BIG DATA CLUB

BDC Seminar

A Brief Introduction to Explainable AI

Nguyễn Thành Hưng -
itcsiu22051@student.hcmiu.edu.vn

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA

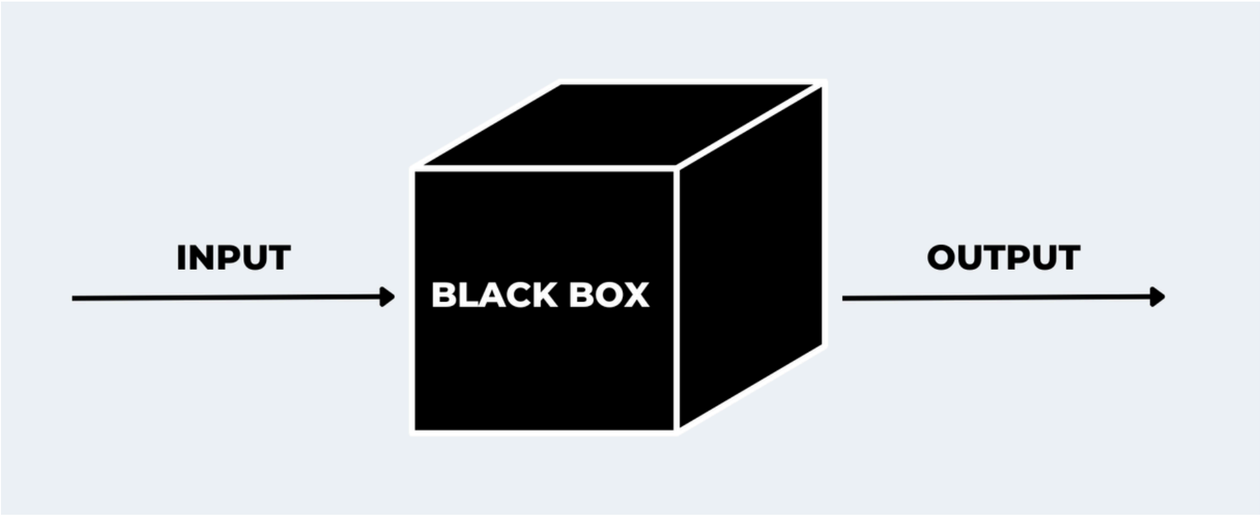
CSE

HPCC
HCMUT, VNU - HCM

HOCHIMINH CITY INTERNATIONAL UNIVERSITY • HCM - IU •

TABLE OF CONTENTS

Problem in modern AI



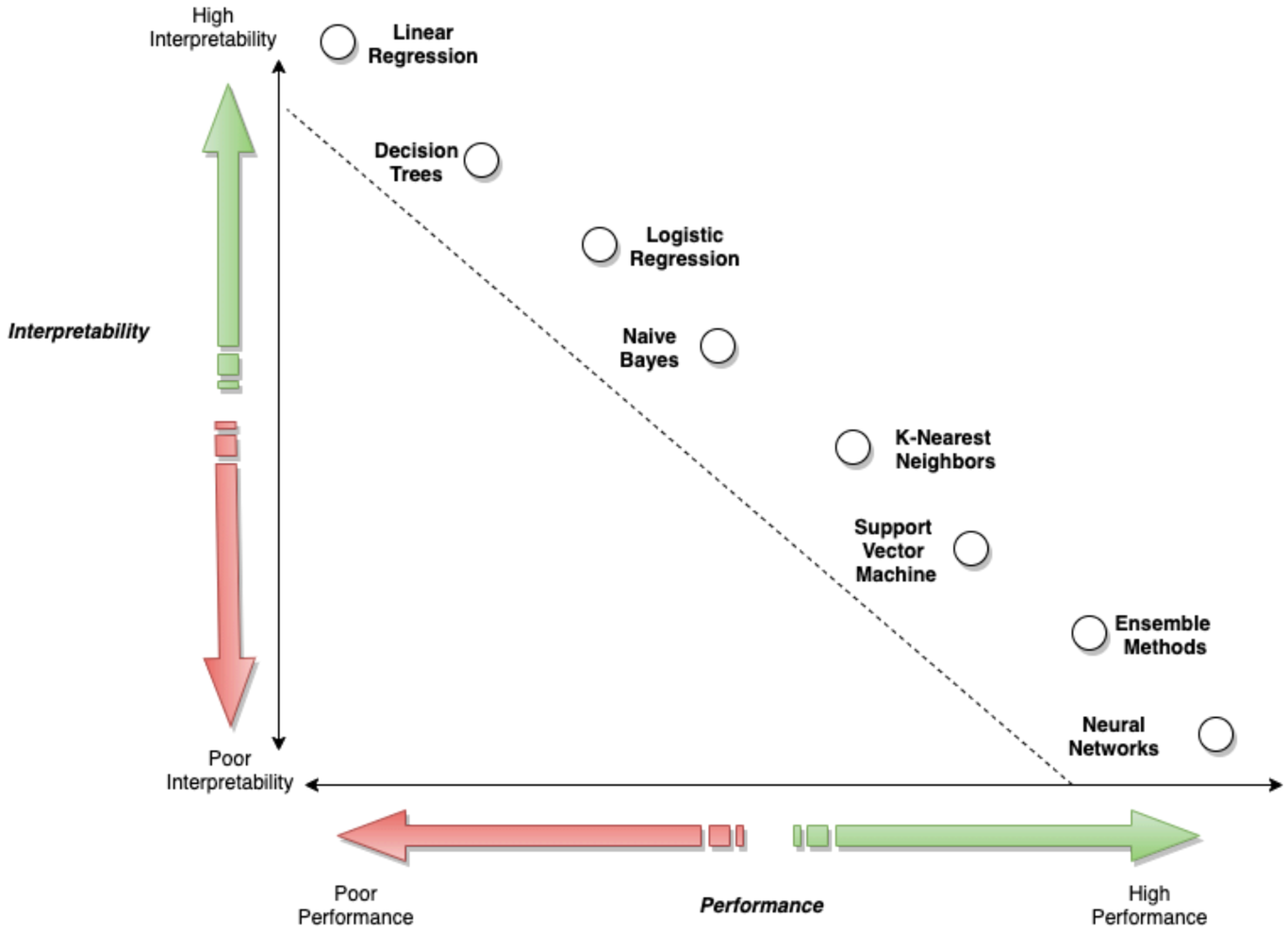
What is Blackbox AI?

It refers to AI complex systems

- **Input/ Output can be observed but the process is incomprehensible**
- **Interpretability Gap: People dont know how actually it works**

As a result, AI Dev makes the “black box” to learn blindly

Motivation



- A trade off for robust ML/AI systems
- Simple AI models seems to be more transparent
- But is the performance always good ?

Motivation

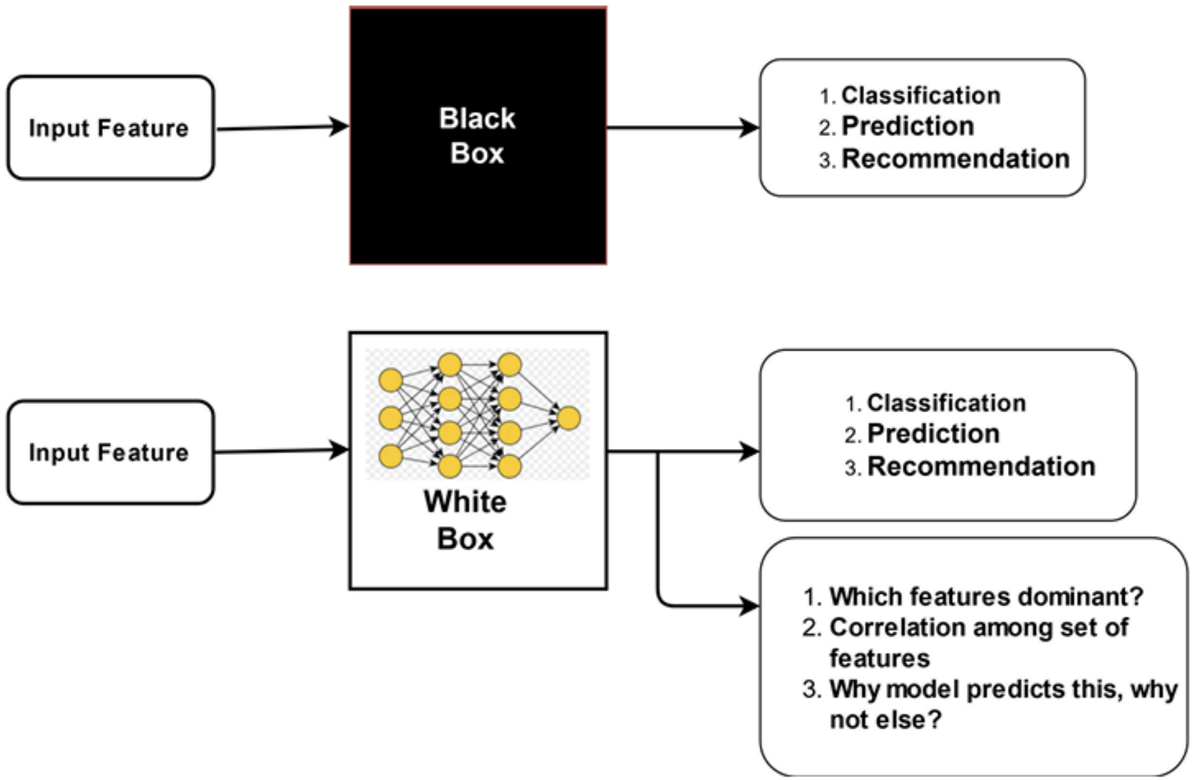
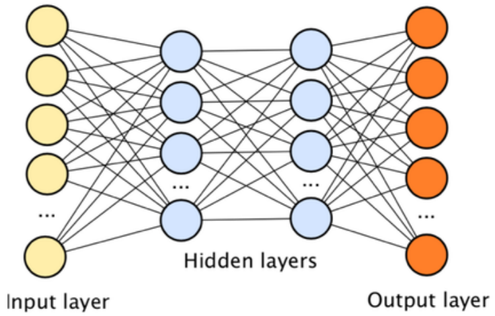


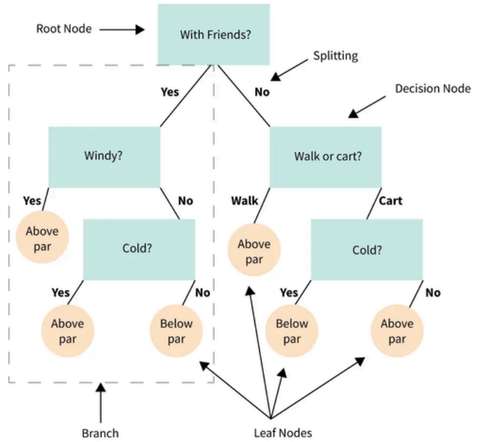
FIGURE 1: AI vs XAI

Achieving Model Understanding

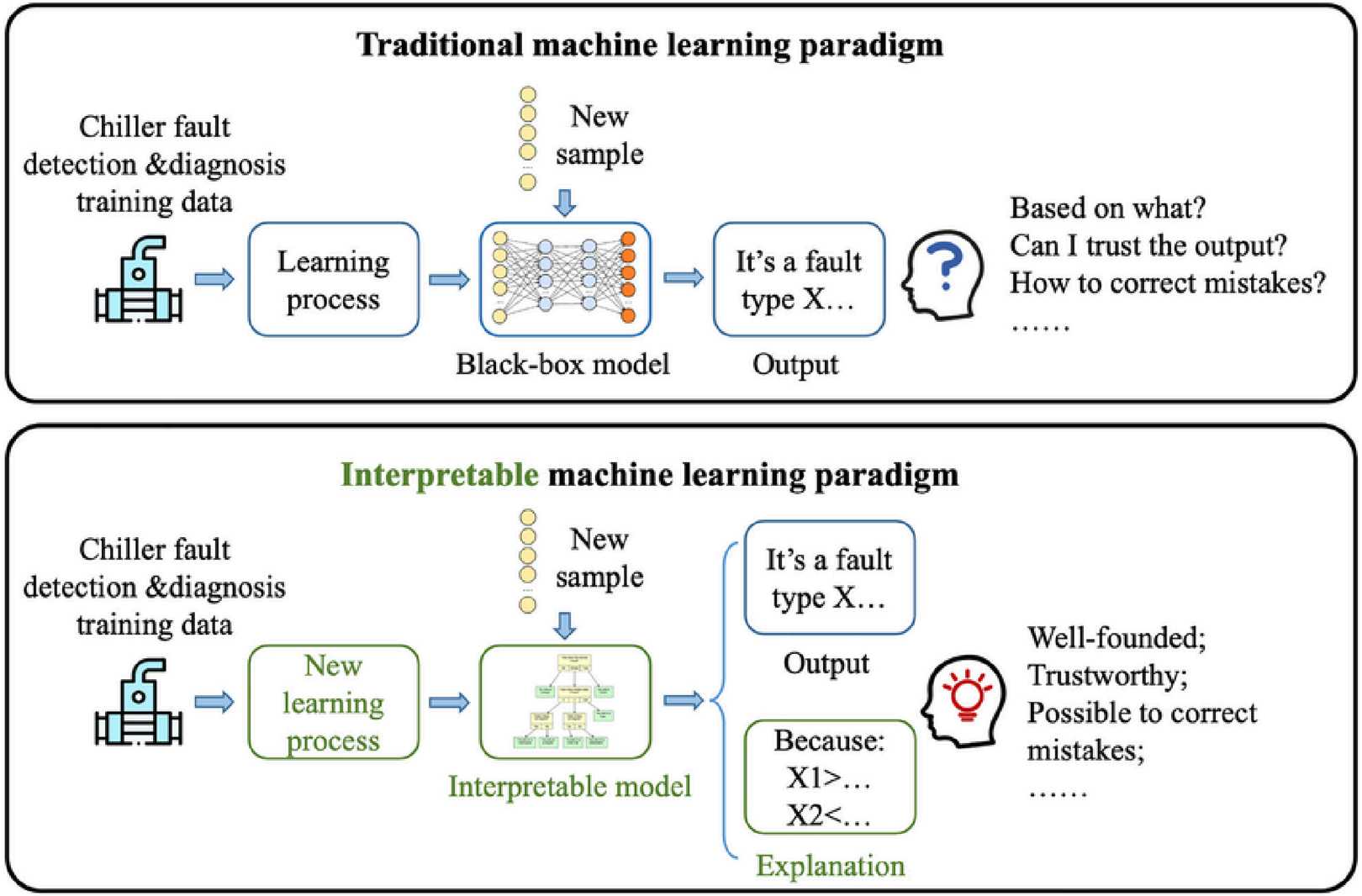


Explainer

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



Why Should I Trust You ?



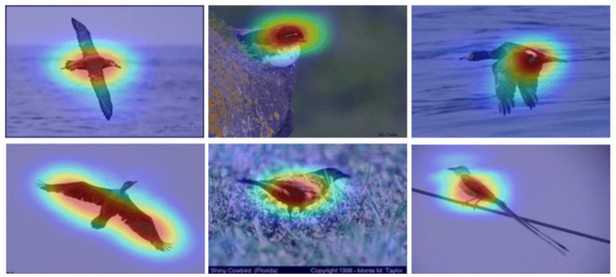
Methodology

Post-hoc methods

Model agnostic

- **Global model agnostic**
  - How feature affect on average
- **Local model agnostic**
  - explain individual agnostic

Model Specific

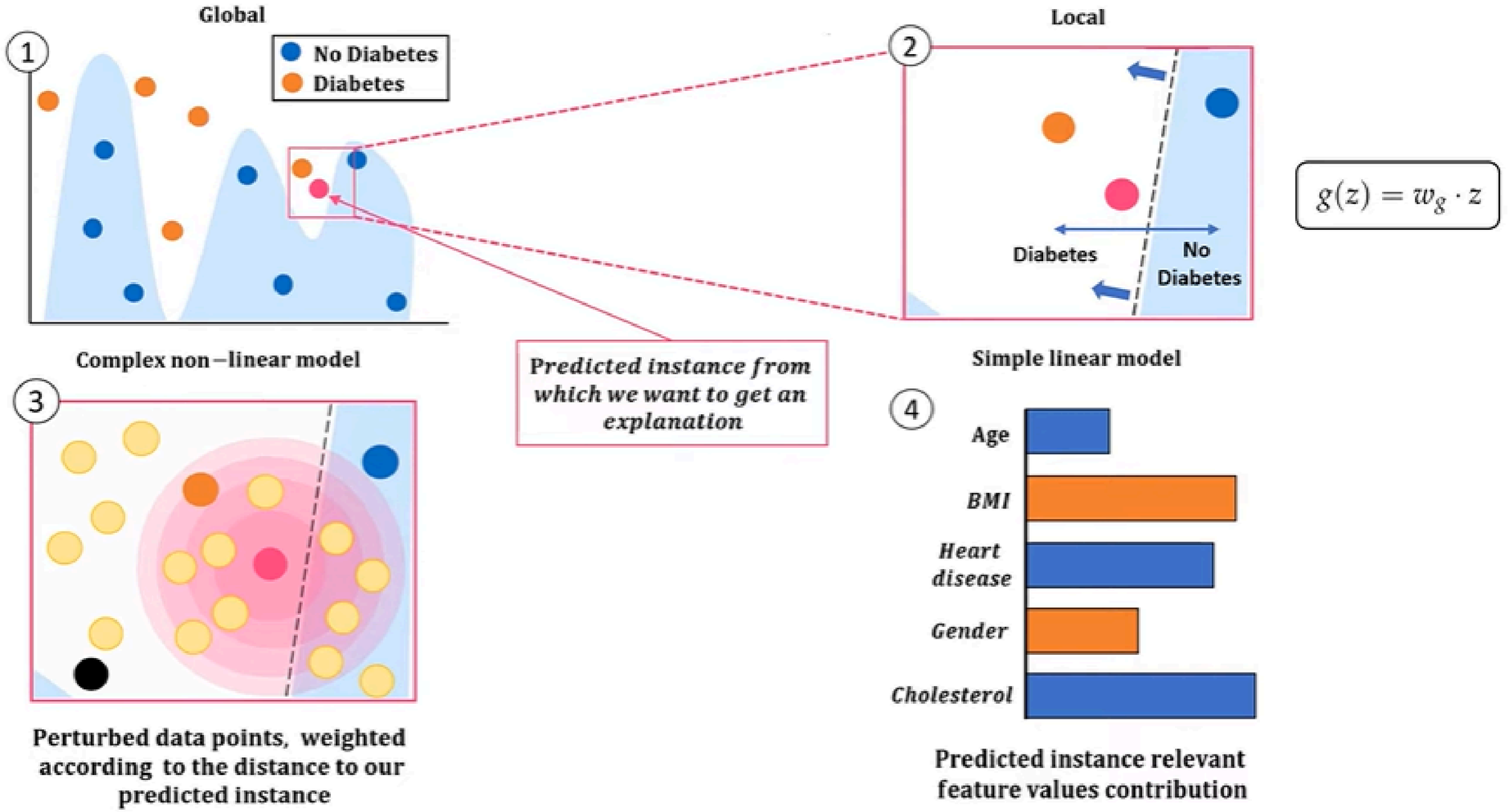
- **Feature Relevance**
  - Based on specific input feature
- **Condition based Explanation**
  - "If feature X > 10, then classify as Y"
- **Rule based learning (using intuitions)**
  - Why learning_rate = 0.01?
- **Saliency map**
  - 

Model Agnostic Technique - LIME

LIME - Local Interpretable-Model agnostic explanations

- **Works on any blackbox model**
- **Interpretability**
- **Works with many data types (text, image, audio)**
- **Explanation for any type of supervised learning model**

LIME Step by Step



The explanation produced by LIME is obtained by the following:

$$f : \mathbb{R}^d \rightarrow \tilde{\mathbb{R}}.$$

Neighbourhood of x

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Num Features (e.g., Age, gender,v.v)

Complex model

Simple interpretable model

Complexity of interpretable model

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

$$g(z) = \omega_g . z$$

$$L(f, g, \pi_x) = \sum_{z \varepsilon Z} \pi_x(z)(f(z) - g(z))^2$$

$\Omega(g)$ **LIME uses sparse linear models (K-LASSO)**

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

Benefits and Limitations

Benefits

- Essentially importance in healthcare or finance
- Provides clear explanations
- Prevent AI engineers to make AI models learn blindly

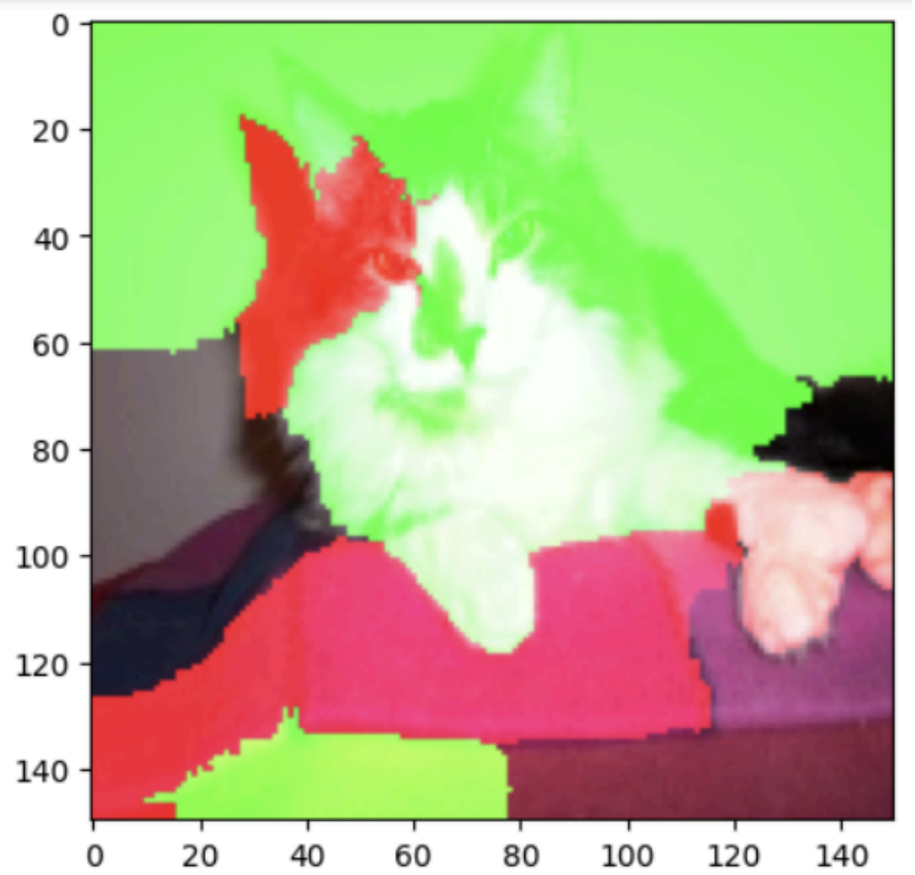
Limitations

- Difficult to interpret Deep Neural Networks (e.g, Transformer,GPT)
- Relying on interpretability can reduce performance
- Depend on Qualitative data

Evaluation

CNN with LIME

LIME demonstrations



Accuracy: 93%

- **Green regions**: postive (associated with cat)

- **Red regions**: negative

Thank you for your Attention

References:

[1] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, February 16). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. ArXiv.org. https://arxiv.org/abs/1602.04938

[2] Gohel, P., Singh, P., & Mohanty, M. (2021). Explainable AI: current status and future directions. ArXiv:2107.07045 [Cs]. https://arxiv.org/abs/2107.07045