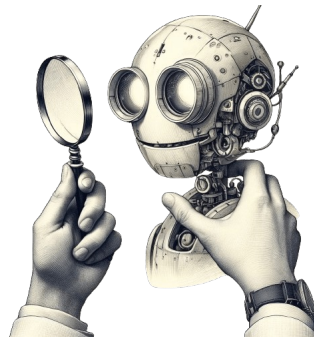


Foundations of Interpretable AI

Tutorial @ **CVPR** *Nashville* JUNE 11-15, 2025



PART I: Motivation and Post-hoc Methods (9:00 – 9:45 am) Aditya Chattopadhyay (Amazon)

PART II: Shapley Values based Methods (9:45 – 10:30 am) Jeremias Sulam (Johns Hopkins)



Coffee break

(10:30 – 11 am)

PART III: Interpretable by Design Methods (11 – 11:45 am) René Vidal (Penn)



Introduction

Aditya Chattopadhyay (Amazon)¹
Jeremias Sulam (Johns Hopkins)
René Vidal (University of Pennsylvania)

¹The content of this tutorial does not relate to Aditya's position at Amazon.



Need for Interpretable AI

- AI models are being increasingly utilized to make decisions that can impact human lives.



Aug 13, 2024 9:06 AM Eastern Daylight Time

InsideTracker Launches Innovative “Ask InsideTracker” AI Chat Feature to Provide Members with Science-Backed Information in Real Time

7 in 10 Companies Will Use AI in the Hiring Process in 2025, Despite Most Saying It’s Biased

Last Updated: October 22, 2024

- Interpretability of these decisions is critical for **reliable** and **responsible** use of AI.

Regulation for AI algorithms

- “Right to Explanation” for AI algorithms being enforced by European Data Protection authorities.

Part of [Chapter IX: Post-Market Monitoring, Information Sharing and Market Surveillance](#) → [Section 4: Remedies](#)

Article 86: Right to Explanation of Individual Decision-Making

Date of entry into force:
2 August 2026

According to:
Article 113

- “Interpretability” of AI algorithms part of FDA guidelines for good ML practices in healthcare.¹

1. <https://www.fda.gov/media/153486/download>

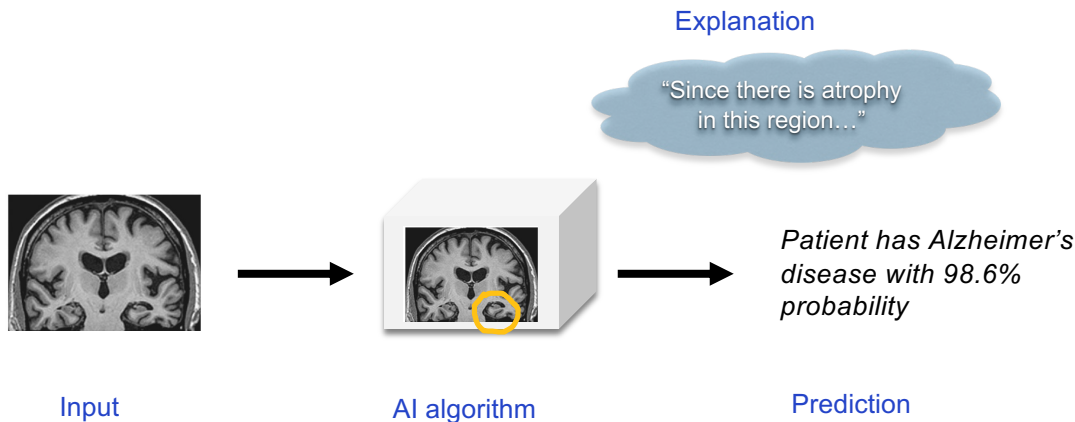
Why Interpretability?

- Mismatch between training objectives (e.g. cross entropy loss) and real world desiderata¹.
 - ❖ Real-world objectives like ethics, bias, and safety are difficult to formalize into mathematical functions.
 - ❖ Real-world environments can change drastically compared to training environments.
- Interpretability helps promote user trust → widespread adoption of AI algorithms.

1. Lipton, Zachary C. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue* 16.3 (2018): 31-57.

What is Interpretable AI?

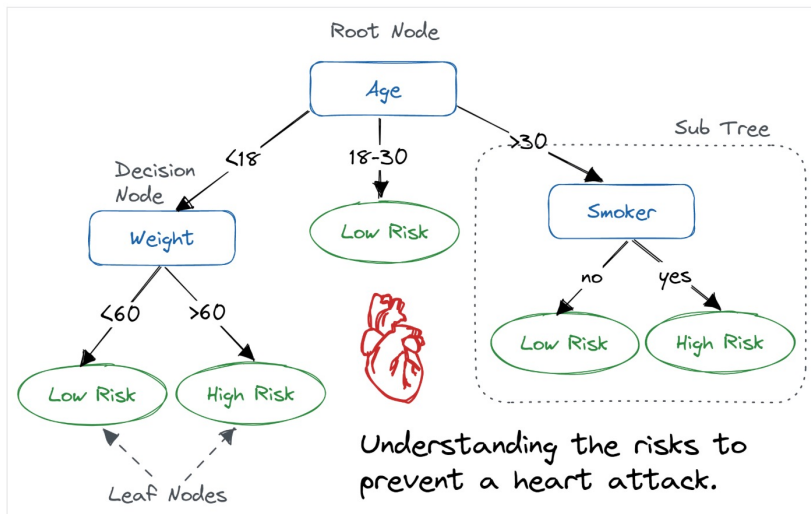
- An AI algorithm that not only makes **accurate predictions** but also provides an **interpretable explanation** for its prediction.



Algorithms for Interpretable AI

➤ Two polarizing approaches

Models that are Interpretable by design



Post-hoc approaches to elicit explanations from black-box models

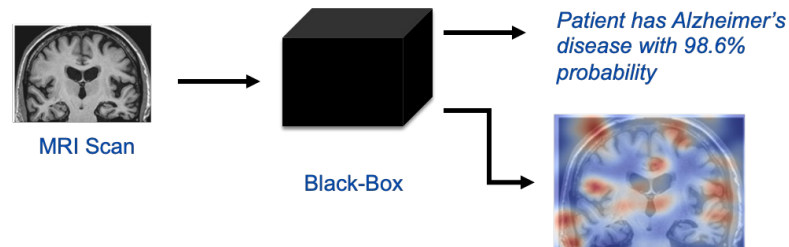


Image: <https://www.datacamp.com/tutorial/decision-tree-classification-python>

Tutorial Outline

1. Post-hoc approaches to model interpretability.

- ❖ Introductory lecture reviewing recent attempts at post-hoc interpretability.

2. Model interpretability with Shapley Coefficients.

- ❖ Deep-dive into Shapley values, a popular post-hoc interpretability method.

3. Interpretability by design.

- ❖ Introductory lecture reviewing recent attempts at developing AI models that are interpretable by design.

Post-hoc Approaches to Model Interpretability

Types of post-hoc interpretability methods

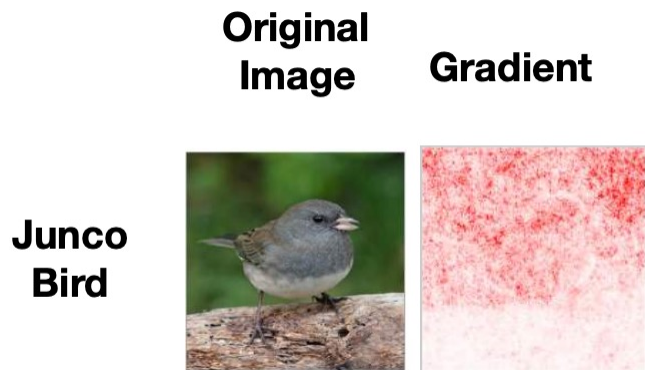
- **Feature attributions:** Explain how different input features affect the model's predictions.
- **Concept-based attributions:** Explain how different high-level semantic concepts affect the model's predictions.
- **Interpretable Model-Agnostic Explanations:** Locally approximate a black-box model with a simpler interpretable-by-design model.

Feature attributions

- **Definition (informal):** Given a model f , and an input x , assign scores to every feature based on how “important” they were for the model’s prediction.
- Different notions of “importance” result in different feature attribution algorithms.
- Most widely used post-hoc interpretability methods.
 - ❖ Popularized as “saliency maps” in vision.

Gradient-based feature attributions

- Gradients measure feature importance in terms of the local sensitivity of the model's output to that feature.



Recall, for a given scalar function f with input x ,

$$\frac{df}{dx} = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

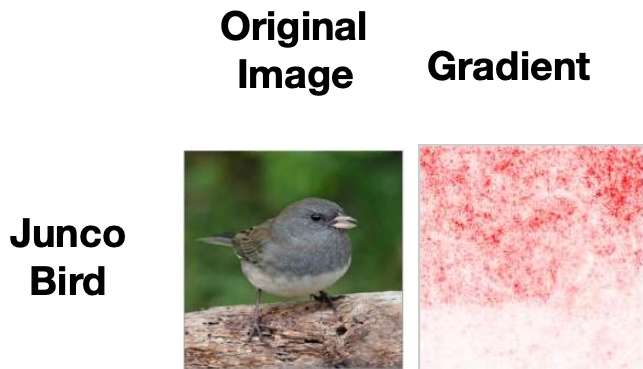
Why did the model predict Junco Bird?

Compute gradient of output logit w.r.t. pixels

Image: Adebayo, Julius, et al. "Sanity checks for saliency maps." *Advances in neural information processing systems* 31 (2018).

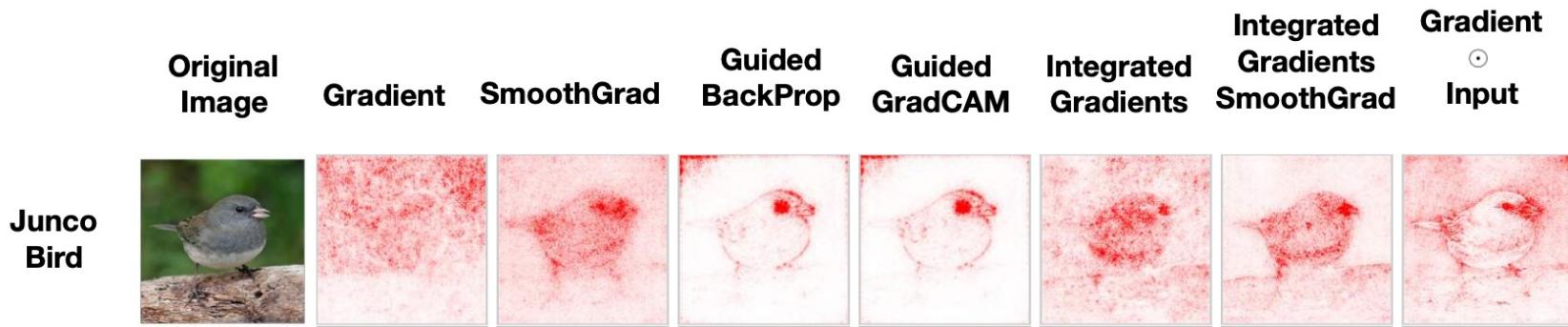
Gradient-based feature attributions

- Individual gradients don't give clear explanations, visually noisy.



Gradient-based feature attributions

- Individual gradients don't give clear explanations, visually noisy.
- Resulted in a flurry of ad-hoc methods that manipulate the gradients to produce "better-looking" feature maps

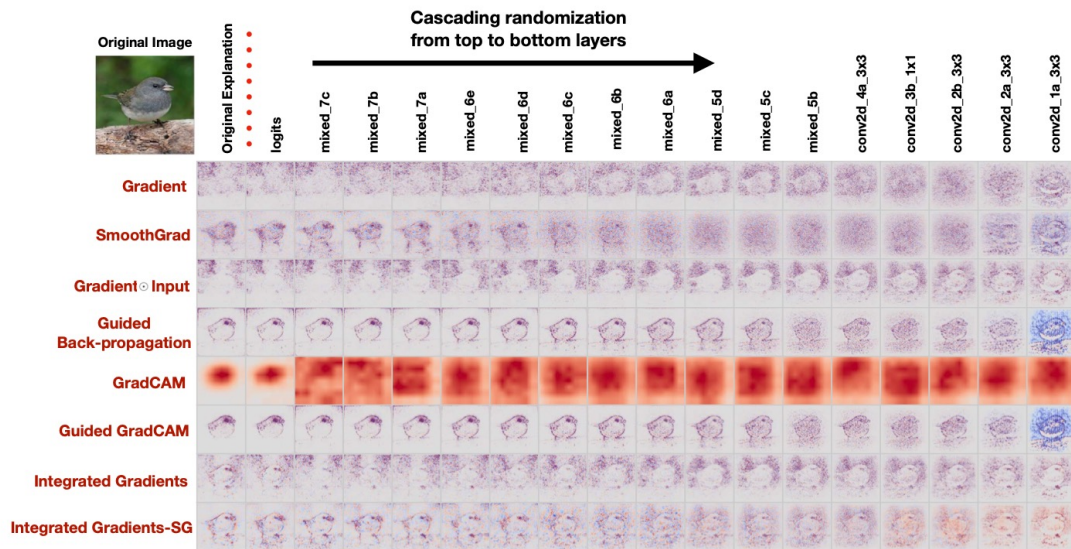


Sanity Checks for Saliency Maps

- Proposes simple diagnostic tests that a “reasonable” explanation method should pass.
 - ❖ Explanation should be sensitive to the weights of the network.
 - ❖ Explanation should help distinguish between “learning” and “memorization”.
- Many existing gradient-based approaches fail these simple tests.

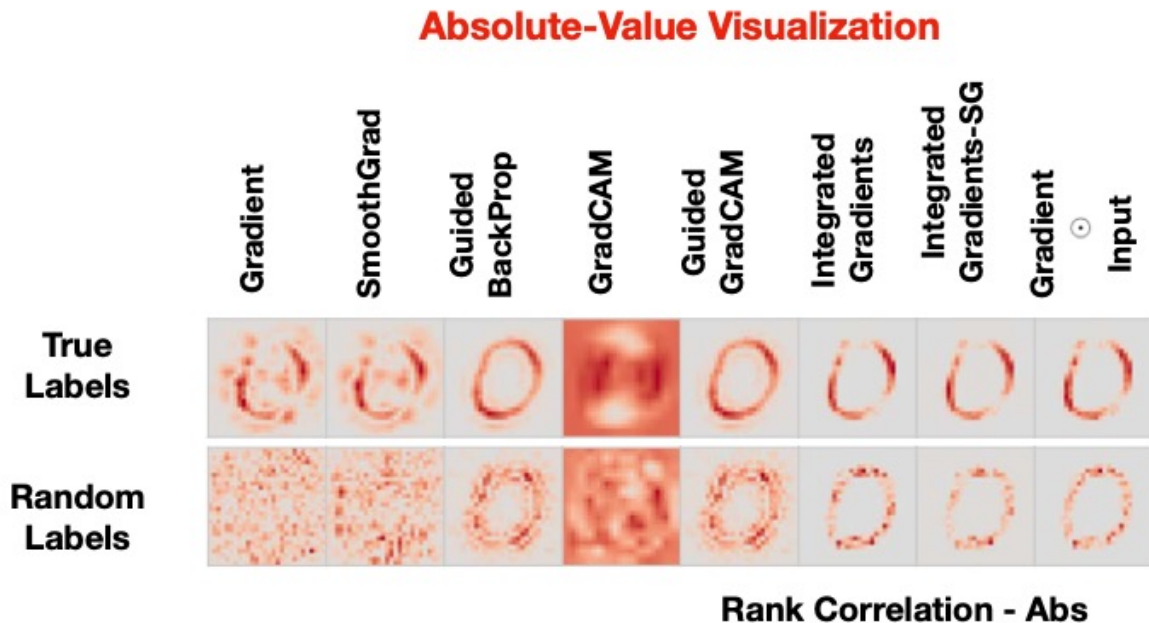
Sanity Checks for Saliency Maps

- ❖ Explanation should be sensitive to the weights of the network.



Sanity Checks for Saliency Maps

- ❖ Explanation should help distinguish between “learning” and “memorization”.



Gradient-based feature attributions

➤ Pro:

- ❖ Simple and efficient to implement, only requires model to be differentiable.
- ❖ Intuitive definition of feature importance.

➤ Cons:

- ❖ Individual feature importance might not be interpretable to the user.
- ❖ Real-world features are often highly correlated, e.g. image pixels. Gradients do not take this into account.

Subset-based feature attributions

- Given an input x and a model f , these methods aim to identify a size-constrained subset of features that contribute most to the prediction $f(x)$.
- Stated formally, for a given budget k , find a subset of feature indices S , such that,

$$\min_{|S| < k} d(f(x_S), f(x))$$

- This is a discrete optimization problem and thus infeasible without approximations.

L2X: a subset-based feature attribution method

- L2X solves the following objective, where I is mutual information.

$$\max_{|S| < k} I(f(x_S), f(x))$$

- Approximations:

- ❖ Mutual Information is intractable → Optimize a variational lower bound.
- ❖ Searching over subsets S is non-differentiable → Continuous relaxation via the Gumbel-SoftMax trick.

¹ Chen, Jianbo, et al. "Learning to explain: An information-theoretic perspective on model interpretation." *International conference on machine learning*. PMLR, 2018.

L2X in action

- Explanations for an LSTM trained for sentiment classification on the IMDB movie review dataset.

Truth	Predicted	Key sentence
positive	positive	There are few really hilarious films about science fiction but this one will knock your sox off. The lead Martians Jack Nicholson take-off is side-splitting. The plot has a very clever twist that has be seen to be enjoyed. This is a movie with heart and excellent acting by all. Make some popcorn and have a great evening.
negative	negative	You get 5 writers together, have each write a different story with a different genre, and then you try to make one movie out of it. Its action, its adventure, its sci-fi, its western, its a mess. Sorry, but this movie absolutely stinks. 4.5 is giving it an awefully high rating. That said, its movies like this that make me think I could write movies, and I can barely write.

- Subsets consist of coherent sentences instead of discontiguous chunks of text!

Subset-based feature attributions

➤ Pro:

- ❖ Accounts for feature correlations by identifying a set.
- ❖ Principled manner of defining “importance” in terms of finding a “minimal” subset required for maintaining performance (accuracy).

➤ Cons:

- ❖ Computational overhead in computing attributions prevent adoption at scale.

Types of post-hoc interpretability methods

- **Feature attributions:** Explain how different input features affect the model's predictions.
- **Concept-based attributions:** Explain how different high-level semantic concepts affect the model's predictions.
- **Interpretable Model-Agnostic Explanations:** Locally approximate a black-box model with a simpler interpretable-by-design model.

Concept-based attributions

- Explanations in terms of raw input features are typically not most intuitive or intelligible to humans.
- Humans reason in terms of high-level concepts.
- How can we assign attributions or importance scores to concepts?

Concept-based attributions



We show saliency maps for predicting brushing teeth.

We see highlighted region near the mouth but what concepts mattered for the decision?

Did the concept “toothbrush” matter?
Did the concept “mouth” matter?

TCAV: Concept-based attributions

- Quantifies how much of a “concept” was important for prediction in a model.

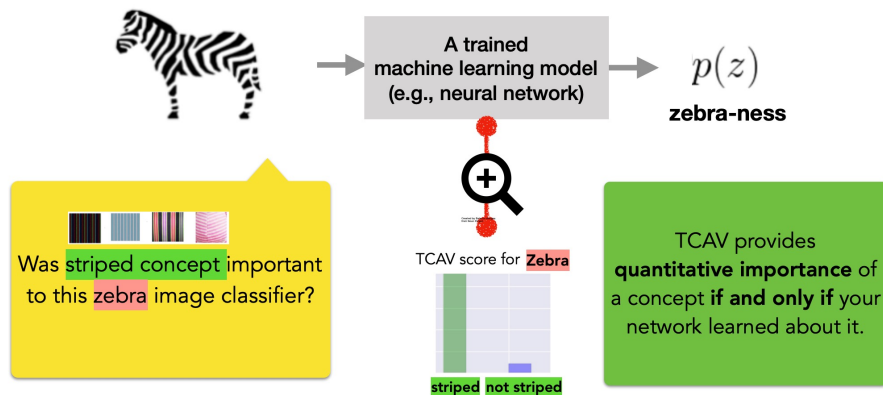


Image: https://beenkim.github.io/slides/TCAV_ICML_pdf.pdf

TCAV: Methodology

1. Get positive and negative training data for concept “stripes”.

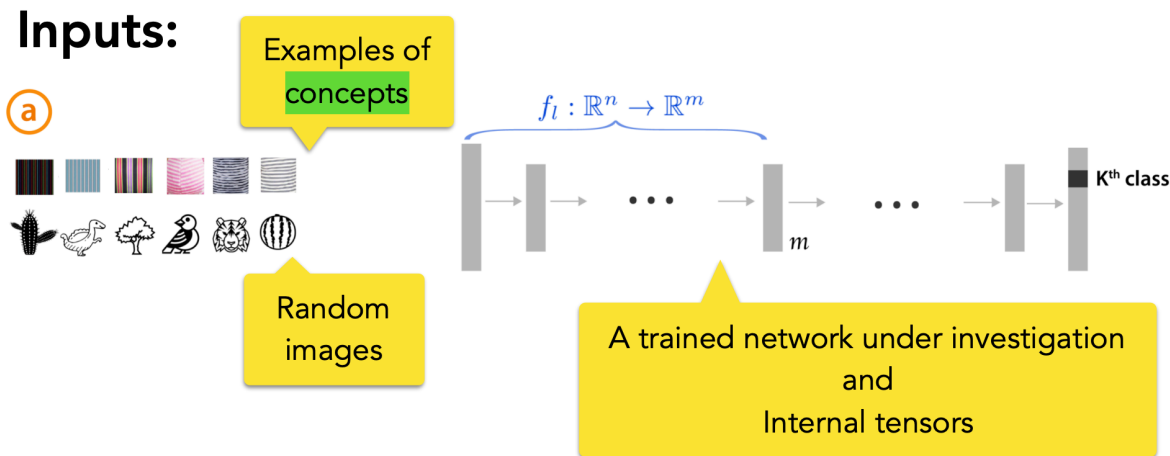


Image: https://beenkim.github.io/slides/TCAV_ICML_pdf.pdf

TCAV: Methodology

2. Learn a concept activation vector for concept “stripes”.

Inputs:

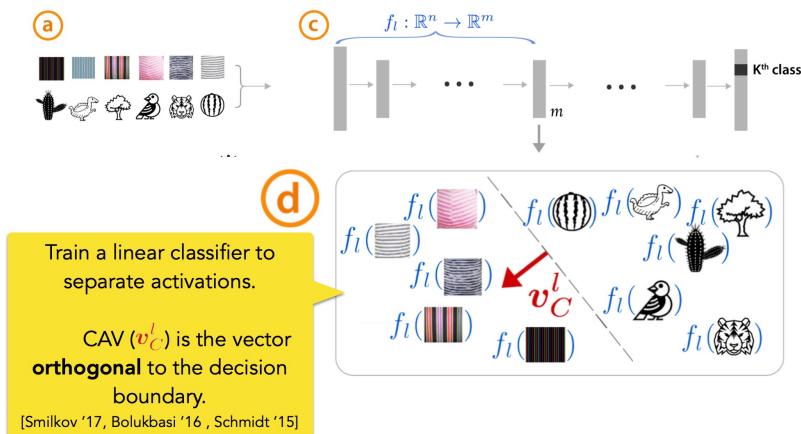
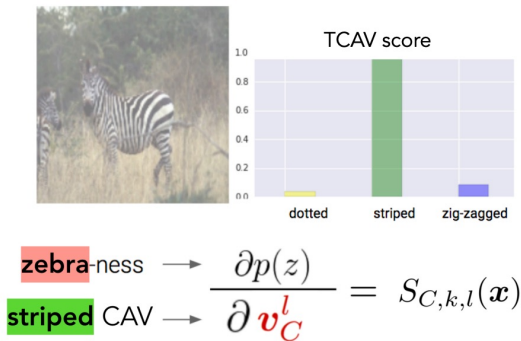


Image: https://beenkim.github.io/slides/TCAV_ICML_pdf.pdf

TCAV: Methodology

3. Get sensitivity of predictions to a concept by directional derivatives, then aggregate over all “Zebra” images to get TCAV score.



Directional derivative with CAV

$$\left. \begin{aligned} S_{C,k,l}(\text{zebra}) \\ S_{C,k,l}(\text{zebra head}) \\ S_{C,k,l}(\text{zebra body}) \\ S_{C,k,l}(\text{zebra tail}) \end{aligned} \right\}$$

$$\text{TCAV}_{Q_{C,k,l}} = \frac{|\{\mathbf{x} \in X_k : S_{C,k,l}(\mathbf{x}) > 0\}|}{|X_k|}$$

Image: https://beenkim.github.io/slides/TCAV_ICML_pdf.pdf

Concept-based attributions

➤ Pro:

- ❖ Provides human-friendly explanations in terms of concept attributions.
- ❖ User can choose to obtain explanations in terms of any concept.

➤ Cons:

- ❖ Burden on user to provide appropriate concepts, which is often not known.
- ❖ Even with known concepts, might not always be feasible to obtain training set of positive and negative images.

Types of post-hoc interpretability methods

- **Feature attributions:** Explain how different input features affect the model's predictions.
- **Concept-based attributions:** Explain how different high-level semantic concepts affect the model's predictions.
- **Interpretable Model-Agnostic Explanations:** Locally approximate a black-box model with a simpler interpretable-by-design model.

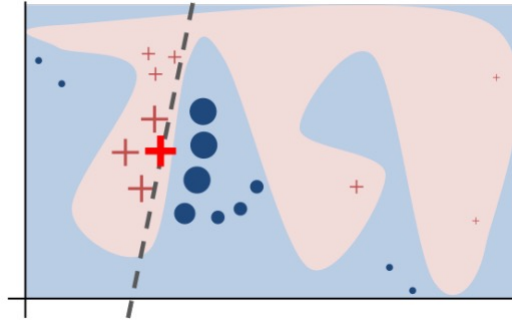
Local Interpretable Model-Agnostic Explanations (LIME)

- So far, we saw post-hoc methods that give explanations in terms of this feature/concept was important for prediction.
- What if the user wants a more holistic understanding of the deep model's decision-making process?
 - ❖ For example, as a Boolean expression or decision tree?
- But these simpler “interpretable” models are not as expressive as a deep network. 😞

¹ Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." *arXiv preprint arXiv:1606.05386* (2016).

LIME: Idea

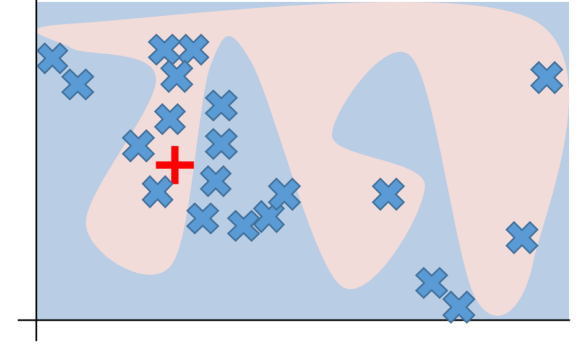
- A non-linear deep network can be locally approximated by a simpler “interpretable” model like logistic regression classifier or a decision tree.



1. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." *arXiv preprint arXiv:1606.05386* (2016).
Image: <https://www.seas.upenn.edu/~obastani/cis7000/spring2024/docs/lecture18.pdf>

LIME: Methodology

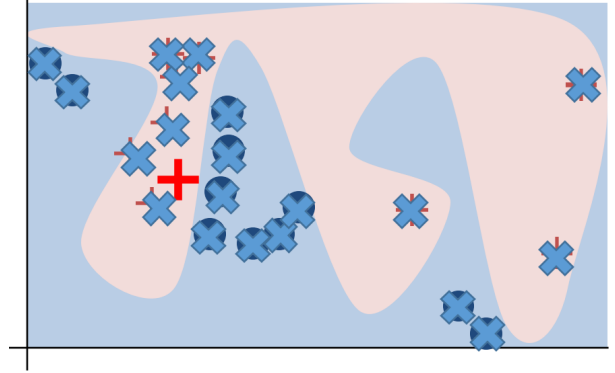
- Red cross x is the point to be explained.
- First, sample points around x.



1. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." *arXiv preprint arXiv:1606.05386* (2016).
Image: <https://www.seas.upenn.edu/~obastani/cis7000/spring2024/docs/lecture18.pdf>

LIME: Methodology

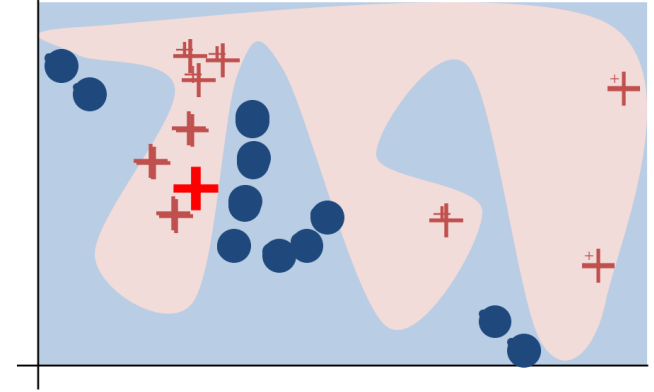
- Red cross x is the point to be explained.
- First, sample points around x.
- Use deep model to predict labels for each sample



1. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." *arXiv preprint arXiv:1606.05386* (2016).
Image: <https://www.seas.upenn.edu/~obastani/cis7000/spring2024/docs/lecture18.pdf>

LIME: Methodology

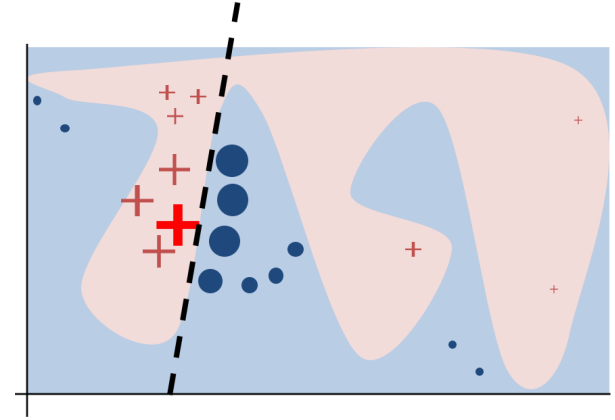
- Red cross x is the point to be explained.
- First, sample points around x.
- Use deep model to predict labels for each sample.
- Weigh samples according to distance to x.



1. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." *arXiv preprint arXiv:1606.05386* (2016).
Image: <https://www.seas.upenn.edu/~obastani/cis7000/spring2024/docs/lecture18.pdf>

LIME: Methodology

- Red cross x is the point to be explained.
- First, sample points around x.
- Use deep model to predict labels for each sample.
- Weigh samples according to distance to x.
- Learn a simple classifier (say linear) on weighted samples



1. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." *arXiv preprint arXiv:1606.05386* (2016).
Image: <https://www.seas.upenn.edu/~obastani/cis7000/spring2024/docs/lecture18.pdf>

LIME in Action

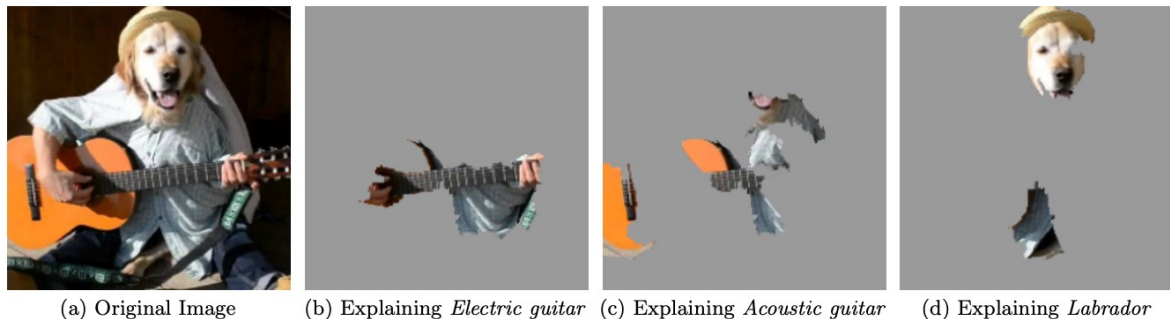


Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Image: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." *arXiv preprint arXiv:1606.05386* (2016).

Local Interpretable Model-Agnostic Explanations

➤ Pro:

- ❖ Is model-agnostic, does not even require access to gradients.
- ❖ User has choice to pick their favorite “interpretable” ML model.

➤ Cons:

- ❖ Computationally expensive since it requires training “interpretable” predictors for every instance to be explained.
- ❖ Explanations very unstable, dependent on the sampling process.

Diagnostics for Post-hoc Methods

- How can we know if a given explanation is correct?
 - ❖ Need explanations to be faithful to the true decision-making process.
- Difficult to evaluate faithfulness since we do not know the true explanation.
- We will describe some proxy methods in next slide.

Diagnostics for Post-hoc Methods

- **Subtractive metrics:** Remove top-k important features and check how much score decreases w.r.t control group of features.
- **Additive metrics:** Keep top-k important features and check how much score increases compared to control of “no features”.
- **Perturbation metrics:** How sensitive are the explanations to small perturbations to the input?

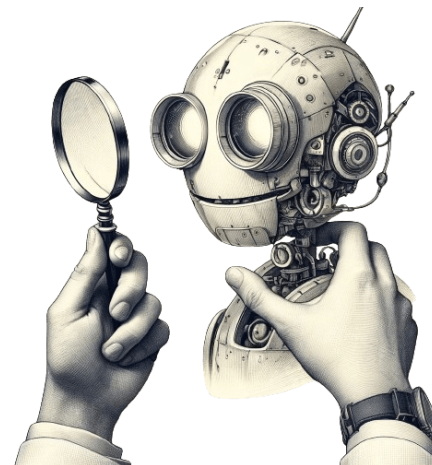
More metrics here: Nauta, Meike, et al. "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai." *ACM Computing Surveys* 55.13s (2023): 1-42.

Thank you!



Foundations of Interpretable AI

Tutorial @ **CVPR** *Nashville* JUNE 11-15, 2025



PART I: Motivation and Post-hoc Methods (9:00 - 9:45 am) **Aditya Chattopadhyay (Amazon)**

PART II: Shapley Value based Methods (9:45 - 10:30 am) **Jeremias Sulam (Johns Hopkins)**

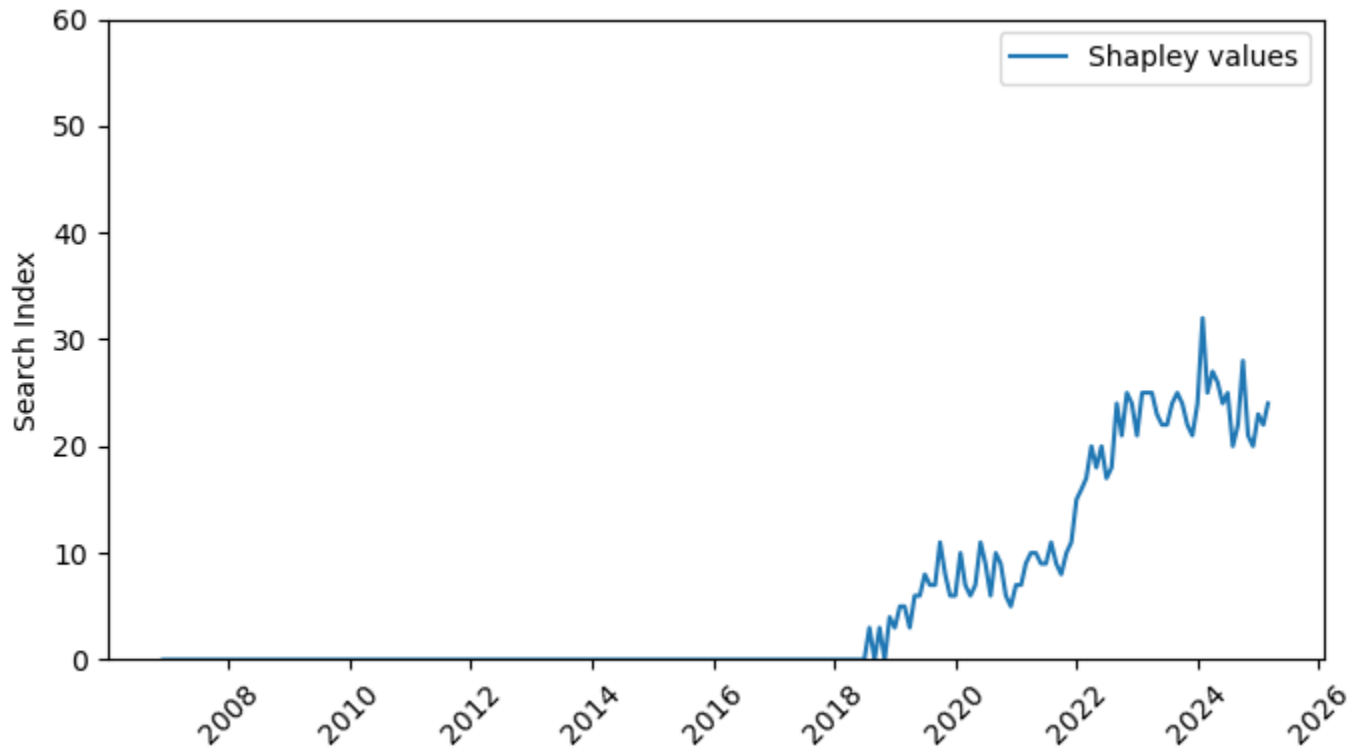
☕ Coffee break ☕ (10:30 - 11:00 am)

PART III: Interpretable by Design Methods (11:00 - 11:45 am) **René Vidal (Penn)**

Shapley Values

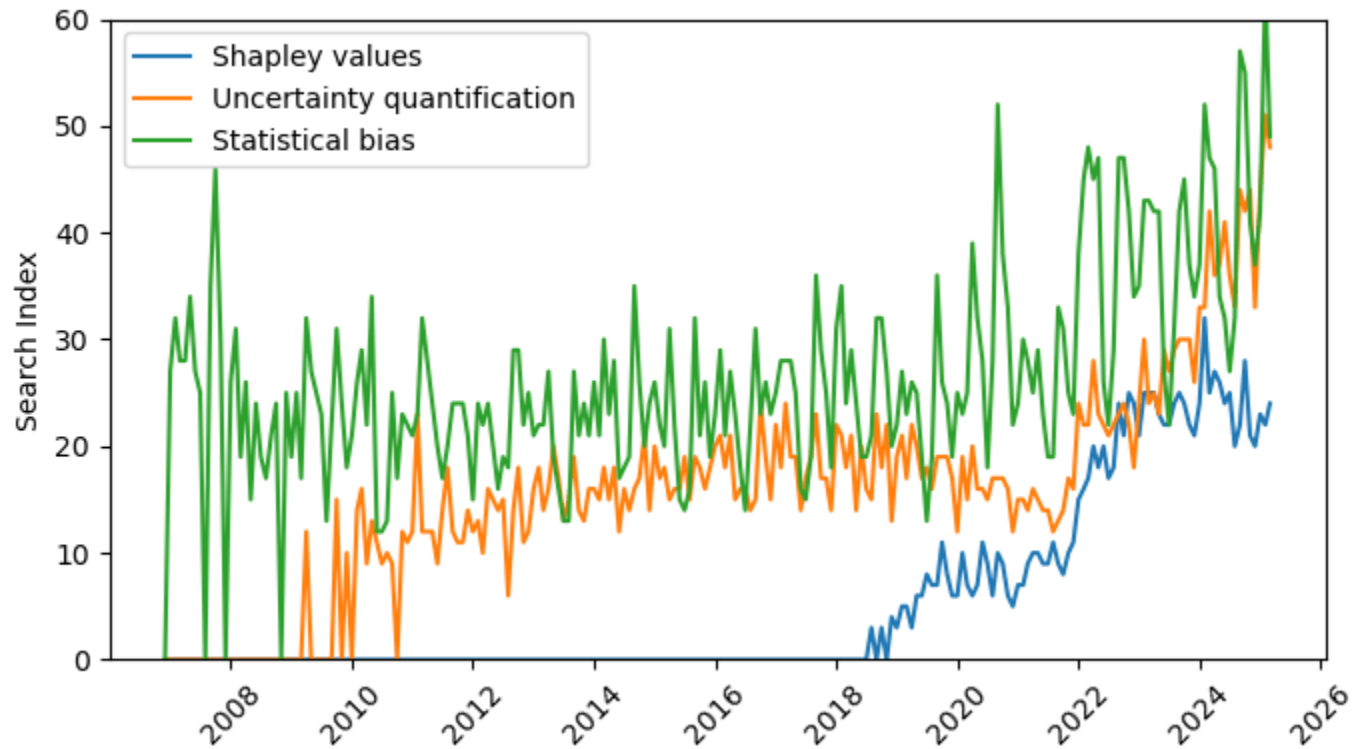
Popularity on

Google Trends



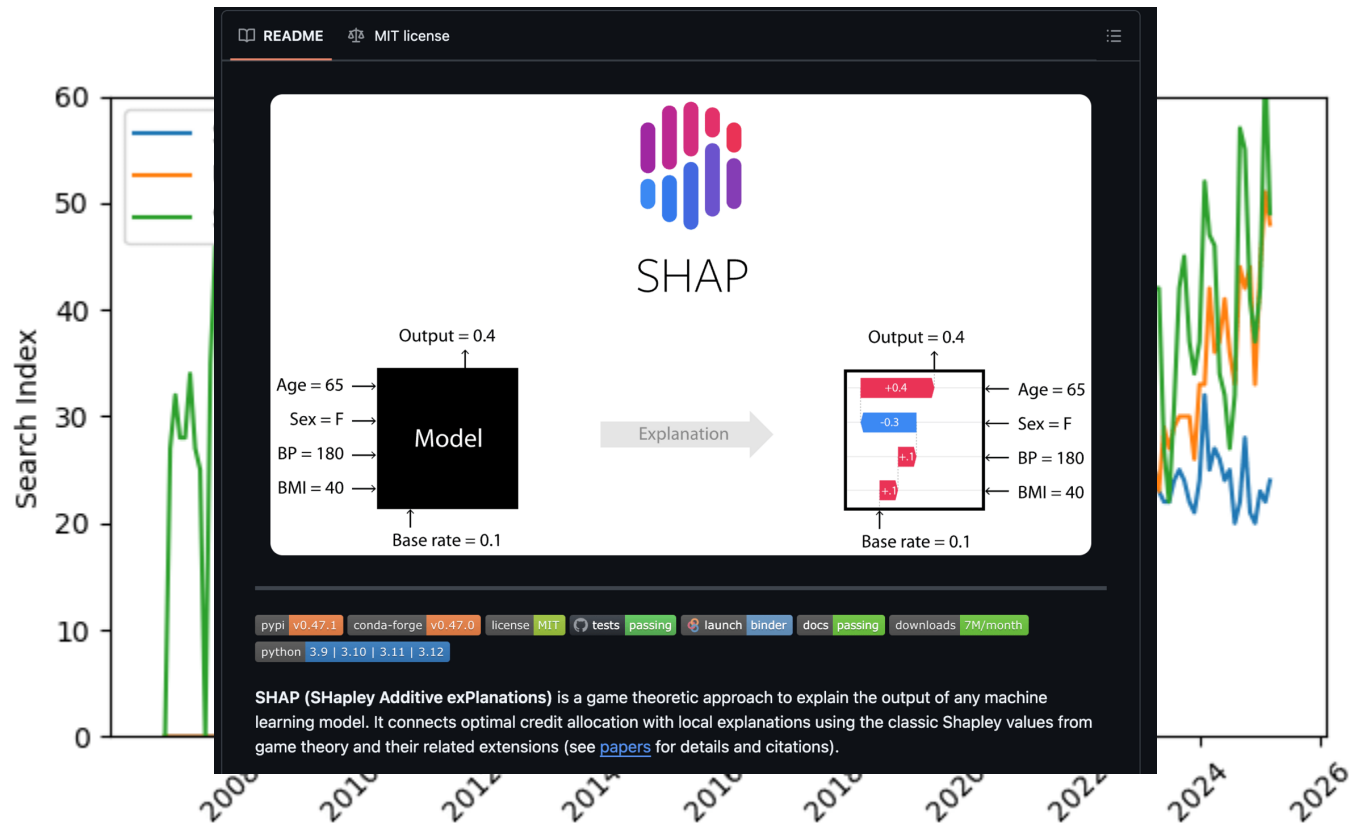
Shapley Values

Popularity on



Shapley Values

Popularity on



Shapley Values

Shapley Values

TODAY

What are they?

How are they computed?

Shapley Values

TODAY

What are they?

How are they computed?

(Shapley for local feature importance)

Shapley Values

TODAY

What are they?

How are they computed?

(Shapley for local feature importance)

- Not an exhaustive literature review
- Not a code & repos review
- Not a demonstration on practical problems

Shapley Values

TODAY

What are they?

How are they computed?

(Shapley for local feature importance)

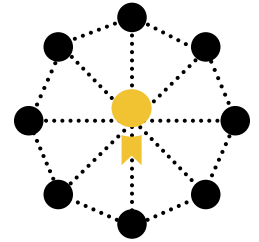
- Not an exhaustive literature review
- Not a code & repos review
- Not a demonstration on practical problems

- Review of **general approaches** and **methodology**
- **Pointers** to where to start looking in **different problem domains**

Shapley Values



Let $G = ([n], f)$ be an n -person cooperative game with characteristic function $v : \mathcal{P}([n]) \mapsto \mathbb{R}$

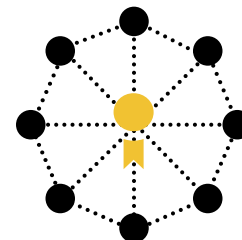


How important is each player for the outcome of the game?

Shapley Values



Let $G = ([n], f)$ be an n -person cooperative game with characteristic function $v : \mathcal{P}([n]) \mapsto \mathbb{R}$



How important is each player for the outcome of the game?

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]$$



marginal contribution of player i with coalition S

Shapley Values

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]$$

Shapley Values

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]$$

- Efficiency

$$v([n]) = v(\emptyset) + \sum_{i=1}^n \phi_i(v)$$

Shapley Values

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]$$

- Efficiency

$$v([n]) = v(\emptyset) + \sum_{i=1}^n \phi_i(v)$$

- Linearity

$$\phi_i(\alpha_1 v_1 + \alpha_2 v_2) = \alpha_1 \phi_i(v_1) + \alpha_2 \phi_i(v_2)$$

for characteristic functions v_1, v_2

Shapley Values

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]$$

- Efficiency

$$v([n]) = v(\emptyset) + \sum_{i=1}^n \phi_i(v)$$

- Symmetry

If $v(S \cup i) = v(S \cup j) \quad \forall S \subseteq [n] \setminus \{i, j\}$
Then $\phi_i(v) = \phi_j(v)$

- Linearity

$$\phi_i(\alpha_1 v_1 + \alpha_2 v_2) = \alpha_1 \phi_i(v_1) + \alpha_2 \phi_i(v_2)$$

for characteristic functions v_1, v_2

Shapley Values

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]$$

- Efficiency

$$v([n]) = v(\emptyset) + \sum_{i=1}^n \phi_i(v)$$

- Symmetry

If $v(S \cup i) = v(S \cup j) \quad \forall S \subseteq [n] \setminus \{i, j\}$
Then $\phi_i(v) = \phi_j(v)$

- Linearity

$\phi_i(\alpha_1 v_1 + \alpha_2 v_2) = \alpha_1 \phi_i(v_1) + \alpha_2 \phi_i(v_2)$
for characteristic functions v_1, v_2

- Nullity

If $v(S \cup i) = v(S) \quad \forall S \subseteq [n] \setminus \{i\}$
Then $\phi_i(v) = 0$

Shapley Explanations for ML

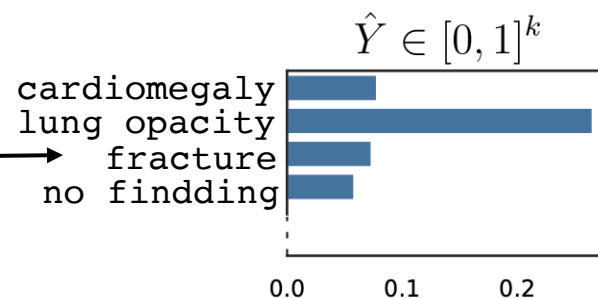
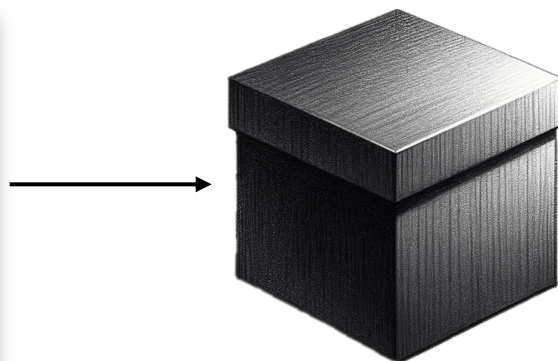
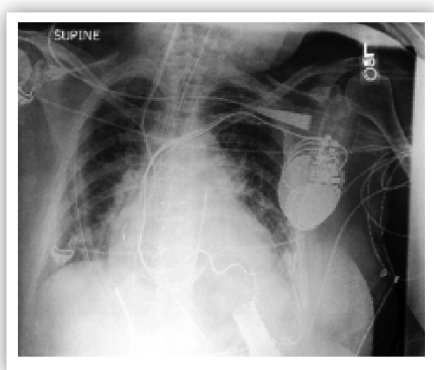
$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]$$

inputs $X \in \mathcal{X} \subset \mathbb{R}^n$

predictor $f : \mathcal{X} \rightarrow \mathcal{Y}$

responses $Y \in \mathcal{Y} = [C]$

$$f(X) = \hat{Y} \approx Y$$



Shapley Explanations for ML

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]$$

inputs $X \in \mathcal{X} \subset \mathbb{R}^n$

predictor $f : \mathcal{X} \rightarrow \mathcal{Y}$

responses $Y \in \mathcal{Y} = [C]$

$$f(X) = \hat{Y} \approx Y$$

Question 1:

How should (can) we choose the function v ?

Shapley Explanations for ML

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]$$

inputs $X \in \mathcal{X} \subset \mathbb{R}^n$

predictor $f : \mathcal{X} \rightarrow \mathcal{Y}$

responses $Y \in \mathcal{Y} = [C]$

$$f(X) = \hat{Y} \approx Y$$

Question 1:

How should (can) we choose the function v ?

Question 2:

How can we (when) compute $\phi_i(v)$?

Shapley Explanations for ML

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]$$

inputs $X \in \mathcal{X} \subset \mathbb{R}^n$

predictor $f : \mathcal{X} \rightarrow \mathcal{Y}$

responses $Y \in \mathcal{Y} = [C]$

$$f(X) = \hat{Y} \approx Y$$

Question 1:

How should (can) we choose the function v ?

Question 2:

How can we (when) compute $\phi_i(v)$?

Question 3:

What do $\phi_i(v)$ say (and don't) about the problem?

Shapley Explanations for ML

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} w(S) [v(S \cup \{i\}) - v(S)]$$

inputs $X \in \mathcal{X} \subset \mathbb{R}^n$

predictor $f : \mathcal{X} \rightarrow \mathcal{Y}$

responses $Y \in \mathcal{Y} = [C]$

$$f(X) = \hat{Y} \approx Y$$

Question 1:

How should (can) we choose the function v ?

Question 2:

How can we (when) compute $\phi_i(v)$?

Question 3:

What do $\phi_i(v)$ say (and don't) about the problem?

Question 1:

How should (can) we choose the function v ?

For any $S \subseteq [n]$, and a sample $x \sim p_X$, we need

$$v_f(S, x) : \mathcal{P}([n]) \times \mathcal{X} \rightarrow \mathbb{R}$$

Question 1:

How should (can) we choose the function v ?

- Fixed reference value

$$v_f(S, x) = f(x_S, x_{\bar{S}}^b)$$

Explicand	Present Features	Baseline	Absent Features
$x_1 x_2 x_3 x_4$	S	$x_1 x_2 x_3 x_4$	$N \setminus S$
x^e 2 9 4 7	$\{1, 2\}$	x^b 0 0 0 0	$\{3, 4\}$
$v(S) = f(x), x =$	2 9	$v(S) = f(x), x =$	2 9 0 0

Question 1:

How should (can) we choose the function v ?

- Fixed reference value

$$v_f(S, x) = f(x_S, x_{\bar{S}}^b)$$

Explicand	Present Features	Baseline	Absent Features
$x_1 x_2 x_3 x_4$	S	$x_1 x_2 x_3 x_4$	$N \setminus S$
x^e 2 9 4 7	$\{1, 2\}$	x^b 0 0 0 0	$\{3, 4\}$
$v(S) = f(x), x =$	2 9	$v(S) = f(x), x =$	2 9 0 0

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} w(S) [f(x_{S \cup i}, x_{\bar{S} \cup i}^b) - f(x_S, x_{\bar{S}}^b)]$$

Question 1:

How should (can) we choose the function v ?

- Fixed reference value

$$v_f(S, x) = f(x_S, x_{\bar{S}}^b)$$

Explicand	Present Features	Baseline	Absent Features
$x_1 x_2 x_3 x_4$	S	$x_1 x_2 x_3 x_4$	$N \setminus S$
x^e 2 9 4 7	$\{1, 2\}$	x^b 0 0 0 0	$\{3, 4\}$
$v(S) = f(x), x =$ 2 9		$v(S) = f(x), x =$ 2 9 0 0	



Easy, cheap




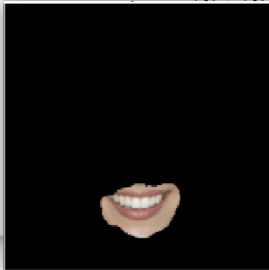
$(x_S, x_{\bar{S}}^b) \not\sim p_X$

Question 1:

How should (can) we choose the function v ?

- Fixed reference value

$$v_f(S, x) = f(x_S, x_{\bar{S}}^b)$$

Explicand	Present Features	Baseline	Absent Features
$x_1 x_2 x_3 x_4$			
x^e	2 9 4 7	0 0	$N \setminus S$ {3, 4}
$v(S) = f(x), x =$		$(x), x =$	2 9 0 0
	x	$(x_S, x_{\bar{S}}^b)$	



Easy, cheap



$(x_S, x_{\bar{S}}^b) \not\sim p_X$

Question 1:

How should (can) we choose the function v ?

- Conditional Data Distribution

$$v_f(S, x) = \mathbb{E}[f(x_S, \tilde{X}_{\bar{S}}) | X_S = x_S]$$

Question 1:

How should (can) we choose the function v ?

- Conditional Data Distribution

$$v_f(S, x) = \mathbb{E}[f(x_S, \tilde{X}_{\bar{S}}) | X_S = x_S]$$

Explicand	Present Features	data																					
$x_1 x_2 x_3 x_4$	S	<table><tr><td>2</td><td>9</td><td>9</td><td>1</td></tr><tr><td>4</td><td>8</td><td>1</td><td>4</td></tr><tr><td>3</td><td>7</td><td>6</td><td>5</td></tr><tr><td>2</td><td>9</td><td>4</td><td>7</td></tr><tr><td>4</td><td>9</td><td>1</td><td>8</td></tr></table>	2	9	9	1	4	8	1	4	3	7	6	5	2	9	4	7	4	9	1	8	
2	9	9	1																				
4	8	1	4																				
3	7	6	5																				
2	9	4	7																				
4	9	1	8																				
x^e	$\{1, 2\}$		$(x_S, \tilde{X}_{\bar{S}}) \sim$																				
			<table><tr><td>2</td><td>9</td><td>9</td><td>1</td></tr><tr><td>2</td><td>9</td><td>4</td><td>7</td></tr></table>	2	9	9	1	2	9	4	7												
2	9	9	1																				
2	9	4	7																				
$v(S) = f(x), x =$	<table><tr><td>2</td><td>9</td><td></td><td></td></tr></table>	2	9																				
2	9																						

Question 1:

How should (can) we choose the function v ?

- Conditional Data Distribution

$$v_f(S, x) = \mathbb{E}[f(x_S, \tilde{X}_{\bar{S}}) | X_S = x_S]$$

Explicand	Present Features	data
$x_1 x_2 x_3 x_4$	S	
x^e 2 9 4 7	$\{1, 2\}$	<div> <div>2 9 9 1</div> <div>4 8 1 4</div> <div>3 7 6 5</div> <div>2 9 4 7</div> <div>4 9 1 8</div> </div>
$v(S) = f(x), x =$ 2 9		

$(x_S, \tilde{X}_{\bar{S}}) \sim$ 2 9 9 1
2 9 4 7



$$(x_S, \tilde{X}_{\bar{S}}) \sim$$

"True ~~to~~ the data"



- Difficult/expensive
- "Breaks" the Null axiom:

$$\text{if } f(x_i, x_{i^c}) = f(x'_i, x_{i^c}) \forall x_{i^c} \not\Rightarrow \phi_i(f) \neq 0$$

Question 1:

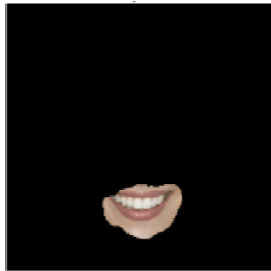
How should (can) we choose the function v ?

- Conditional Data Distribution

$$v_f(S, x) = \mathbb{E}[f(x_S, \tilde{X}_{\bar{S}}) | X_S = x_S]$$



x



x_S



$(x_S, \tilde{X}_{\bar{S}}) \sim$

"True ~~to~~ the data"



- Difficult/expensive
- "Breaks" the Null axiom:

if $f(x_i, x_{i^c}) = f(x'_i, x_{i^c}) \forall x_{i^c} \not\Rightarrow \phi_i(f) \neq 0$

Question 1:

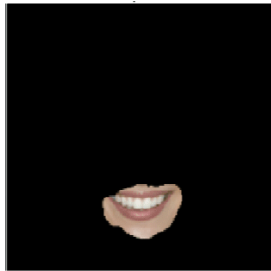
How should (can) we choose the function v ?

- Conditional Data Distribution

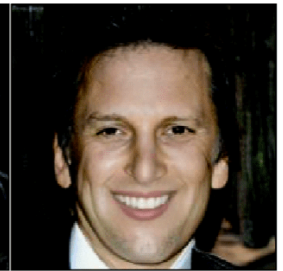
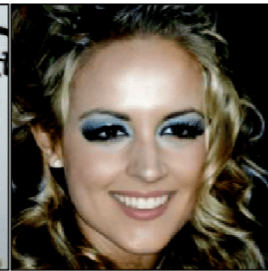
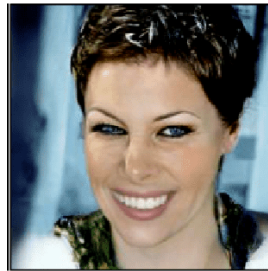
$$v_f(S, x) = \mathbb{E}[f(x_S, \tilde{X}_{\bar{S}}) | X_S = x_S]$$



x



x_S



$(x_S, \tilde{X}_{\bar{S}})$



$(x_S, \tilde{X}_{\bar{S}}) \sim$

"True ~~to~~ the data"



- Difficult/expensive
- "Breaks" the Null axiom:

if $f(x_i, x_{i^c}) = f(x'_i, x_{i^c}) \forall x_{i^c} \not\Rightarrow \phi_i(f) \neq 0$

Question 1:

How should (can) we choose the function v ?

- Conditional Data Distribution

$$v_f(S, x) = \mathbb{E}[f(x_S, \tilde{X}_{\bar{S}}) | X_S = x_S]$$

Alternative: learn a model g_θ for the conditional expectation

$$v_f(S, x) = \mathbb{E}[f(x_S, \tilde{X}_{\bar{S}}) | X_S = x_S] \approx g_\theta(x, S)$$

[Frye et al, 2021]



$$(x_S, \tilde{X}_{\bar{S}}) \sim$$

"True ~~to~~ the data"



- Difficult/expensive
- "Breaks" the Null axiom:

$$\text{if } f(x_i, x_{i^c}) = f(x'_i, x_{i^c}) \forall x_{i^c} \not\Rightarrow \phi_i(f) \neq 0$$

Question 1:

How should (can) we choose the function v ?

- Marginal Data Distribution

$$v_f(S, x) = \mathbb{E}[f(x_S, \tilde{X}_{\bar{S}})]$$

Explicand	Present Features	data				
$x_1 x_2 x_3 x_4$	S	2	9	9	1	
x^e	$\{1, 2\}$	4	8	1	4	
		3	7	6	5	
		2	9	4	7	
		4	9	1	8	
$v(S) = f(x), x =$	2 9					$(x_S, \tilde{X}_{\bar{S}}) \sim$
						2 9 \times
						9 1
						1 4
						6 5
						4 7
						1 8

Question 1:

How should (can) we choose the function v ?

- Marginal Data Distribution (interventional expectation)

$$v_f(S, x) = \mathbb{E}[f(x_S, \tilde{X}_{\bar{S}})] = \mathbb{E}[f(X) | do(S)]$$

Explicand	Present Features	data																					
$x_1 x_2 x_3 x_4$	S	<table><tr><td>2</td><td>9</td><td>9</td><td>1</td></tr><tr><td>4</td><td>8</td><td>1</td><td>4</td></tr><tr><td>3</td><td>7</td><td>6</td><td>5</td></tr><tr><td>2</td><td>9</td><td>4</td><td>7</td></tr><tr><td>4</td><td>9</td><td>1</td><td>8</td></tr></table>	2	9	9	1	4	8	1	4	3	7	6	5	2	9	4	7	4	9	1	8	
2	9	9	1																				
4	8	1	4																				
3	7	6	5																				
2	9	4	7																				
4	9	1	8																				
x^e	$\{1, 2\}$		$(x_S, \tilde{X}_{\bar{S}}) \sim$																				
$v(S) = f(x), x =$			<table><tr><td>2</td><td>9</td><td>×</td><td><table><tr><td>9</td><td>1</td></tr><tr><td>1</td><td>4</td></tr><tr><td>6</td><td>5</td></tr><tr><td>4</td><td>7</td></tr><tr><td>1</td><td>8</td></tr></table></td></tr></table>	2	9	×	<table><tr><td>9</td><td>1</td></tr><tr><td>1</td><td>4</td></tr><tr><td>6</td><td>5</td></tr><tr><td>4</td><td>7</td></tr><tr><td>1</td><td>8</td></tr></table>	9	1	1	4	6	5	4	7	1	8						
2	9	×	<table><tr><td>9</td><td>1</td></tr><tr><td>1</td><td>4</td></tr><tr><td>6</td><td>5</td></tr><tr><td>4</td><td>7</td></tr><tr><td>1</td><td>8</td></tr></table>	9	1	1	4	6	5	4	7	1	8										
9	1																						
1	4																						
6	5																						
4	7																						
1	8																						

Question 1:

How should (can) we choose the function v ?

- Marginal Data Distribution (interventional expectation)

$$v_f(S, x) = \mathbb{E}[f(x_S, \tilde{X}_{\bar{S}})] = \mathbb{E}[f(X) | do(S)]$$

Explicand	Present Features	data	
$x_1 x_2 x_3 x_4$	S		
x^e	$\{1, 2\}$		
$v(S) = f(x), x =$			



- Easier than conditional
- “true to the model”
maintains Null axiom



- $(x_S, \tilde{X}_{\bar{S}}) \not\sim p_X$
- can hide correlations in the data

Question 1:

How should (can) we choose the function v ?

Example (mortality prediction)

[Chen et al, True to the model or true to the data? 2020]

- National Health and Nutrition Examination Survey (NHANES)

$X = (\text{Age}, \text{IR}, \text{WaistC}, \text{BP}, \text{BMI})$

$$f(x) = \beta^\top x, \text{ with } \beta_5 = 0$$

Question 1:

How should (can) we choose the function v ?

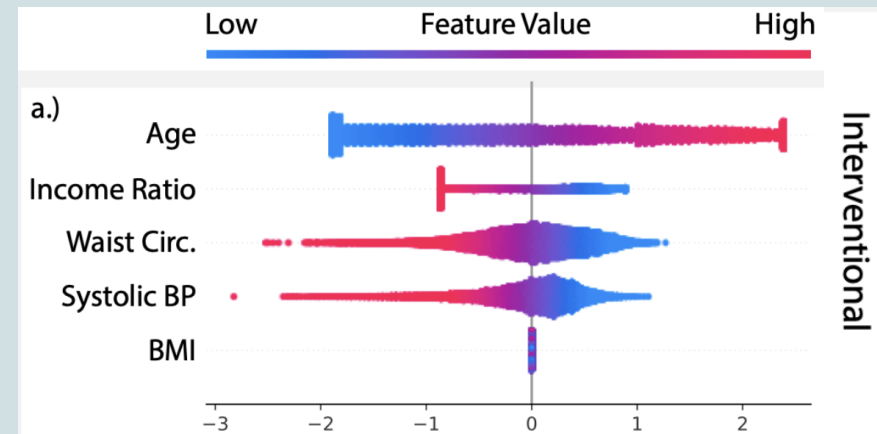
Example (mortality prediction)

[Chen et al, True to the model or true to the data? 2020]

- National Health and Nutrition Examination Survey (NHANES)

$X = (\text{Age}, \text{IR}, \text{WaistC}, \text{BP}, \text{BMI})$

$$f(x) = \beta^\top x, \text{ with } \beta_5 = 0$$



Question 1:

How should (can) we choose the function v ?

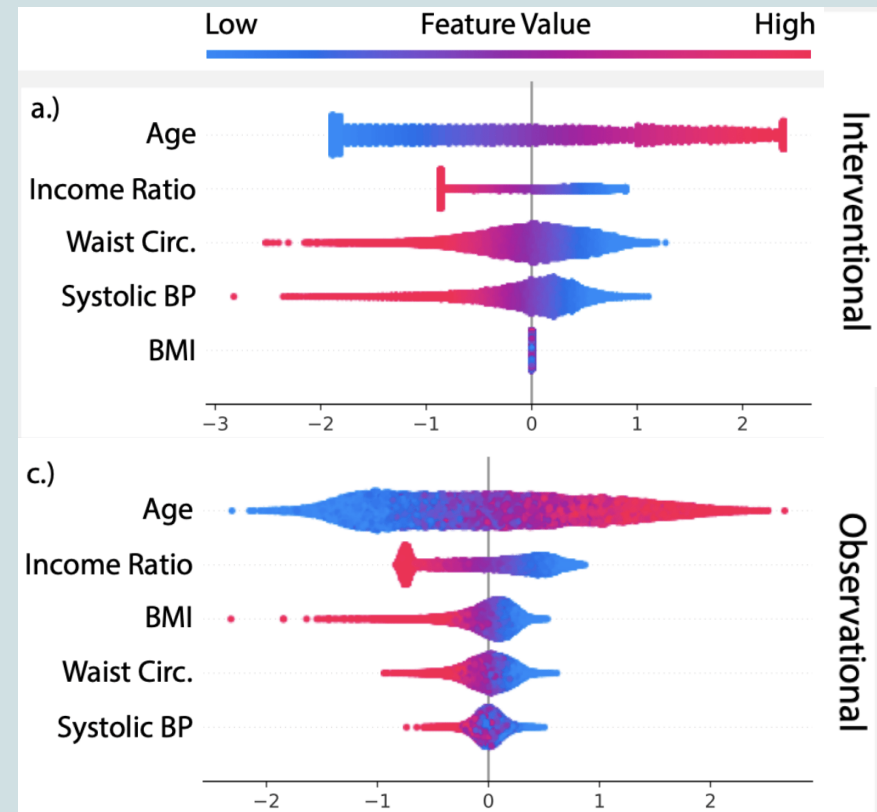
Example (mortality prediction)

[Chen et al, True to the model or true to the data? 2020]

- National Health and Nutrition Examination Survey (NHANES)

$X = (\text{Age}, \text{IR}, \text{WaistC}, \text{BP}, \text{BMI})$

$f(x) = \beta^\top x$, with $\beta_5 = 0$



Question 1:

How should (can) we choose the function v ?

Example (mortality prediction)

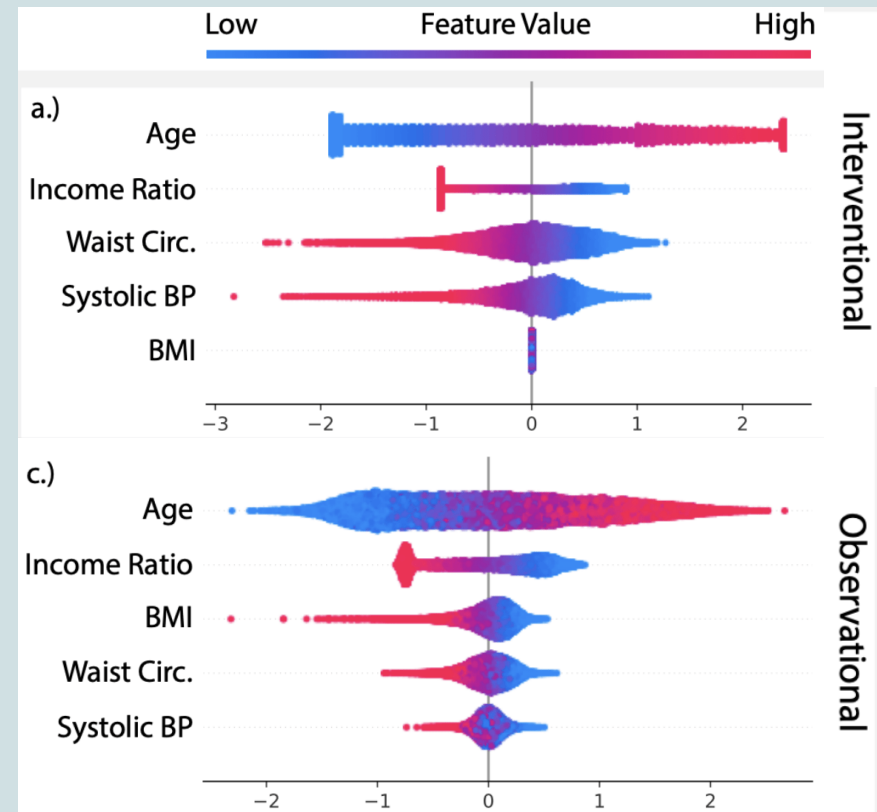
[Chen et al, True to the model or true to the data? 2020]

- National Health and Nutrition Examination Survey (NHANES)

$X = (\text{Age}, \text{IR}, \text{WaistC}, \text{BP}, \text{BMI})$

$f(x) = \beta^\top x$, with $\beta_5 = 0$

- Auditing a predictor?
(e.g. loan recommendations?)
- Feature discovery / bias analysis?



Question 1:

How should (can) we choose the function v ?

- Linear model (approximation)

$$v_f(S, x) = \mathbb{E}[f(x_S, \tilde{X}_{\bar{S}})] \approx f(x_S, \mathbb{E}[\tilde{X}_{\bar{S}}])$$



- Easiest, popular in practice



- $(x_S, \tilde{X}_{\bar{S}}) \not\sim p_X$
except in linear models
(and feature independence)

Shapley Explanations for ML

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} w(S) [v(S \cup \{i\}) - v(S)]$$

inputs $X \in \mathcal{X} \subset \mathbb{R}^n$

predictor $f : \mathcal{X} \rightarrow \mathcal{Y}$

responses $Y \in \mathcal{Y} = [C]$

$$f(X) = \hat{Y} \approx Y$$

Question 1:

How should (can) we choose the function v ?

Question 2:

How can we (when) compute $\phi_i(v)$?

Question 3:

What do $\phi_i(v)$ say (and don't) about the problem?

Question 2:

How can we (when) compute $\phi_i(v)$?

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} w(S) [v(S \cup \{i\}) - v(S)] \quad \text{intractable.. } \mathcal{O}(2^n)$$

Question 2:

How can we (when) compute $\phi_i(v)$?

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} w(S) [v(S \cup \{i\}) - v(S)] \quad \text{intractable.. } \mathcal{O}(2^n)$$

- Monte Carlo Sampling

$$\phi_i(v) = \sum_{S' \sim w(S)} [v(S' \cup \{i\}) - v(S')]$$

Question 2:

How can we (when) compute $\phi_i(v)$?

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} w(S) [v(S \cup \{i\}) - v(S)] \quad \text{intractable.. } \mathcal{O}(2^n)$$

- Monte Carlo Sampling

$$\phi_i(v) = \sum_{S' \sim w(S)} [v(S' \cup \{i\}) - v(S')]$$

$$\phi_i(v) = \frac{1}{n!} \sum_{\pi \subseteq \Pi(n)} [v(\text{Pre}^i(\pi) \cup \{i\}) - v(\text{Pre}^i(\pi))]$$

Question 2:

How can we (when) compute $\phi_i(v)$?

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} w(S) [v(S \cup \{i\}) - v(S)] \quad \text{intractable.. } \mathcal{O}(2^n)$$

- Monte Carlo Sampling

$$\phi_i(v) = \sum_{S' \sim w(S)} [v(S' \cup \{i\}) - v(S')]$$

$$\phi_i(v) = \frac{1}{n!} \sum_{\pi \subseteq \Pi(n)} [v(\text{Pre}^i(\pi) \cup \{i\}) - v(\text{Pre}^i(\pi))]$$

- Weighted Least Squares (*kernelSHAP*)

$$\phi_i(v) = \arg \min_{\beta \in \mathbb{R}^{n+1}} \sum_{S \subseteq [n]} \omega(S) (v(S) - \beta_0 - \sum_{j \in S} \beta_j)^2$$

Question 2:

How can we (when) compute $\phi_i(v)$?

- Weighted Least Squares (*kernelSHAP*)

$$\phi_i(v) = \arg \min_{\beta \in \mathbb{R}^{n+1}} \sum_{S \subseteq [n]} \omega(S) (v(S) - \beta_0 - \sum_{j \in S} \beta_j)^2$$

Question 2:

How can we (when) compute $\phi_i(v)$?

- Weighted Least Squares (*kernelSHAP*)

$$\phi_i(v) = \arg \min_{\beta \in \mathbb{R}^{n+1}} \sum_{S \subseteq [n]} \omega(S) (v(S) - \beta_0 - \sum_{j \in S} \beta_j)^2$$

- Weighted Least Squares, *amortized* (*FastSHAP*)

$$\Phi_{\text{fastShap}} = \arg \min_{\phi_\theta: \mathcal{X} \rightarrow \mathbb{R}^n} \mathbb{E}_X \sum_{y \in [k]} \sum_{S \subseteq [n]} \omega(S) (v(S, y) - \sum_{j \in S} \phi_\theta(X, y)_j)^2$$

... and stochastic versions [Covert et al, 2024]

Question 2:

How can we (when) compute $\phi_i(v)$ **if we know more?**

(about the model)

Question 2:

How can we (when) compute $\phi_i(v)$ **if we know more?**

(about the model)

- **Linear models** $f(x) = \beta^\top x$

Closed-form expressions (*for marginal distributions and baselines*)

$$\phi_i(f, x) = \beta_i(x_i - \mu_i)$$

(*also for conditional if assuming Gaussian features*)

Question 2:

How can we (when) compute $\phi_i(v)$ if we know more?

(about the model)

- **Linear models** $f(x) = \beta^\top x$

Closed-form expressions (*for marginal distributions and baselines*)

$$\phi_i(f, x) = \beta_i(x_i - \mu_i)$$

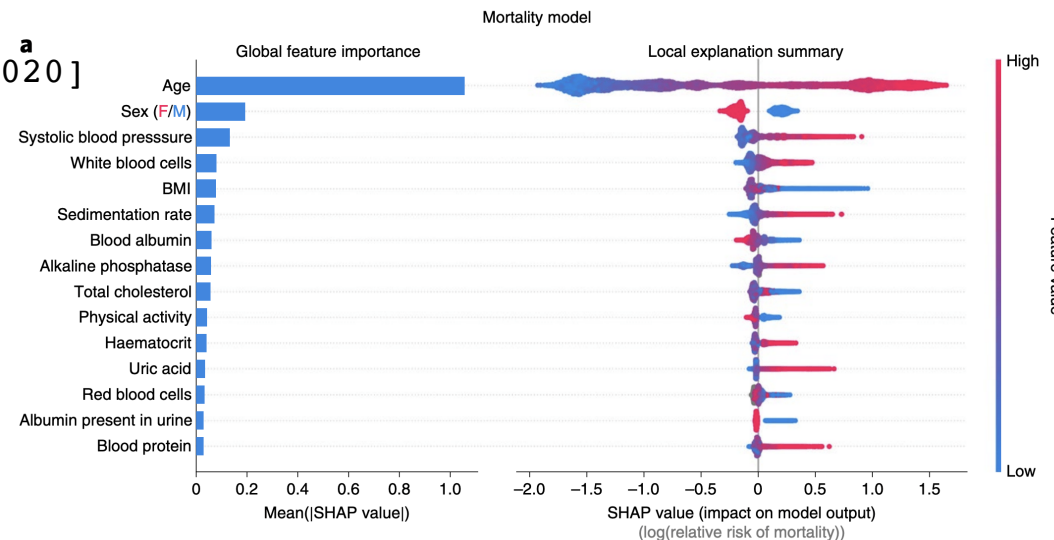
(*also for conditional if assuming Gaussian features*)

- **Tree models** [Lundberg et al, 2020]

Polynomial time algorithm (exact)

(TreeSHAP) for $\phi_i(f)$

$$\mathcal{O}(N_{\text{trees}} N_{\text{leaves}} \text{Depth}^2)$$

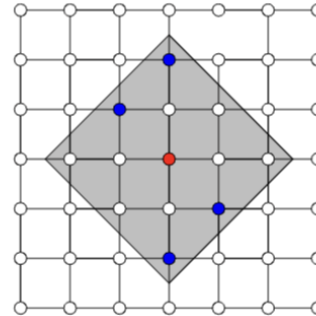
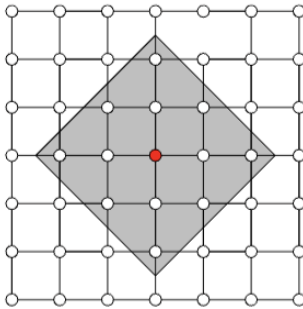


Question 2:

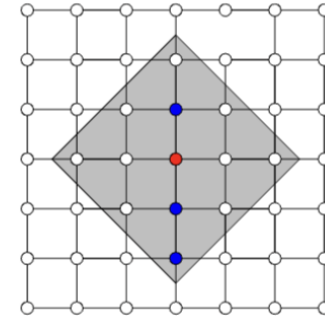
How can we (when) compute $\phi_i(v)$ if we know more?

(about the model)

- **Local models** [Chen et al, 2019]



L-Shap



C-Shap

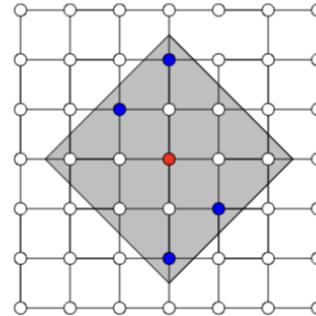
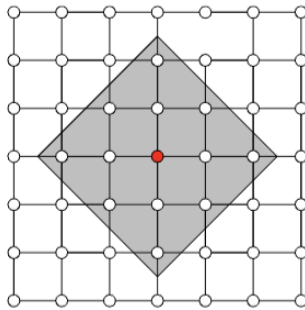
Observation: Restrict computation of $\phi_i(f)$ to local areas of influence given by a graph structure

Question 2:

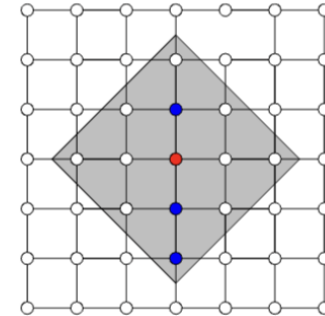
How can we (when) compute $\phi_i(v)$ if we know more?

(about the model)

- **Local models** [Chen et al, 2019]



L-Shap



C-Shap

Observation: Restrict computation of $\phi_i(f)$ to local areas of influence given by a graph structure

$$\hat{\phi}_i^k(v) = \frac{1}{|\mathcal{N}_k(i)|} \sum_{S \subseteq \mathcal{N}_k(i) \setminus i} w(S) [v(S \cup \{i\}) - v(S)]$$

\Rightarrow complexity $\mathcal{O}(2^k n)$

Question 2:

How can we (when) compute $\phi_i(v)$ if we know more?

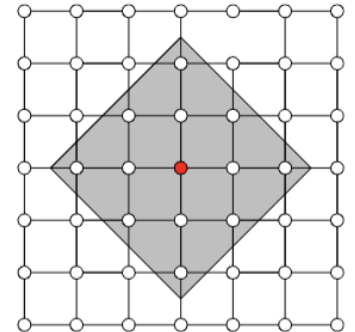
(about the model)

- **Local models** [Chen et al, 2019]

Observation: Restrict computation of $\phi_i(f)$ to local areas of influence given by a graph structure

$$\hat{\phi}_i^k(v) = \frac{1}{|\mathcal{N}_k(i)|} \sum_{S \subseteq \mathcal{N}_k(i) \setminus i} w(S) [v(S \cup \{i\}) - v(S)]$$

\Rightarrow complexity $\mathcal{O}(2^k n)$



Question 2:

How can we (when) compute $\phi_i(v)$ if we know more?

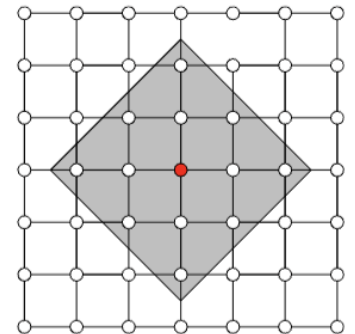
(about the model)

- **Local models** [Chen et al, 2019]

Observation: Restrict computation of $\phi_i(f)$ to local areas of influence given by a graph structure

$$\hat{\phi}_i^k(v) = \frac{1}{|\mathcal{N}_k(i)|} \sum_{S \subseteq \mathcal{N}_k(i) \setminus i} w(S) [v(S \cup \{i\}) - v(S)]$$

\Rightarrow complexity $\mathcal{O}(2^k n)$



Correct approximations (*informal statement*)

Let $S \subset \mathcal{N}_k(i)$. If, for any $T \subseteq S \setminus \{i\}$, $(X_i \perp\!\!\!\perp X_{[n] \setminus S} | X_T)$ and $(X_i \perp\!\!\!\perp X_{[n] \setminus S} | X_T, Y)$

Then $\hat{\phi}_i^k(v) = \phi_i(v)$

(and approximately bounded otherwise, controlled)

Question 2:

How can we (when) compute $\phi_i(v)$ if we know more?

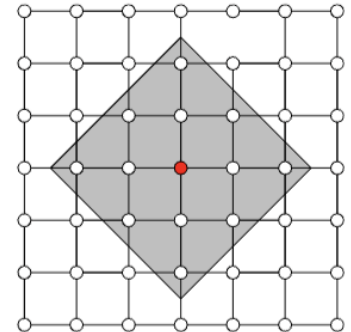
(about the model)

- **Local models** [Chen et al, 2019]

Observation: Restrict computation of $\phi_i(f)$ to local areas of influence given by a graph structure

$$\hat{\phi}_i^k(v) = \frac{1}{|\mathcal{N}_k(i)|} \sum_{S \subseteq \mathcal{N}_k(i) \setminus i} w(S) [v(S \cup \{i\}) - v(S)]$$

\Rightarrow complexity $\mathcal{O}(2^k n)$



Method	Explanation									
Shapley	It	is	not	heartwarming	or	entertaining	.	It	just	sucks
C-Shapley	It	is	not	heartwarming	or	entertaining	.	It	just	sucks
L-Shapley	It	is	not	heartwarming	or	entertaining	.	It	just	sucks
KernelSHAP	It	is	not	heartwarming	or	entertaining	.	It	just	sucks
SampleShapley	It	is	not	heartwarming	or	entertaining	.	It	just	sucks

Question 2:

How can we (when) compute $\phi_i(v)$ **if we know more?**

(about the model)

- **Hierarchical Shapley (h-Shap)** [Teneggi et al, 2022]

$$\text{Observation: } f(x) = 1 \Leftrightarrow \exists i : f(x_i, \tilde{X}_{-i}) = 1 \quad (\text{A1})$$

Question 2:

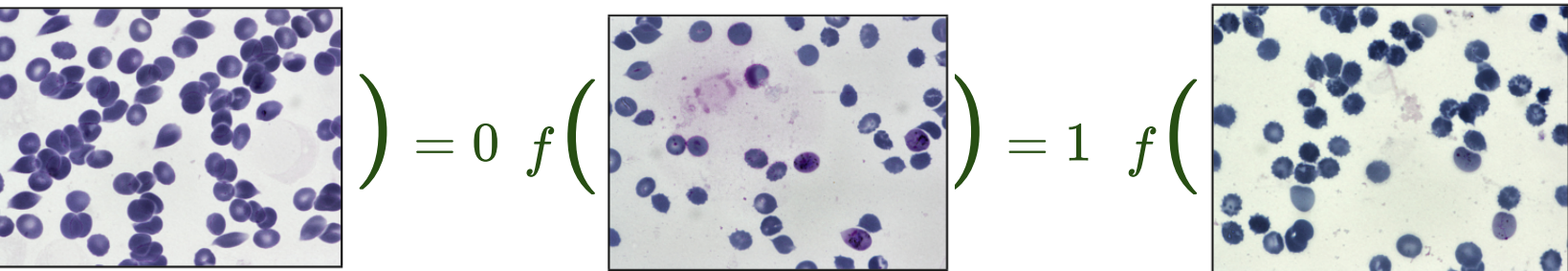
How can we (when) compute $\phi_i(v)$ if we know more?

(about the model)

- **Hierarchical Shapley (h-Shap)** [Teneggi et al, 2022]

$$\text{Observation: } f(x) = 1 \Leftrightarrow \exists i : f(x_i, \tilde{X}_{-i}) = 1 \quad (\text{A1})$$

Example: $f(x) = 1$ if x contains a sick cell

$$f\left(\text{img}_1\right) = 0 \quad f\left(\text{img}_2\right) = 1 \quad f\left(\text{img}_3\right) = 0$$


Question 2:

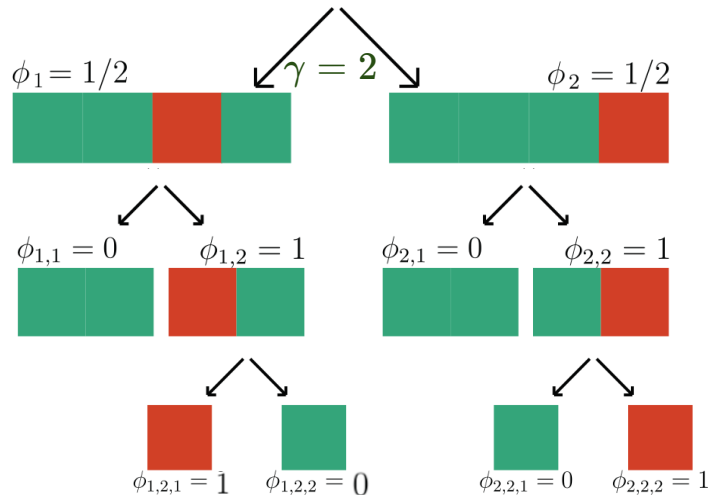
How can we (when) compute $\phi_i(v)$ if we know more?

(about the model)

- **Hierarchical Shapley (h-Shap)** [Teneggi et al, 2022]

$$\text{Observation: } f(x) = 1 \Leftrightarrow \exists i : f(x_i, \tilde{X}_{-i}) = 1 \quad (\text{A1})$$

$$f(\text{■} \text{■} \text{■} \text{■} \text{■} \text{■} \text{■} \text{■}) = 1$$



Question 2:

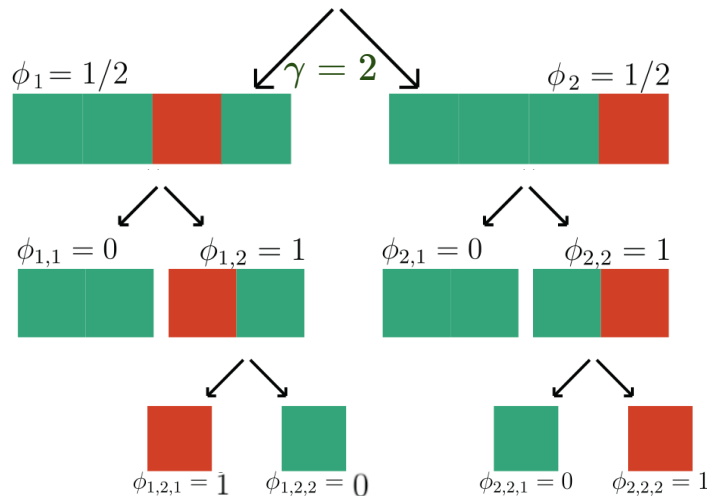
How can we (when) compute $\phi_i(v)$ if we know more?

(about the model)

- **Hierarchical Shapley (h-Shap)** [Teneggi et al, 2022]

$$\text{Observation: } f(x) = 1 \Leftrightarrow \exists i : f(x_i, \tilde{X}_{-i}) = 1 \quad (\text{A1})$$

$$f(\text{■} \text{■} \text{■} \text{■} \text{■} \text{■} \text{■}) = 1$$



1. Complexity $\mathcal{O}(2^\gamma k \log n)$

2. Correct approximation (informal)

- Under A1, $\phi_i^{\text{h-Shap}}(f) = \phi_i(f)$
- Bounded approximation as deviating from A1

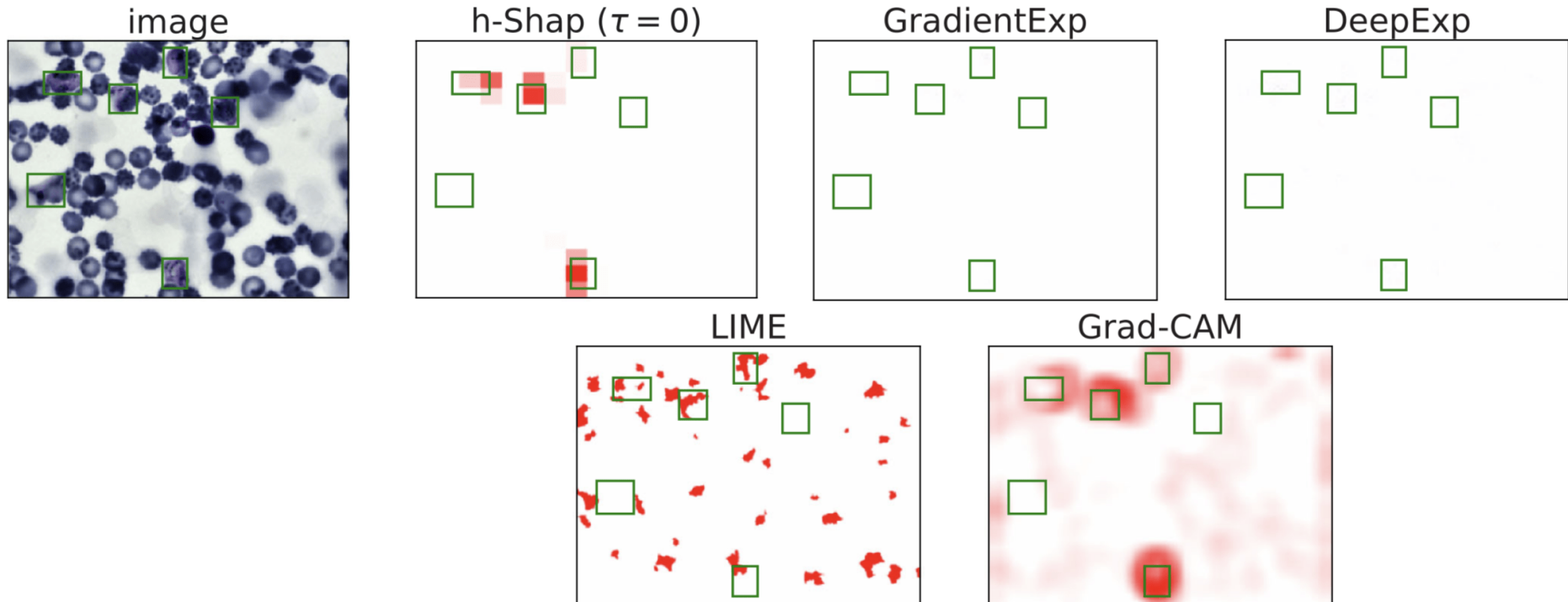
Question 2:

How can we (when) compute $\phi_i(v)$ if we know more?

(about the model)

- **Hierarchical Shapley (h-Shap)** [Teneggi et al, 2022]

$$\text{Observation: } f(x) = 1 \Leftrightarrow \exists i : f(x_i, \tilde{X}_{-i}) = 1 \quad (\text{A1})$$



Question 2:

How can we (when) compute $\phi_i(v)$ **if we know more?**

(about the model)

- **Shapley Approximations for Deep Models**

DeepLift (Shrikumar et al, 2017): biased estimation of **baseline** Shap

DeepShap (Chen et al, 2021): biased estimation of **marginal** Shap

DASP (Ancona et al, 2019) Uncertainty propagation for **baseline (zero)** Shap
assuming Gaussianity and independence of features

Shapnets [Wang et al, 2020]: Computation for **small-width networks**

Transformers (ViTs) [Covert et al, 2023]: leveraging **attention** to fine-tune a surrogate model for Shap estimation

... not an exhaustive list!

Shapley Explanations for ML

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} w(S) [v(S \cup \{i\}) - v(S)]$$

inputs $X \in \mathcal{X} \subset \mathbb{R}^n$

predictor $f : \mathcal{X} \rightarrow \mathcal{Y}$

responses $Y \in \mathcal{Y} = [C]$

$$f(X) = \hat{Y} \approx Y$$

Question 1:

How should (can) we choose the function v ?

Question 2:

How can we (when) compute $\phi_i(v)$?

Question 3:

What do $\phi_i(v)$ say (and don't) about the problem?

Question 3:

What do $\phi_i(v)$ say (and don't) about the problem?

Question 3:

What do $\phi_i(v)$ say (and don't) about the problem?

Interpretability as Conditional Independence

- Explaining uncertainty via Shapley Values [Watson et al, 2023]

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} w(S) [v_{\text{KL}}(S \cup \{i\}) - v_{\text{KL}}(S)]$$

$$\text{with } v_{\text{KL}}(S, x) = -D_{\text{KL}}(p_{Y|x} \parallel p_{Y|x_s})$$

Question 3:

What do $\phi_i(v)$ say (and don't) about the problem?

Interpretability as Conditional Independence

- Explaining uncertainty via Shapley Values [Watson et al, 2023]

$$\phi_i(v) = \sum_{S \subseteq [n] \setminus \{i\}} w(S) [v_{\text{KL}}(S \cup \{i\}) - v_{\text{KL}}(S)]$$

$$\text{with } v_{\text{KL}}(S, x) = -D_{\text{KL}}(p_{Y|x} \parallel p_{Y|x_s})$$

Theorem (informal)

$$Y \perp\!\!\!\perp X_j \mid X_s = x_s \Rightarrow v_{\text{KL}}(S \cup \{i\}, x) - v_{\text{KL}}(S, x) = 0$$

Question 3:

What do $\phi_i(v)$ say (and don't) about the problem?

Interpretability as Conditional Independence

Question 3:

What do $\phi_i(v)$ say (and don't) about the problem?

Interpretability as Conditional Independence

- SHAP-XRT: Shapley meets Hypothesis Testing [Teneggi et al, 2023]

$$H_{i,S}^0 : (f(x_{S \cup \{i\}}, \tilde{X}_{\overline{S \cup \{i\}}})) \stackrel{d}{=} (f(x_S, \tilde{X}_{\bar{S}}))$$

$\hat{p}_{i,S} \leftarrow \text{XRT} : \text{eXplanation Randomization Test, via access to } X_{\bar{S}} \sim p_{X_{\bar{S}}|x_s}$

Question 3:

What do $\phi_i(v)$ say (and don't) about the problem?

Interpretability as Conditional Independence

- SHAP-XRT: Shapley meets Hypothesis Testing [Teneggi et al, 2023]

$$H_{i,S}^0 : (f(x_{S \cup \{i\}}, \tilde{X}_{\overline{S \cup \{i\}}})) \stackrel{d}{=} (f(x_S, \tilde{X}_{\bar{S}}))$$

$\hat{p}_{i,S} \leftarrow \text{XRT} : \text{eXplanation Randomization Test, via access to } X_{\bar{S}} \sim p_{X_{\bar{S}}|x_S}$

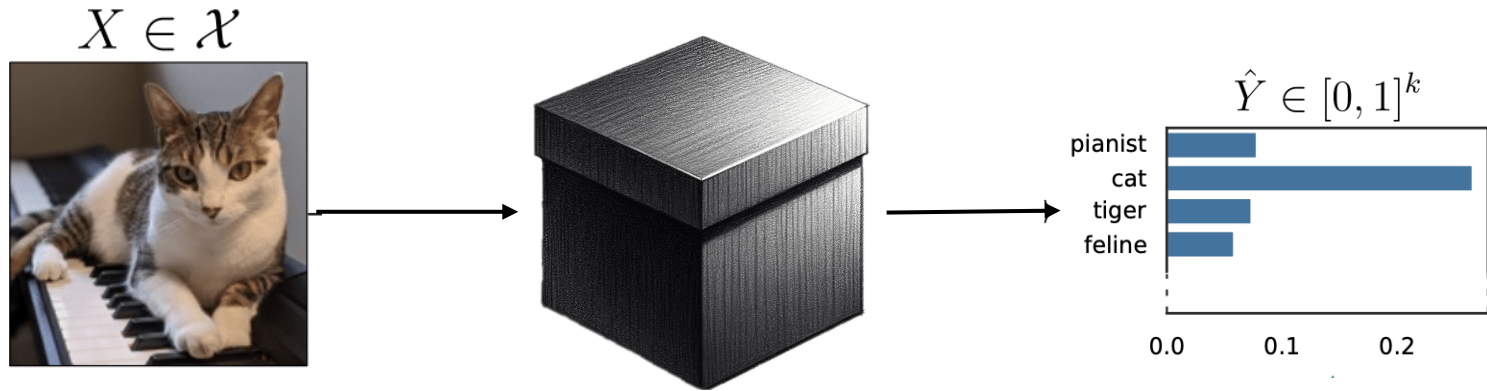
Theorem (informal)

For $f : \mathcal{X} \rightarrow [0, 1]$, $\mathbb{E}[\hat{p}_{i,S}] \leq 1 - \mathbb{E}[v(S \cup i) - v(S)]$

Thus, large $\mathbb{E}[v(S \cup i) - v(S)] \Rightarrow \text{reject } H_{i,S}^0$

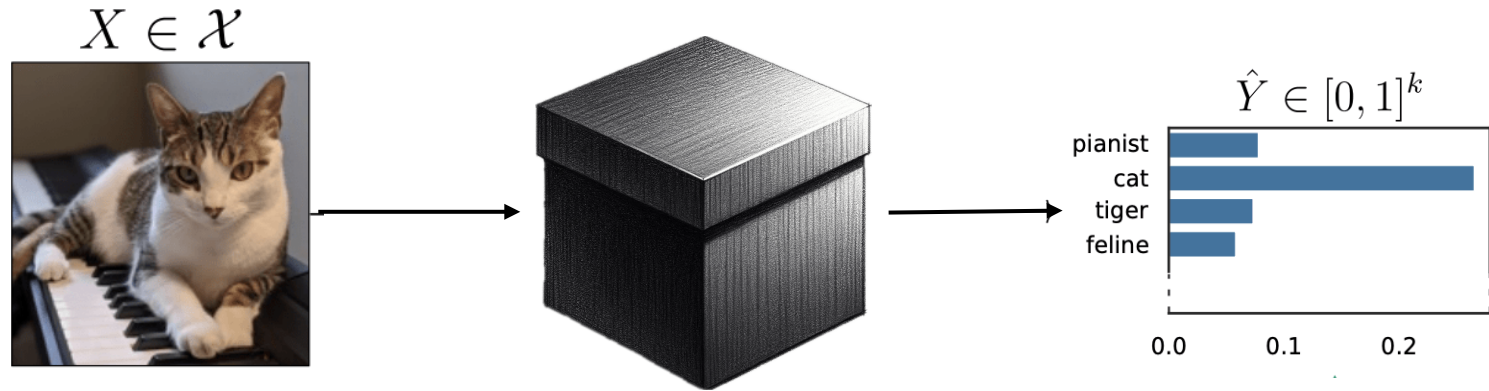
Last Question 3': Should we just focus on CIT?

[Teneggi et al, *Testing semantic importance via betting*, 2024]



Last Question 3': Should we just focus on CIT?

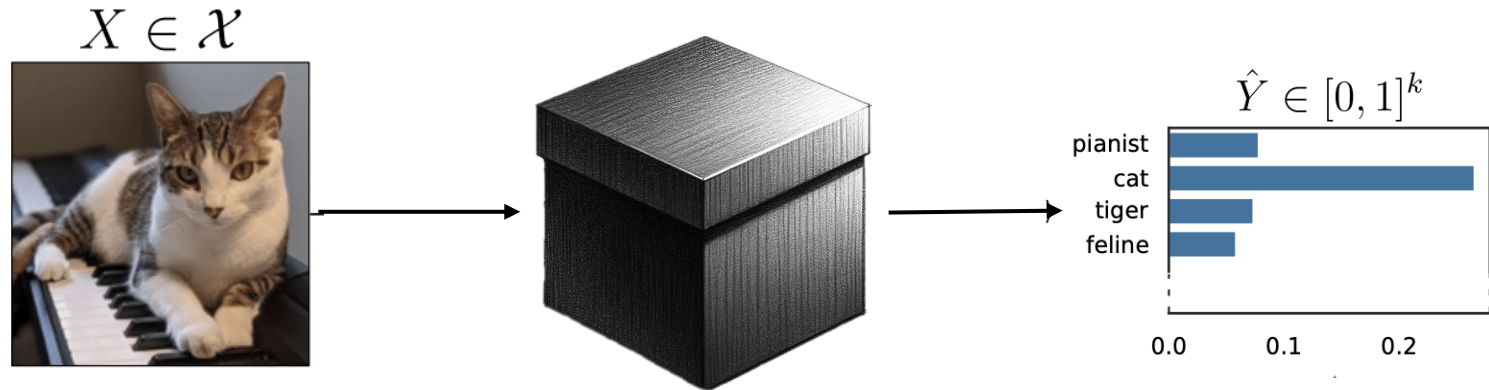
[Teneggi et al, *Testing semantic importance via betting*, 2024]



Is the **piano** important for $\hat{Y} = \text{cat}$, given that there is a **cute mammal** in the image?

Last Question 3' : Should we just focus on CIT?

[Teneggi et al, *Testing semantic importance via betting*, 2024]



Is the **piano** important for $\hat{Y} = \text{cat}$, given that there is a **cute mammal** in the image?

Local Conditional Importance

$$H_0^{j,S} : f(\tilde{H}_{S \cup \{j\}}) \stackrel{d}{=} f(\tilde{H}_S)$$

features with concepts $S \cup \{i\}$

features with concepts S

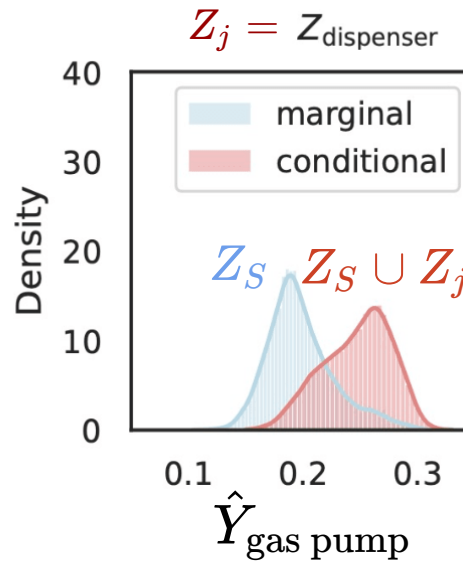
Last Question 3': Should we just focus on CIT?

[Teneggi et al, *Testing semantic importance via betting*, 2024]

Local Conditional Importance

$$H_0^{j,S} : f(\tilde{H}_{S \cup \{j\}}) \stackrel{d}{=} f(\tilde{H}_S)$$

gas pump



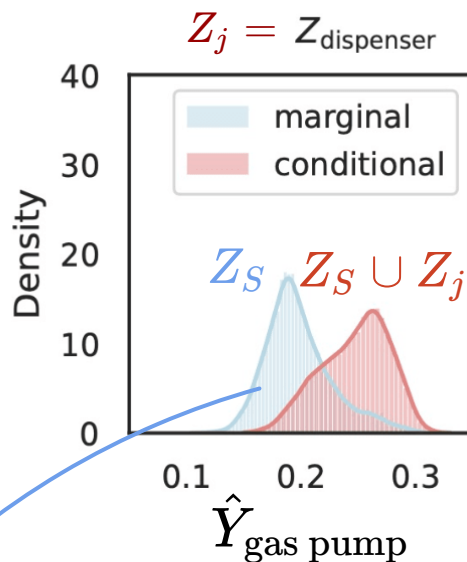
Last Question 3': Should we just focus on CIT?

[Teneggi et al, *Testing semantic importance via betting*, 2024]

Local Conditional Importance

$$H_0^{j,S} : f(\tilde{H}_{S \cup \{j\}}) \stackrel{d}{=} f(\tilde{H}_S)$$

gas pump



$$\tilde{Z}_S = [\underbrace{z_{\text{text}}, z_{\text{old}}}_S, Z_{\text{dispenser}}, Z_{\text{trumpet}}, Z_{\text{fire}}, \dots]$$

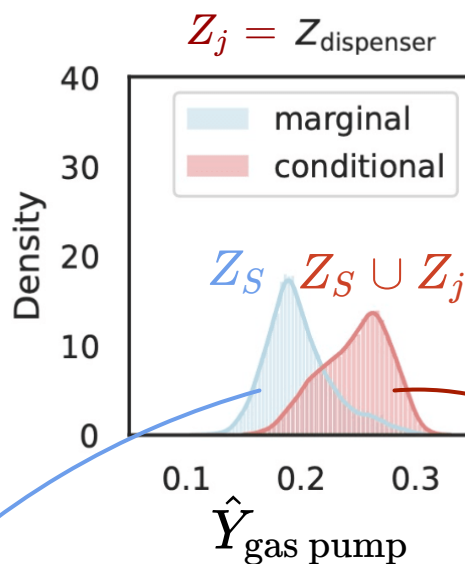
Last Question 3': Should we just focus on CIT?

[Teneggi et al, *Testing semantic importance via betting*, 2024]

Local Conditional Importance

$$H_0^{j,S} : f(\tilde{H}_{S \cup \{j\}}) \stackrel{d}{=} f(\tilde{H}_S)$$

gas pump



$$\tilde{Z}_S = [\underbrace{z_{\text{text}}, z_{\text{old}}, Z_{\text{dispenser}}, Z_{\text{trumpet}}, Z_{\text{fire}}, \dots}_S]$$

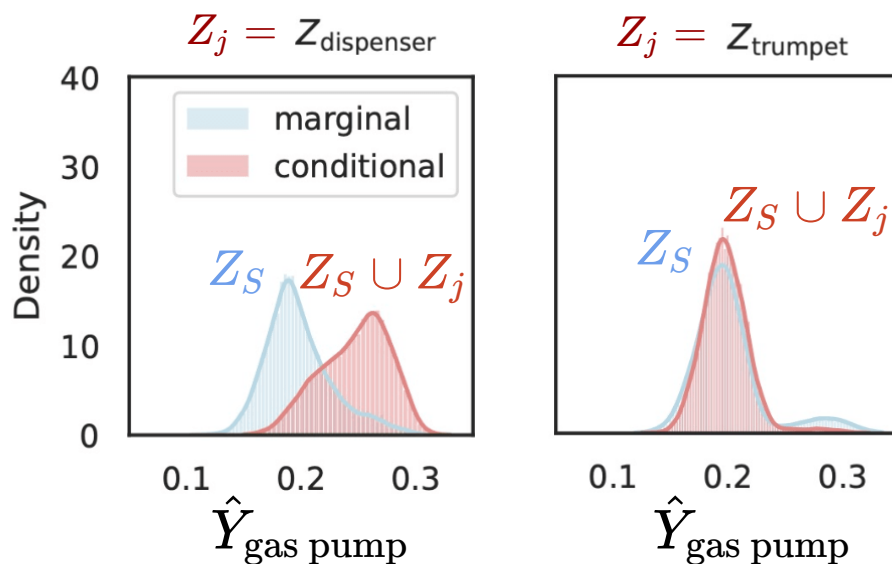
$$\tilde{Z}_{S \cup j} = [\underbrace{z_{\text{text}}, z_{\text{old}}, z_{\text{dispenser}}, Z_{\text{trumpet}}, Z_{\text{Fire}}, \dots}_S, \underbrace{\phantom{z_{\text{text}}, z_{\text{old}}, z_{\text{dispenser}}, Z_{\text{trumpet}}, Z_{\text{Fire}}, \dots}}_j]$$

Last Question 3': Should we just focus on CIT?

[Teneggi et al, *Testing semantic importance via betting*, 2024]

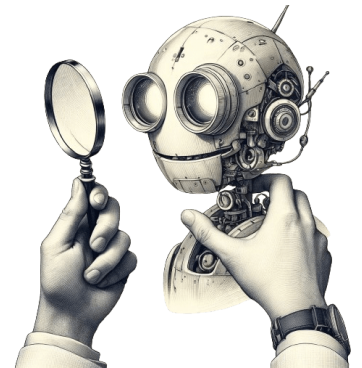
Local Conditional Importance

$$H_0^{j,S} : f(\tilde{H}_{S \cup \{j\}}) \stackrel{d}{=} f(\tilde{H}_S)$$



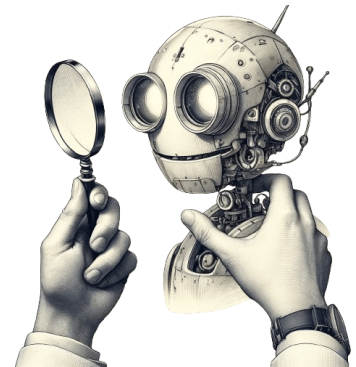
$$\tilde{Z}_S = \underbrace{[z_{\text{text}}, z_{\text{old}}, Z_{\text{dispenser}}, Z_{\text{trumpet}}, Z_{\text{fire}}, \dots]}_S \quad \tilde{Z}_{S \cup j} = \underbrace{[z_{\text{text}}, z_{\text{old}}, Z_{\text{dispenser}}, z_{\text{trumpet}}, Z_{\text{Fire}}, \dots]}_S \underbrace{\quad}_j$$

Conclusions



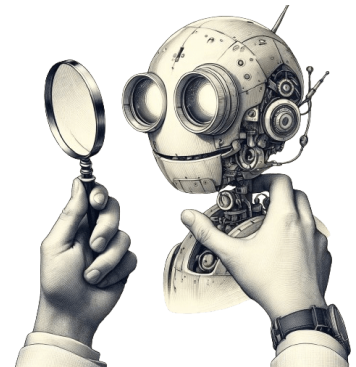
Conclusions

- Shapley Values are one of the **most popular** wrapper-explanation methods



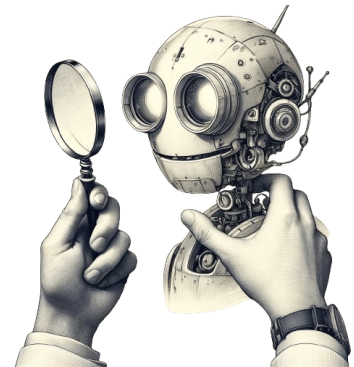
Conclusions

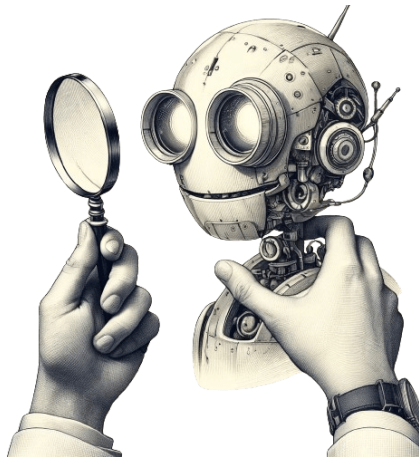
- Shapley Values are one of the **most popular** wrapper-explanation methods
- Requires **care** when choosing what **distributions** to sample from, dependent on the setting
"true to the model" vs "true to the data"



Conclusions

- Shapley Values are one of the **most popular** wrapper-explanation methods
- Requires **care** when choosing what **distributions** to sample from, dependent on the setting
"true to the model" vs "true to the data"
- While proposed in a different context, they can be used to test for **specific statistical** claims





Foundations of Interpretable AI

PART I: Motivation and Post-hoc Methods (9:00 - 9:45 am) Aditya Chattopadhyay (Amazon)

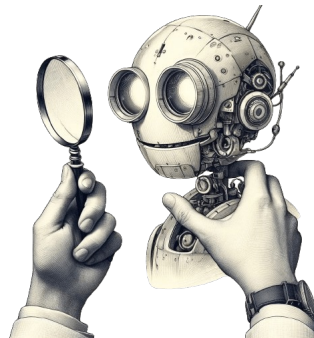
PART II: Shapley Values Based Methods (9:45 - 10:30 am) Jeremias Sulam (Johns Hopkins)

☕ Coffee break ☕ (10:30 - 11:00 am)

PART III: Interpretable by Design Methods (11:00 - 11:45 am) René Vidal (Penn)

Foundations of Interpretable AI

Tutorial @ **CVPR** *Nashville* JUNE 11-15, 2025



PART I: Motivation and Post-hoc Methods

(9:00 – 9:45 am) Aditya Chattopadhyay (Amazon)

PART II: Shapley Value Based Methods

(9:45 – 10:30 am) Jeremias Sulam (Johns Hopkins)

☕ Coffee break ☕

(10:30 – 11 am)

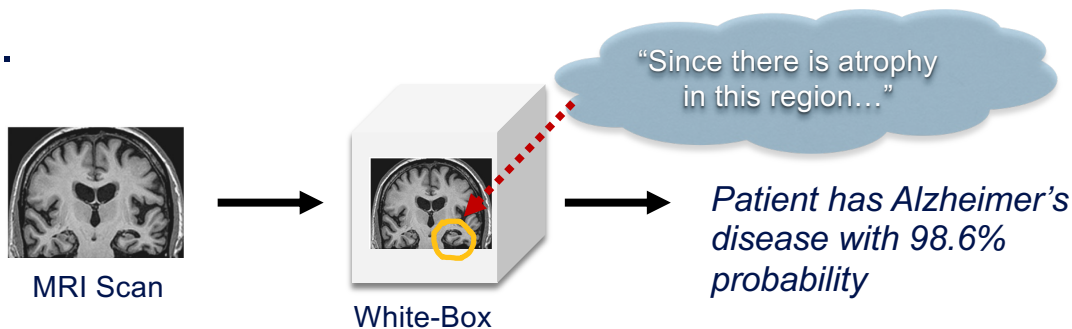
PART III: Interpretable by Design Methods

(11 – 11:45 am)

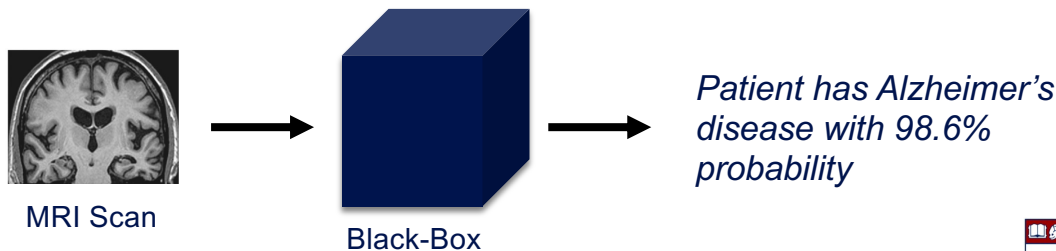
René Vidal (Penn)

Interpretability Crisis

- As deep learning is widely used in safety critical applications, there is a need for developing **trustworthy and interpretable models**.
- Ideally we desire...



- But in reality



Main Trend: Post-hoc Explanations

- Most method interpret black-box models **post-hoc** using **importance scores** based on the **sensitivity of the model output to the input features**:

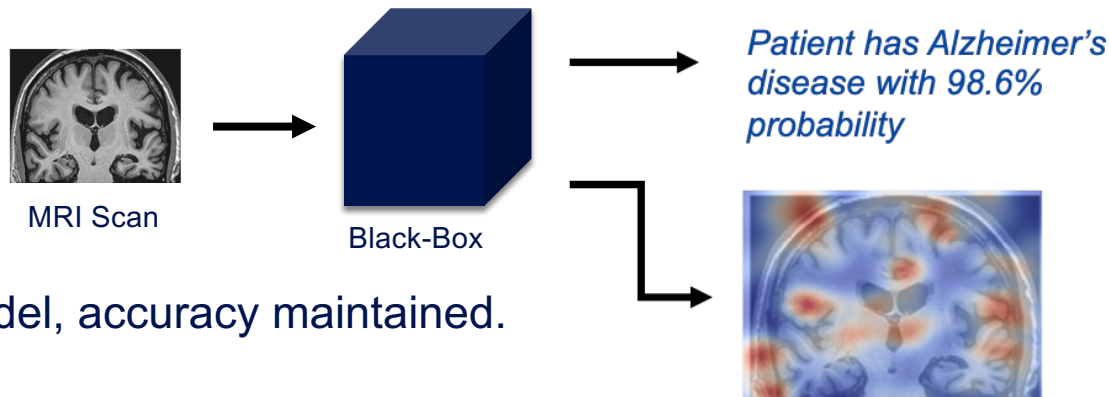
- LIME [1]
- Grad-CAM [2]
- SHAP [3]

- The Good:**

- No need to retrain model, accuracy maintained.

- The Bad:**

- Explanations are unreliable; not faithful to the model it tries to explain [4].
- Feature importance scores might not be interpretable to end-users [5].



[1] Ribeiro, Singh, Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. KDD, 2016.

[2] Selvaraju, Cogswell, Das, Vedantam, Parikh, Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. ICCV 2017.

[3] Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. NIPS, pp 4765–4774, 2017.

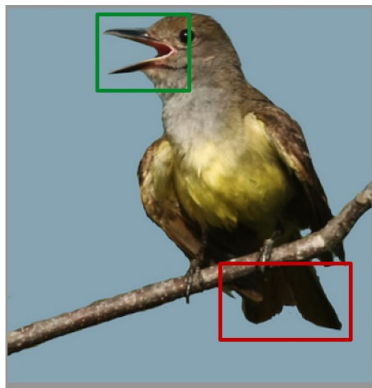
[4] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. Sanity checks for saliency maps. NeurIPS, 2018

[5] Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 2019.

Need for Interpretable-by-Design Models

- **Explanations** are **user/task/domain dependent** and best described in terms of words/attributes/facts that support the **decision's reasoning**.
- We can capture this via a **user/task/domain dependent query set Q** .

(a) **Task:** bird classification
Queries: parts, attributes



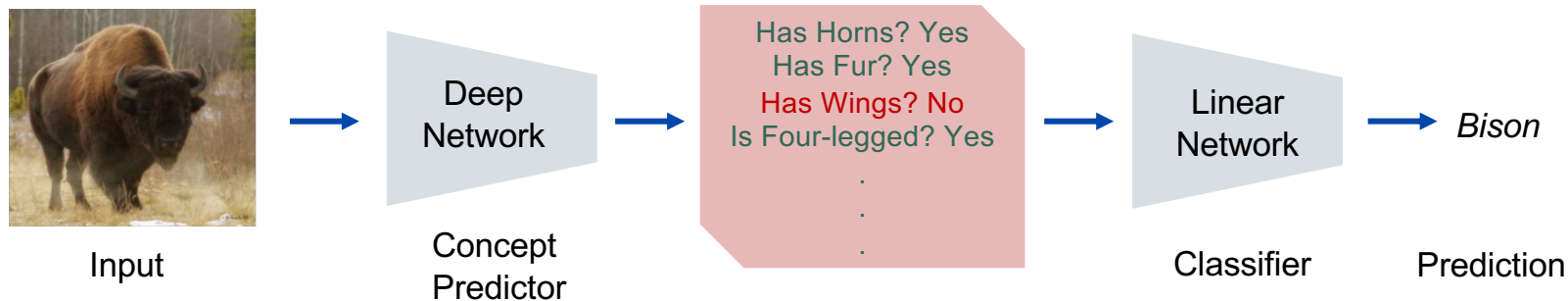
(b) **Task:** scene interpretation
Queries: objects, relationships



(c) **Task:** medical diagnosis
Queries: symptoms

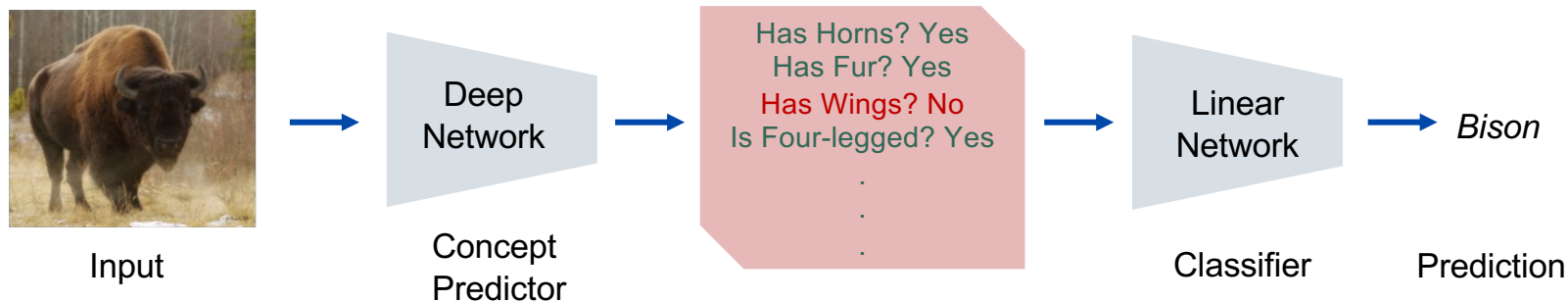
0. Ear pain
1. Sore throat
2. Fever
3. Cough
4. Nasal congestion
5. Allergic reaction
6. Shortness of breath
7. Painful sinuses

Concept Bottleneck Models (CMBs)



- Concept Bottleneck Models (CBMs) [1].
 - **Specify a query set**: define a set of task-relevant concepts Q .
 - **Answer queries**: train deep network to predict concepts from Q in image x .
 - **Make prediction**: train linear classifier on predicted concepts.
- **Explain prediction via weights** of linear layer for different concepts.

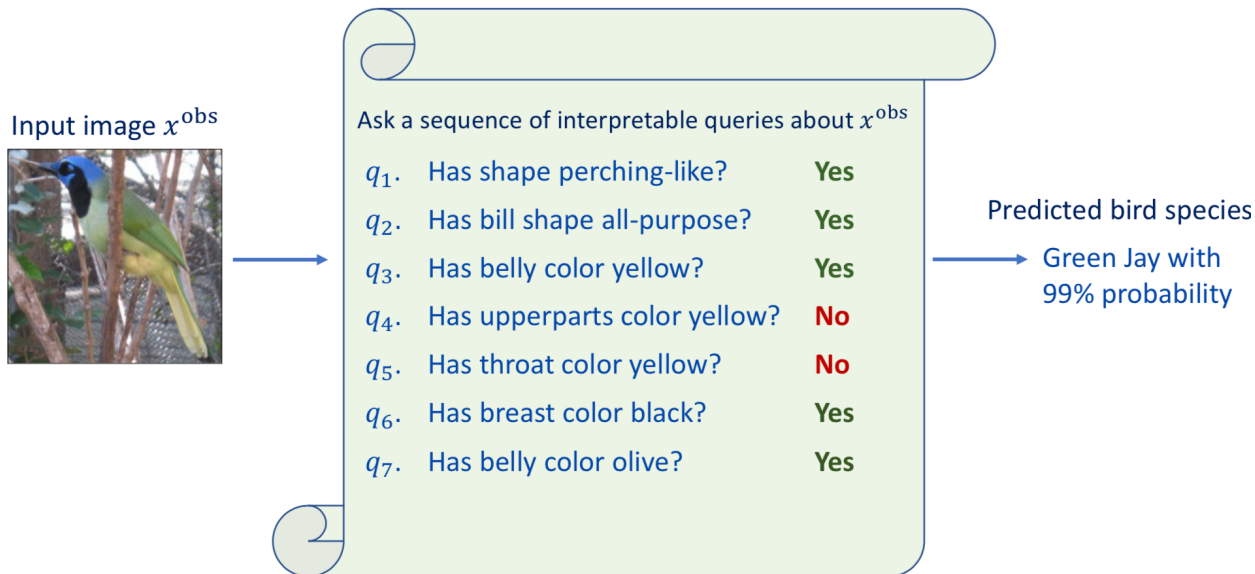
Are Concept Bottleneck Models Enough?



- **Limited expressivity:** linear classification layer limits expressivity of CBMs when “concept answers \rightarrow class prediction” map is non-linear.
- **Limited interpretability:** explanations in terms of coefficients of linear weights not always desirable to end-users, especially non-AI experts.
- **Limited flexibility:** same explanations for all inputs in the same class.

Information Pursuit Framework

- **Information Pursuit:** interpretable-by-design framework based on:
 - Selecting the **smallest number of queries** that are **sufficient** for prediction.
 - Making a prediction based only on the chain of query-answer pairs.



Ingredients Needed to Implement this Framework

- **Q1: How do we define the set of queries?**
- **Q2: Given an input and a query, how do we answer the query?**
- **Q3: How do we select queries that form the explanation?**

[1] Chattopadhyay, Slocum, Haeffele, Vidal, Geman. Interpretable by design: Learning predictors by composing interpretable queries. TPAMI 2022.

[2] Chattopadhyay, Chan, Haeffele, Geman, Vidal. Variational Information Pursuit for Interpretable Predictions, ICLR 2023.

[3] Chattopadhyay, Pilgrim, Vidal. Information Maximization Perspective of Orthogonal Matching Pursuit with Applications to Explainable AI. NeurIPS 2023.

[4] Chattopadhyay, Chan, Vidal. Bootstrapping Variational Information Pursuit with Foundation Models for Interpretable Image Classification. ICLR 2024.

[5] Chattopadhyay, Haeffele, Vidal, Geman. Performance Bounds for Active Binary Testing with Information Maximization. ICML 2024.

Q1: How to define the set of queries?

Q1: How do we Define the Set of Queries?

- Defined by **domain experts** [1,2]

- Assume queries have similar **semantic resolution**.

- **CUB dataset**

- 200+ bird classes
- 300+ bird attributes

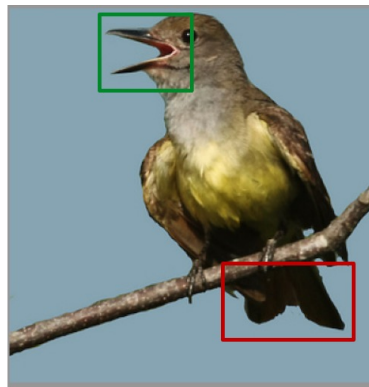
- **SymCAT-200 dataset**

- 200 disease diagnosis
- 326 patient symptoms

- **Challenge**

- **Annotating queries is very costly**

(a) **Task:** bird classification
Queries: parts, attributes



(c) **Task:** medical diagnosis
Queries: symptoms

- 0. Ear pain
- 1. Sore throat
- 2. Fever
- 3. Cough
- 4. Nasal congestion
- 5. Allergic reaction
- 6. Shortness of breath
- 7. Painful sinuses

[1] Koh, P. W., Nguyen, T., Tang, Y. S., Musmann, S., Pierson, E., Kim, B., & Liang, P. Concept bottleneck models. ICML, 2020.
[2] Chattopadhyay, Slocum, Haeffele, Vidal, Geman. Interpretable by design: Learning predictors by composing interpretable queries. TPAMI 2022.
[3] Oikarinen, T., Das, S., Nguyen, L. M., & Weng, T. W. (2023). Label-free concept bottleneck models. ICLR 2023
[4] Chattopadhyay, Chan, Vidal. Bootstrapping Variational Information Pursuit with Foundation Models for Interpretable Image Classification. ICLR 2024.

Q1: How do we Define the Set of Queries?

- Defined by **large language models** [3,4].
 - E.g., ask LLM for list of attributes of all relevant categories.

For every {class}:

PROMPT to GPT-3: List the useful visual attributes (and their values) of the bird image category '{class = Blue Jay}'.

RESPONSE:

1. Color: Blue, White, Black
 2. Size: Medium
 3. Shape: Long Tail, Crested Head
 4. Pattern: Spotted, Striped
- ⋮
- N. <attr>: <value>

Convert to queries:

$q_1 = \text{blue color}$
 $q_2 = \text{black color}$
 $q_3 = \text{medium size}$
 $q_4 = \text{spotted pattern}$
⋮
 $q_L = \text{value } \langle \text{attr} \rangle$

Union over all classes

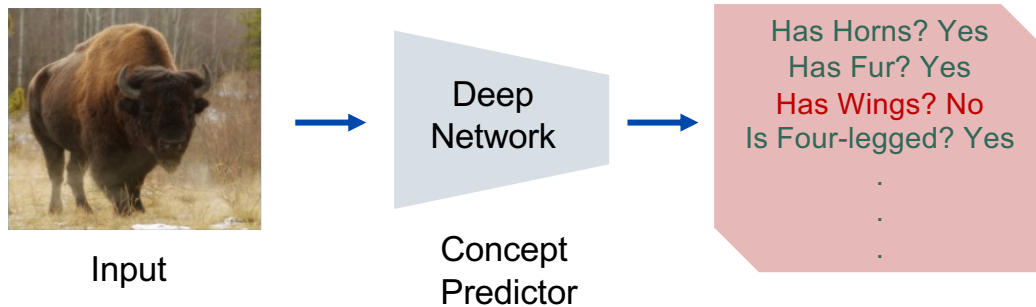
Query Set Q

[1] Koh, P. W., Nguyen, T., Tang, Y. S., Musmann, S., Pierson, E., Kim, B., & Liang, P. Concept bottleneck models. ICML, 2020.
[2] Chattopadhyay, Slocum, Haeffele, Vidal, Geman. Interpretable by design: Learning predictors by composing interpretable queries. TPAMI 2022.
[3] Oikarinen, T., Das, S., Nguyen, L. M., & Weng, T. W. (2023). Label-free concept bottleneck models. ICLR 2023
[4] Chattopadhyay, Chan, Vidal. Bootstrapping Variational Information Pursuit with Foundation Models for Interpretable Image Classification. ICLR 2024.

**Q2: Given an input and a query,
how do we answer the query?**

Q2: How do we Answer a Query for a given Input?

- **Train classifiers** on data annotated with query answers [1].

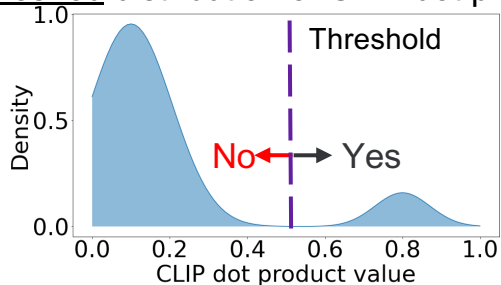


- **Challenge:** need **tons of data annotated with all concepts/attributes/facts**
=> few datasets have such detailed annotations.
- **Challenge:** cannot handle **new queries** that have not been annotated.

Q2: How do we Answer a Query for a given Input?

- Use **Vision Language Models (VLMs)** to answer queries
 - **Challenge:** State-of-the-art VLMs like **Llama [1]** and **BLIP [2]** are too slow to be used in an online manner.
 - **Challenge:** CLIP [3] is relatively light-weight, but **CLIP dot products** between query and image **are inadequate**: they are **not interpretable**.

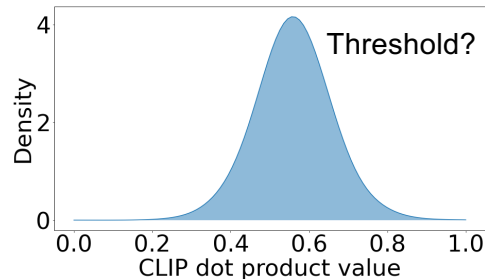
Desired distribution of CLIP dot products



Input image x^{obs}



Observed distribution of CLIP dot products



[1] Touvron, Lavril, Izacard, Martinet, Lachaux, Lacroix, Rozière et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971, 2023.

[2] Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." ICML 2023.

[3] Radford, Kim, Hallacy, Ramesh, Goh, Agarwal, Sastry et al. "Learning transferable visual models from natural language supervision." ICML 2021

Q2: Can we Improve CLIP without Annotations?

- In image classification, **most query answers are known to be false based on the class alone.**
 - **Example:** Know class is dog → “does the subject have fins?” is false → no need to see the image.

Yes! Use
LLMs

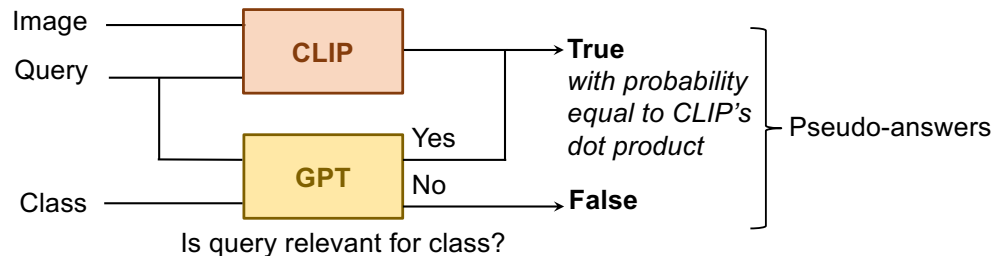
Input image x^{obs}



- We need to look at the image only for queries relevant to the class.
 - **Example:** “Does the subject have a leash?”. Need to see image since not all dogs have a leash.

Concept Question Answering System [1]

- **Pseudo-labeling:** Use GPT to determine class-relevant queries and use CLIP to determine probability of being true based on image.

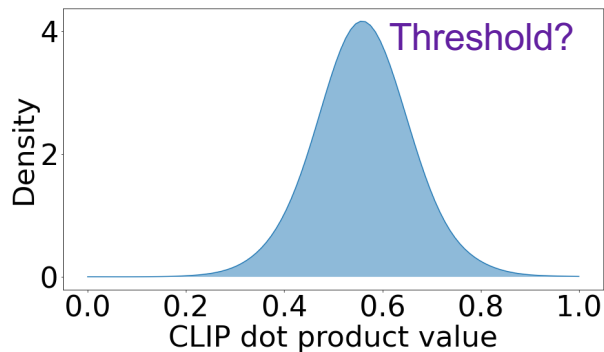


- **Concept-QA:** Train a **lightweight visual question answering (VQA)** system using pseudo-answers as we don't know class at test time.

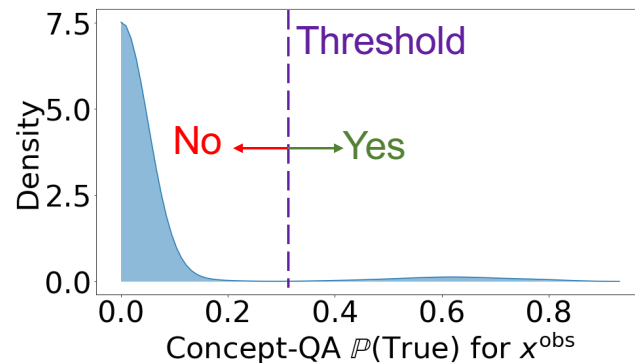


Interpretability of Concept-QA answers

- Concept-QA is more interpretable than CLIP!



Input image x^{obs}



Accuracy of Concept-QA answers

- Concept-QA is **more accurate** than CLIP & **more efficient** than BLIP2:
 - Concept-QA takes 0.04s per query vs 1.52s per query for BLIP2 FlanT5 model!

Model	ImageNet		Places365		CUB-200		CIFAR-10		CIFAR-100	
	Acc.	F ₁	Acc.	F ₁	Acc.	F ₁	Acc.	F ₁	Acc.	F ₁
CLIP-Bin _{std}	0.55	0.39	0.58	0.42	0.56	0.48	0.58	0.47	0.51	0.21
CLIP-Bin _{norm}	0.50	0.27	0.49	0.26	0.56	0.45	0.66	0.53	0.54	0.24
BLIP2 ViT-g OPT _{2.7B}	0.55	0.31	0.76	0.18	0.53	0.35	0.73	0.13	0.86	0.07
BLIP2 ViT-g FlanT5 _{XL}	0.86	0.56	0.87	0.62	0.70	0.40	0.83	0.59	0.87	0.41
Concept-QA (Ours)	0.87	0.56	0.83	0.45	0.80	0.54	0.80	0.62	0.80	0.38

Manually annotated 2.5K randomly sampled image-query pairs for each dataset.

**Q3: How do we select the queries
that form an explanation?**

Information Pursuit (IP)

- **Q3: How do we select queries that form the explanation?**
 - Shorter chains are easier to interpret.
 - Select **smallest number of queries** that are **sufficient for prediction**.

Generative-IP (G-IP) [1]

Learn **deep generative model** and use it to select **most informative queries**.

Variational-IP (V-IP) [2]

Train deep network to **select the next optimal query** given answers thus far.

IP-OMP [3]

Use **orthogonal matching pursuit** and **large vision and language models**.

[1] Chattopadhyay, Slocum, Haeffele, Vidal, Geman. Interpretable by design: Learning predictors by composing interpretable queries. TPAMI 2022.

[2] Chattopadhyay, Chan, Haeffele, Geman, Vidal. Variational Information Pursuit for Interpretable Predictions, ICLR 2023.

[3] Chattopadhyay, Pilgrim, Vidal. Information Maximization Perspective of Orthogonal Matching Pursuit with Applications to Explainable AI. NeurIPS 2023.

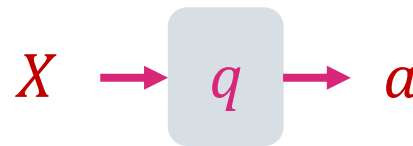
[4] Chattopadhyay, Chan, Vidal. Bootstrapping Variational Information Pursuit with Foundation Models for Interpretable Image Classification. ICLR 2024.

[5] Chattopadhyay, Haeffele, Vidal, Geman. Performance Bounds for Active Binary Testing with Information Maximization. ICML 2024.

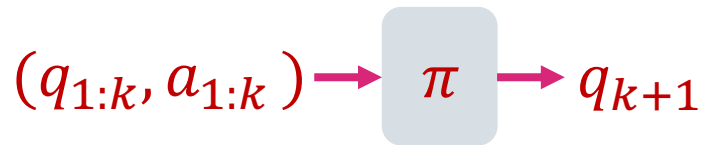
Information Pursuit: Problem Formulation

- Notation

- $X \in \mathcal{X}$: input variable (data).
- $Y \in \mathcal{Y}$: prediction variable (label).
- $Q = \{q: \mathcal{X} \rightarrow \mathcal{A}\}$: query set.



- **Querier π** : a function that selects the next question given history.



- **Code $^{\pi}_Q(X)$** : chain of query-answers selected by the querier for input X .

$$(q_{1:k}, a_{1:k})$$

Information Pursuit: Optimal Querier

- What properties should an **ideal querier** have?
 - **Minimality**: shorter explanations are easier to interpret and thus preferred.
 - **Sufficiency**: explanations (query-answer chains) should be a sufficient statistic for Y .
- Balance minimality of explanation with sufficiency via the objective:

$$\begin{aligned} \min_{\pi} \quad & \mathbb{E} [|\text{Code}_{\pi_Q}^{\pi}(X)|] && \text{(Minimality)} \\ \text{s. t.} \quad & \mathbb{P}(Y | \text{Code}_{\pi_Q}^{\pi}(X)) = \mathbb{P}(Y | X) && \text{(Sufficiency)} \end{aligned}$$

- Above problem is NP-Hard to solve [1], thus need for approximations.

Generative Information Pursuit (G-IP)

- Given query set Q , Information Pursuit (IP) **selects queries sequentially and adaptively in order of information gain [1]**.

Information Pursuit Algorithm

Queries are chosen according to observed x .

- First query and prediction:

$$q_1 = \arg \max_{q \in Q} I(q(X); Y)$$

$$y_1 = \arg \max_{y \in Y} \mathbb{P}(y \mid q_1(x))$$

- Next query and prediction:

$$q_{k+1} = \arg \max_{q \in Q} I(q(X); Y \mid q_{1:k}(x))$$

$$y_{k+1} = \arg \max_{y \in Y} \mathbb{P}(y \mid q_{1:k+1}(x))$$

- Termination and prediction:

$$q_{L+1} = q_{STOP} \quad \text{if} \quad \max_{q \in Q} I(q(X); Y \mid q_{1:L}(x)) = 0$$

$$y_{L+1} = \arg \max_{y \in Y} \mathbb{P}(y \mid q_{1:L}(x))$$

$q_{1:k}(x)$ is the event that contains all realizations of X that agree on the first k query-answers for x .

Generative Information Pursuit (G-IP)

- Selecting the first query requires computing

$$\operatorname{argmax}_{q \in Q} I(q(X); Y)$$

- Later queries need computing

$$\operatorname{argmax}_{q \in Q} I(q(X); Y \mid q_{1:k}(x))$$

History



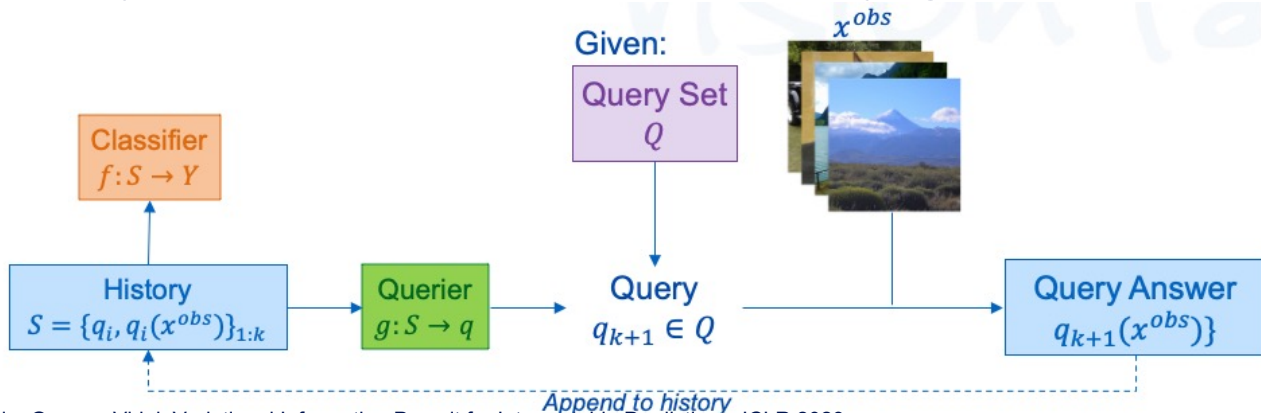
- **Generative IP:** learn **deep generative model** for $\mathbb{P}(q(X); Y)$ and use it to compute mutual information (via sampling) and select best query.
- **Challenge:** estimating mutual information in high dimensions is hard.

Variational Information Pursuit (V-IP)

- Train **querier** g_η to select the most informative query for **classifier** f_θ .

$$\min_{\theta, \eta} \mathbb{E}_{X, S} [D_{KL}(\mathbb{P}(Y | X) || \mathbb{P}_\theta(Y | q_\eta, S))]$$
$$\text{s.t. } q_\eta = g_\eta(S), \quad \mathbb{P}_\theta(Y | q_\eta, S) = f_\theta(q_\eta \cup S)$$

- Theorem:** selecting the most informative query given history \equiv finding query that, when added to the history, gives best prediction.



IP vs Orthogonal Matching Pursuit (OMP)

- **IP:** Given queries selected thus far, IP selects query that is most informative for Y

$$q_{k+1} = \operatorname{argmax}_{q \in Q} I(q(X); Y | q_{1:k}(x))$$

- **OMP:** given atoms selected thus far, OMP selects atom that is most correlated with x

$$\min_{\beta} \|\beta\|_0 \text{ s.t. } D\beta = x$$

$$i_{k+1} = \operatorname{argmax}_{d \in D} |\langle d, x - D\beta_k \rangle|$$

- **CLIP-IP-OMP [1]:** decompose image as sparse linear combination of semantic dictionary

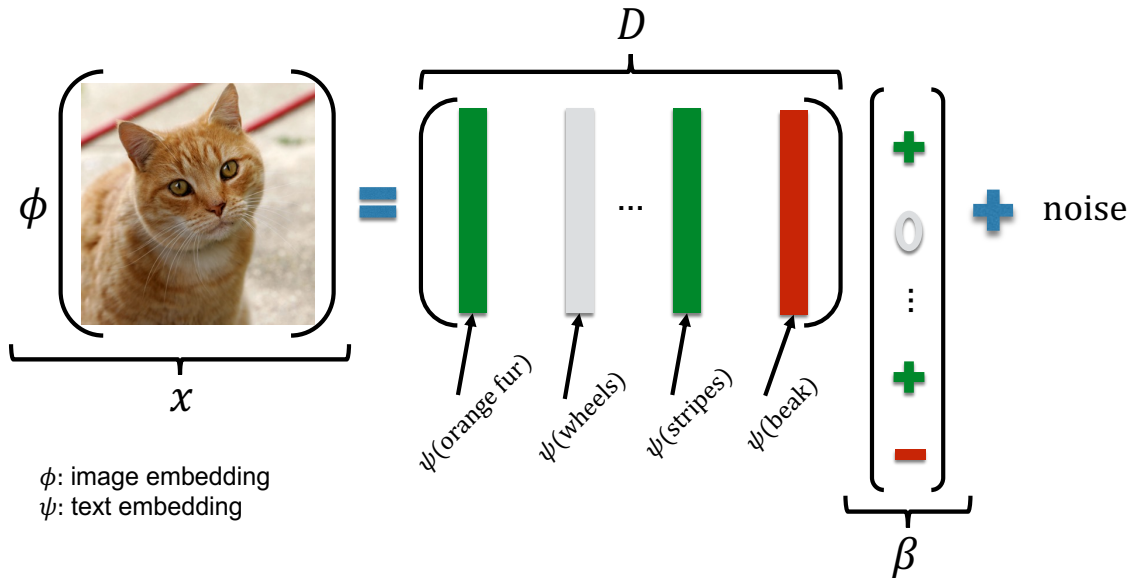
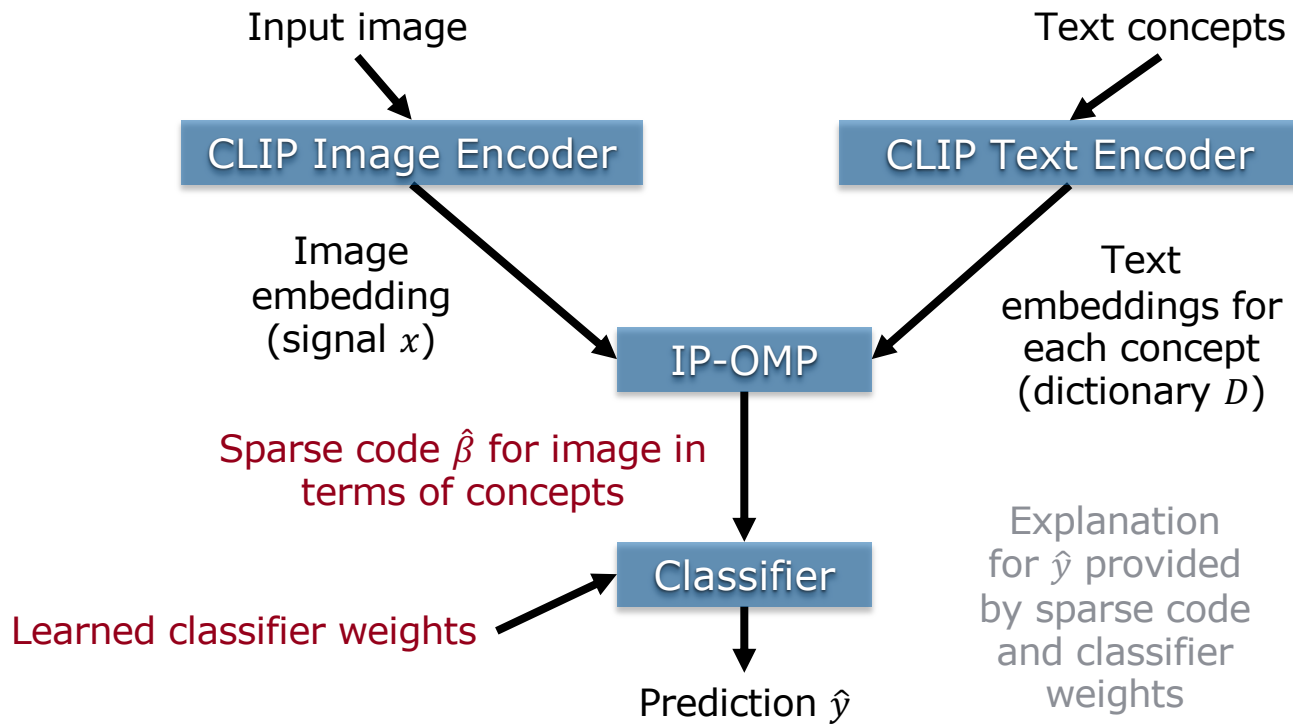


Image credit: <https://en.wiktionary.org/wiki/cat#/media/File:Cat03.jpg>

CLIP-IP-OMP: Details



Summary of the Information Pursuit Framework

- **Q1: How do we define the set of queries?**
 - Defined by **domain experts** [1].
 - Defined by **large language models** [4].
- **Q2: Given an input and a query, how do we answer the query?**
 - **Train classifiers** on data annotated with query answers by task experts [1].
 - Use domain-specific **pre-trained large vision language models** [4].
- **Q3: How do we select queries that form the explanation?**
 - **Information Pursuit:** Select **smallest number of queries** that are sufficient for prediction using Generative IP [1], Variational IP [2], and OMP [3].

[1] Chattopadhyay, Slocum, Haeffele, Vidal, Geman. Interpretable by design: Learning predictors by composing interpretable queries. TPAMI 2022.

[2] Chattopadhyay, Chan, Haeffele, Geman, Vidal. Variational Information Pursuit for Interpretable Predictions, ICLR 2023.

[3] Chattopadhyay, Pilgrim, Vidal. Information Maximization Perspective of Orthogonal Matching Pursuit with Applications to Explainable AI. NeurIPS 2023.

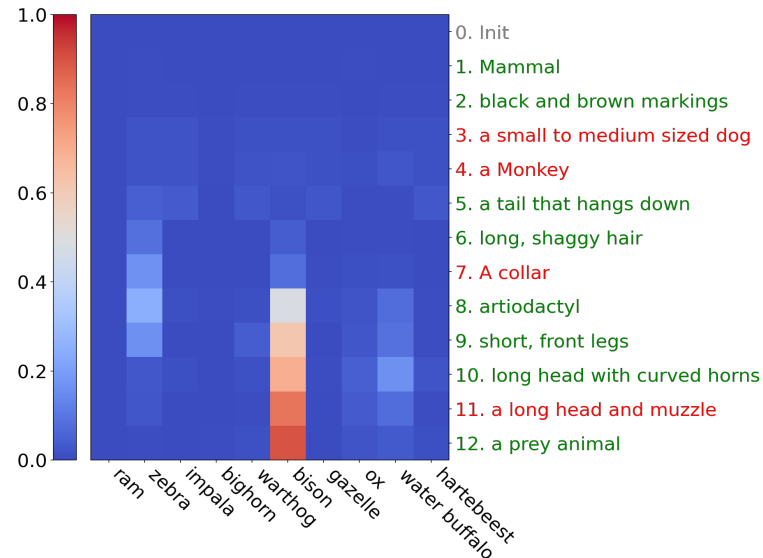
[4] Chattopadhyay, Chan, Vidal. Bootstrapping Variational Information Pursuit with Foundation Models for Interpretable Image Classification. ICLR 2024.

[5] Chattopadhyay, Haeffele, Vidal, Geman. Performance Bounds for Active Binary Testing with Information Maximization. ICML 2024.

Applications

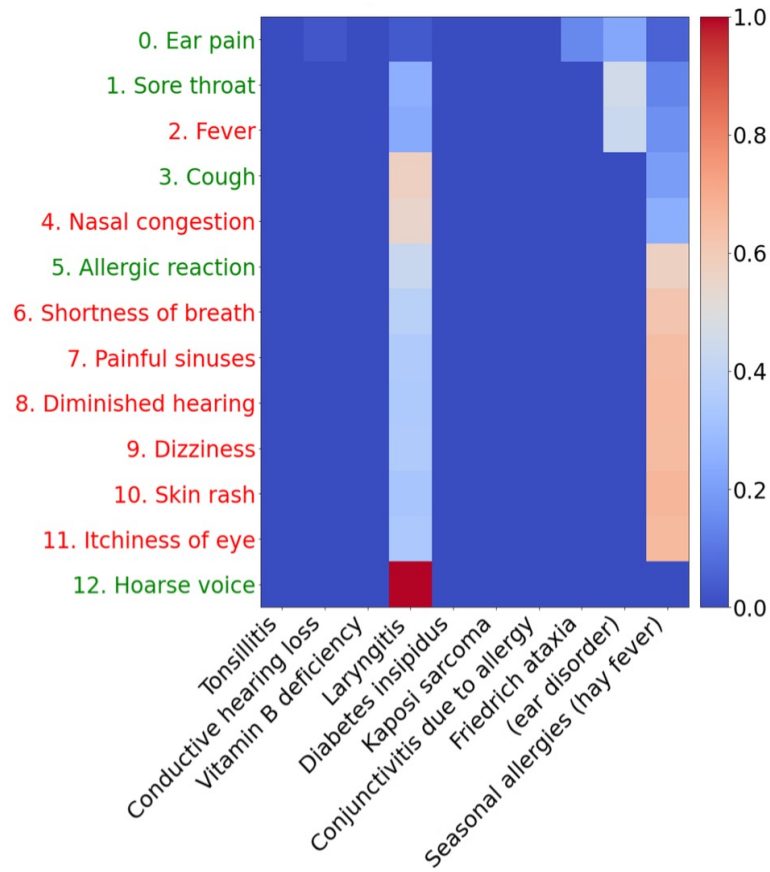
Interpretable Image Classification by V-IP

- **Task:** Image classification.
- **Query set:** Queries about presence or absence of different semantic concepts.
- **Dataset:** ImageNet
 - 1000 classes

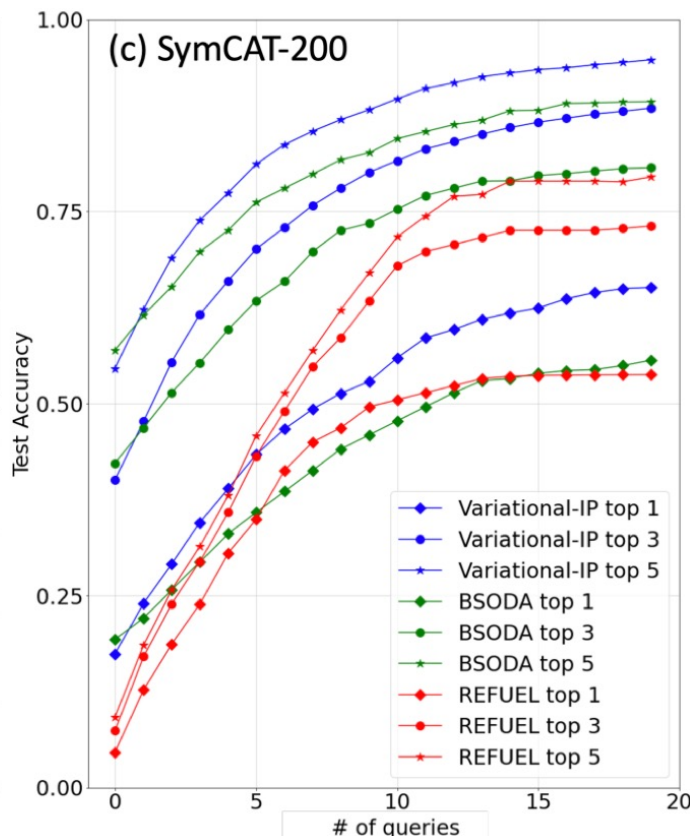
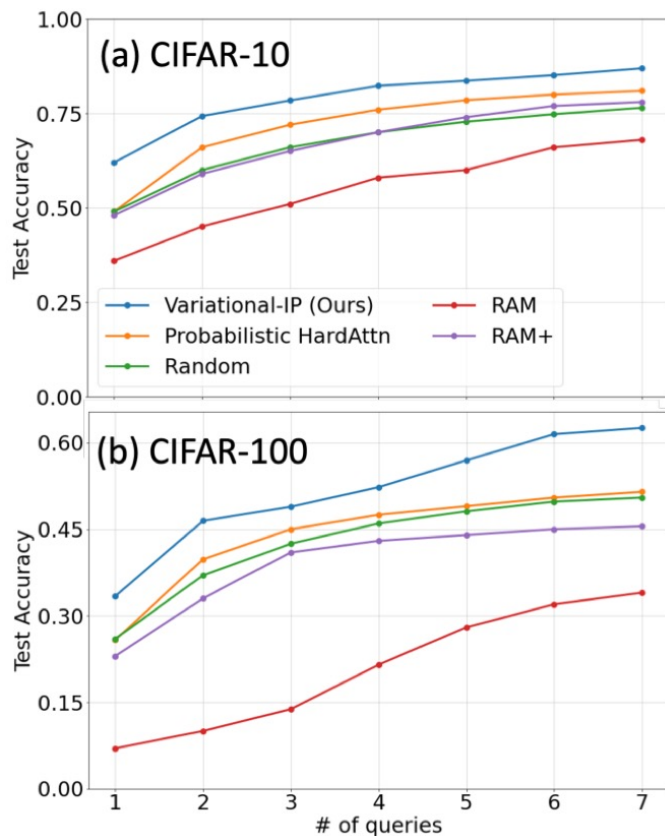


Interpretable Medical Diagnosis by VI-P

- **Task:** Disease diagnosis.
- **Query set:** Queries about presence or absence of different symptoms.
- **Dataset:** SymCAT-200
 - 1.1M doctor-patient dialogues about 326 symptoms indicative of 200 diseases.
 - Each dialogue: 2-3 symptoms per patient.
 - 326 binary queries, one per symptom.

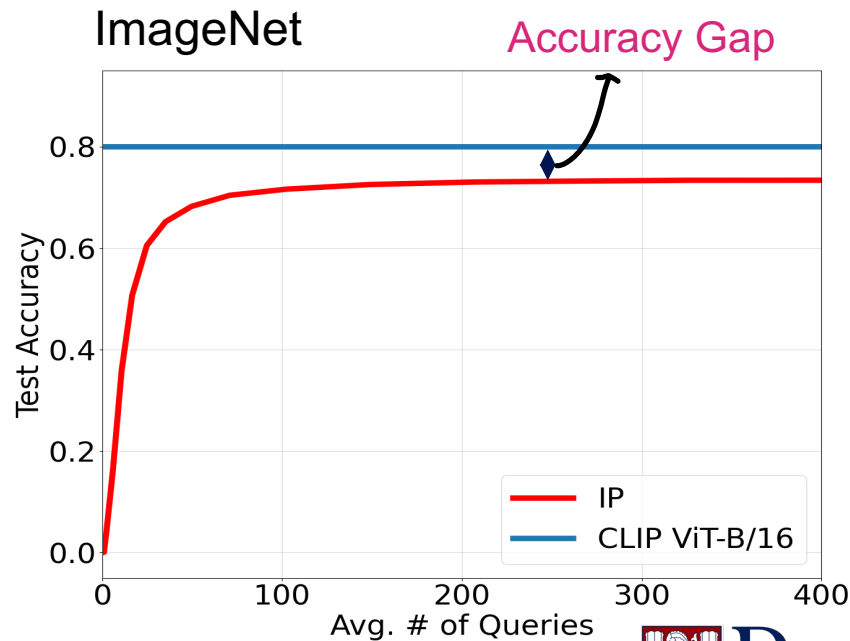
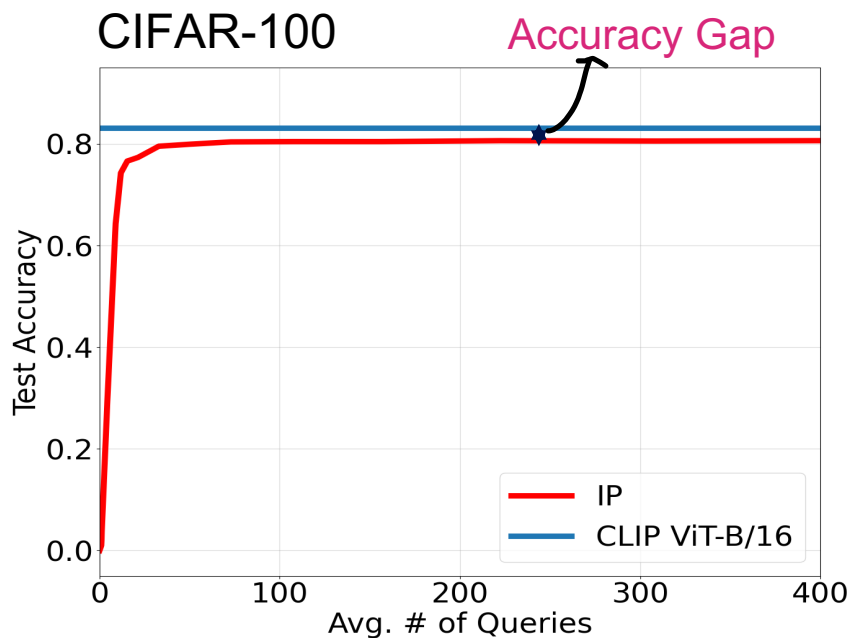


Accuracy Versus Number of Queries



Accuracy-Explainability Tradeoff

- How far is interpretable-by-design from black-box model performance

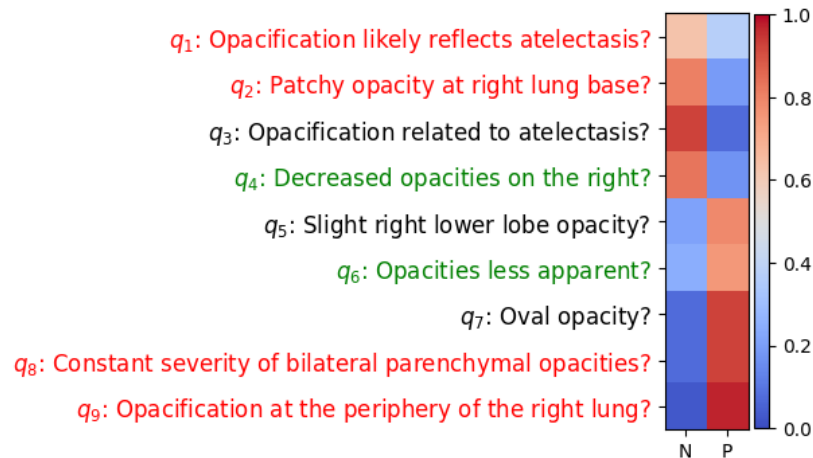


Interpretable Radiological Report Classification

- **Task:** Predict disease label in a radiological report.
- **Query set:** Queries about presence or absence of facts in a radiology report.
- **Dataset:** MIMIC-CXR
 - Data: 227,827 reports.
 - Queries are binary questions, one for each possible fact.
 - The task is to predict the disease label.

Radiology report: As compared to the previous radiograph, the vertebral fixation devices in the right-sided Port-A-Cath are in constant position. *The pre-existing parenchymal opacities have slightly decreased in extent and severity.* Also, a component of minimal interstitial fluid overload is less severe than on the previous image. Unchanged cardiac silhouette. No pleural effusions.

Disease to classify: Lung opacity



Interpretable Radiological Report Classification

- **Q1: How do we define the set of queries?**
 - Leverage LLMs and medical knowledge to **extract 591,920 facts** from 227,827 reports in the MIMIC-CXR dataset [1].
- **Q2: How do we answer a query for a given input?**
 - Leverage LLMs and medical knowledge to **verify if a fact is present** in a radiology report [2].
- **Q3: How do we select the best queries to form an explanation?**
 - Select **smallest number of facts** that are **sufficient for disease prediction** [2] using Variational IP [3,4].

[1] Messina, Vidal, Parra, Soto, Araujo. Extracting and Encoding: Leveraging LLMs and Medical Knowledge to Enhance Radiological Text Representation. ACL 2024.

[2] Ge, Chan, Messina, Vidal. Information Pursuit for Interpretable Classification of Chest Radiology Reports. ArXiv 2025.

[3] Chattopadhyay, Chan, Haeffele, Geman, Vidal. Variational Information Pursuit for Interpretable Predictions. ICLR 2023.

[4] Chattopadhyay, Chan, Vidal. Bootstrapping Variational Information Pursuit with Foundation Models for Interpretable Image Classification. ICLR 2024.

Interpretable Radiological Report Classification

- Average precision (AP) and F1 score of IP-CRR on six binary prediction tasks:
 - Lung Opacity (LO), Calcification of the Aorta (CA), Support Devices(SD),
 - Cardiomegaly(CM), Pleural Effusion(PE), and Pneumonia(PN).

Methods	AP						F1					
	LO	CA	SD	CM	PE	PN	LO	CA	SD	CM	PE	PN
CXR-BERT (FT-Last)	0.900	0.361	0.969	0.864	0.945	0.449	0.829	0.223	0.912	0.789	0.887	0.449
CXR-BERT (FT-All)	0.984	0.992	0.970	0.964	0.962	0.641	0.987	0.991	0.978	0.982	0.953	0.541
Flan-T5-large	0.527	0.073	0.445	0.380	0.616	0.190	0.663	0.139	0.321	0.543	0.754	0.299
CBM	0.947	0.345	0.934	0.791	0.874	0.432	0.884	0.241	0.853	0.738	0.801	0.431
IP-CRR	0.972	0.578	0.959	0.892	0.925	0.468	0.918	0.350	0.889	0.811	0.860	0.451

Summary

- **Information Pursuit**: an interpretable-by-design prediction framework.
- **Generative model**: use LLMs to define queries, VLMs to answer queries, and G-IP, V-IP, OMP to select queries and make predictions.

Input image x^{obs}



Ask a sequence of interpretable queries about x^{obs}

q_1 .	Has shape perching-like?	Yes
q_2 .	Has bill shape all-purpose?	Yes
q_3 .	Has belly color yellow?	Yes
q_4 .	Has upperparts color yellow?	No
q_5 .	Has throat color yellow?	No
q_6 .	Has breast color black?	Yes
q_7 .	Has belly color olive?	Yes

Predicted bird species

Green Jay with
99% probability

Open Questions

- **How to define interpretability?**
 - Hypothesis tests on the importance of a feature?
 - Minimum set of interpretable features that are sufficient for prediction?
 - What about causality-based explanations?
- **How to evaluate if a model is interpretable?**
 - Human evaluations?
 - Can humans predict a class based on explanation?
 - Benchmarks

Thank you



Penn
UNIVERSITY of PENNSYLVANIA