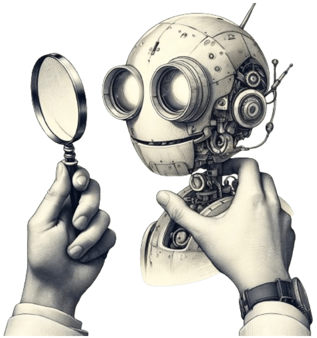
Foundations of Interpretable AI

Tutorial @ CVPR Nashville JUNE 11-15, 2025



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

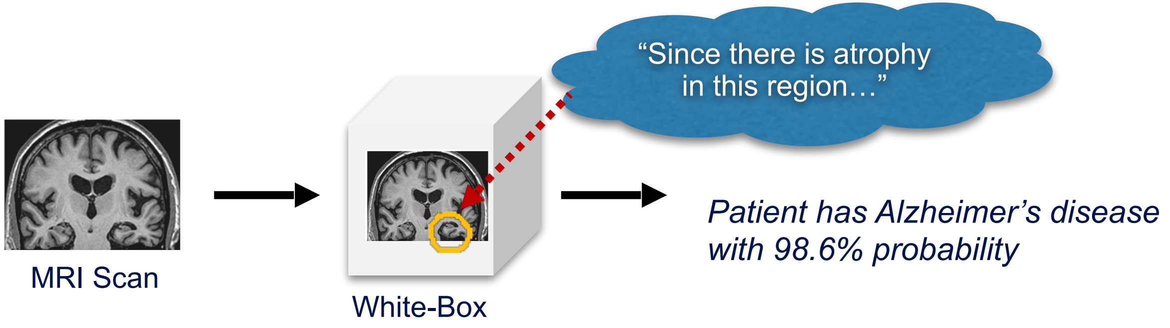
PART II: Shapley Value Based Methods (9:45 – 10:30 am) Jeremias Sulam (Johns Hopkins)

☕ Coffee break ☕ (10:30 – 11 am)

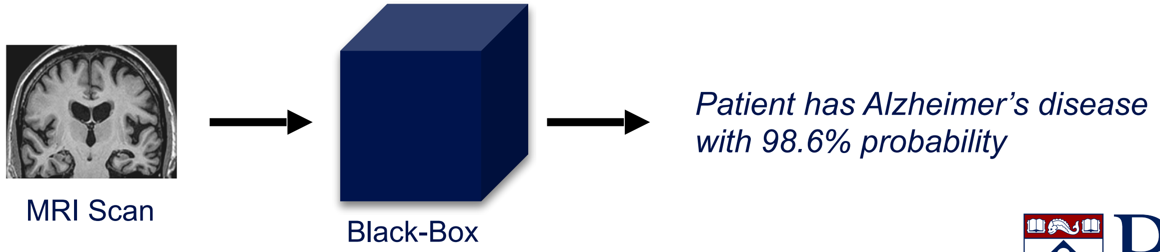
PART III: Interpretable by Design Methods (11 – 11:45 am) René Vidal (Penn)

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...

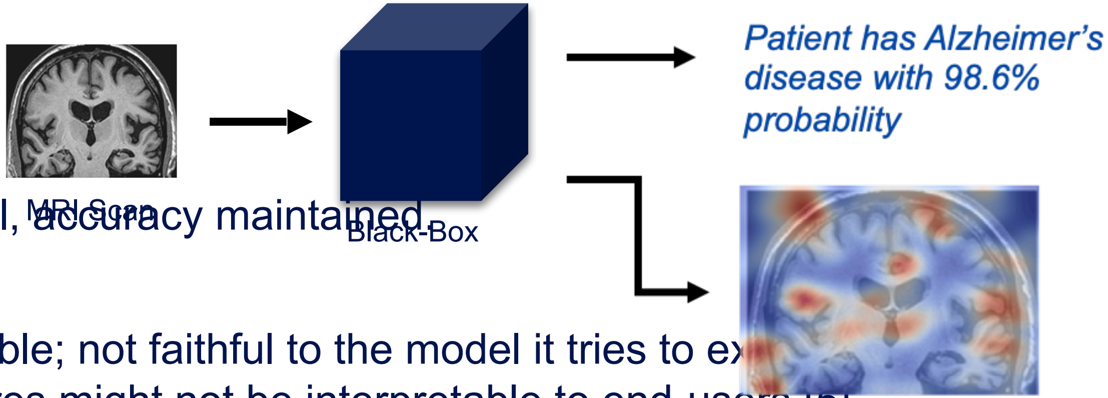


- But in reality



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 ex
  - Feature importance scores might not be interpretable to end-users [5].

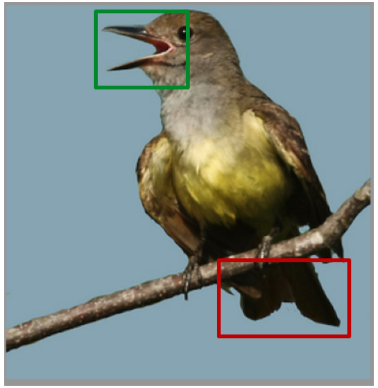
[1] Ribeiro, Singh, Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. KDD, 2016.
[2] Selvaraju, Cogswell, Das, Vedantam, Parikh, Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. ICCV 2017.
[3] Lundberg and Su-In Lee. A Unified Approach 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.

Penn
UNIVERSITY of PENNSYLVANIA

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

Penn
UNIVERSITY of PENNSYLVANIA

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.

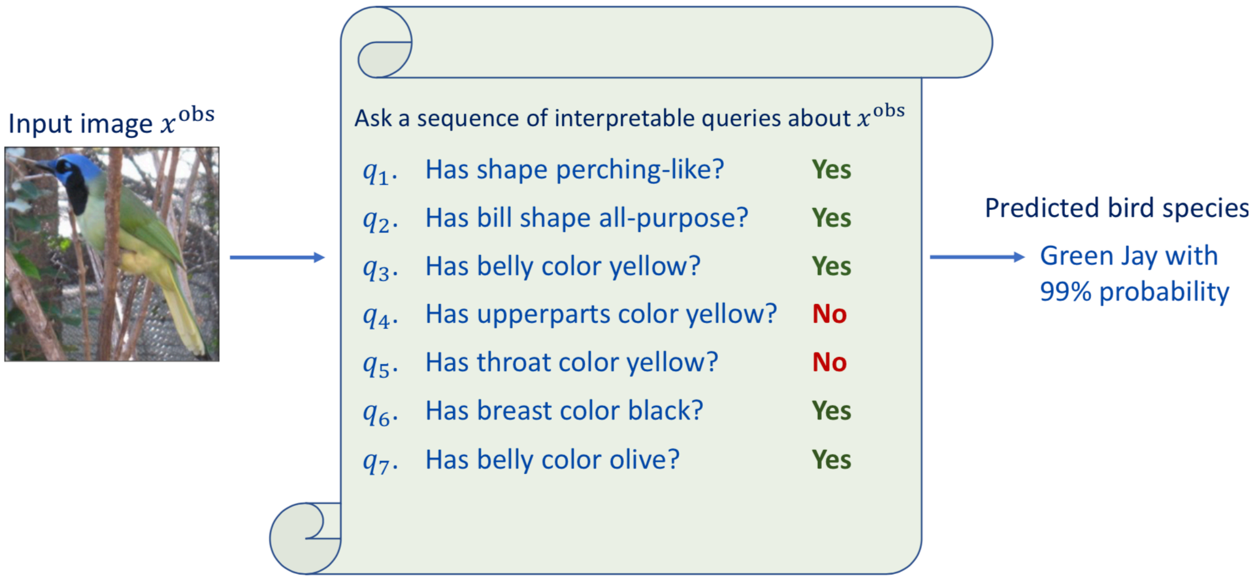
[1] Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. Concept bottleneck models. ICML, 2020.

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

- **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.



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

Ingredients Needed to Implement this Framework

- **Q1: How do we define the set of queries?**
  - Defined by domain experts [1]

- **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].

-
  - **Information Pursuit**: Select smallest number of queries that are sufficient

[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.

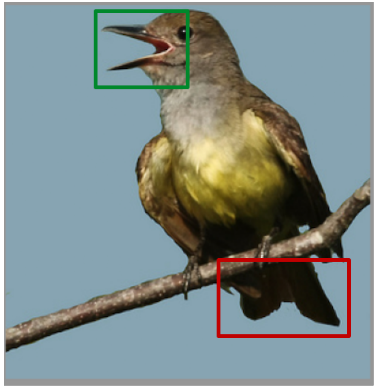
Penn
University of Pennsylvania

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



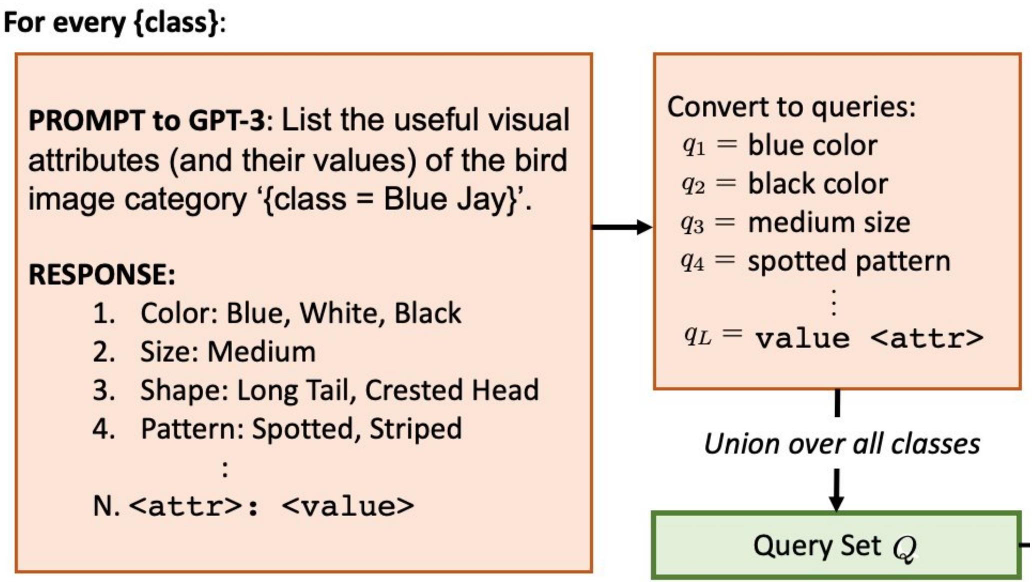
(c) **Task**: medical diagnosis
Queries: symptoms

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., Mussmann, 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.

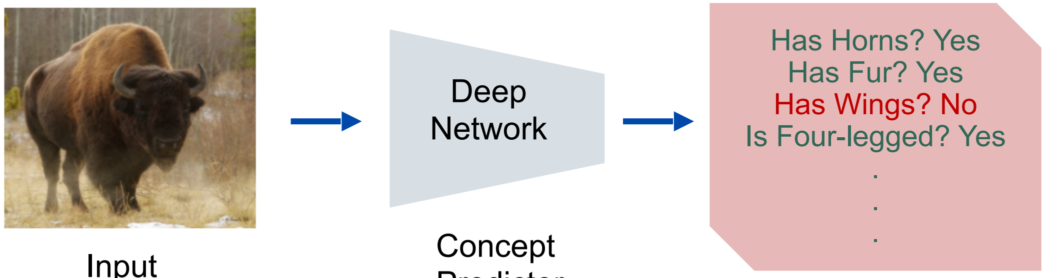


[1] Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. Concept bottleneck models. ICML, 2020.
[2] Chattopadhyay, Slocum, Haeffele, Vidal, Geman. Interpretable by design: Learning predictors by composing interpretable queries. TPAMI 2022.
[3] Oikarinen, T., Das, S., Nguyen, 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].



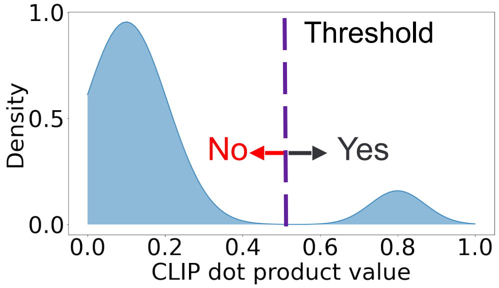
  - **Challenge:** need tons of data annotated with all concepts/attributes/facts => few datasets have such detailed annotations.
  - **Challenge:** cannot handle new queries that have not been annotated.

[1] Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. Concept bottleneck models. ICML, 2020.

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

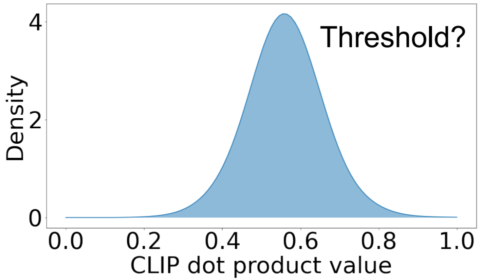
- Use Vision Language Models (VLMs) to answer queries
  - **Challenge:** State-of-the-art VLMs like Llama [1] and BLIP [2] are too slow to be used in an online manner.
  - **Challenge**: CLIP [3] is relatively light-weight, but CLIP dot products between query and image are inadequate: they are not interpretable.

<u>Desired</u> distribution of CLIP dot products



Input image x^{obs}

<u>Observed</u> 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.

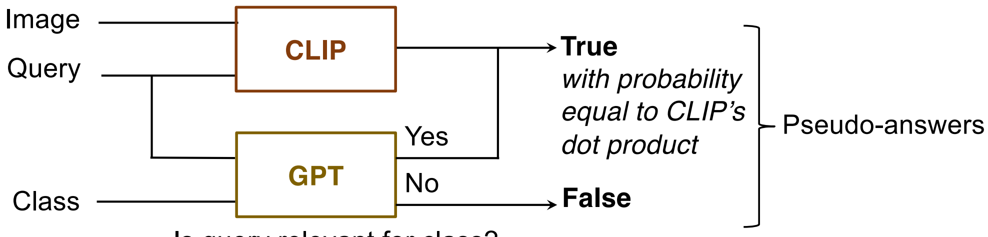
Yes! Use LLMs

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.

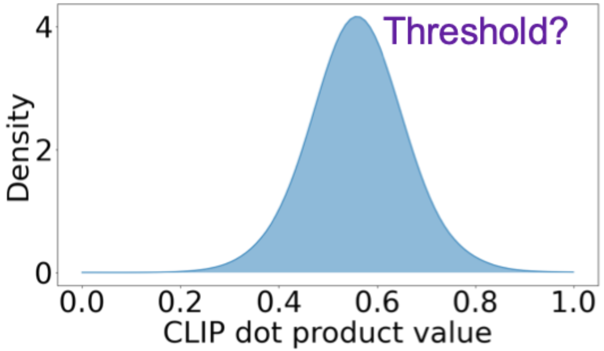


- **Concept-QA**: Train a lightweight visual question answering (VQA) system using pseudo-answers as we don't know class at test time.

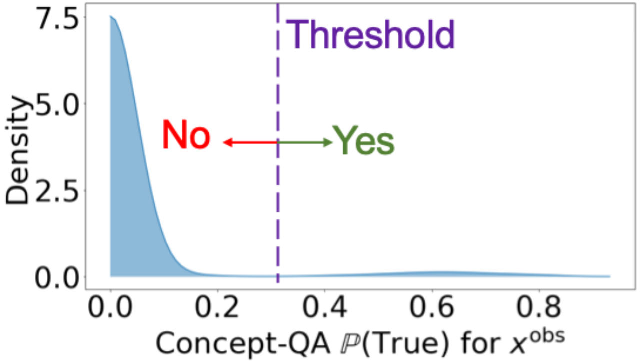
[1] Chattopadhyay, Chan, Vidal. "Bootstrapping Variational Information Pursuit with Foundation Models for Interpretable Image Classification." ICLR 2024

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 $\text{OPT}_{2.7\text{B}}$	0.55	0.31	0.76	0.18	0.53	0.35	0.73	0.13	0.86	0.07
BLIP2 ViT-g $\text{FlanT5}_{\text{XL}}$	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
  - $X \in \mathscr{X}$: input variable (data).
  - $Y \in \mathscr{Y}$: prediction variable (label).
  - $Q = \{q : \mathscr{X} \to \mathscr{A}\}$: query set.

$$X \xrightarrow{\quad} \boxed{q} \xrightarrow{\quad} a$$

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

$$(q_{1:k}, \ a_{1:k}) \xrightarrow{\quad} \boxed{\pi} \xrightarrow{\quad} q_{k+1}$$

- $\mathrm{Code}^{\pi}_{Q}(X)$: 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:

$$\min_{\pi} \mathbb{E}\left[\left|\mathrm{Code}_Q^{\pi}(X)\right|\right] \qquad \text{(Minimality)}$$

$$\text{s.t.} \quad \mathbb{P}\left(Y \middle| \mathrm{Code}_Q^{\pi}(X)\right) = \mathbb{P}(Y \mid X) \qquad \text{(Sufficiency)}$$

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

[1] H. Laurent and R. L. Rivest, "Constructing optimal binary decision trees is np-complete", Inf. Process. Lett., vol. 5, no. 1, pp. 15-17, 1976.

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:

$$q_1 = \arg\max_{q \in Q} I(q(X); Y) \qquad y_1 = \arg\max_{y \in Y} \mathbb{P}(y \mid q_1(x))$$

- Next query and prediction:

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

- Termination and prediction:

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

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

[1] Geman and Jedynak, An active testing model for tracking roads from satellite images, TPAMI, 1996.

Generative Information Pursuit (G-IP)

- Selecting the first query requires computing

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

History

- Later queries need computing

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

- **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.

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

Penn
UNIVERSITY of PENNSYLVANIA

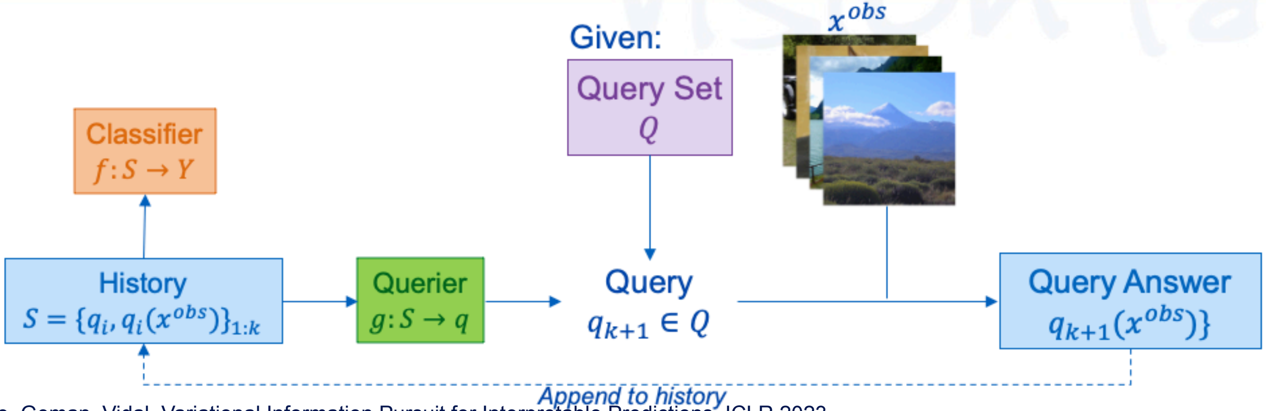
Variational Information Pursuit (V-IP)

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

$$\min_{\theta,\eta} \mathbb{E}_{X,S}\left[D_{KL}\left(\mathbb{P}(Y \mid X) \,||\, \mathbb{P}_\theta\left(Y \mid q_\eta, S\right)\right)\right]$$

$$s.t. \quad q_\eta = g_\eta(S), \quad \mathbb{P}_\theta\left(Y \mid q_\eta, S\right) = f_\theta\left(q_\eta \cup S\right)$$

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



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

IP vs Orthogonal Matching Pursuit (OMP)

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

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

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

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

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

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

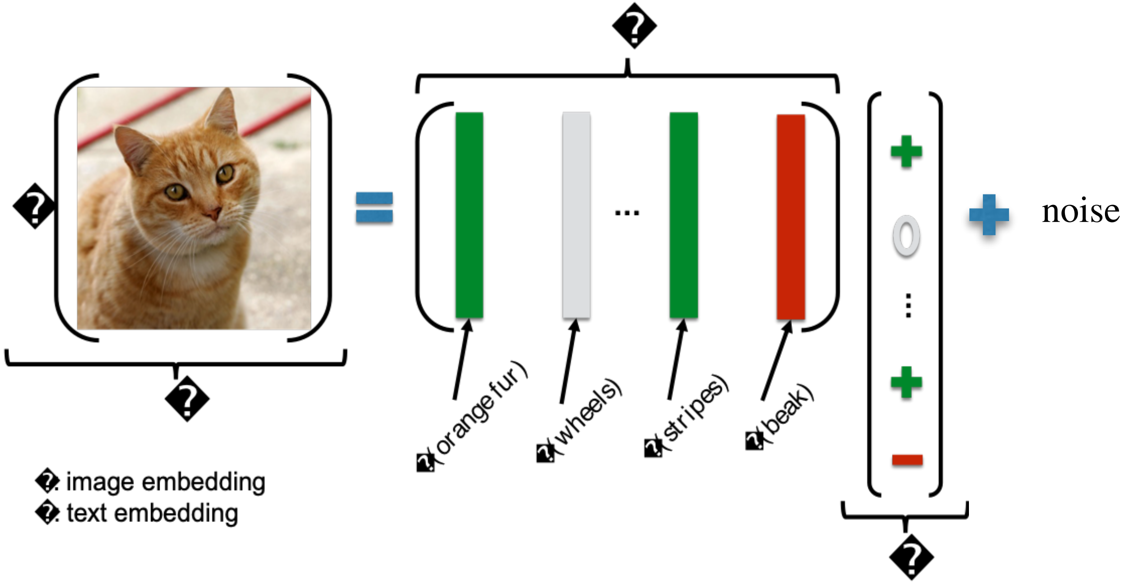
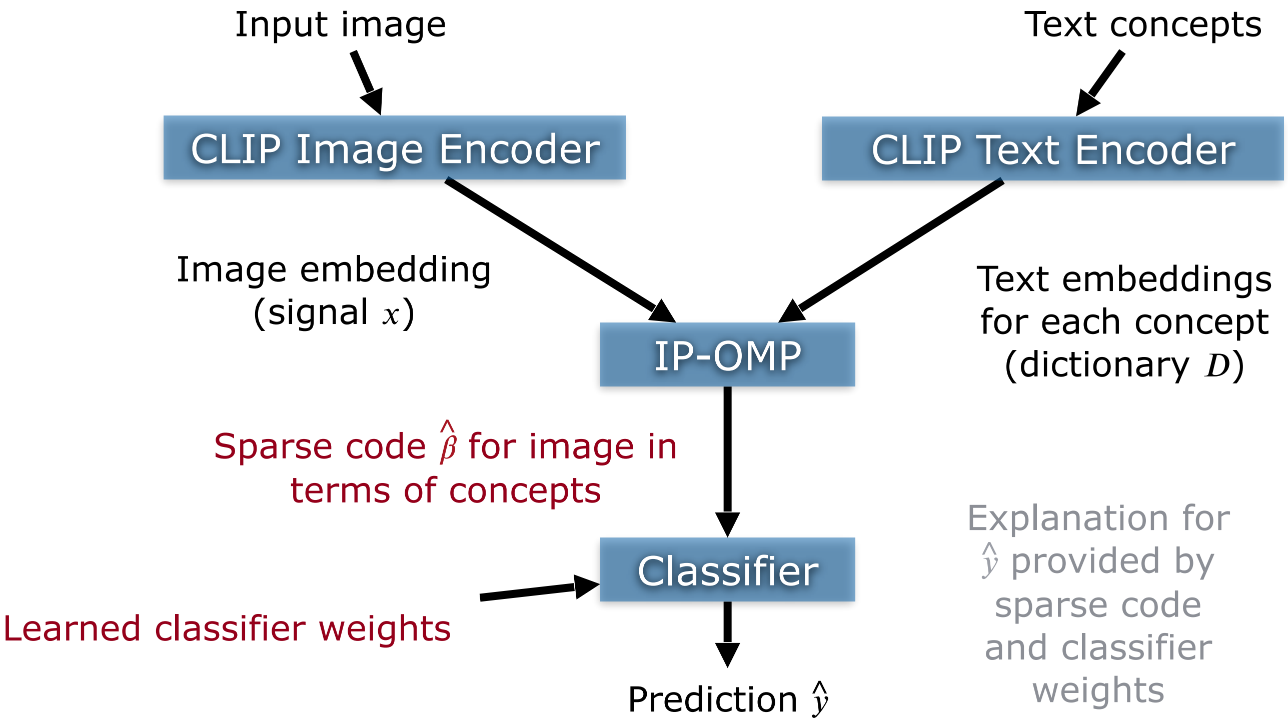


Image credit: https://en.wiktionary.org/wiki/cat#/media/File:Cat03.jpg

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

CLIP-IP-OMP: Details



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

Summary of the Information Pursuit Framework

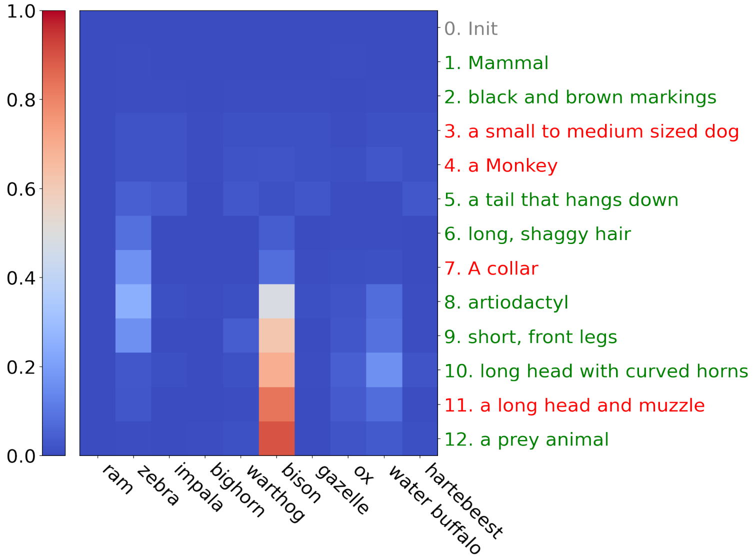
- **Q1: How do we define the set of queries?**
  - Defined by domain experts [1].
  - Defined by large language models [4].
- **Q2: Given an input and a query, how do we answer the query?**
  - Train classifiers on data annotated with query answers by task experts [1].
  - Use domain-specific pre-trained 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

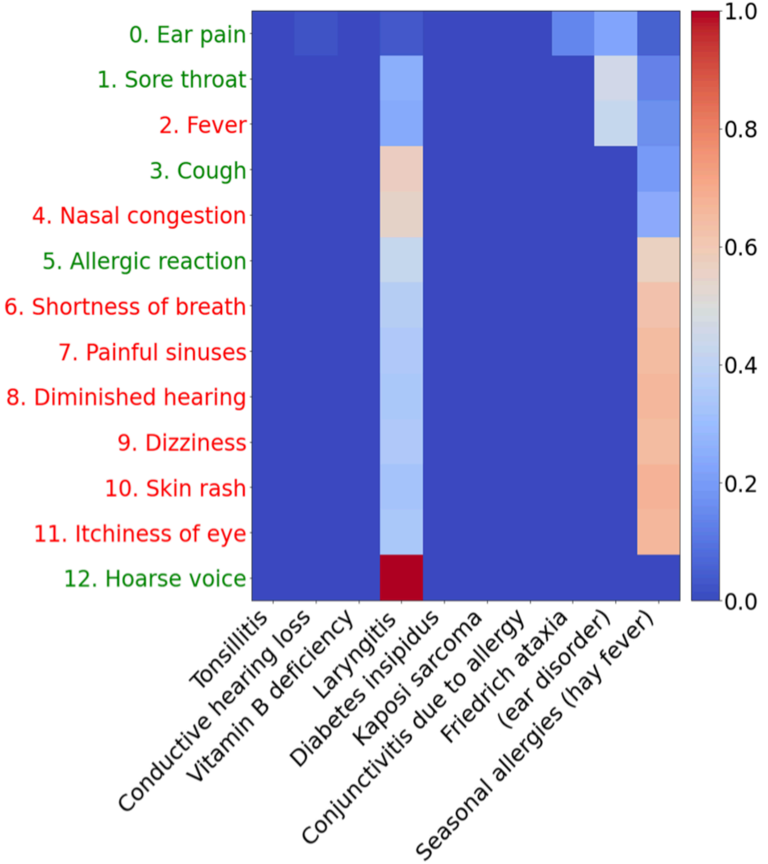
Interpretable Image Classification by V-IP

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

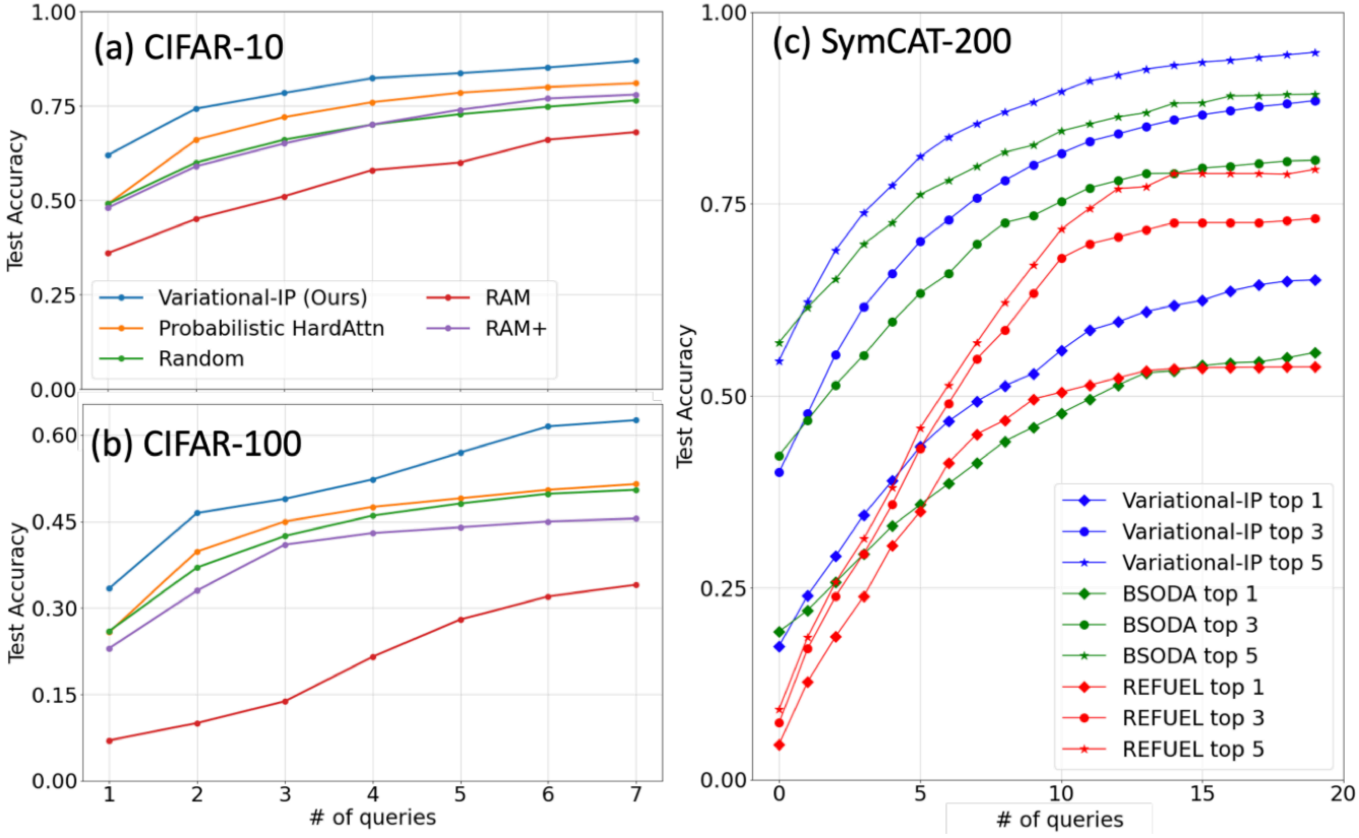


Interpretable Medical Diagnosis by VI-P

- **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.



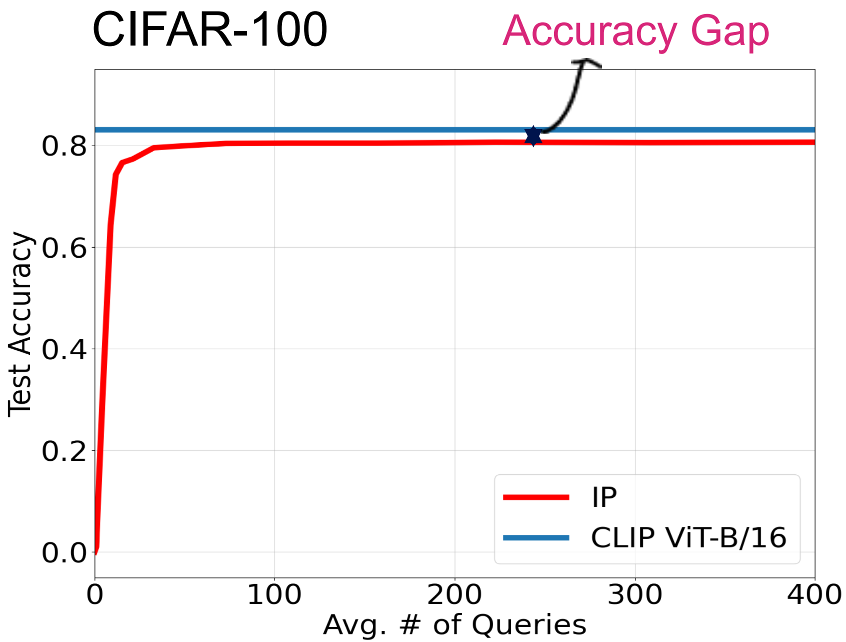
Accuracy Versus Number of Queries

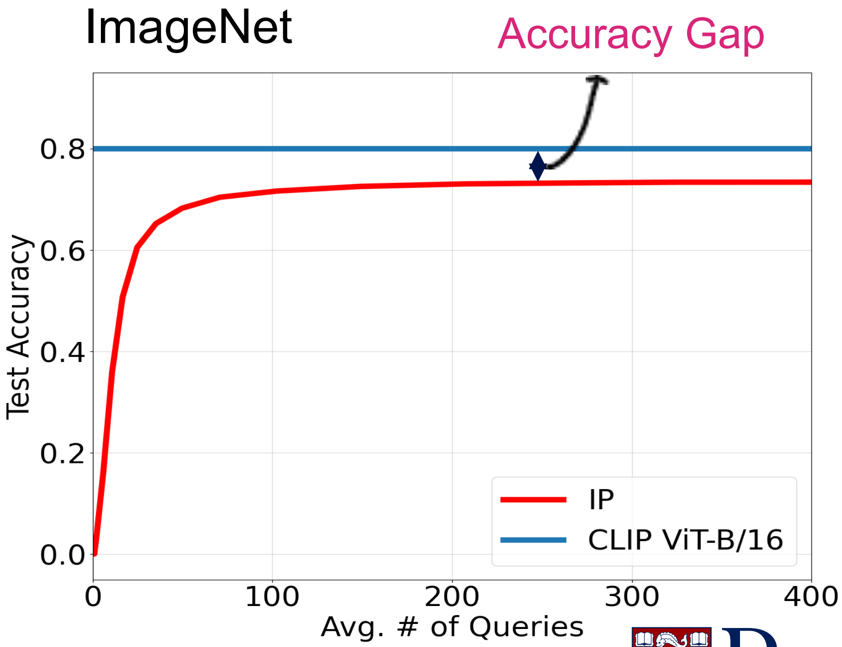


Penn
University of Pennsylvania

Accuracy-Explainability Tradeoff

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





Penn
University of Pennsylvania

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.

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.

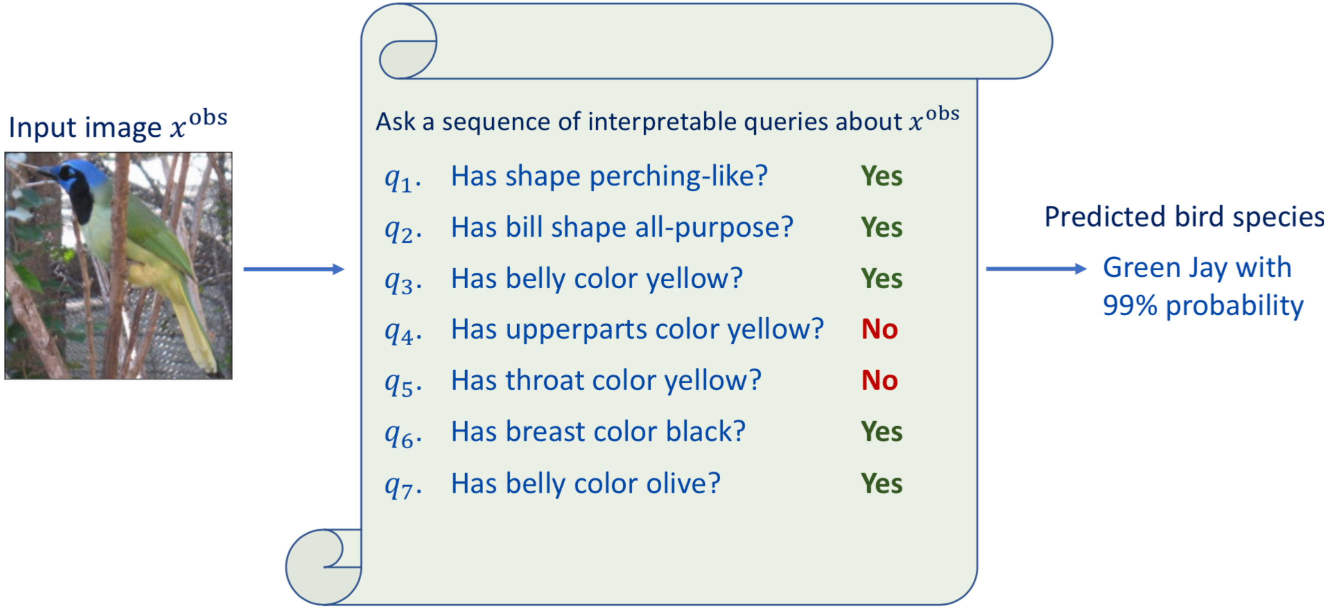
Interpretable Radiological Report Classification

- 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.



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

- Add Text here
- Add Text here

Penn
UNIVERSITY of PENNSYLVANIA