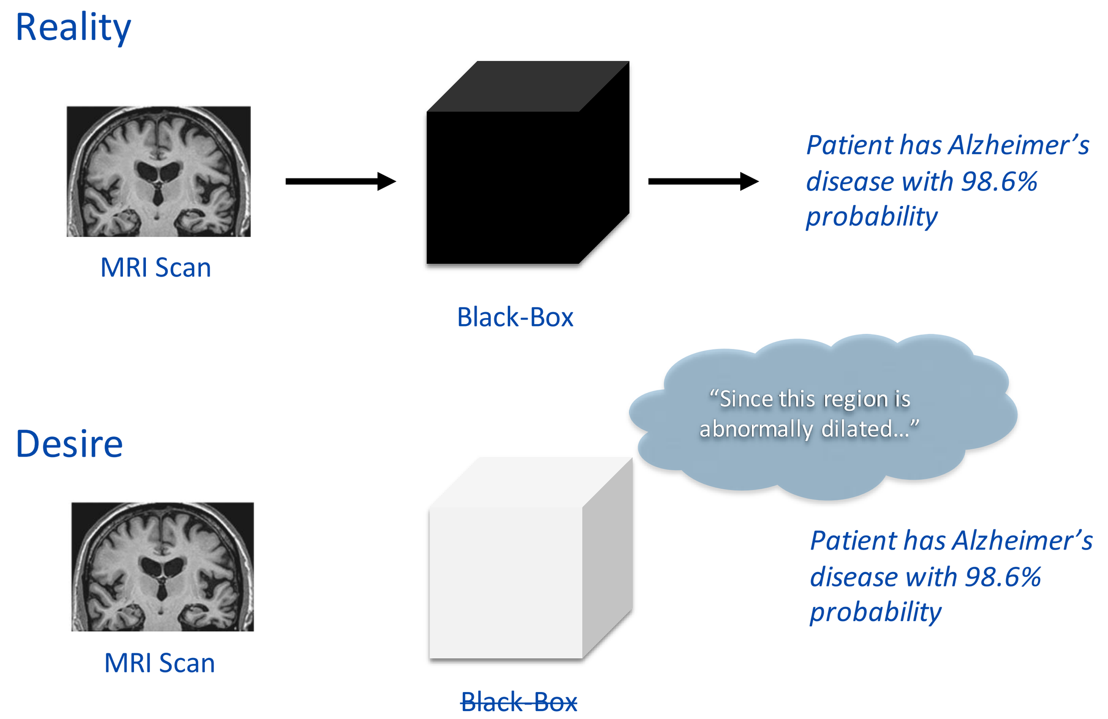
Interpretable by Design: Learning Predictors by Composing Interpretable Queries

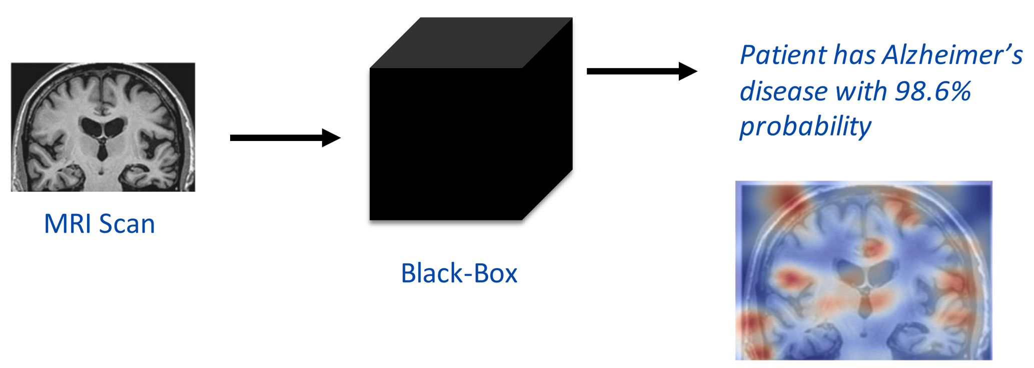
Aditya Chattopadhyay, Stewart Slocum, Benjamin D. Haeffele, René Vidal, Donald Geman

JOHNS HOPKINS MATHEMATICAL INSTITUTE for DATA SCIENCE

Interpretability Crisis



Prior Work: Post-Hoc interpretability



- Current trend is to interpret black-box models post-hoc.
- **The Good:** No need to retrain model, accuracy maintained.
- **The Bad:**
  - Explanations generated are unreliable; not faithful to the model it tries to explain.[1]
  - Salient parts of image might not be most informative to end-users.[2]

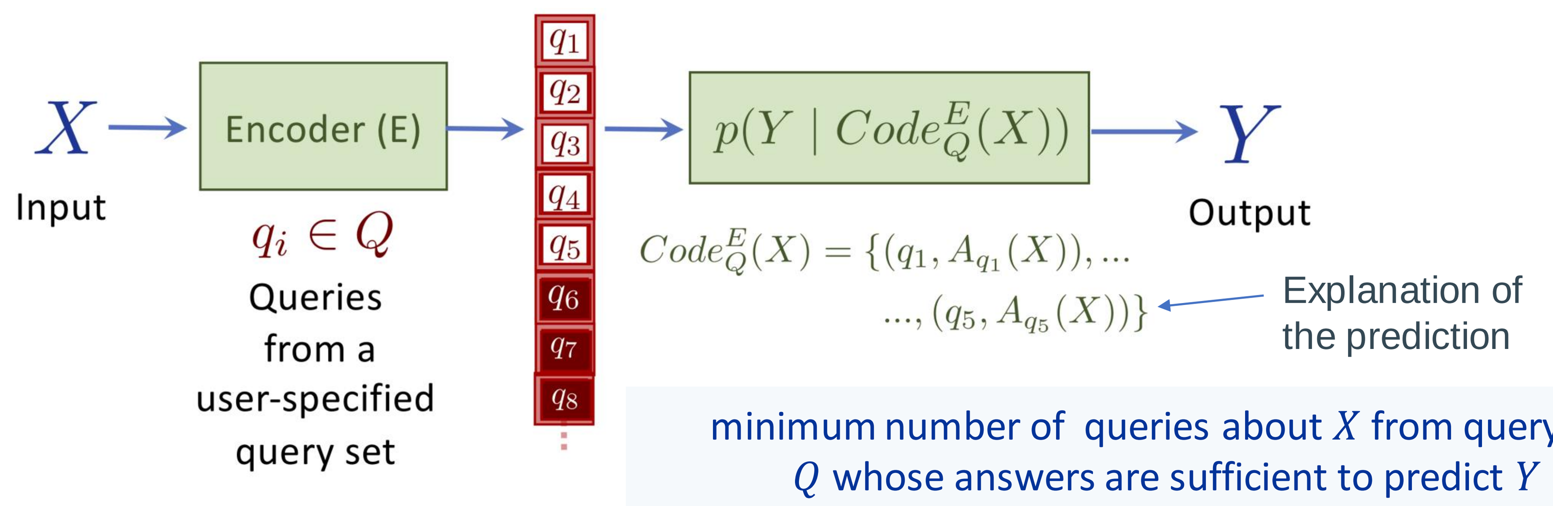
Models should be Interpretable by Design

- Learning models that are interpretable by design solves all the shortcomings of post-hoc interpretability methods. However, there are key challenges.
- **Challenge 1:** *An ideal* interpretable explanation of a model's prediction is highly *task-dependent* and *end-user* dependent.
  - A model for image classification is often considered interpretable if its decision can be explained in terms of patterns occurring in salient parts of the image.
  - In a medical task explanations in terms of causality and mechanism could be desired
- **Challenge 2:** Desirable interpretations are often *compositional* and can be constructed and explained from a set of *elementary units*. For instance, words, parts of an image, or domain-specific concepts.
- **Challenge 3:** Following the principle of Occam's razor we would like the explanations to be composed of the smallest number of queries.

Proposed Framework

- We propose the concept of a query set Q which is a set of user-defined task-dependent functions of data. Each with a specific interpretation to the user.
- We propose an information-theoretic framework to compose these queries to form concise explanations of model predictions.

Idea: Given Q, propose the following optimization problem.

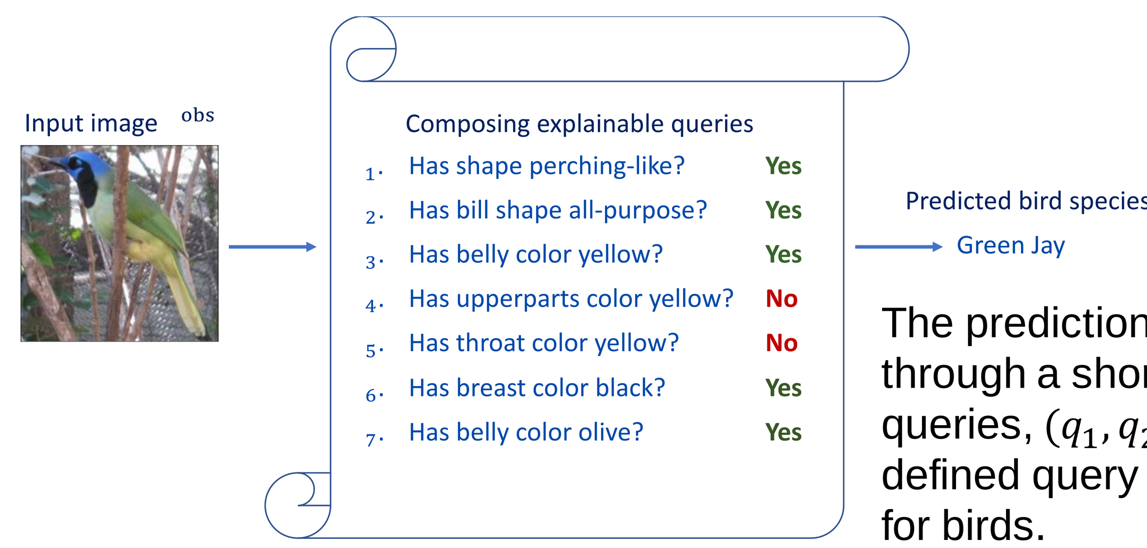


minimum number of queries about X from query set Q whose answers are sufficient to predict Y

Minimal: $$\min_E \mathbb{E}_X [|Code^E_Q(X)|] \quad (1)$$

Sufficient: s.t. $p(y \mid Code^E_Q(x)) = p(y \mid x) \; \forall x \in \mathcal{X}, y \in \mathcal{Y}$

How does this make decisions interpretable?



The prediction of a bird species is explained through a short sequence of interpretable queries, $(q_1, q_2, \ldots, q_7)$, derived from a user-defined query set of domain-specific attribute for birds.

Information Pursuit: a greedy approximation

- Unfortunately solving the objective in (1) is NP-Hard. We propose to use a greedy approximation called Information Pursuit (IP).[3]
- IP selects queries in order of information gain.

Definition: IP Encoder

Queries are chosen according to observed x.

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

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

Computational Challenge: How do we compute the mutual information terms required for carrying out IP on high-dimensional data like images?

Making IP tractable with Deep Generative Models

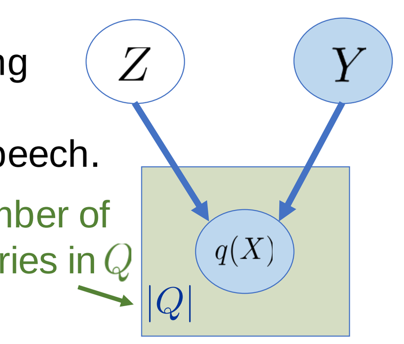
Computational Challenges

- Selecting the first query requires computing $I(q(X); Y)$
  - Need a joint distribution of $q(X)$ and Y.
- Later queries require computing $I(q(X); Y \mid q_{1:k}(x))$ (History)
  - Need a joint distribution of $(q(X), Y)$ given History.
  - As histories get longer, we run out of samples that match History.
- The above two problems need to be solved $\forall q \in Q$, which scales linearly with the number of queries.

Modelling Assumption: Assume query answers are conditionally independent given target variable Y and "some" latent variable Z.

Examples,

- Z = pose and lighting conditions.
- Z = phonemes in speech.



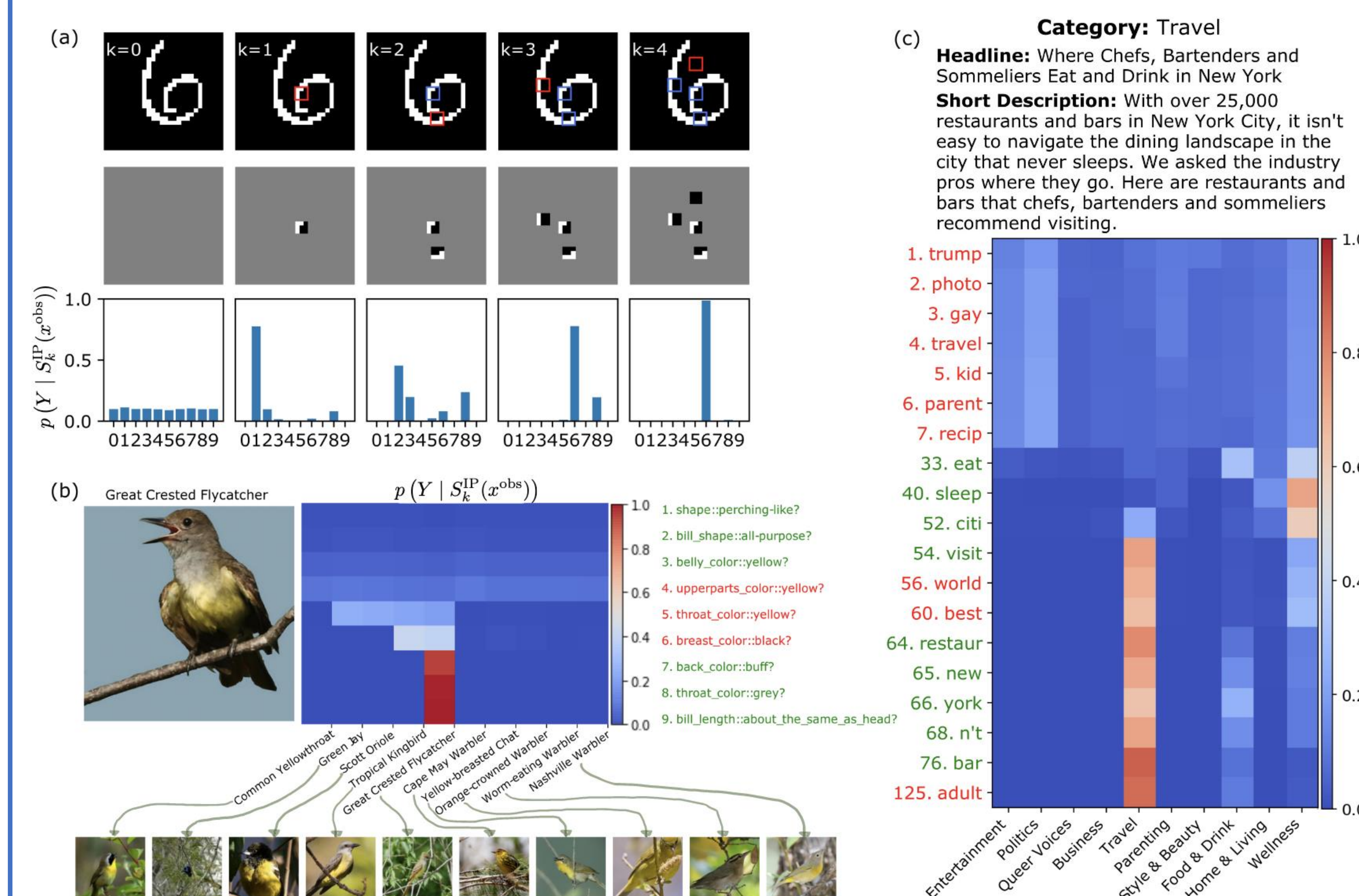
$Q(X) = \{q(x) : q \in Q\}$

$$p(Q(X), Z, Y) = \prod_q p(q(X) \mid Z, Y) p(Z) p(Y)$$

Proposed Solution

- We learn this joint distribution of all query-answers $Q(X)$ and labels Y using a Variational Autoencoder.
- Our modelling assumption of conditional independence makes estimating $I(q(X); Y \mid q_{1:k}(x))$ tractable using Markov Chain Monte Carlo (MCMC) sampling.
  - In particular, we employ the Unadjusted Langevin Algorithm (ULA) to carry out MCMC and get samples from the required posterior distributions.

Experiments: IP in action



References

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.*
2. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215.
3. D. Geman and B. Jedynak, "An active testing model for tracking roads in satellite images," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 1, pp. 1–14, 1996.