

# Anchors: High Precision Model-Agnostic Explanations

Marco Ribeiro

University of Washington

Sameer Singh

University of California, Irvine

Carlos Guestrin

University of Washington

Explanations (Ribeiro et al., 2018)

## Abstract

We introduce a novel model-agnostic system that explains the behavior of complex models with high-precision rules called *anchors*, representing local, “sufficient” conditions for predictions. We propose an algorithm to efficiently compute these explanations for any black-box model with high probability. We show that anchors enable users to explain a myriad of different models for different domains and tasks. In a user study, we show that anchors enable users to predict how a model would behave on unseen instances with less effort and higher precision, as compared to existing linear explanations or no explanations.

## Introduction

Sophisticated machine learning models such as deep neural networks have been shown to be highly accurate for many applications, even though their complexity virtually makes them black-boxes. As a consequence of the need for users to understand the behavior of these models, *interpretable machine learning* has seen a resurgence in recent years, ranging from the design of novel *globally*-interpretable machine learning models (Lakkaraju, Bach, and Leskovec 2016; Ustun and Rudin 2015; Wang and Rudin 2015) to local explanations (for individual predictions) that can be computed for any classifier (Bachrens et al. 2010; Ribeiro, Singh, and Guestrin 2016b; Strumbelj and Kononenko 2010).

+ This movie is not bad.    — This movie is not very good.

(a) Instances



["not", "bad"] → Positive    ["not", "good"] → Negative

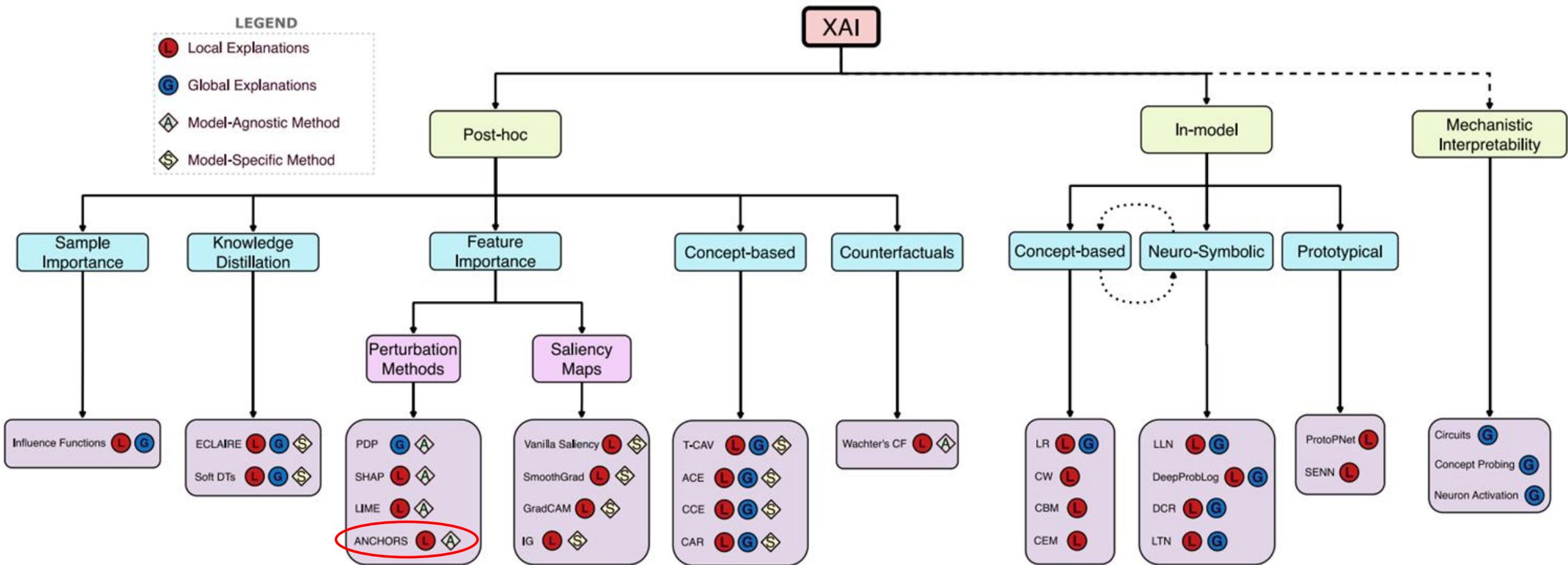
(c) Anchor explanations

Figure 1: Sentiment predictions, LSTM

explanations are in some way local, it is not clear whether they *apply* to an unseen instance. In other words, their coverage (region where explanation applies) is unclear. Unclear coverage can lead to low human precision, as users may think an insight from an explanation applies to unseen instances even when it does not. When combined with the arithmetic involved in computing the contribution of the features in linear explanations, the human effort required can be quite high.

Take for example LIME (Ribeiro, Singh, and Guestrin

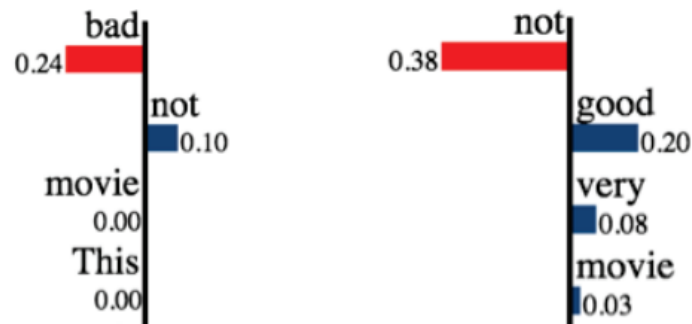
L193 Explainable AI, MPhil Candidate Sonia Koszut



## What are Anchors? Why do we need them?

+ This movie is not bad.      — This movie is not very good.

(a) Instances



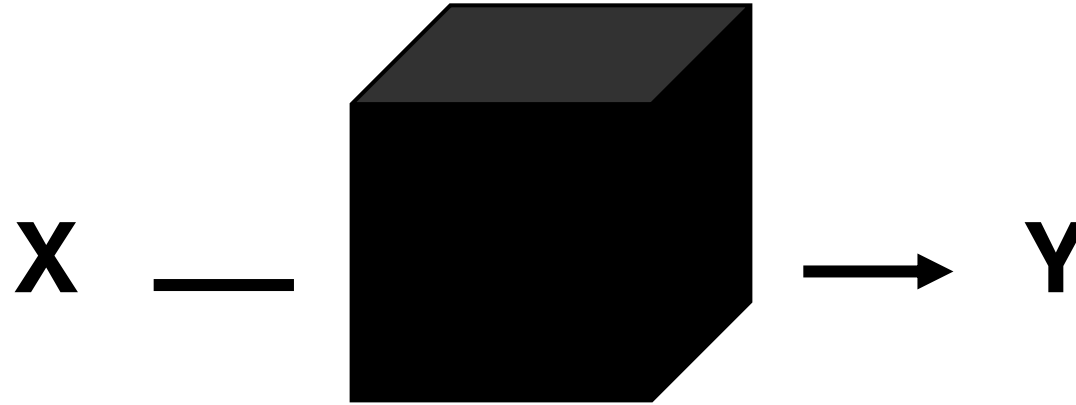
(b) LIME explanations

{"not", "bad"} → Positive      {"not", "good"} → Negative

- Linear explanations might not generalise well to unseen examples (their coverage is unclear).
- Significant human effort is required to understand linear explanations.
- Linear explanations assume the model is locally linear or close to linear, which may not be the case.

**Answer:** Anchors as simple if-then rules

# How does it work?



$x$  = „not bad”

This movie is not bad

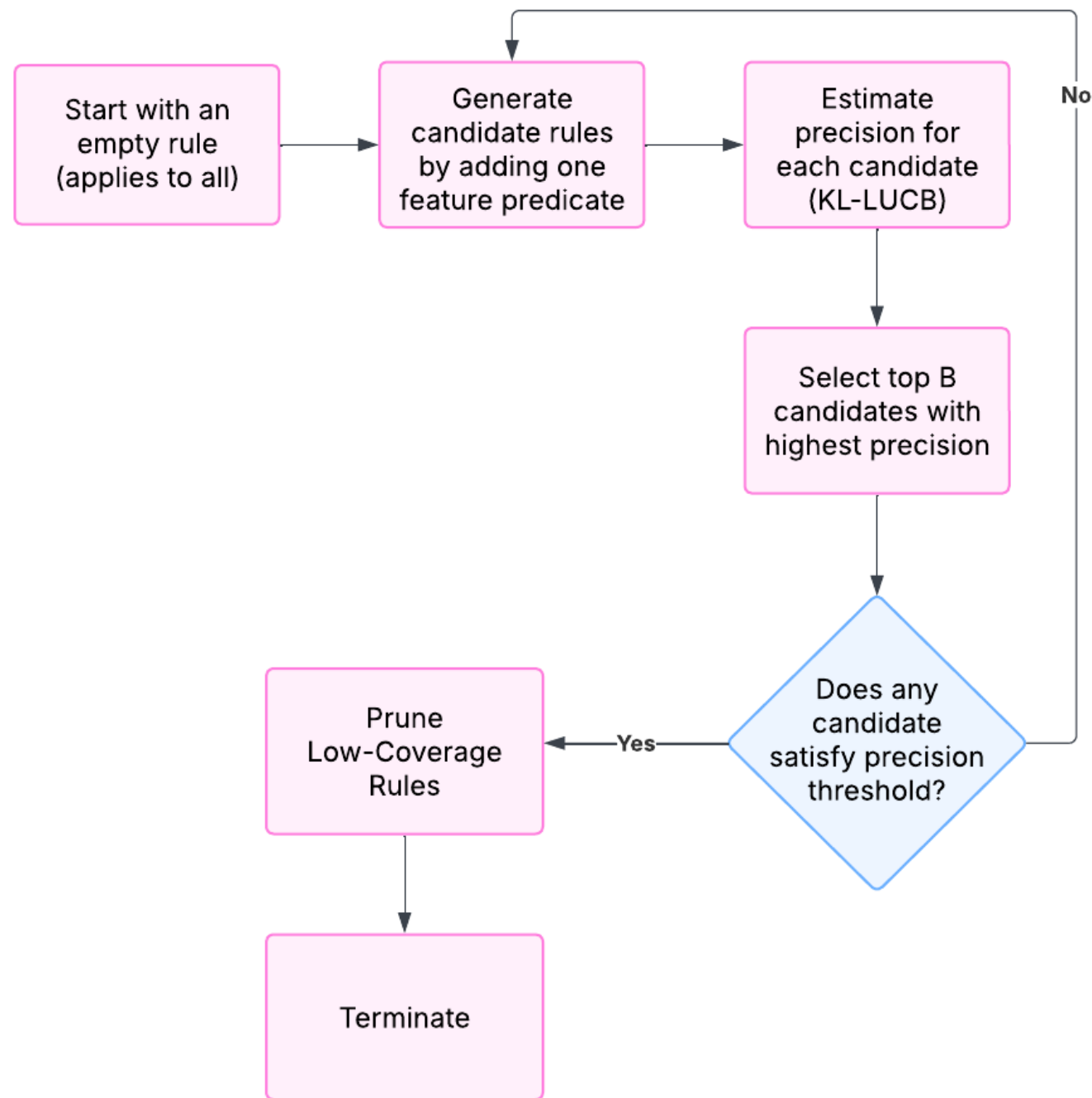
This audio is not bad

This novel is not bad

This footage is not bad

# Make it efficient

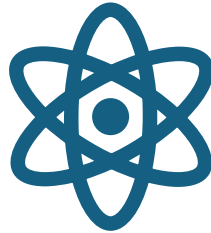
$$\max_{A \text{ s.t. } P(\text{prec}(A) \geq \tau) \geq 1 - \delta} \text{cov}(A).$$



# Technical Problems



Computational  
inefficiencies



Sensitivity to  
perturbations



Unclear conflict  
resolution

# Fixing anchors

- Perturbations should preserve real-world dependencies and incorporate domain knowledge,
- Introduce tie-breaking rule for conflicting anchors.



## Anchors: how far do they take us?

- Rules might not be best described using input tokens only. Human-made rules are often descriptive.

English	Portuguese
<b>This is</b> the <b>question</b> we must address	<b>Esta</b> é a questão que temos que enfrentar
<b>This is</b> the <b>problem</b> we must address	<b>Este</b> é o problema que temos que enfrentar
<b>This is</b> <b>what</b> we must address	É <b>isso</b> que temos de enfrentar

Table 2: Anchors (in bold) of a machine translation system for the Portuguese word for “This” (in pink).



## — Anchors: how far do they take us?

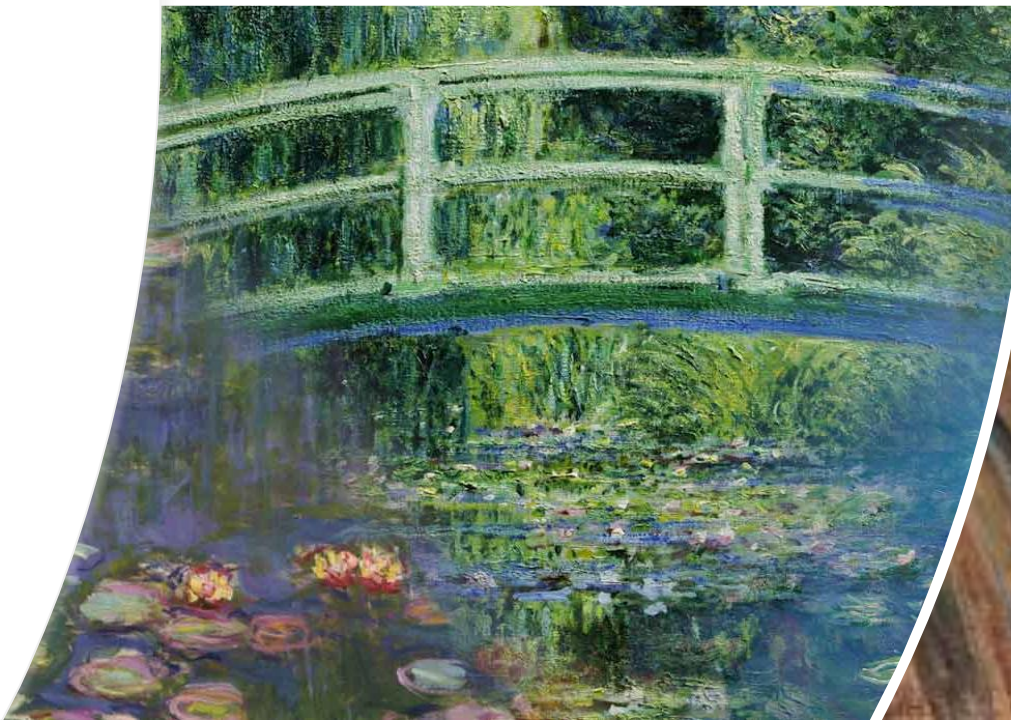
- Anchors can only describe concatenative relations between tokens, no other logical operators.





## Anchors: how far do they take us?

- Sometimes there are no rules with high coverage and all anchors we can get are specific, long, and difficult to comprehend.





# How do humans do it?

- Pre-defined rules + experience (intuition).
- Refined, more generalised anchors can give us a springboard into understanding a model, but we need something more to understand the nuances.



# Summary of Strengths and Limitations

## Strengths

High human precision

Clear rule coverage

Works well for rule-based reasoning tasks

Does not require mental arithmetics

More interpretable than linear approximations

## Limitations

Overly specific rules

Conflict resolution is unclear

Difficult to apply to image and text models

Low coverage in some cases

Poor performance on complex dependencies



No free lunch



# Conclusion and My Thoughts



# Questionable claims

Definitions, definitions.... Predictability vs. Explainability

Name \_\_\_\_\_

Signature \_\_\_\_\_

Date \_\_\_\_\_

# Questionable claims

- Metrics used to compare LIME and anchors: human precision, and coverage

		Precision		Coverage	
		anchor	lime-n	anchor	lime-t
adult	logistic	<u>95.6</u>	<u>81.0</u>	<u>10.7</u>	<u>21.6</u>
	gbt	<u>96.2</u>	<u>81.0</u>	<u>9.7</u>	<u>20.2</u>
	nn	<u>95.6</u>	<u>79.6</u>	<u>7.6</u>	<u>17.3</u>
rcdv	logistic	<u>95.8</u>	<u>76.6</u>	<u>6.8</u>	<u>17.3</u>
	gbt	<u>94.8</u>	<u>71.7</u>	<u>4.8</u>	<u>2.6</u>
	nn	<u>93.4</u>	<u>65.7</u>	<u>1.1</u>	<u>1.5</u>
lending	logistic	<u>99.7</u>	<u>80.2</u>	<u>28.6</u>	<u>12.2</u>
	gbt	<u>99.3</u>	<u>79.9</u>	<u>28.4</u>	<u>9.1</u>
	nn	<u>96.7</u>	<u>77.0</u>	<u>16.6</u>	<u>5.4</u>

Table 4: Average precision and coverage with **simulated users** on 3 tabular datasets and 3 classifiers. *lime-n* indicates direct application of LIME to unseen instances, while *lime-t* indicates a threshold was tuned using an oracle to achieve the same precision as the anchor approach. The anchor approach is able to maintain very high precision, while a naive use of linear explanations leads to varying degrees of precision.





## Questionable claims

Authors limited VQA system from 1000 to 5 possible outcomes for the user studies to reduce visual overload.



Questions?