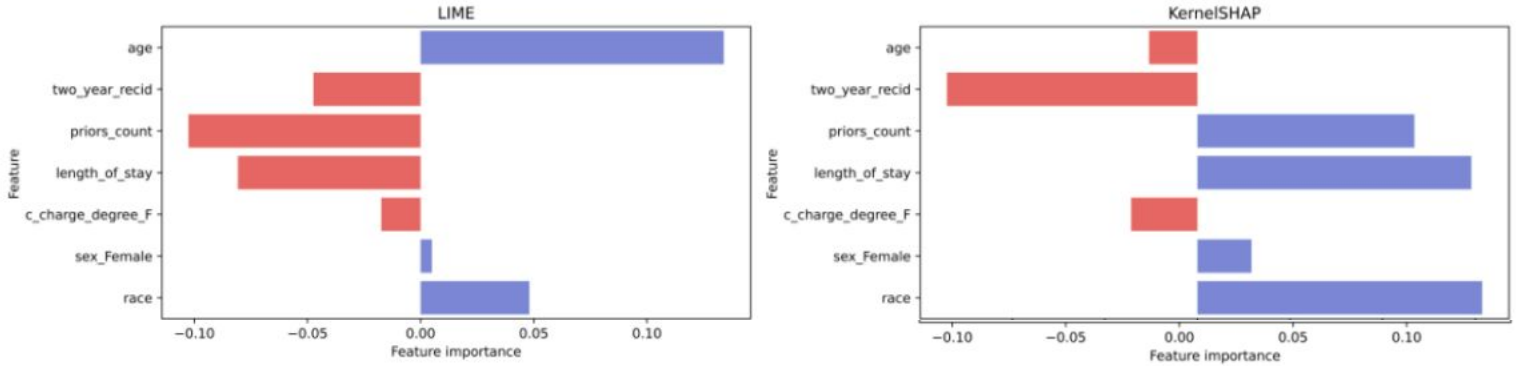


COMPAS data, NN

- We carried out empirical analysis with 6 post hoc explanation methods, 4 real world datasets (tabular, NLP, images), 8 model classes, and found several disagreements between explanation methods

How do Practitioners Resolve Disagreements?

Below, you see a data point, as well as its explanation using methods **LIME** and **KernelSHAP**.



As a reminder, the 7 features of the COMPAS dataset are **age**, **two_year_recid** (whether the defendant recidivated after 2 years of the original crime, **priors_count** (number of prior crimes committed), **length_of_stay** (length the defendant stayed in jail), **c_charge_degree** (whether the previous charge was a Misdemeanor or Felony), **sex**, and **race**

To what extent do you think the two explanations shown above agree or disagree with each other?

○ Completely agree ◉ Mostly agree ○ Mostly disagree ○ Completely disagree

Please explain why you chose the above answer.

Since you believe that the above explanations disagree (to some extent), which explanation would you rely on?

○ LIME explanation ◉ KernelSHAP explanation ○ It depends

Please explain why you chose the above answer.

How do Practitioners Resolve Disagreements?

- Online user study where 25 users were shown explanations that disagree and asked to make a choice, and explain why

- Practitioners are choosing methods due to:
  - Associated theory or publication time (33%)
  - Explanations matching human intuition better (32%)
  - Type of data (23%)
    - E.g., LIME or SHAP are better for tabular data

How do Practitioners Resolve Disagreements?

Algorithm	Reasons that algorithm was chosen in disagreement
KernelSHAP	• [36%] SHAP is better for tabular data (*"SHAP is more commonly used [than Gradient] for tabular data"*) • [25%] SHAP is more familiar (*"More information present + more familiarity"*) • [14%] SHAP is a better algorithm overall (*"SHAP seems more methodical than LIME"*, *"SHAP is a more rigorous approach [than LIME] in theory"*)
SmoothGrad	• [33%] SmoothGrad paper is newer or better (*"SmoothGrad is apparently more robust"*, *"SmoothGrad is often considered improved verison of grad"*) • [58%] Reasons based on the explainability map shown (*"directionality of the attributions ... [agree] with intuition"*, *"gradient has unstability problems [, so] smoothgrad"*)
LIME	• [54%] LIME is better for tabular data (*"I use LIME for structured data."*) • [15%] LIME is more familiar/easier to interpret (*"I am more familiar with LIME"*, *"LIME is easy to interpret"*)
Integrated Gradients	• [86%] Integrated Gradients paper is better (*"IG came after gradients and paper shows improvements"*, *"integrated gradients paper showed improvements [over Gradient × Input]"*

Insights and Moving Forward

- Feature attribution methods often disagree in practice w.r.t. basic insights, and practitioners adopt ad hoc heuristics to resolve those disagreements!

- Why do feature attribution methods disagree?

- Given that feature attribution methods disagree, which explanation method should we choose for different kinds of data and applications?

Why do Feature Attribution Methods Disagree?

- Various feature attribution methods (e.g., LIME, C-LIME, KernelSHAP, Occlusion, Vanilla Gradients, Gradient times Input, SmoothGrad, Integrated Gradients) are essentially local function approximations.

$$g^* = \arg\min_{g \in \mathcal{G}} \mathbb{E}_{\xi \sim \mathcal{Z}} \ell(f, g, \mathbf{x}_0, \xi)$$

- But…

Han et. al., NeurIPS 2022

Why do Feature Attribution Methods Disagree?

- But, they adopt different loss functions, and local neighborhoods

Explanation Method	**Local Neighborhood $\mathcal{Z}$ around $\mathbf{x}_0$**	**Loss Function ℓ**
C-LIME	$\mathbf{x}_0 + \xi;\ \xi(\in \mathbb{R}^d) \sim \text{Normal}(0, \sigma^2)$	Squared Error
SmoothGrad	$\mathbf{x}_0 + \xi;\ \xi(\in \mathbb{R}^d) \sim \text{Normal}(0, \sigma^2)$	Gradient Matching
Vanilla Gradients	$\mathbf{x}_0 + \xi;\ \xi(\in \mathbb{R}^d) \sim \text{Normal}(0, \sigma^2), \sigma \to 0$	Gradient Matching
Integrated Gradients	$\xi\mathbf{x}_0;\ \xi(\in \mathbb{R}) \sim \text{Uniform}(0, 1)$	Gradient Matching
Gradients $\times$ Input	$\xi\mathbf{x}_0;\ \xi(\in \mathbb{R}) \sim \text{Uniform}(a, 1), a \to 1$	Gradient Matching
LIME	$\mathbf{x}_0 \odot \xi;\ \xi(\in \{0,1\}^d) \sim$ Exponential kernel	Squared Error
KernelSHAP	$\mathbf{x}_0 \odot \xi;\ \xi(\in \{0,1\}^d) \sim$ Shapley kernel	Squared Error
Occlusion	$\mathbf{x}_0 \odot \xi;\ \xi(\in \{0,1\}^d) \sim$ Random one-hot vectors	Squared Error

Han et. al., 2022

Why Do Feature Attribution Methods Disagree?

- *No Free Lunch Theorem for Explanation Methods*: No single method can perform optimally across all neighborhoods

Theorem 3 (No Free Lunch for Explanation Methods). *Consider the scenario where we explain a black-box model f around point $\mathbf{x}_0$ using an interpretable model g from class $\mathcal{G}$ and a valid loss function ℓ where the distance between f and $\mathcal{G}$ is given by $d(f, \mathcal{G}) = \min_{g \in \mathcal{G}} \max_{\mathbf{x} \in \mathcal{X}} \ell(f, g, 0, \mathbf{x})$. Then, for any explanation g^* on a neighborhood distribution $\xi_1 \sim \mathcal{Z}_1$ such that $\max_{\xi_1} \ell(f, g^*, \mathbf{x}_0, \xi_1) \leq \epsilon$, we can always find another neighborhood $\xi_2 \sim \mathcal{Z}_2$ such that $\max_{\xi_2} \ell(f, g^*, \mathbf{x}_0, \xi_2) \geq d(f, \mathcal{G})$.*

Which Method Should We Choose?: Take 1

- A guiding principle based on function approximation: choose a method which recovers the underlying model when the model is a member of the explanation function class

- For continuous data, use additive continuous noise methods (e.g. SmoothGrad, Vanilla Gradients, C-LIME) or multiplicative continuous noise methods (e.g. Integrated Gradients, Gradient x Input). For binary data, use binary noise methods (e.g. LIME, KernelSHAP, Occlusion).

Which Method Should We Choose?: Take 2

- OpenXAI: open-source framework to readily evaluate and benchmark post hoc explanation methods
- Systematic, efficient, and reproducible evaluations of post hoc explanation methods on various datasets
- Assessing reliability of post hoc explanation methods from diverse perspectives (e.g., faithfulness, stability, fairness)
- (Customizable) dashboards to compare existing and new methods across various datasets easily

Agarwal et. al., NeurIPS 2022; OpenXAI: Towards a Transparent Evaluation of Model Explanations; open-xai.github.io

Conclusions and Summary

- Several methods proposed to “explain” machine learning models in prior research
- Important to characterize these methods, and understand which methods can be useful under what circumstances
- Critical to bridge the gaps between researchers and practitioners

Thank You!

- Email: hlakkaraju@hbs.edu; hlakkaraju@seas.harvard.edu;
- Webpage: https://himalakkaraju.github.io
- Course on interpretability and explainability: https://interpretable-ml-class.github.io/
- Multiple tutorials on explaining ML models (ranging from 1 hour to 3 hours): explainml-tutorial.github.io
- Trustworthy ML Initiative: https://www.trustworthyml.org/
  - Lots of resources and seminar series on topics related to explainability, fairness, adversarial robustness, differential privacy, causality etc.