Discovering Unknown Unknowns of Predictive Models

Hima Lakkaraju*, Ece Kamar+, Rich Caruana+, Eric Horvitz+

*Stanford University, +Microsoft Research
Unknown Unknowns

- Unknown unknowns of a predictive model are data points for which the model makes a confident prediction but is incorrect.

- They arise due to mismatch between training and test data distributions.

- Goal:
  - The predictive model is a black box
  - No access to the training data
Unknown Unknowns

- Unknown unknowns of a predictive model are data points for which the model makes a confident prediction but is incorrect.

- They arise due to mismatch between training and test data distributions.
Unknown Unknowns

- Unknown unknowns of a predictive model are data points for which the model makes a confident prediction but is incorrect.

- They arise due to mismatch between training and test data distributions.

- **Goal:** Discover unknown unknowns assuming
  - The predictive model is a black box
  - No access to the training data
Example

Training Data
- Dogs
- Cats

Learning Algorithm

Predictive Model

Cat
(Conf. = 0.96)
How do we detect *unknown unknowns* of this black-box model?
Inputs labeled as cats “confidently” by the model.
Our Framework

Inputs labeled as cats “confidently” by the model

Interpretable & Meaningful Partitions

White Dogs
White Cats
Brown Cats
Brown Dogs
Interpretable & Meaningful Partitions

Inputs labeled as cats “confidently” by the model

White Dogs

White Cats

Brown Cats

Brown Dogs

K-armed Bandit
Our Framework

Interpretable & Meaningful Partitions

White Dogs

White Cats

Brown Cats

Brown Dogs

K-armed Bandit

Identified Unknown Unknowns