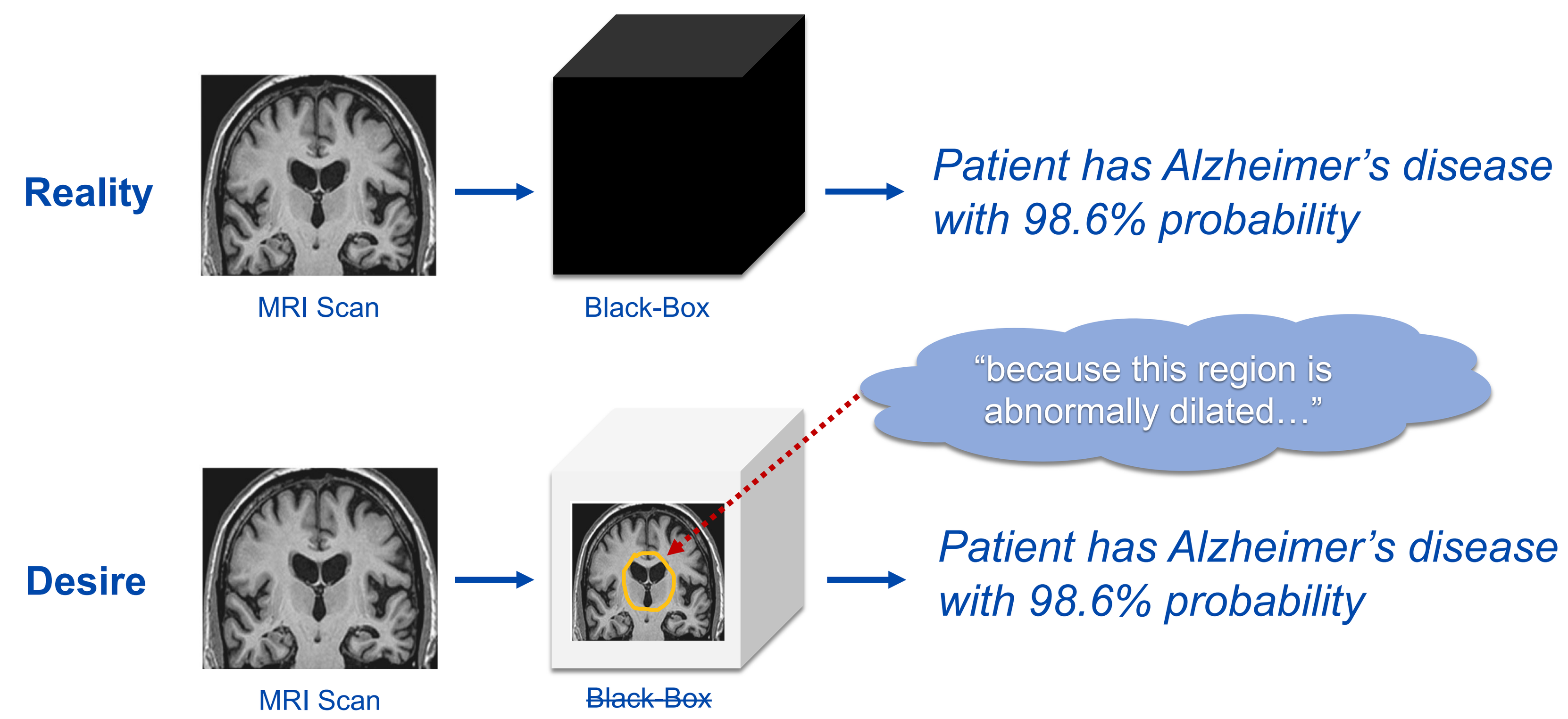
Variational Information Pursuit For Interpretable Predictions

Aditya Chattopadhyay† Kwan Ho Ryan Chan † Benjamin D. Haeffele† Donald Geman † René Vidal ‡
†Johns Hopkins University ‡ University of Pennsylvania

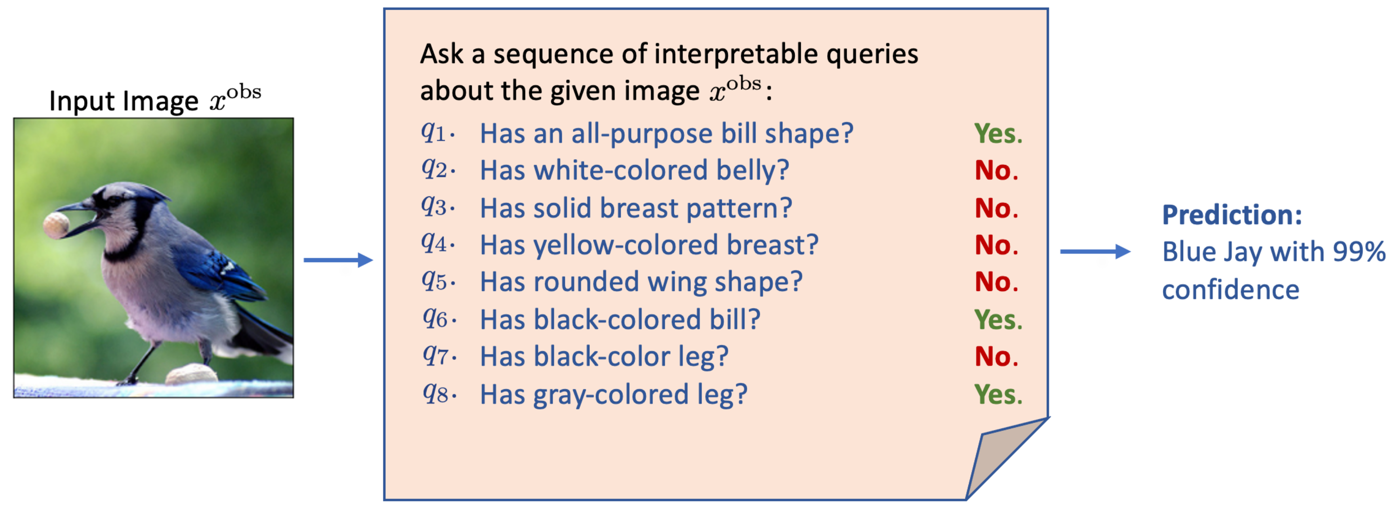
Need For Interpretable Machine Learning



Interpretable By Design

- Recent work introduced **Information Pursuit (IP)**[1] as a framework for making interpretable decisions in machine learning.
- User defines a set of queries Q, which are functions of the data interpretable to the user.
- IP sequentially and adaptively selects queries from Q, until the answers are sufficient for prediction.
  - The sequence of query-answer pairs obtained serves as an explanation for the prediction.

How Does This Make Decisions Interpretable?



- **Task:** Bird species identification.
- **Query set:** Queries about presence of visual attributes of birds.
- The prediction of a bird species is explained through a short sequence of interpretable queries, $(q_1, q_2, ..., q_9)$ derived from a user-defined query set of domain-specific attribute for birds.

Information Pursuit: Algorithm

- **Information Pursuit (IP)**: greedy strategy where queries are chosen sequentially in order of information gain[2].

IP: ALGORITHM

Queries are chosen according to observed input x^{obs}.

- First query: $q_1 = \arg\max_{q \in Q} I(q(X); Y)$
- Next query: $q_{k+1} = \arg\max_{q \in Q} I(q(X); Y \mid q_{1:k}(x^{\text{obs}}))$ ← History
- Termination: $q_{L+1} = q_{\text{STOP}}$ if $\max_{q \in Q} I(q(X); Y \mid q_{1:L}(x^{\text{obs}})) \approx 0$

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

- X, Y: random variables pertaining to data and labels respectively.
- $q(X)$: answer to query q evaluated at X.

Generative-IP: Prior Approach

- Generative-IP (G-IP)[1] carries out IP by learning a generative model for the joint distribution of query-answers and labels.
- **Limitation:** Need efficient inference and sampling techniques to compute the argmax in IP using the learnt model.

This Work: Variational Characterization Of IP

- Generative models are only a means to an end.
  - What we really want is the most informative next query, not really in actual values of mutual information.
- We show that, given history $q_{1:k}(x^{\text{obs}})$, the most informative query

$$q_{k+1} = \arg\min_{q \in Q} D_{\text{KL}}\left(P\left(Y \mid X, q_{1:k}(x^{\text{obs}})\right) \,\|\, P\left(Y \mid q(X), q_{1:k}(x^{\text{obs}})\right)\right)$$

- This motivates the following stochastic objective called **Variational Information Pursuit (V-IP)**,

$$\min_{\theta, \eta} \mathbb{E}_{X,S}[D_{\text{KL}}(P(Y \mid X) \,\|\, P_\theta(Y \mid q_\eta(X), S)]$$ ← Random History

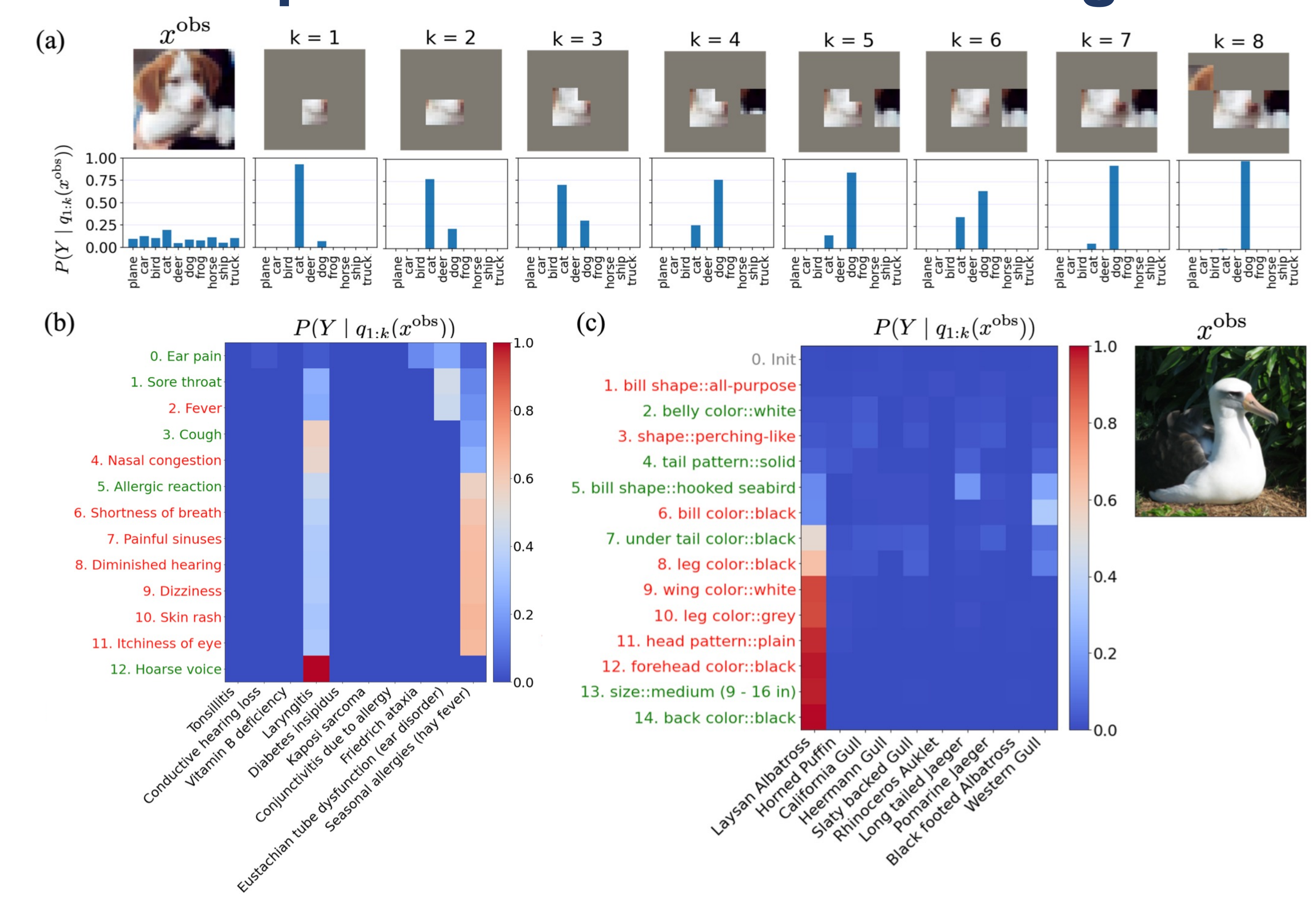
where $q_\eta := g_\eta(S)$ → V-IP querier

$P_\theta(Y \mid q_\eta(X), S) := f_\theta(\{q_\eta, q_\eta(X)\} \cup S)$ — g_η and f_θ are parameterized by deep networks.

  - The V-IP querier is a deep network that takes as input a random history (random set of query-answer pairs) and outputs a query from Q.

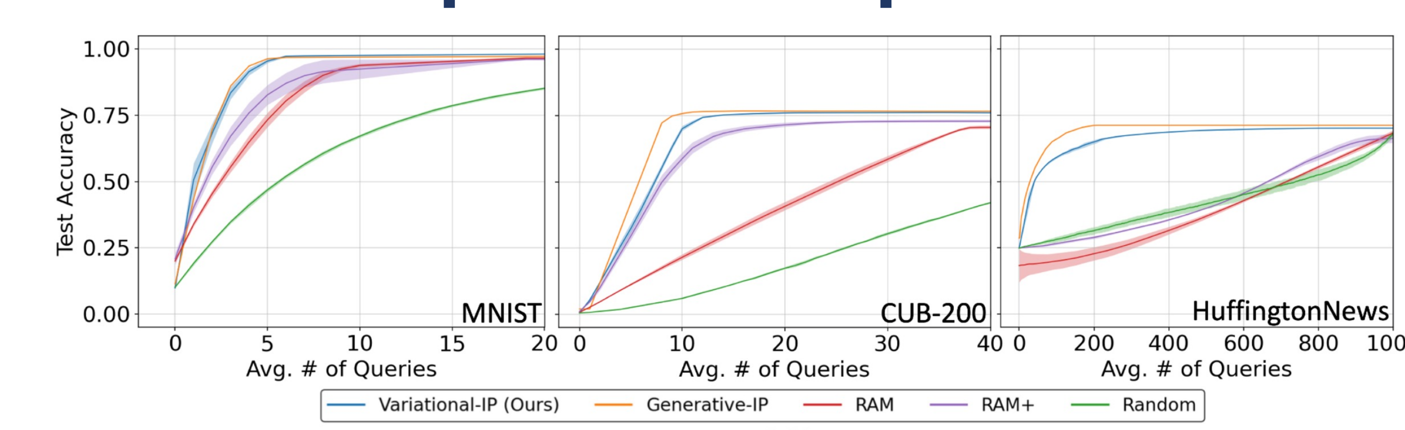
Theorem (Informal): *The optimal querier to the V-IP objective is the function that maps any given history (set of query-answer pairs) to the most informative next query about* Y.

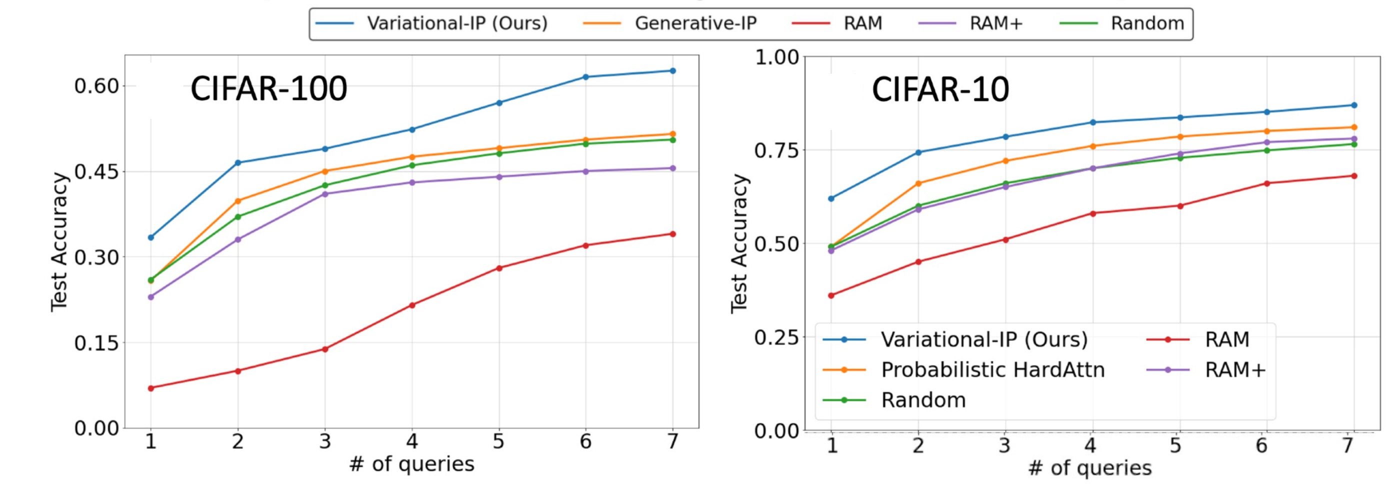
Interpretable Predictions Using V-IP



Each figure illustrates one run of the V-IP algorithm, depicting the sequence of query-answer chains obtained for a randomly chosen test sample from the (a) CIFAR-10, (b) SymCAT-200, and (c) CUB-200 datasets respectively.

Empirical Comparisons





- On datasets like MNIST, where good generative models are available, G-IP performs *slightly* better than V-IP in terms of avg. # queries needed to reach a certain level of test accuracy.
- On complex datasets like RGB images (CIFAR-{10,100}), V-IP outshines all baselines.
- V-IP inference is **10-100x faster** than G-IP in all cases!

References

1. Chattopadhyay, Aditya, et al., "Interpretable by design: Learning predictors by composing interpretable queries", TPAMI, 2022.
2. Geman and Jedynak, "An active testing model for tracking roads from satellite images", TPAMI, 1996.

Supported by ARO under the MURi Research Initiative W911NF-17-1-0304, NSF Grant 2031985, NSF GRFP Grant DGE2139757, and by Simons Foundation Mathematical and Scientific Foundations of Deep Learning (MoDL) grant 135615.