

# XAI Introduction: LIME và ANCHOR

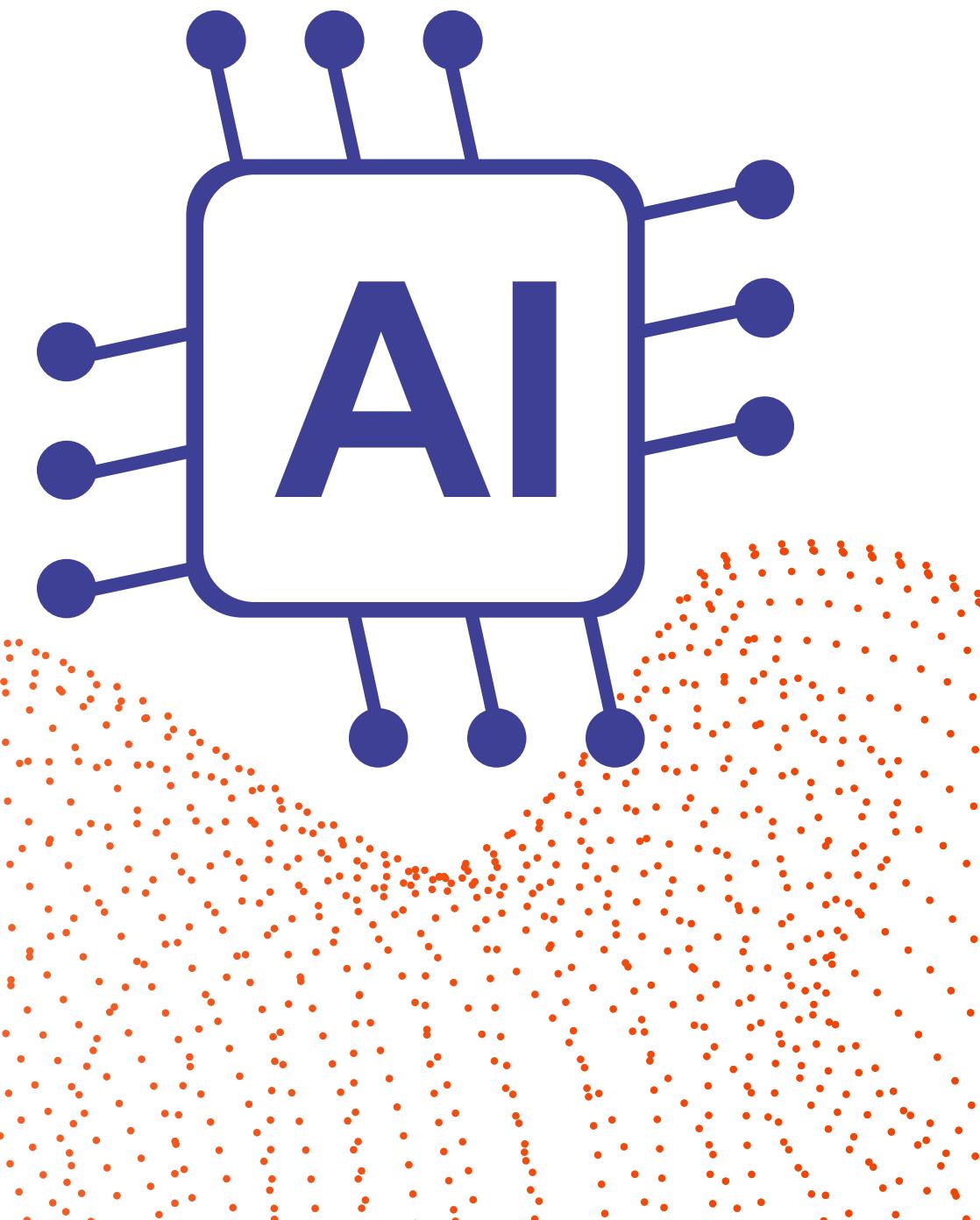
Trần Trung Hậu MSV:23020061  
Mai Minh Tùng MSV:23020432

# AI: Hộp đen hay người bạn đồng hành đáng tin cậy



# AI đang thay đổi thế giới.... nhưng chúng ta có hiểu nó?

- AI ngày càng được ứng dụng rộng rãi trong  
nhiều lĩnh vực : ý tế, tài chính, giao thông vận  
tải,...



# AI đang thay đổi thế giới.... nhưng chúng ta có hiểu nó?

- Các mô hình AI hiện đại (đặc biệt là mạng nơ-ron sâu) thường rất phức tạp và khó hiểu (mô hình hộp đen)



# AI đang thay đổi thế giới.... nhưng chúng ta có hiểu nó?

- Việc thiếu hiểu biết về cách AI đưa ra quyết định có thể gây rủi ro (ví dụ: thiên vị, thiếu trách nhiệm giải trình) và thiếu tin tưởng



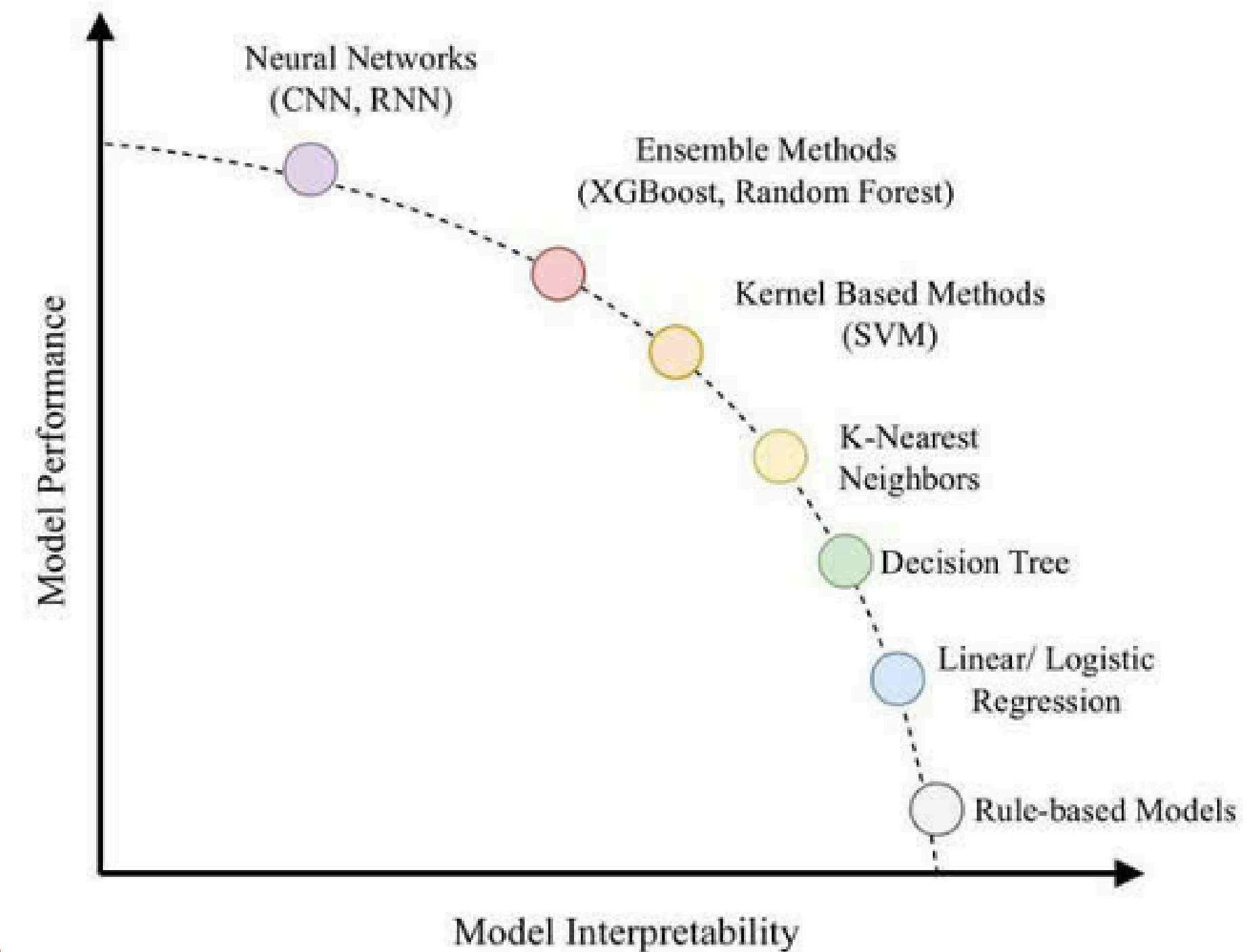
Hình 3: Minh họa sự thiên vị của hệ thống COMPAS trong đánh giá nguy cơ tái phạm tội. Nguồn

# Interpretability và Explainability

Tính chất	INTERPRETABILITY (Khả năng diễn giải)	EXPLAINABILITY (Khả năng giải thích)
Mục tiêu	Hiểu cách mô hình hoạt động	Giải thích lý do đưa ra quyết định
Tập trung vào	Cấu trúc và cơ chế mô hình	Kết quả và bằng chứng
Tính chất	Trực quan, đơn giản, minh bạch	Dễ hiểu, chính xác, đầy đủ
Phạm vi	Thường cho mô hình đơn giản	Cho cả mô hình đơn giản và phức tạp

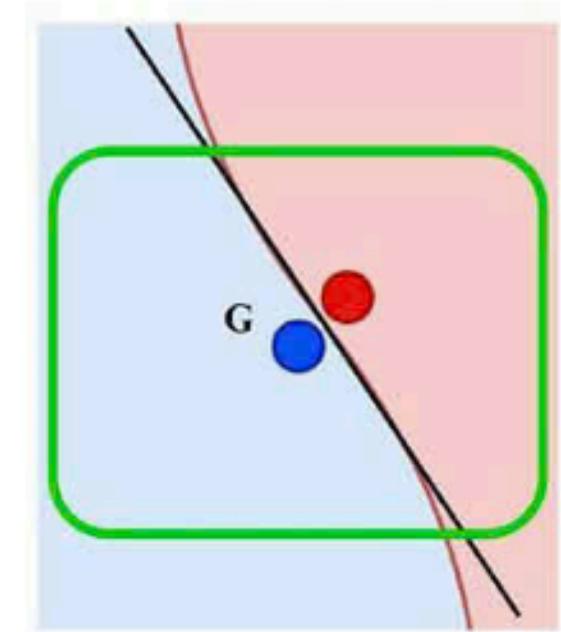
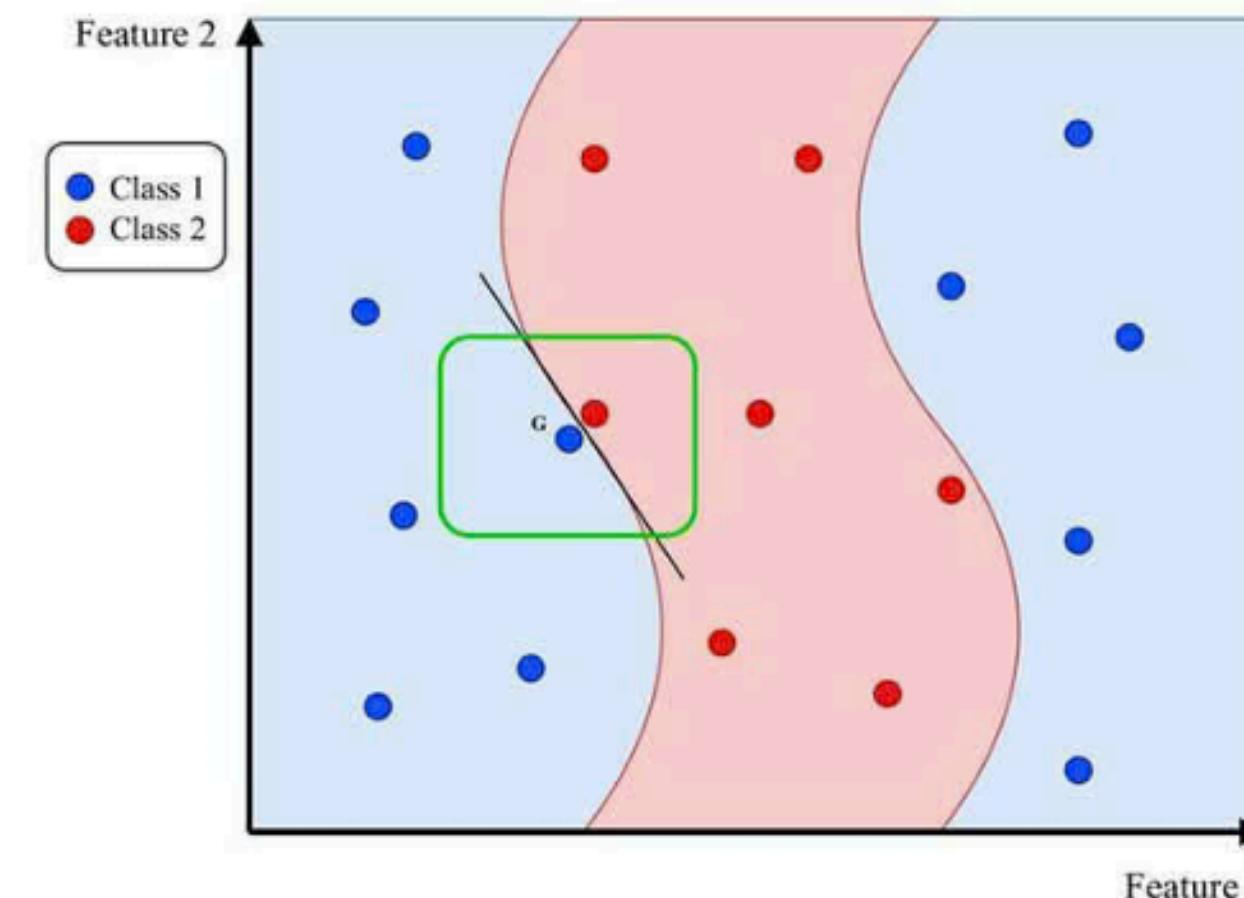
# Đánh đổi giữa hiệu suất và khả năng diễn giải

- Mô hình đơn giản thì dễ hiểu minh bạch nhưng không hiệu quả
- Mô hình phức tạp thì hiệu suất cao nhưng khó giải thích, diễn đạt



# LIME: Giải thích các dự đoán cục bộ

- LIME (Local Interpretable Model-Agnostic Explanations) là gì ?
- Ý tưởng: Xấp xỉ hành vi của mô hình phức tạp bằng một mô hình đơn giản trong vùng lân cận của điểm dữ liệu cần giải thích

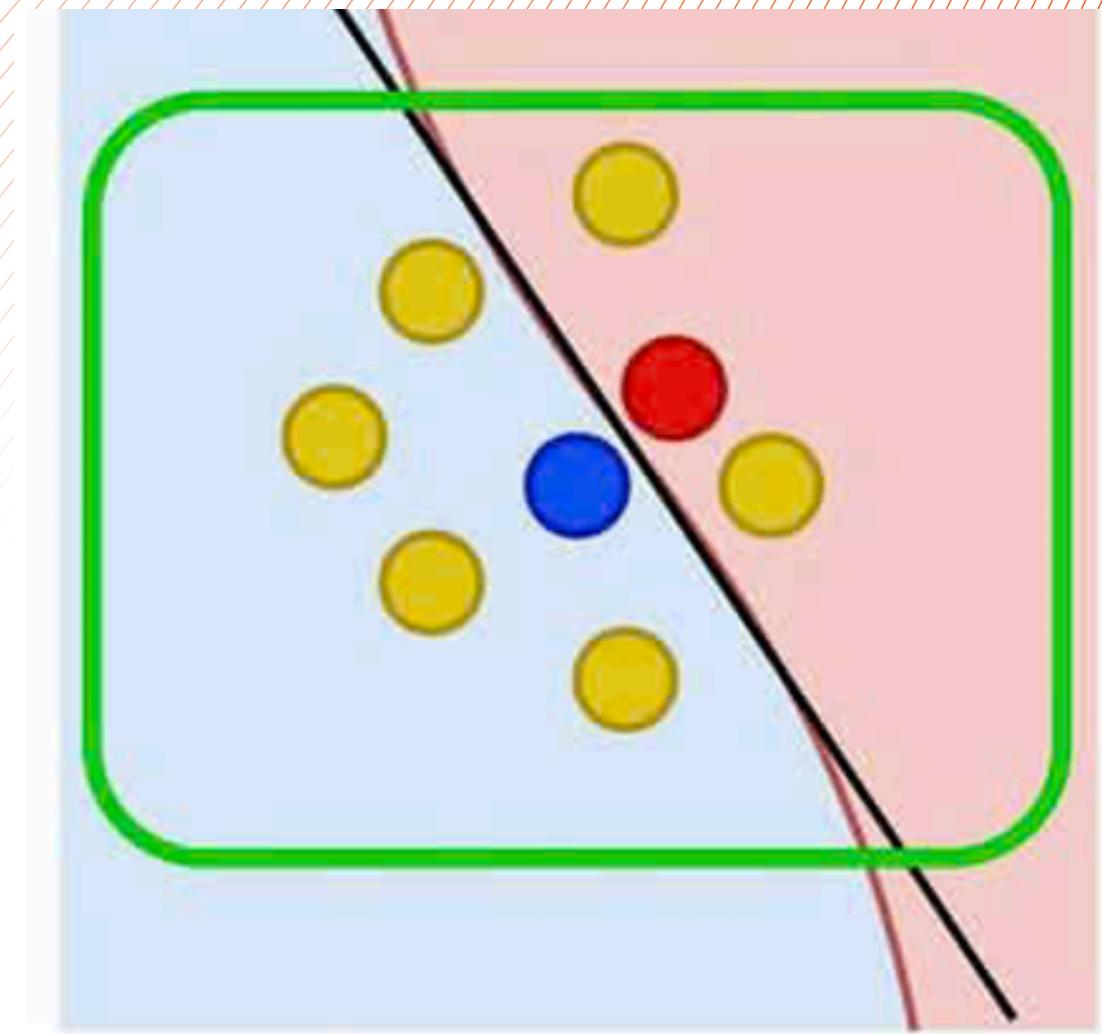


Hình 5: Minh họa nguyên lý hoạt động của LIME.

# LIME: Giải thích từng bước

## Buộc 1: Tạo mẫu dữ liệu lân cận (Perturbation)

Tạo ra các điểm dữ liệu mới xung quanh điểm dữ liệu cần giải thích bằng cách thay đổi giá trị đặc trưng



Hình 6: Các điểm dữ liệu giả (màu vàng) được tạo quanh G.

# LIME: Giải thích từng bước

## Bước 2: Dự đoán bằng mô hình gốc (Make prediction)

Sử dụng mô hình phức tạp ban đầu để dự đoán nhãn cho các điểm dữ liệu giả đã tạo

## Bước 3: Tính khoảng cách và trọng số (Calculate Distance and Weights

Đo khoảng cách giữa mỗi điểm dữ liệu giả và điểm gốc để sử dụng làm trọng số trong bước tiếp theo. Điểm càng gần điểm gốc sẽ có trọng số càng cao

# LIME: Giải thích từng bước

## Bước 4: Chọn các đặc trưng quan trọng (Select Features)

Lựa chọn một tập hợp nhỏ các đặc trưng có ảnh hưởng lớn nhất đến kết quả dự đoán.

## Bước 5: Huấn luyện mô hình đơn giản (Fit Simple Model)

Sử dụng các mẫu dữ liệu giả đã tạo ra để huấn luyện mô hình đơn giản.

# Hạn chế của LIME

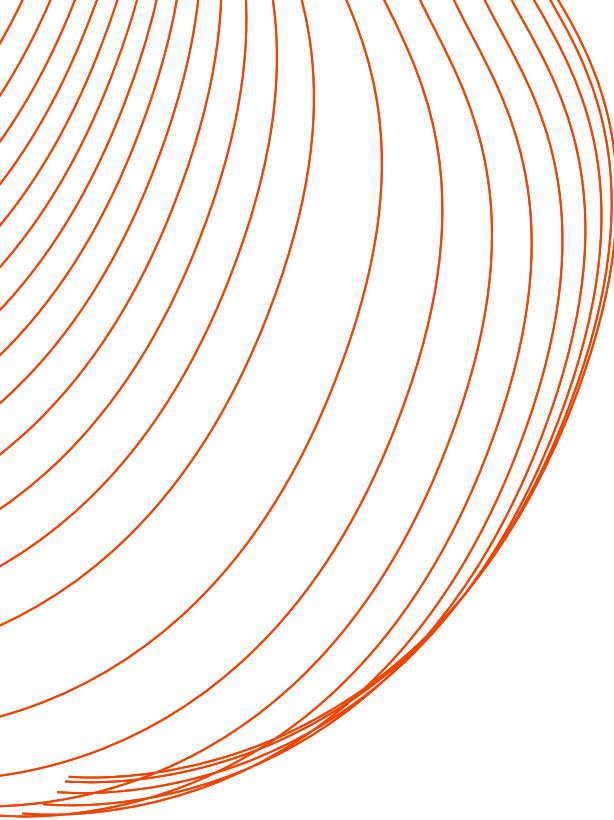
- Không ổn định  
(Inconsistency)

- Phụ thuộc vào siêu tham số

- Khó khăn trong việc chuyển đổi biểu diễn đặc trưng

- Giới hạn cục bộ

- Hiệu năng tính toán



# Giới thiệu về ANCHOR

## Giải thích quyết định trong AI

- Giới thiệu ANCHOR và cách hoạt động động và triển khai.
- Đặt vấn đề: Làm sao để "giải thích" một quyết định cụ thể của AI một cách đáng tin cậy?

# Động lực của ANCHOR và Hạn chế của LIME

Hạn chế của LIME	Động lực của ANCHOR
<ul style="list-style-type: none"><li>Tính không ổn định: Kết quả thay đổi khi chạy lại.</li><li>Giới hạn cục bộ: Chỉ giải thích được trong phạm vi nhỏ</li><li>Phụ thuộc vào siêu tham số: Kết quả dễ bị ảnh hưởng bởi tham số.</li><li>Khó diễn dải: Phải chuyển đổi dữ liệu về dạng dễ hiểu.</li></ul>	<ul style="list-style-type: none"><li>Giải quyết các hạn chế của LIME.</li><li>Tạo ra lời giải thích đáng tin cậy và phạm vi lớn hơn.</li><li>Tập trung vào việc tìm các bộ luật có độ bao phủ cao và chính xác cao.</li></ul>

Dặt vấn đề: Làm sao để ta có được lời giải thích đáng tin cậy và phạm vi rộng?

# Khái niệm cơ bản về ANCHOR

- ANCHOR: Tập hợp các quy tắc hoặc điều kiện dưới dạng IF - THEN.
- Example:

Dữ liệu đầu vào và dự đoán	Giải thích của ANCHOR
$28 < \text{Age} \leq 37$ $\text{Workclass} = \text{Private}$ $\text{Education} = \text{High School grad}$ $\text{Marital Status} = \text{Married}$ $\text{Occupation} = \text{Blue-Collar}$ $\text{Relationship} = \text{Husband}$ $\text{Race} = \text{White}$ $\text{Sex} = \text{Male}$ $\text{Capital Gain} = \text{None}$ $\text{Capital Loss} = \text{Low}$ $\text{Hours per week} \leq 40.00$ $\text{Country} = \text{United-States}$	<b>IF</b> $28 < \text{Age} \leq 37$ <b>AND</b> $\text{Workclass} = \text{Private}$ <b>AND</b> $\text{Education} = \text{High School grad}$ <b>AND</b> $\text{Marital Status} = \text{Married}$ <b>AND</b> $\text{Occupation} = \text{Blue-Collar}$ <b>AND</b> $\text{Relationship} = \text{Husband}$ <b>AND</b> $\text{Race} = \text{White}$ <b>AND</b> $\text{Sex} = \text{Male}$ <b>AND</b> $\text{Capital Gain} = \text{None}$ <b>AND</b> $\text{Capital Loss} = \text{Low}$ <b>AND</b> $\text{Hours per week} \leq 40.00$ <b>AND</b> $\text{Country} = \text{United-States}$ <b>THEN PREDICT</b> $\text{Salary} > 50K$

# Cách tính độ chính xác và độ bao phủ:

Độ chính xác (Precision):	Độ bao phủ (Coverage):
<ul style="list-style-type: none"><li>Tỷ lệ các mẫu dữ liệu thỏa mãn luật A, có cùng kết quả dự đoán với mẫu gốc.</li><li>Công thức: <math>\text{prec}(A) = (\text{Số mẫu có cùng dự đoán}) / (\text{Tổng số mẫu thỏa mãn } A)</math>.</li></ul>	<ul style="list-style-type: none"><li>Tỷ lệ các mẫu trong tập dữ liệu thỏa mãn luật A.</li><li>Công thức: <math>\text{cov}(A) = (\text{Số mẫu trong tập kiểm tra thỏa mãn } A) / (\text{Tổng số mẫu trong tập kiểm tra})</math>.</li></ul>

**Mục tiêu: Tìm ra bộ luật có độ chính xác và độ bao phủ tốt nhất**

# Quy trình hoạt động của ANCHOR

## Bước 1: Sinh ra các luật ứng viên (Generate Candidate Rules):

- Khởi tạo tập hợp luật ứng viên.
- Dựa vào đặc trưng của dữ liệu để tạo các predicate (điều kiện) có thể.
- Ví dụ: "Tuổi > 30", "Màu đỏ", "Từ 'khuyến mãi' xuất hiện".

# Quy trình hoạt động của ANCHOR

## Bước 2: Chia thuật toán tìm kiếm thành nhiều "beam":

- Để tìm kiếm hiệu quả hơn, tránh bị mắc kẹt, chia thành nhiều luồng tìm kiếm song song.
- Trong mỗi beam, áp dụng vòng lặp tìm kiếm (xem bước 3).

# Quy trình hoạt động của ANCHOR

## Bước 3: Vòng lặp tìm kiếm (KL-LUCB):

- Tạo mẫu dữ liệu lân cận: Thay đổi dữ liệu một cách khéo léo, nhưng vẫn đảm bảo tuân thủ luật hiện tại.
- Dự đoán bằng mô hình gốc: Áp dụng mô hình gốc lên dữ liệu lân cận.
- Tính toán độ chính xác và độ bao phủ: Xác định mức độ tin cậy và phạm vi áp dụng của luật hiện tại.
- Thuật toán KL-LUCB: Tìm luật tối ưu dựa trên độ chính xác, độ bao phủ, và khoảng tin cậy.

Kết quả: Nhận được bộ luật có độ tin cậy và độ bao phủ cao nhất.

# Thuật toán KL-LUCB - Tìm kiếm luật tối ưu

Thuật toán cân bằng giữa việc khám phá và khai thác.

Cách hoạt động	Kiểm tra điều kiện:
<ul style="list-style-type: none"><li><b>Khởi tạo:</b> Bắt đầu với một số mẫu nhỏ.</li><li><b>Ước lượng:</b> Tính độ chính xác của luật dựa trên các mẫu.</li><li><b>Tính khoảng tin cậy:</b> Xác định khoảng tin cậy cho độ chính xác.</li></ul>	<ul style="list-style-type: none"><li>Nếu lower bound của khoảng tin cậy vượt ngưỡng (7): Chọn luật.</li><li>Nếu upper bound của khoảng tin cậy dưới ngưỡng: Loại luật.</li></ul>

- Chọn luật có độ chính xác cao nhất và độ bao phủ lớn nhất.
- Nếu không tìm thấy luật nào thỏa mãn, mở rộng luật (thêm điều kiện) và lặp lại từ đầu.

# Ưu và nhược điểm của ANCHOR (Tổng kết)

Ưu điểm	Nhược điểm
<ul style="list-style-type: none"><li>• Cung cấp lời giải thích đáng tin cậy và có phạm vi rộng.</li><li>• Giải thích theo các luật dễ hiểu.</li><li>• Cho biết các điều kiện quan trọng nhất ảnh hưởng đến dự đoán.</li><li>• Có thể áp dụng cho nhiều loại dữ liệu khác nhau.</li></ul>	<ul style="list-style-type: none"><li>• Khó cài đặt và điều chỉnh.</li><li>• Thời gian chạy có thể chậm (do cần nhiều lần gọi mô hình).</li><li>• Có thể phụ thuộc vào cách tạo mẫu dữ liệu lân cận.</li><li>• Độ bao phủ có thể thấp (luật chỉ đúng cho một phần nhỏ dữ liệu).</li><li>• Thiếu sự tường minh trong một số loại dữ liệu.</li></ul>