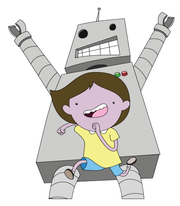
Explainable Artificial Intelligence via Bayesian Teaching

Scott Cheng-Hsin Yang & Patrick Shafto
Department of Mathematics & Computer Science, Rutgers University–Newark

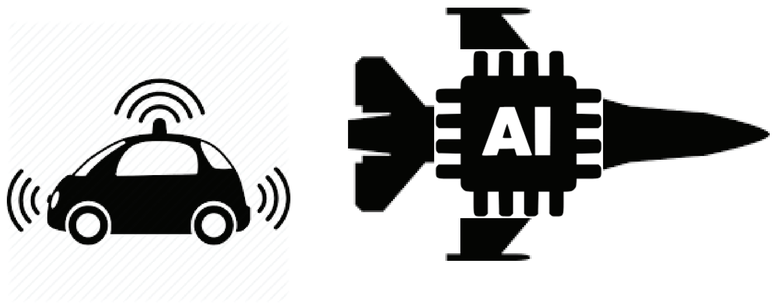
RUTGERS UNIVERSITY | NEWARK



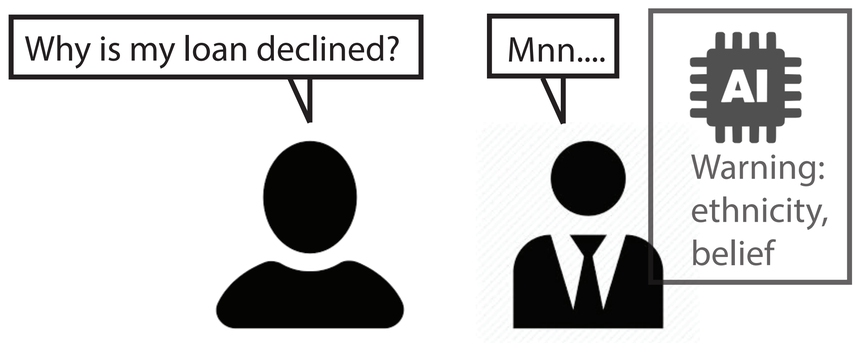
shaftolab.com

The problem of Explainable Artificial Intelligence

Why did the model predict this as a rifle?
Why did it not predict this as other things?



Can I sign my name off this autonomous vehicle and aircraft for operation Z?



EU regulations on algorithmic decision-making and a "right to explanation" (Goodman & Flaxman, 2016).

Bayesian Teaching

Popular approaches (XAI with more AI):

Use interpretable models to explain opaque models: visualization, shallow model, logic model, tree model, causal model, etc.

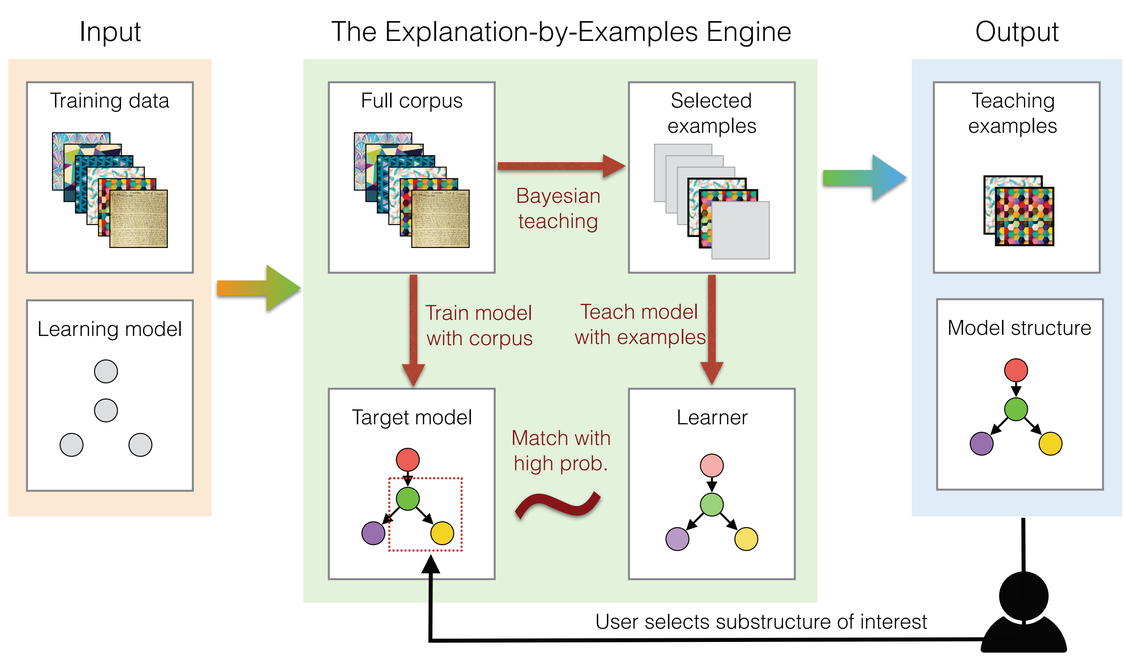
Explain with interpretable modality: attention map, text generation.

The Bayesian teaching approach:

Explanation as the inverse of model learning.

A model-agnostic system that samples data subsets to explain model inferences to a domain (but not necessary technical) expert.

Use data, the natural common language between users and models, to explain the model's inferences.



$$P_T(x|\Theta) \propto P_L(\Theta|x)$$

Supervised learning
x examples and labels
Θ parameters, boundaries

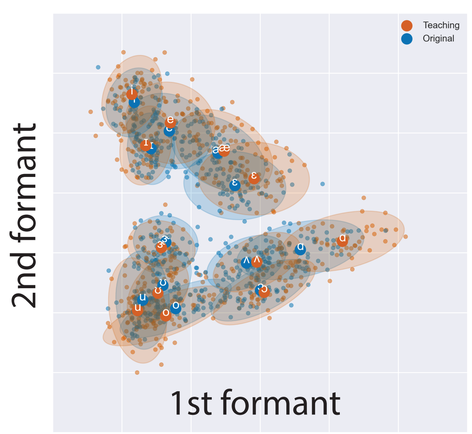
Unsupervised learning
x examples
Θ latent structures

Reinforcement learning
x actions, observations, rewards
Θ learned policy & world model

Deep learning
x training examples
Θ network weights

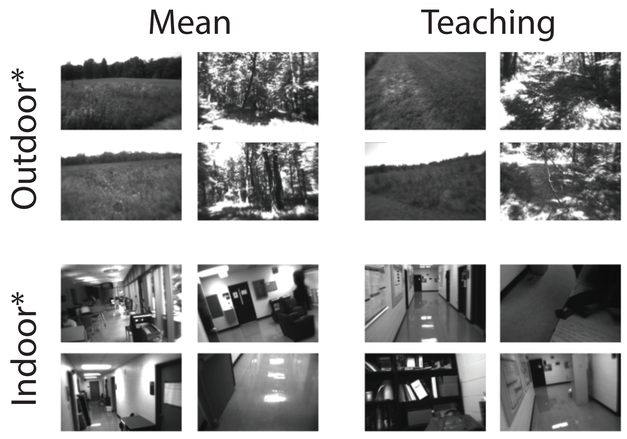
Empirical Support

Infant-directed speech (IDS)



IDS is different from adult-directed speech and is consistent with what Bayesian teaching would produce to teach adult phonetic categories (Eaves et al. 2016).

Teaching the models' (IGMM) categories



Teaching examples help people learn the categories extracted by the unsupervised model better than examples that convey just the category means (Schweinhart et al. 2016).

Teaching the models' (pLDA) category (angry)

The best examples are better than random examples at helping people learn the supervised model's predictions. They also span category-irrelevant diversity and avoid highly atypical examples.

Challenges

Extend to **more expressive models**, such as deep probabilistic models and probabilistic programming. This is hard because inference is still hard.

Give **more expressive explanation** by giving multiple examples from large datasets conditioned on substructures of expressive models. The sampling of Bayesian teaching is hard because the number of subsets that can be chosen explodes combinatorially and the sampling landscape is highly multi-modal.

Design and implement an intuitive **user interphase** that supports different types of data and interactions to help user explore the model.

Test whether an XAI framework can help user (DARPA XAI BAA, 2016): 1) predict the model's predictions, 2) understand why the model makes predictions the way they do, 3) understand when the model would fail, 4) develop trust toward the model, and 5) know how to correct the model.

Related Work

Pedagogical reasoning (Shafto & Goodman 2008; Shafto et al. 2014): alternately iterate Bayesian learning and teaching until convergence.

Machine teaching (Zhu 2013, 2015): an optimization framework for finding the subset of data that makes the learning model's induced inference *closest* to the target model.

Coreset (Feldman 2010; Bachem et al. 2017): a computational geometry framework for finding subsets of data that make the induced inference *close enough* to the target model.

Algorithmic teaching (Zilles et al. 2008; Doliwa et al. 2014): teaching set and teaching dimension in deterministic cooperative setting.

Inverse reinforcement learning applied to education (Libby et al. 2016; Rafferty et al. 2016): guided inquiry, personalization, strategy planning.