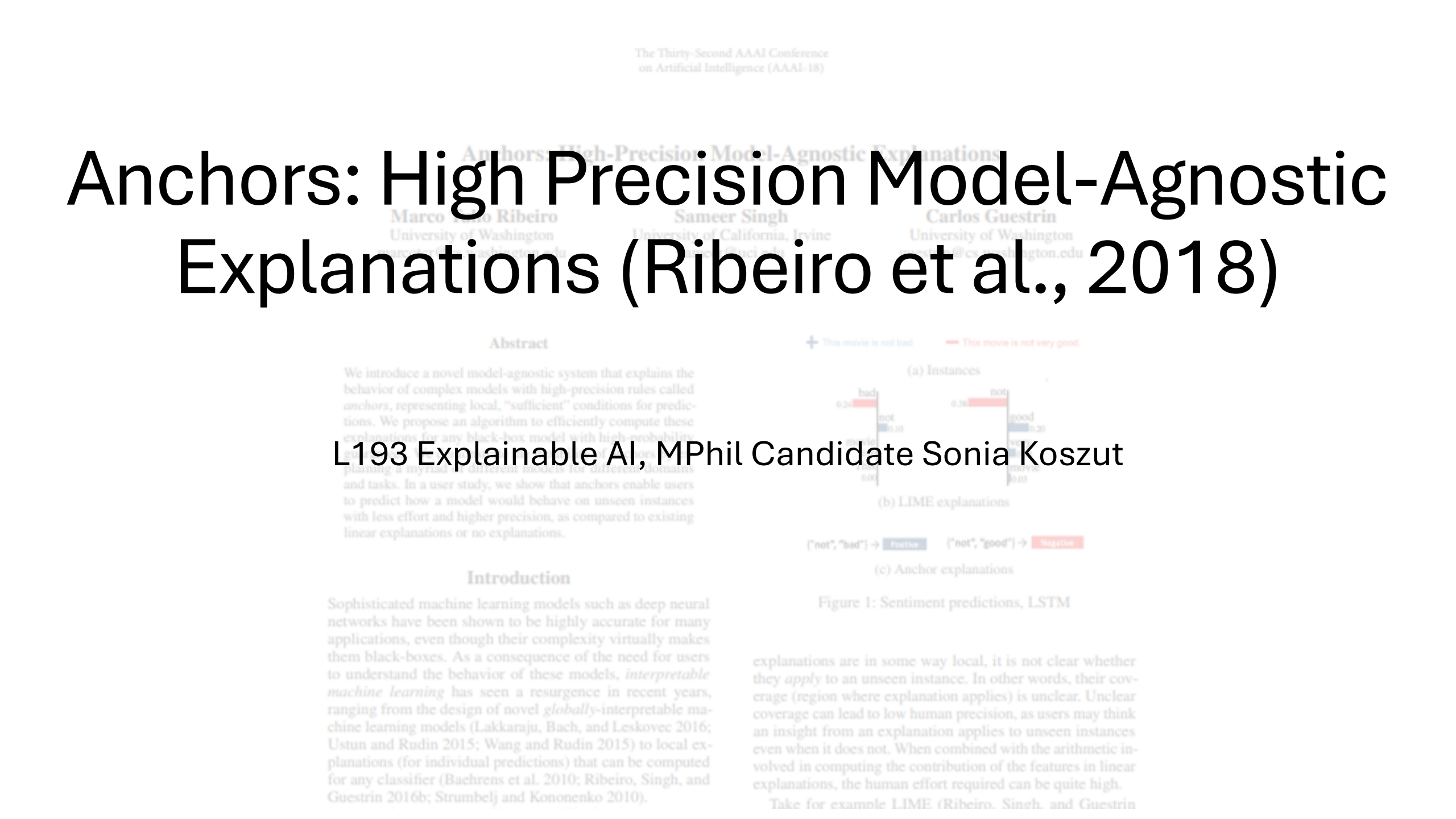
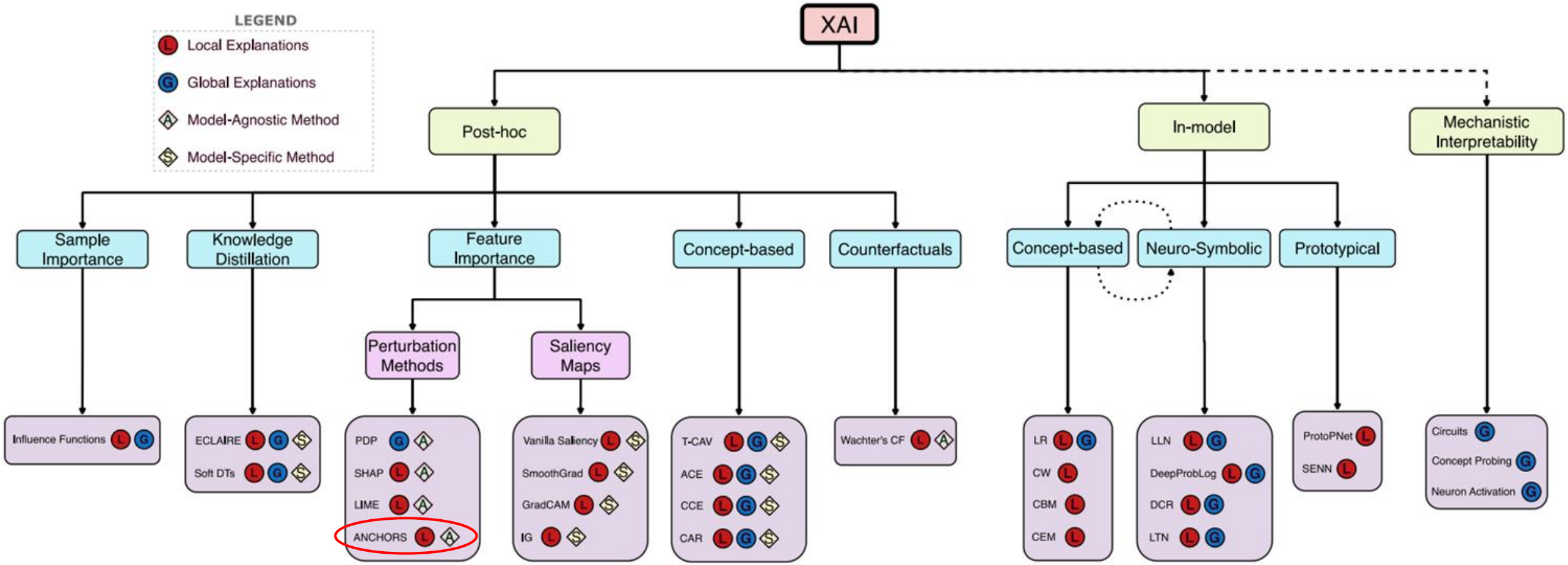
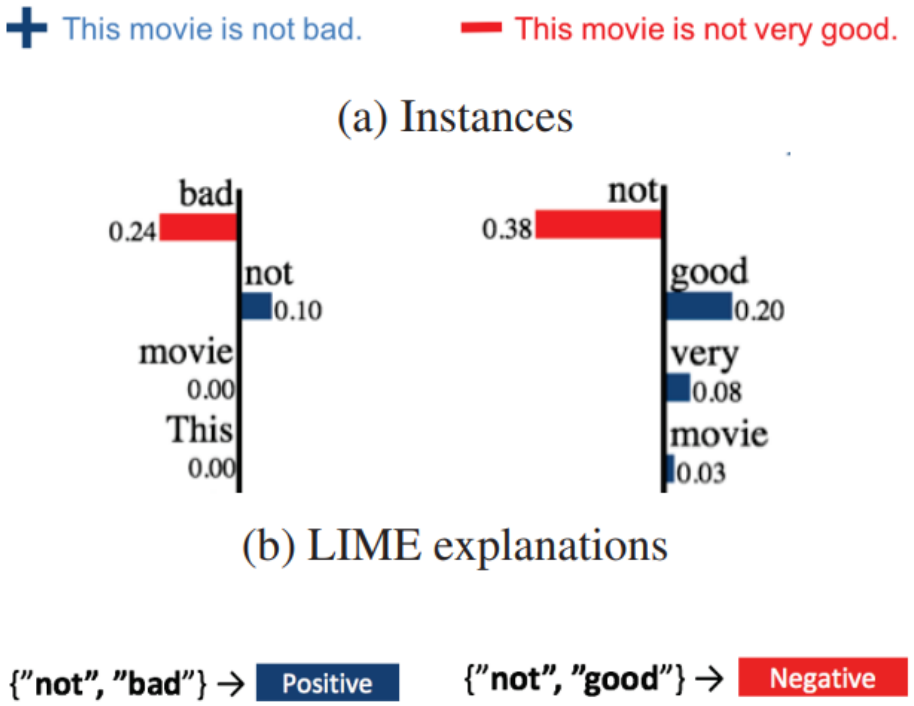
Anchors: High Precision Model-Agnostic Explanations (Ribeiro et al., 2018)

L193 Explainable AI, MPhil Candidate Sonia Koszut




LEGEND
Local Explanations
Global Explanations
Model-Agnostic Method
Model-Specific Method
XAI
Post-hoc
In-model
Mechanistic Interpretability
Sample Importance
Knowledge Distillation
Feature Importance
Concept-based
Counterfactuals
Concept-based
Neuro-Symbolic
Prototypical
Perturbation Methods
Saliency Maps
Influence Functions
ECLAIRE
Soft DTs
PDP
SHAP
LIME
ANCHORS
Vanilla Saliency
SmoothGrad
GradCAM
IG
T-CAV
ACE
CCE
CAR
Wachter's CF
LR
CW
CBM
CEM
LLN
DeepProbLog
DCR
LTN
ProtoPNet
SENN
Circuits
Concept Probing
Neuron Activation

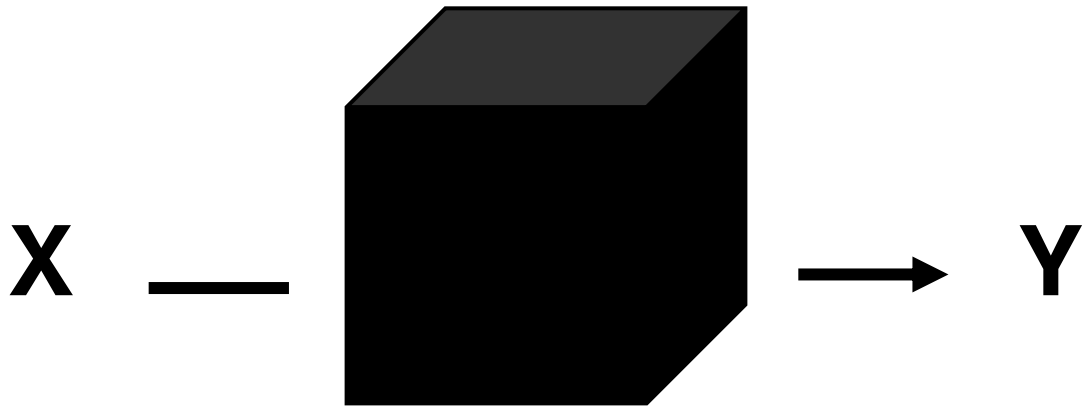


What are Anchors? Why do we need them?

- 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

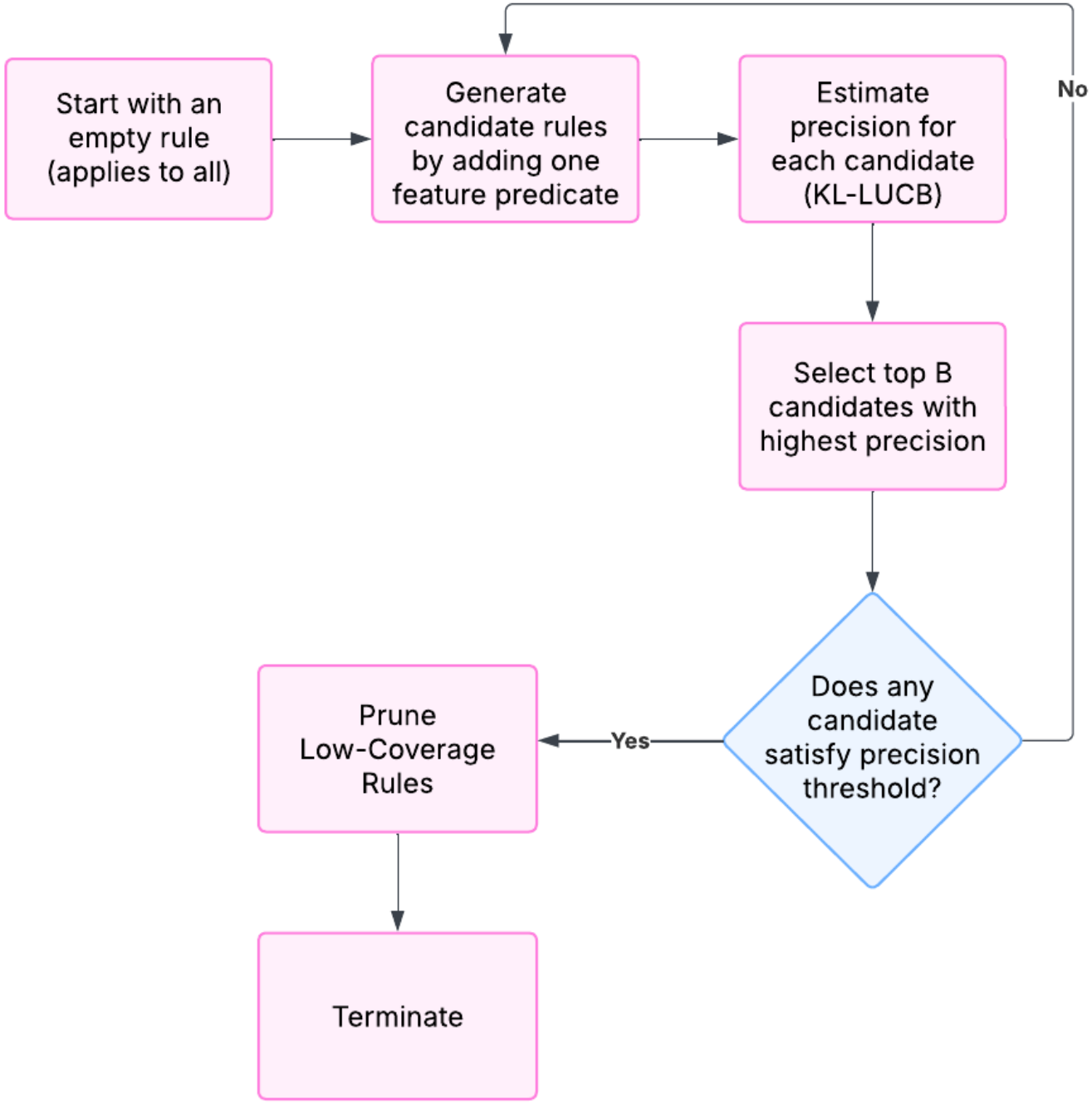
This audio is not bad

This novel is not bad

This footage is not bad

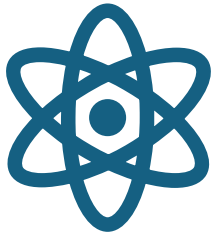
Make it efficient

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



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
This is the **question** we must address	Esta é a questão que temos que enfrentar
This is the **problem** we must address	Este é o problema que temos que enfrentar
This is what we must address	É isso 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?

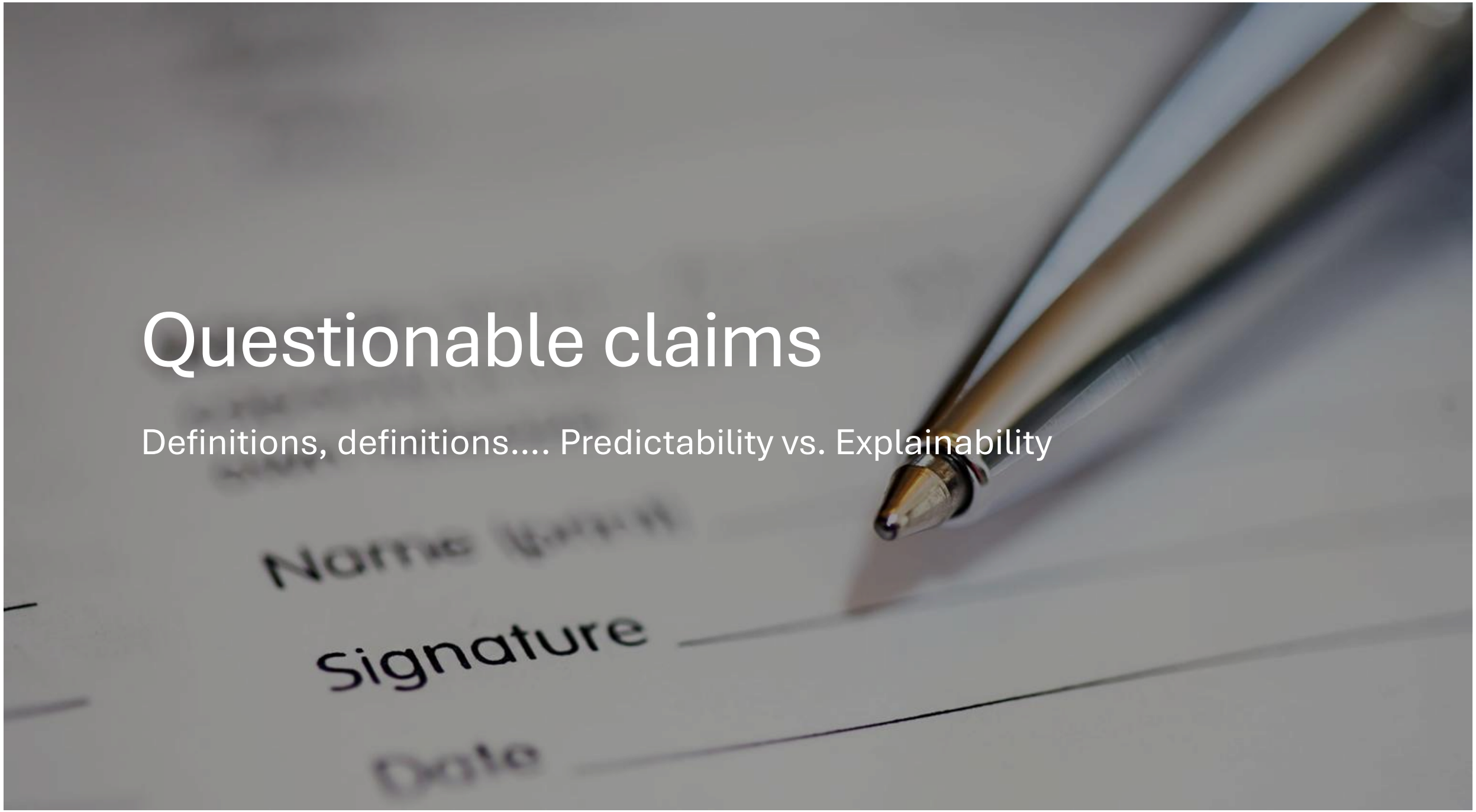
- Pre-defined rules + experience (intuition).
- Refined, more generalised anchors can give us a springboard into understanding a model, but we need something more to understand the nuances.

Summary of Strengths and Limitations

Strengths	Limitations
High human precision	Overly specific rules
Clear rule coverage	Conflict resolution is unclear
Works well for rule-based reasoning tasks	Difficult to apply to image and text models
Does not require mental arithmetics	Low coverage in some cases
More interpretable than linear approximations	Poor performance on complex dependencies

No free lunch

Conclusion and My Thoughts



Questionable claims

Definitions, definitions…. Predictability vs. Explainability

Questionable claims

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

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

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?