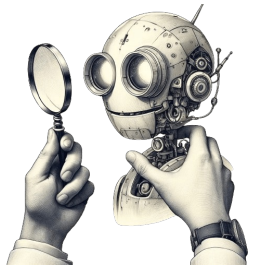
ICCV HONOLULU HAWAII
OCT 19-23, 2025

Foundations of Interpretable AI



PART I: Motivation, Post-hoc Methods, Explainable-by-Design Methods	**(8:00 - 9:00 am)**	**Aditya Chattopadhyay (Amazon)**
☕ Short break ☕	(9:00 – 9:15 am)	
PART II: Shapley Value Based Methods	(9:15 – 10:15 am)	Jeremias Sulam (Johns Hopkins)
☕ Short break ☕	(10:15 – 10:30 am)	
PART III: Information Pursuit	(10:30 – 11:30 am)	René Vidal (Penn)
QA Session	(11:30 – 12:00 noon)	

Foundations of Interpretable AI

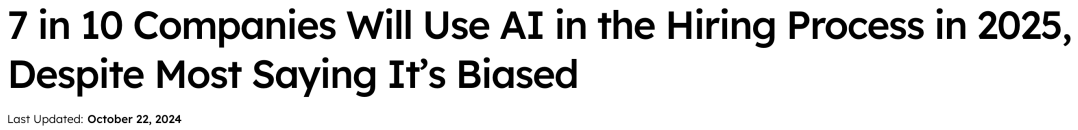
Introduction

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

[1]*The 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.



- 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.

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

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

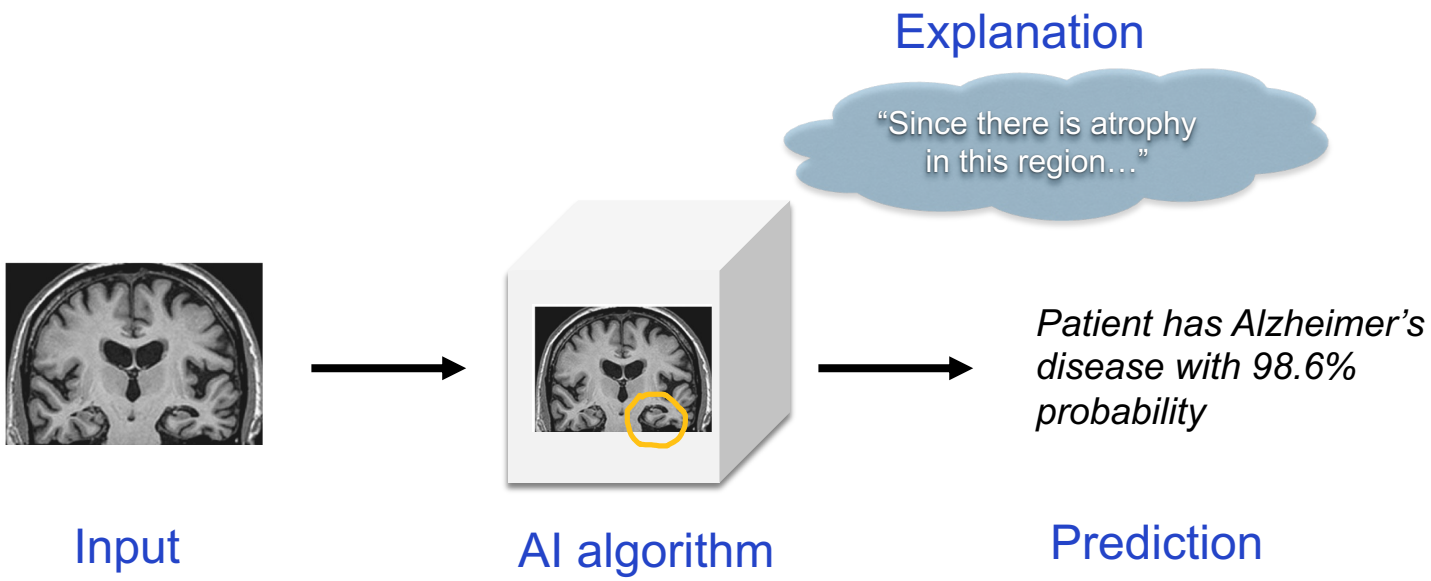
Why Interpretability?

- Mismatch between training objectives (e.g. cross entropy loss) and real world desiderata[1].
  - 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 explainable-by-design

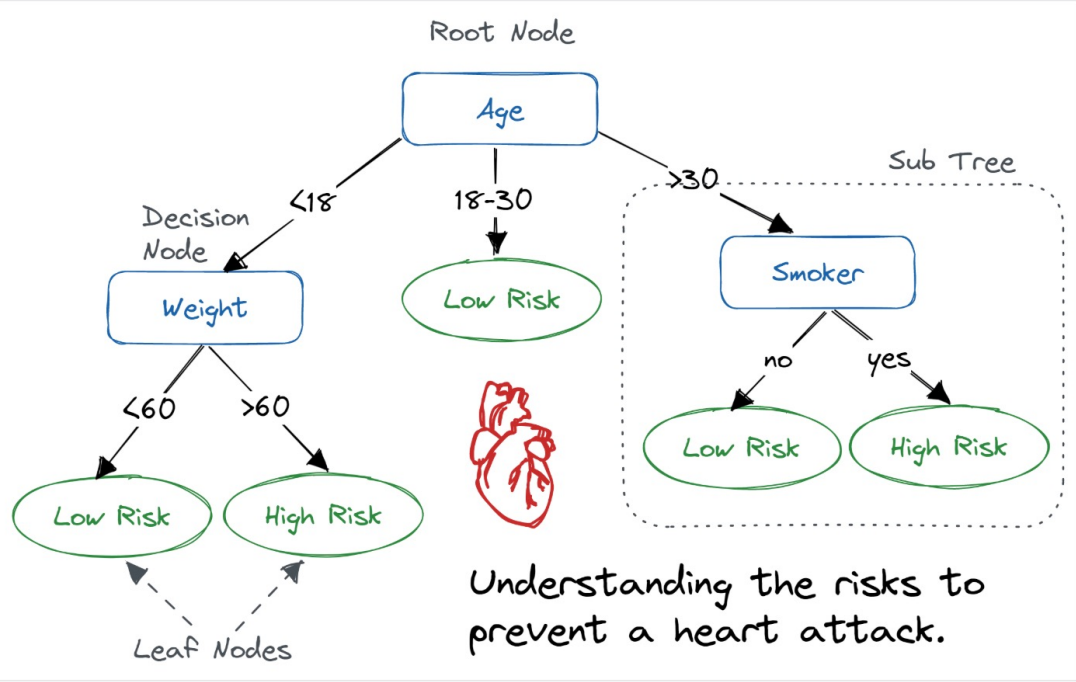
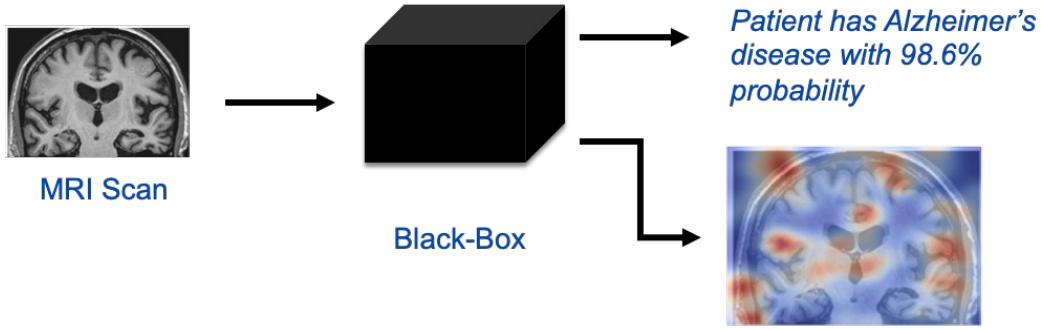


Image: https://www.datacamp.com/tutorial/decision-tree-classification-python

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



Tutorial Outline

1. Recent approaches to model interpretability.
   - ❖ Introductory lecture reviewing recent attempts at post-hoc interpretability and explainable-by-design approaches.

2. Model interpretability with Shapely Coefficients.
   - ❖ Deep-dive into Shapley values, a popular post-hoc interpretability method.

3. Information Pursuit: a framework for explainable-by-design ML
   - ❖ Deep-dive into Information Pursuit, a popular explainable-by-design interpretability method.

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

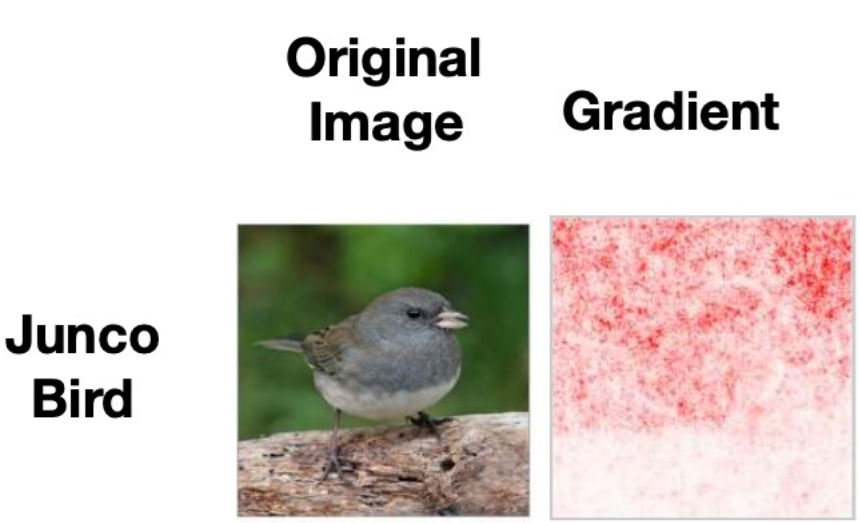
- **Definition (informal):** Given a model f, and an input x, assign scores to every feature based on how “important” they were for the model’s prediction.

- 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.



function f with input x,

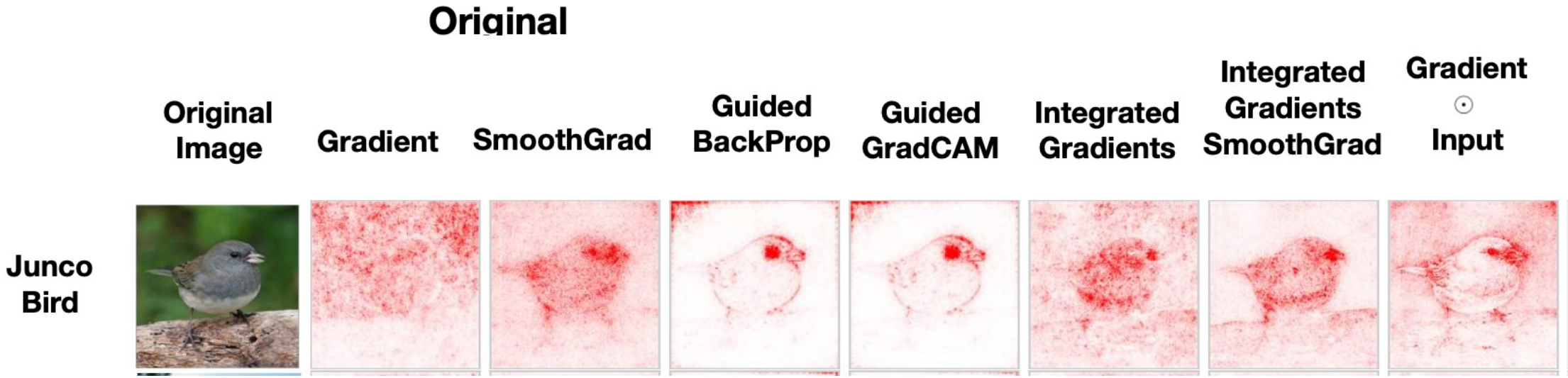
Why did the model predict Junco Bird?

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.
- Resulted in a flurry of ad-hoc methods that manipulate the gradients to produce "better-looking" feature maps



Sanity Checks for Saliency Maps[1]

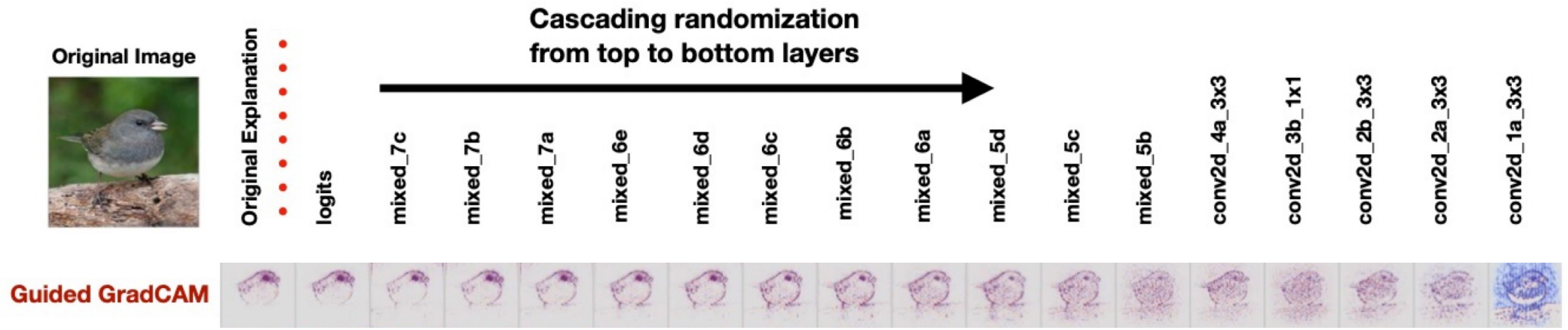
- Proposes simple diagnostic tests that a “reasonable” explanation method should pass.
  - Explanation should be sensitive to the weights of the network.
  - Explanations should help distinguish between “learning” and “memorization”.

- Many existing gradient-based approaches fail these simple tests.

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in neural information processing systems*, *31*.

Sanity Checks for Saliency Maps[1]

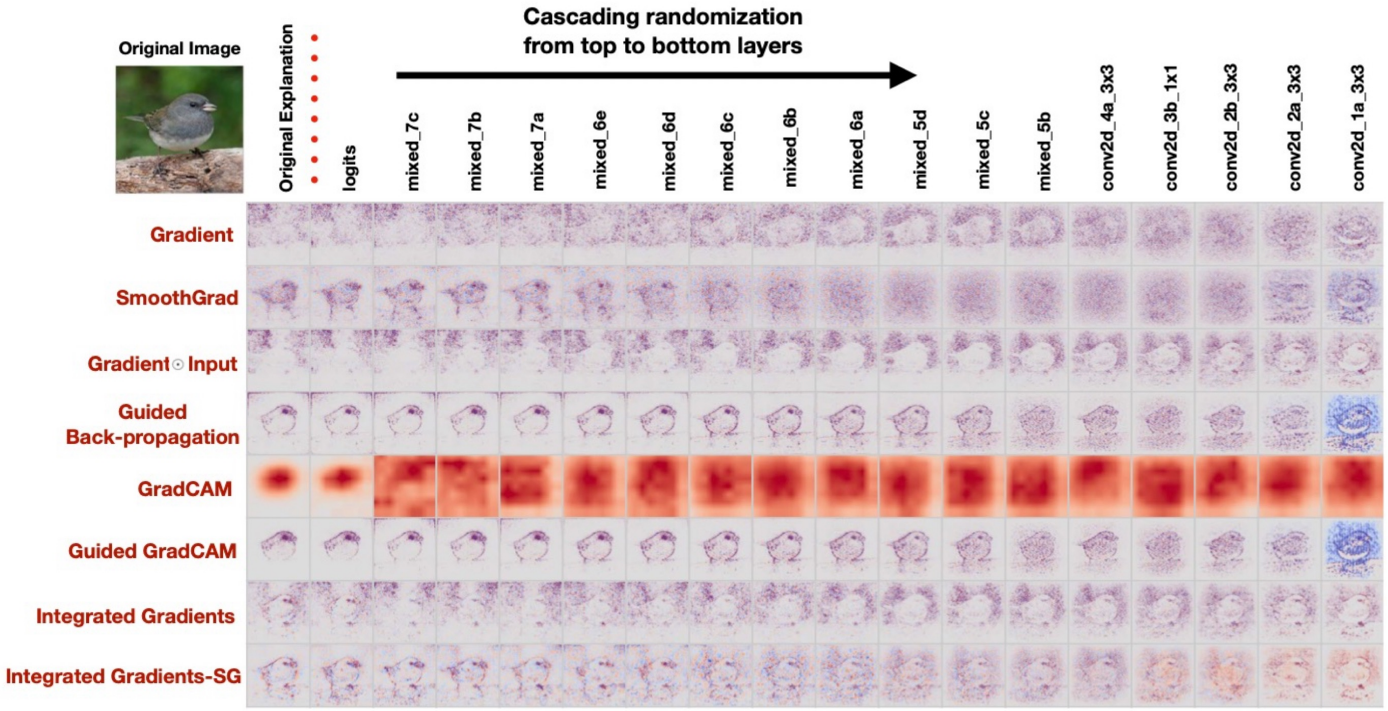
- Explanation should be sensitive to the weights of the network.



1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in neural information processing systems*, *31*.

Sanity Checks for Saliency Maps[1]

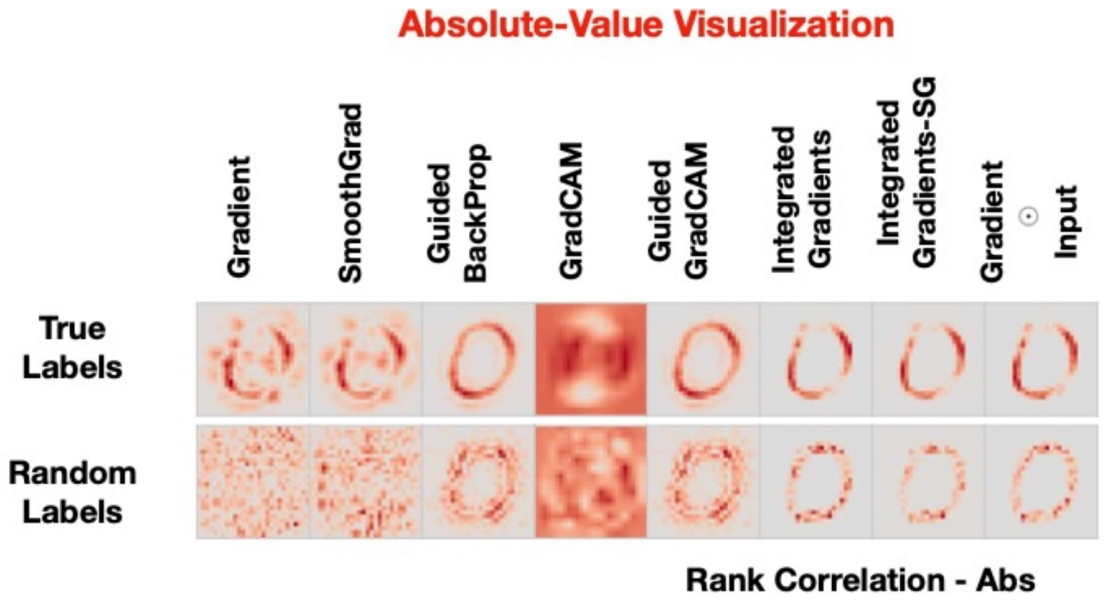
- Explanation should be sensitive to the weights of the network.



1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in neural information processing systems*, *31*.

Sanity Checks for Saliency Maps[1]

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



1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in neural information processing systems*, *31*.

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 Activation Vectors

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

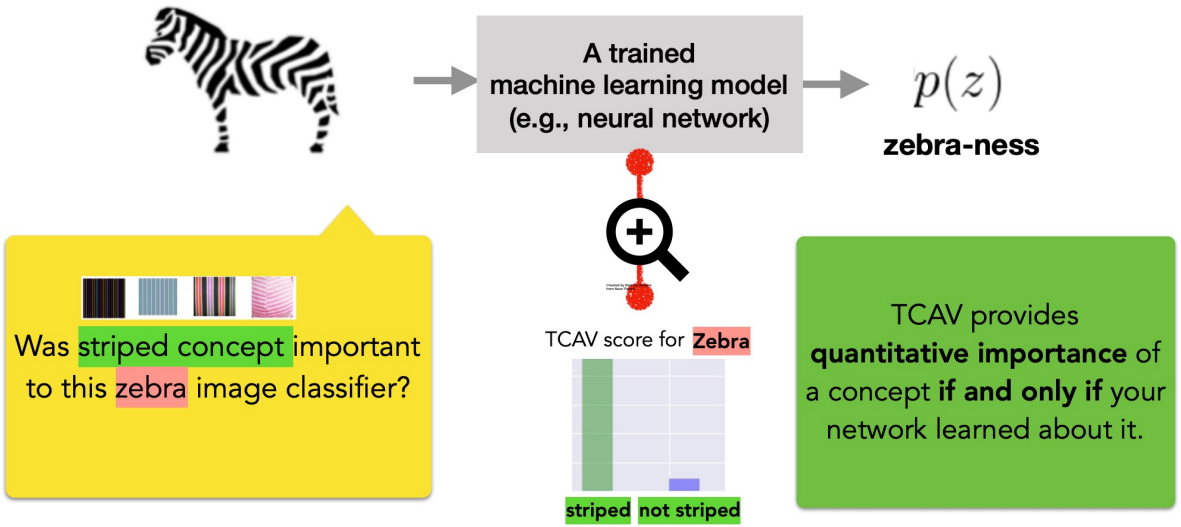


Image: https://beenkim.github.io/slides/TCAV_ICML_pdf.pdf

TCAV: Methodology

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

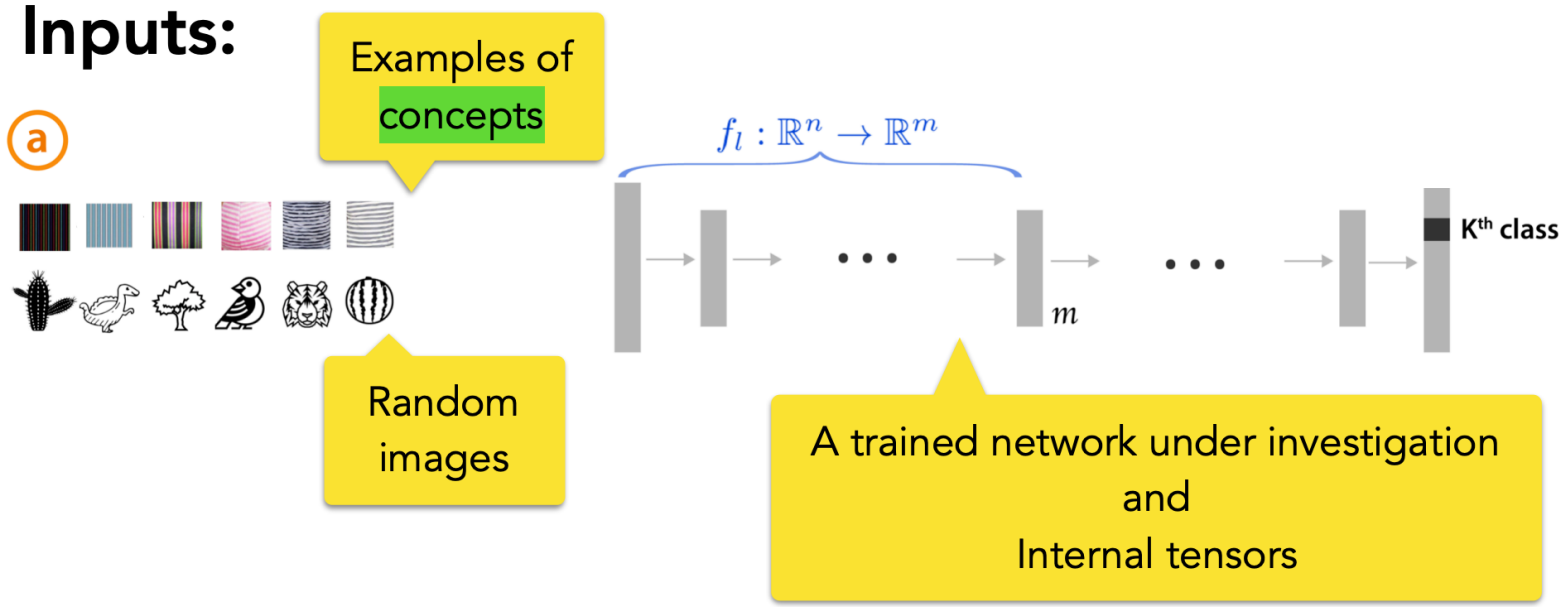


Image: https://beenkim.github.io/slides/TCAV_ICML_pdf.pdf

TCAV: Methodology

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

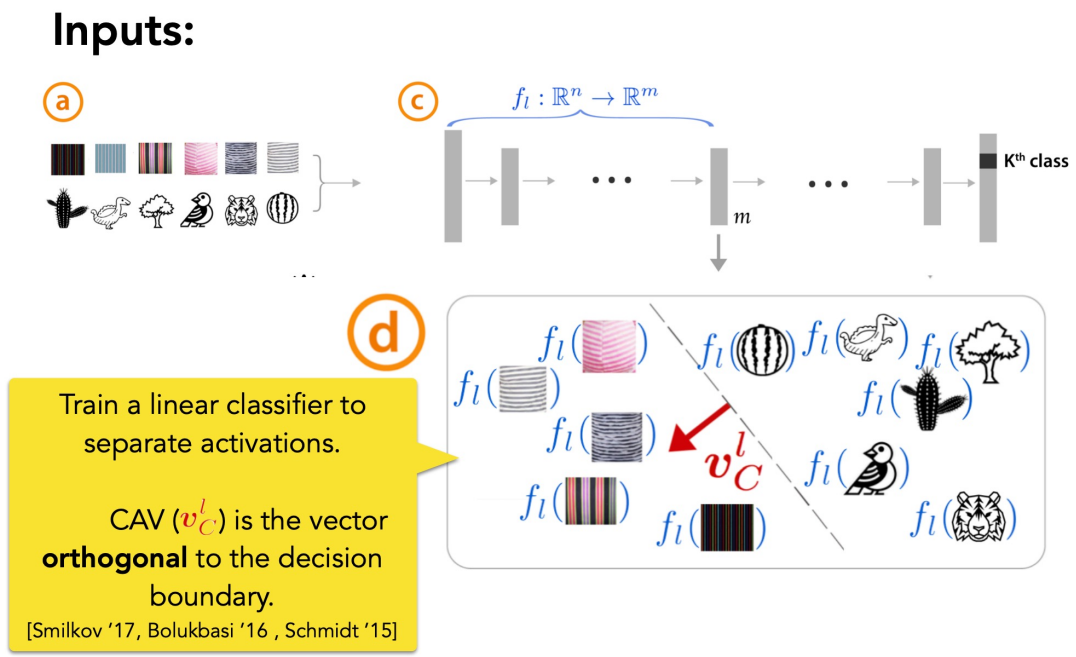
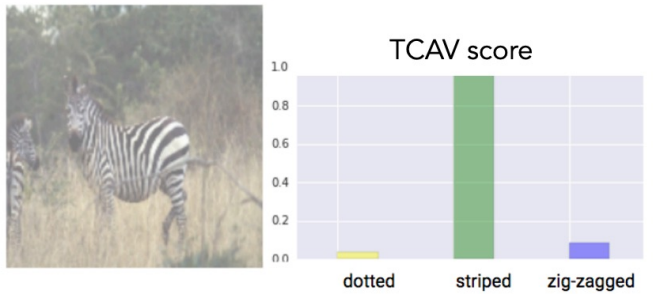


Image: https://beenkim.github.io/slides/TCAV_ICML_pdf.pdf

TCAV: Methodology

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



zebra-ness → $\frac{\partial p(z)}{\partial \boldsymbol{v}_C^l} = S_{C,k,l}(\boldsymbol{x})$

striped CAV →

Directional derivative with CAV

$S_{C,k,l}($  $)$

$S_{C,k,l}($  $)$

$S_{C,k,l}($  $)$

$S_{C,k,l}($  $)$

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

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.

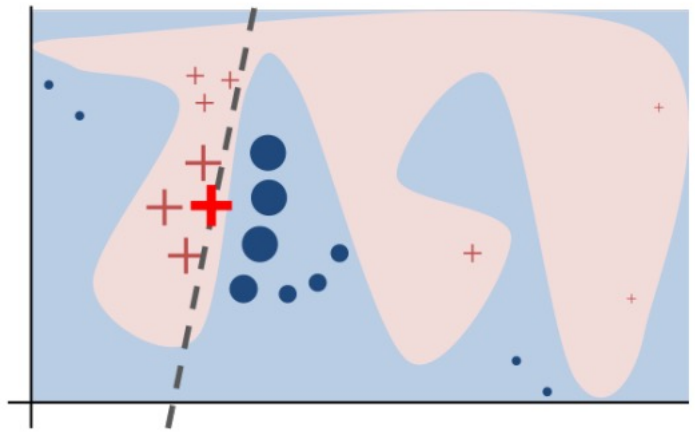
- What if the user wants a more holistic understanding of the deep model's decision-making process?
  - For example, as a Boolean expression or decision tree?

- 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

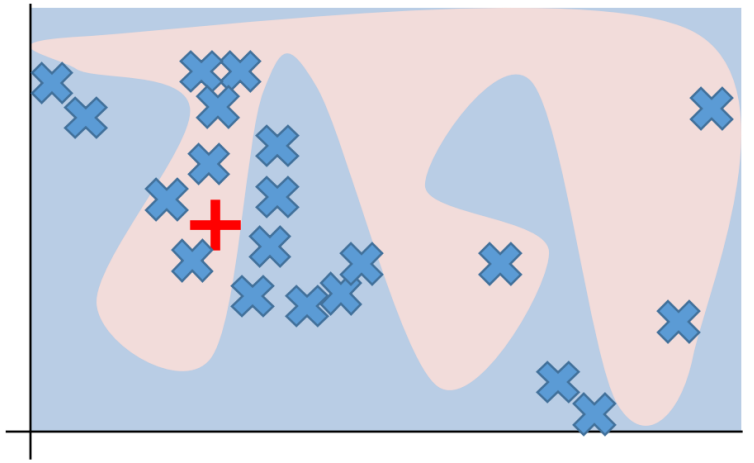
- A non-linear deep network can be locally approximated by a simpler “interpretable” model like logistic regression classifier or a decision tree.



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

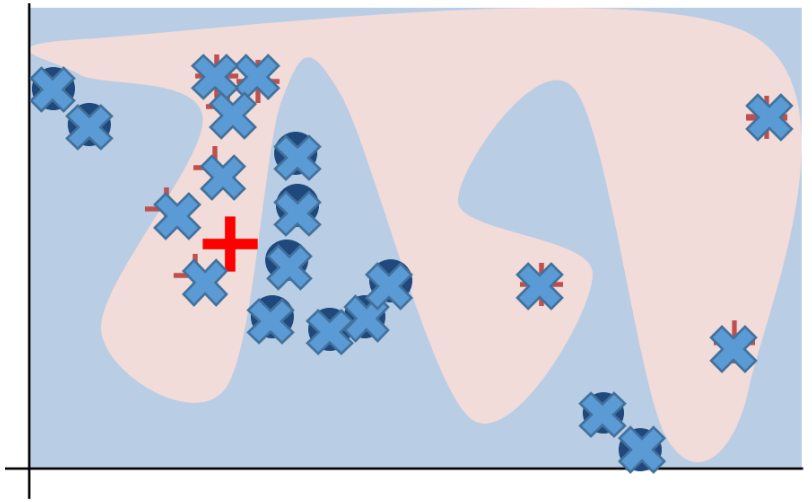
- Red cross x is the point to be explained.
- First, sample points around x.



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

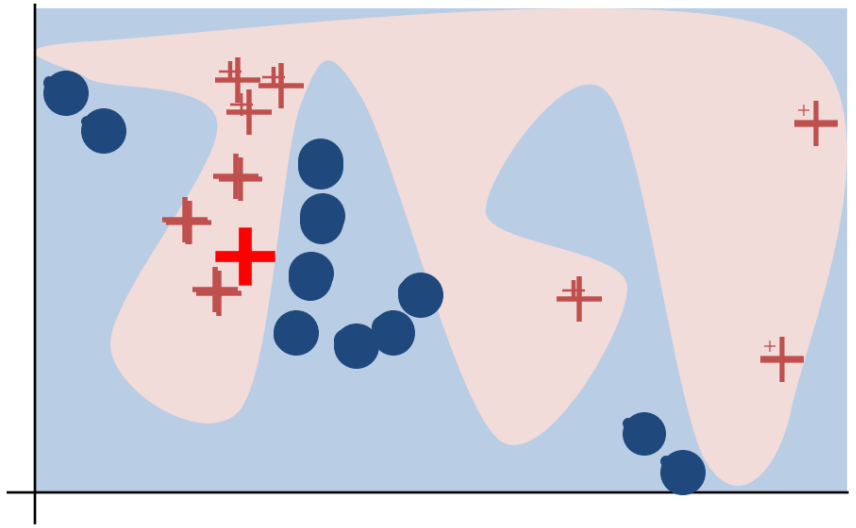
- Red cross x is the point to be explained.
- First, sample points around x.
- Use deep model to predict labels for each sample



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

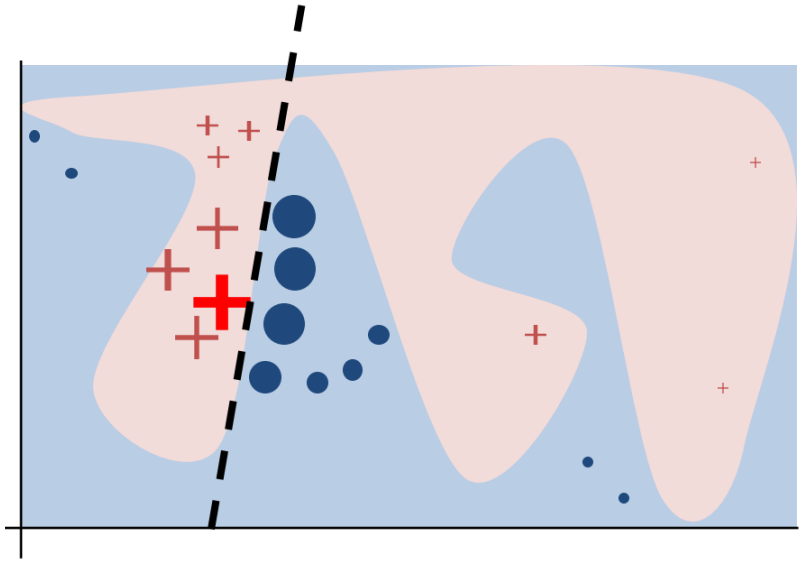
LIME: Methodology

- 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

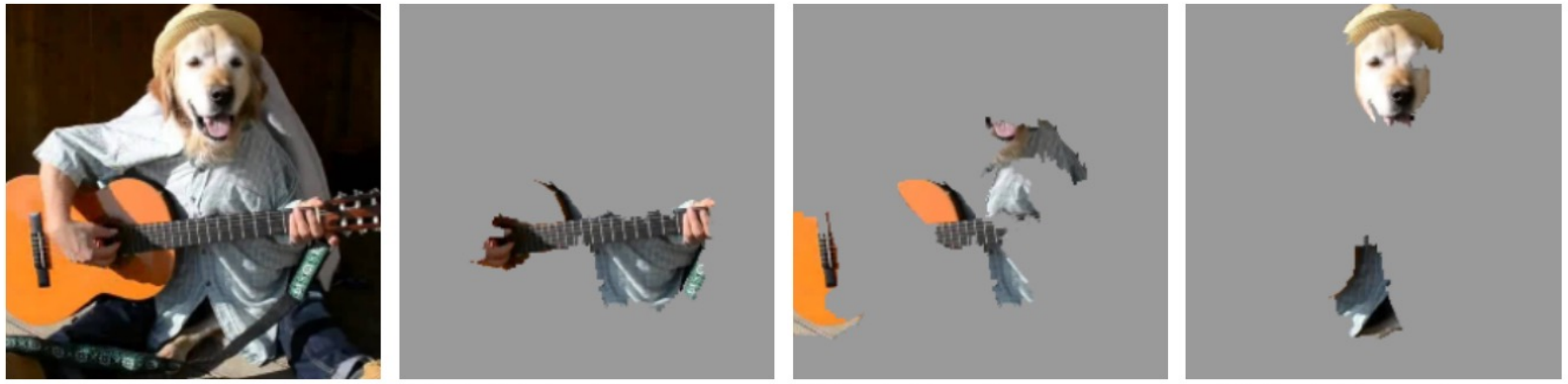
LIME: Methodology



- 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



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

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).

Locally 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.

Explainable-by-design Approaches to Model Interpretability

Explainable-by-design approaches to interpretability.

- Recall no ground truth available to measure quality of post-hoc explanations!
  - Very difficult to guarantee faithfulness of these explanations.

- **Main idea:** Design models that incorporate explanation as part of their forward function (decision-making process).

Classical explainable models

Decision Trees

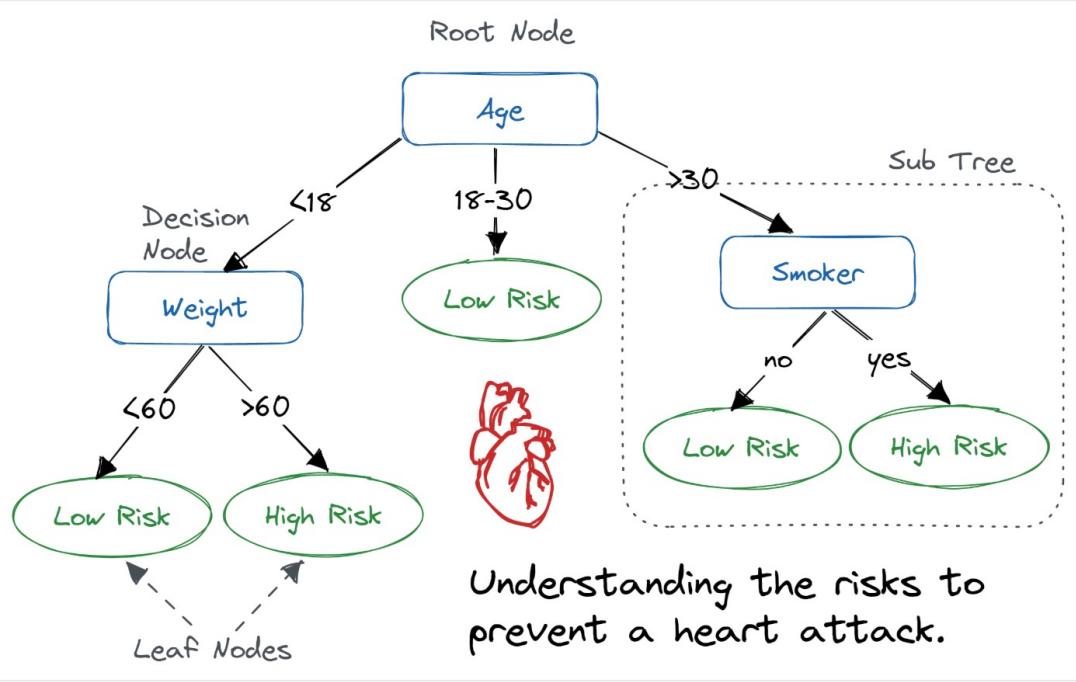


Image: https://www.datacamp.com/tutorial/decision-tree-classification-python

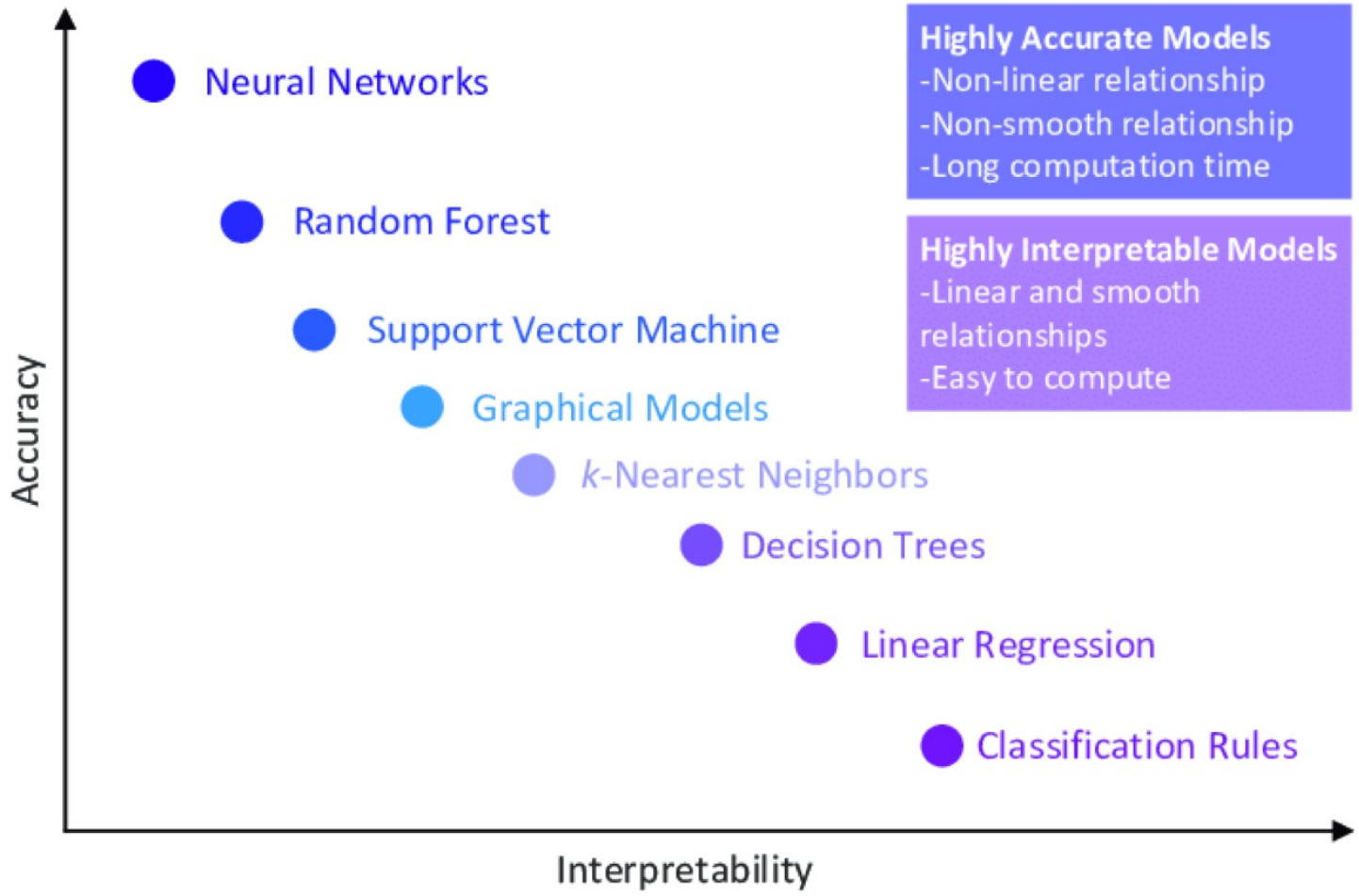
Linear Regression

$$y = \sum_{\{i\}} w_i \, x_i$$

$w_i : coefficients$

$x_i : Interpretable \; features$

Accuracy vs Interpretability Tradeoff



Classical to deep explainable models

Decision Trees

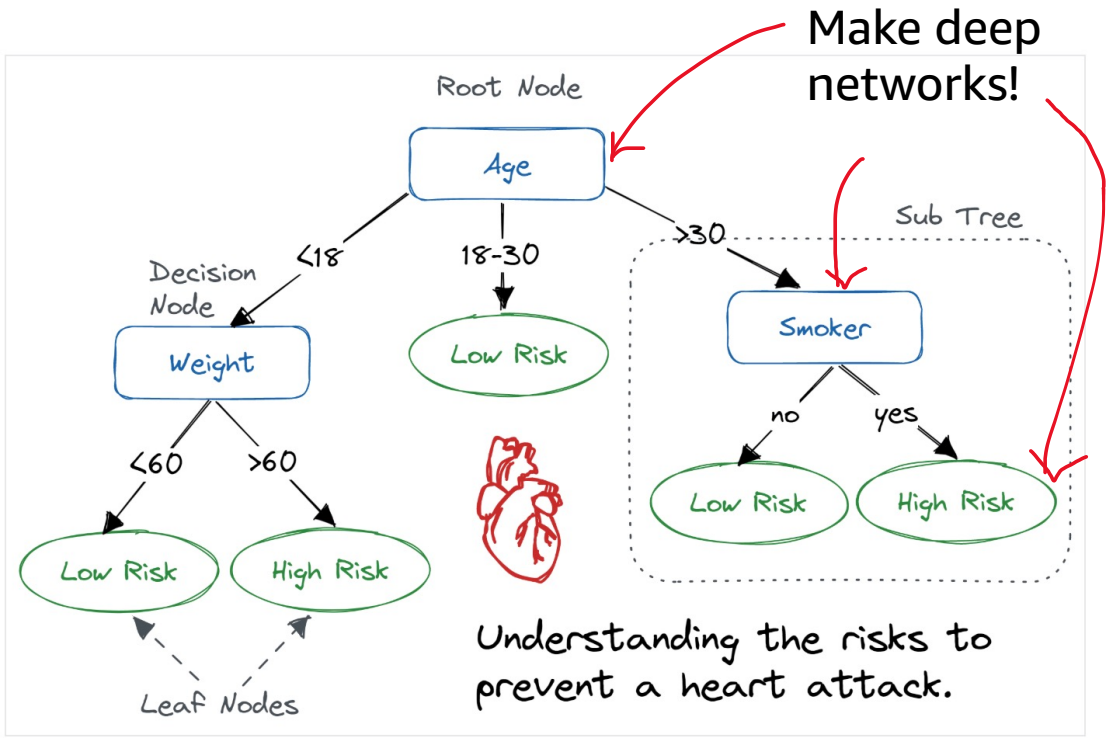


Image: https://www.datacamp.com/tutorial/decision-tree-classification-python

Linear Regression

Make deep networks!

$$y = \sum_{\{i\}} w_i \, x_i$$

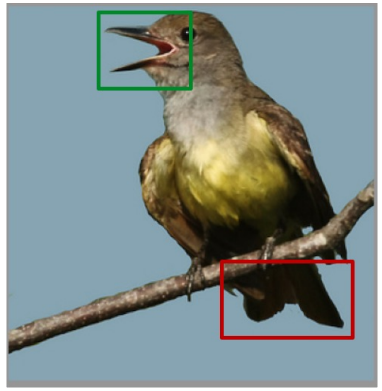
$w_i : coefficients$

$x_i : Interpretable\ features$

Concept-based explainable models

- Design networks that make predictions based on **a task-specific set of semantic concepts**.
  - These concepts support the prediction's explanation/reasoning.

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

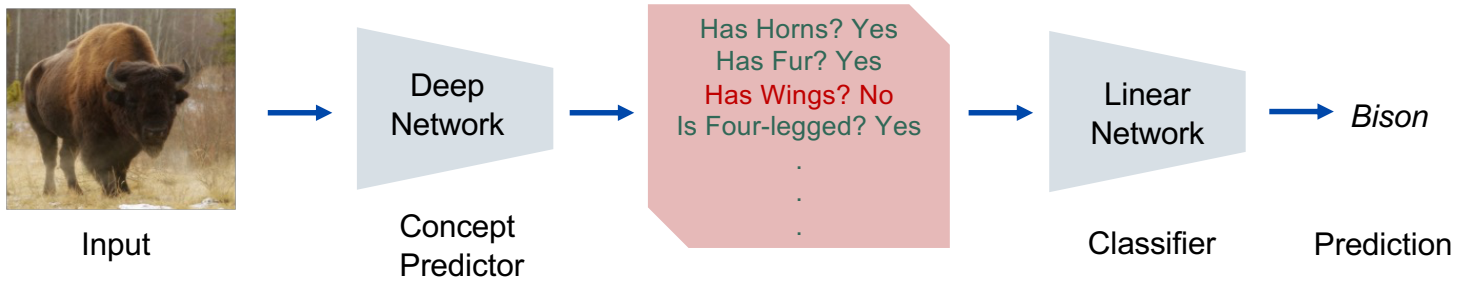


(b) **Task**: scene interpretation
Concepts: objects, relationships

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

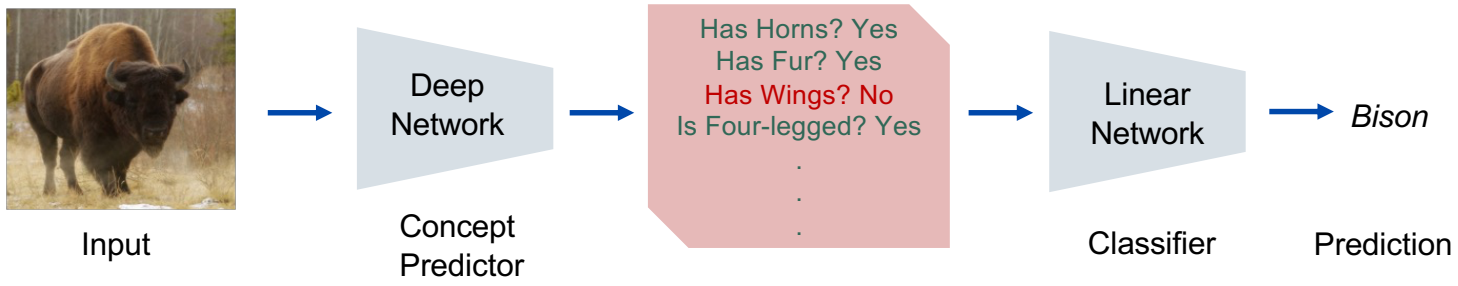
0. Ear pain
1. Sore throat
2. Fever
3. Cough
4. Nasal congestion
5. Allergic reaction
6. Shortness of breath
7. Painful sinuses

Concept Bottleneck Models



- 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.

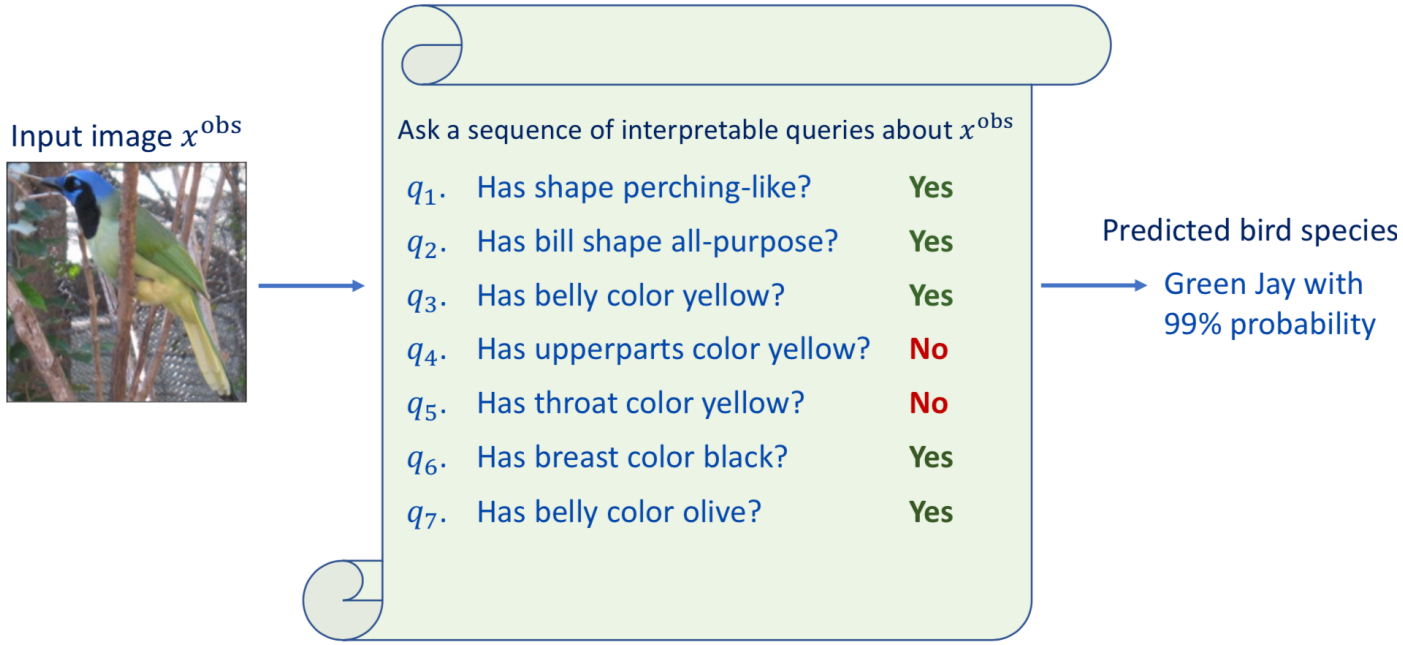
Are Concept Bottleneck Models Enough?



- **Limited expressivity**: linear classification layer limits expressivity of CBMs when "concept answers → 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.

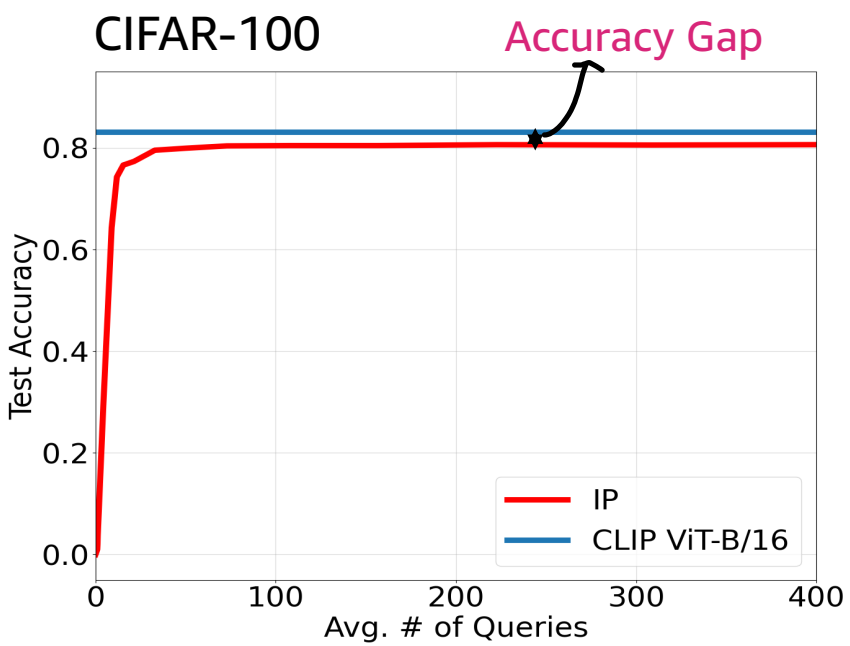
Information Pursuit Framework

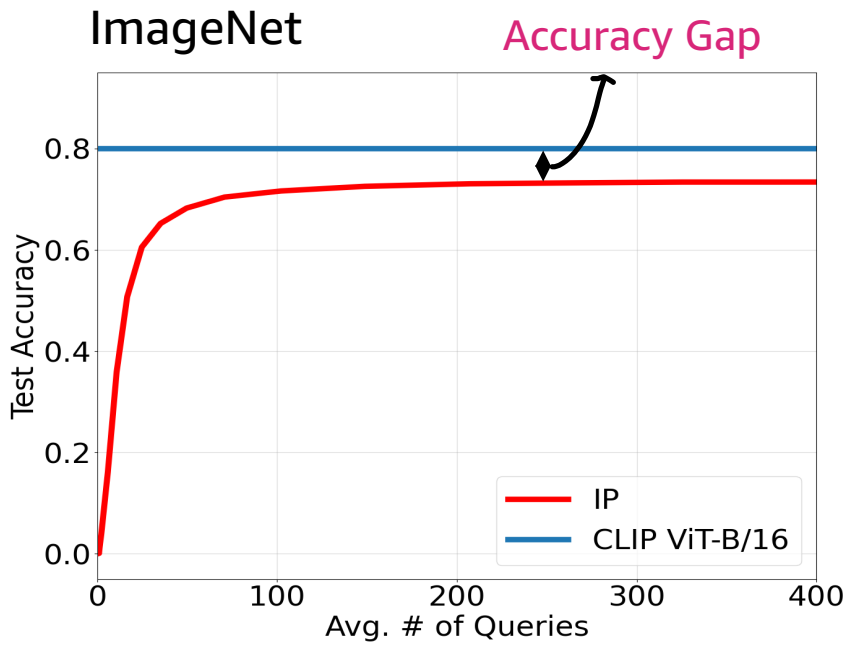
- An information-theoretic framework based on 20Q parlor game.
- Make prediction based on **smallest number of queries** that are **sufficient** for prediction.



Accuracy-Interpretability Tradeoff

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





Concept-based explainable models

- **Pro:**
  - Explanations in terms of concepts more amenable to humans.
  - User can decide the language of explanations by choosing the concepts.

- **Cons:**
  - User needs to define set of concepts for every task.
    - How to know concepts are sufficient for the task?
  - Need a mechanism to predict the presence of a concept given data.

Thank you!