Logistics: Project

- Project presentation
  - Dec 7/9 during lectures
  - Everyone is expected to attend both lectures
  - 10 minutes for presentation + 1~2 minutes for QA and transition

- Project reports
  - Due: Dec 12 (no late submissions)
  - Up to 6 pages (plus additional pages for only references/citations)
  - No strict format requirements
    - You are encouraged to use standard templates

- Check Piazza posts for details/updates

Assignment 4

- No class on Nov 23, Tuesday (happy thanksgiving!)

- Instead, I'll give a list of talks/tutorials relevant to this course.
  - Choose one of them, watch the content, and write a report
  - The report needs to be no less than 2 pages, with any reasonable format
  - The report will serve as your assignment 4

- Check the list of talks on Piazza

Peer Review

- Please submit the peer review by 6pm

Lecture 22
Interpretable/Intelligible Machine Learning

Instructor: Chien-Ju (CJ) Ho

Who Need Explanations and Why?

- Developers:
  - Debug the machine learning models and improve robustness

- Users:
  - More likely to trust the models and act based on the predictions

- Government:
  - By law, many consequential decisions (medical, financial, etc) need to be explainable

- Society:
  - Help uncover the biases hidden underneath the predictions

How to Achieve Interpretability?

**Before building
any model**

[Kim and Doshi-Velez. 2017]

How to Achieve Interpretability?
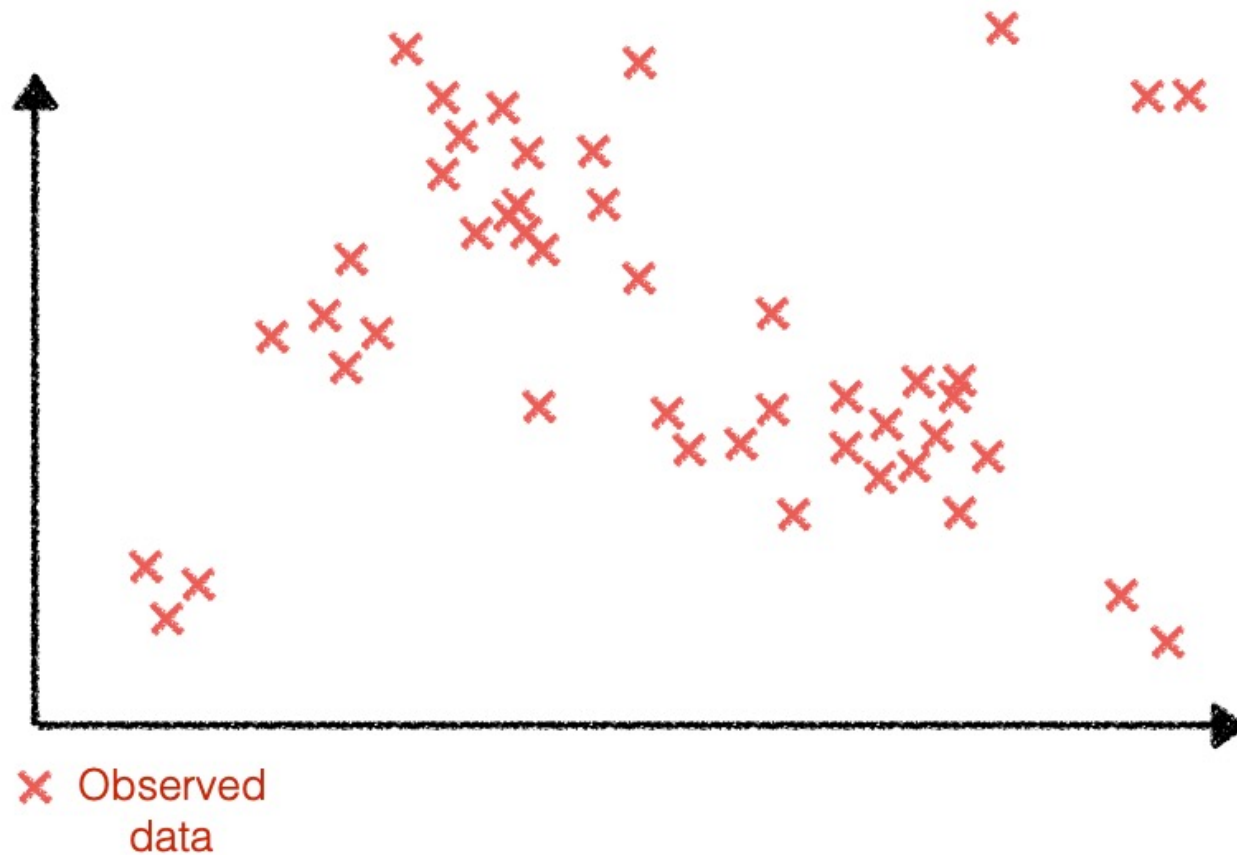
**Before building
any model**

**Building
a new model**

[Kim and Doshi-Velez. 2017]

How to Achieve Interpretability?

**Before building
any model**
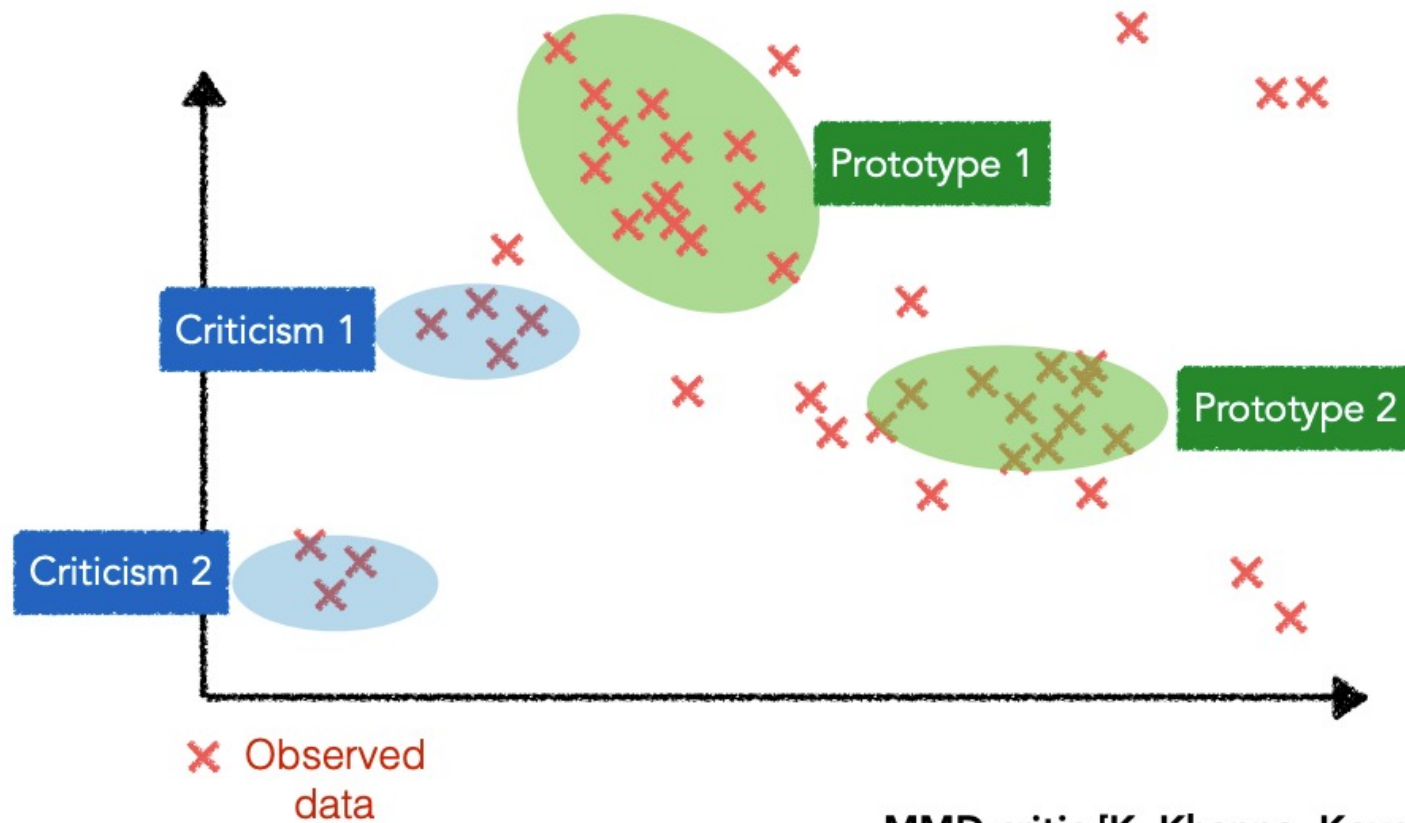
**Building
a new model**

**After
building a model**

[Kim and Doshi-Velez. 2017]

Before Building a Model

- Exploratory Data Analysis

✗ Observed
data

Before Building a Model

- Exploratory Data Analysis

MMD-critic [K. Khanna, Koyejo '16]
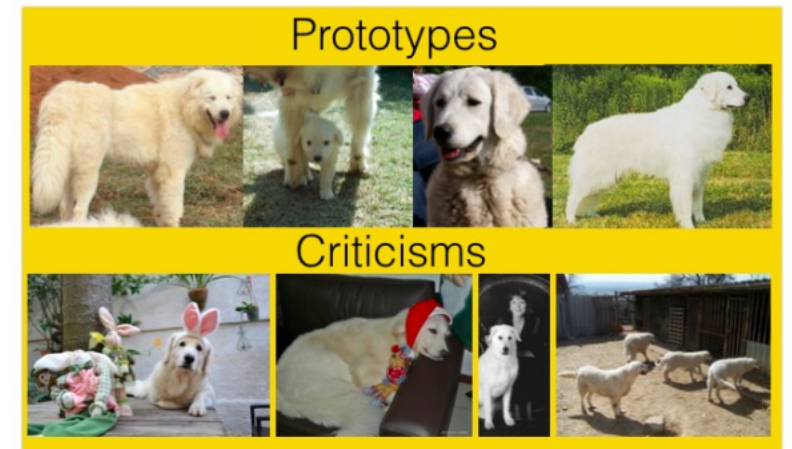
Before Building a Model

- Datasheets for datasets [Gebru et al. 2018]

Datasheet for 🛶 (QuAC)

1 Motivation for Datasheet Creation

Why was the dataset created?
We collected 🛶 to facilitate designing and evaluating models for information-seeking dialog, a sequential QA task that involves resolving coreferences, dealing with unanswerable questions, and leveraging world knowledge.

Has the dataset been used already?
All papers reporting on 🛶 are required to submit their results to http://quac.ai.

Who funded the dataset?
🛶 was co-funded by the Allen Institute of Artificial Intelligence and the DARPA CwC program through ARO (W911NF-15-1-0543).

2 Dataset Composition

What are the instances?
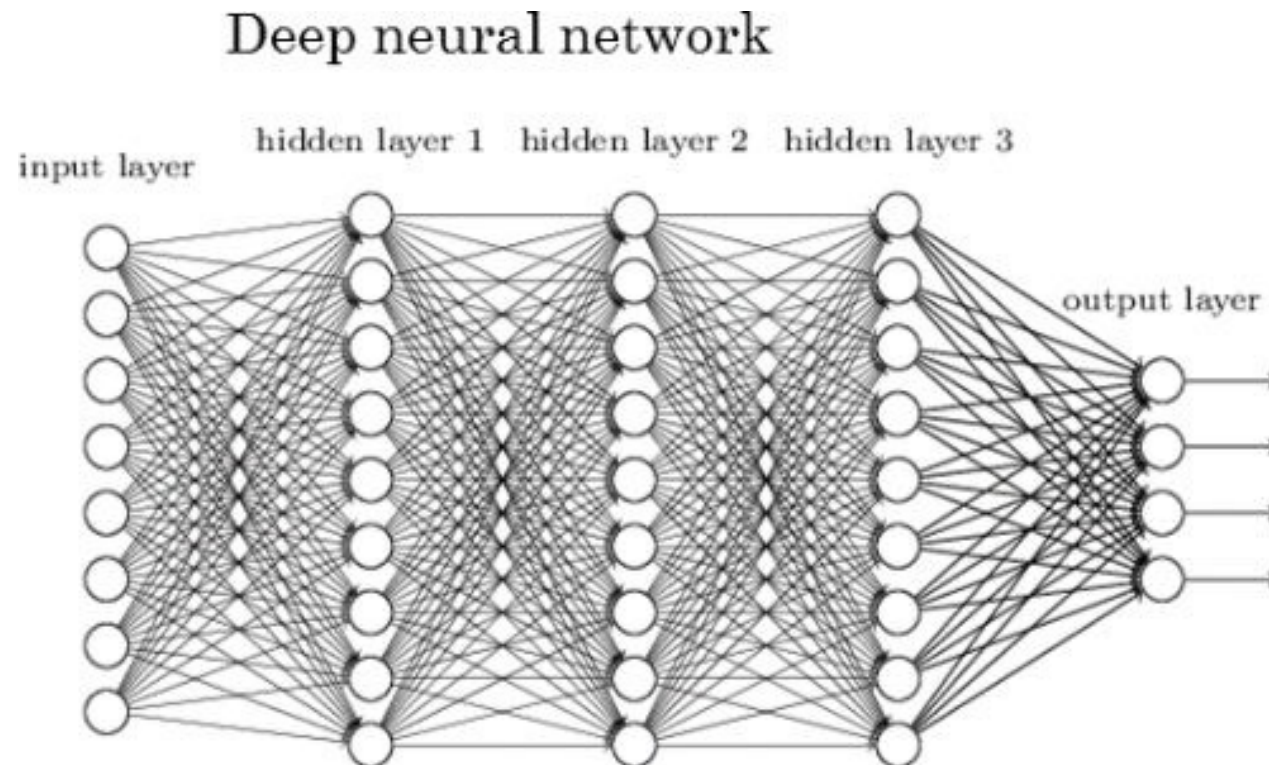The core problem involves predicting a text span

Are there recommended data splits or evaluation measures?
The release comes with a train/dev split such that there is no overlap in sections across splits. Furthermore, the dev and test sets only include one dialog per section, in contrast to the training set which can have multiple dialogs per section. Dev and test instances come with five reference answers instead of just one as in the training set; we obtain the extra references to improve the reliability of our evaluations, as questions can have multiple valid answer spans. The test set is not publicly available; instead, researchers must submit their models to the 🛶 leaderboard at http://quac.ai, which will run the model on our hidden test set.

We provide an official evaluation script for 🛶 used by our leaderboard for test set evaluation. The script computes two metrics: word-level F1 and human equivalence (HEQ). If a particular in-
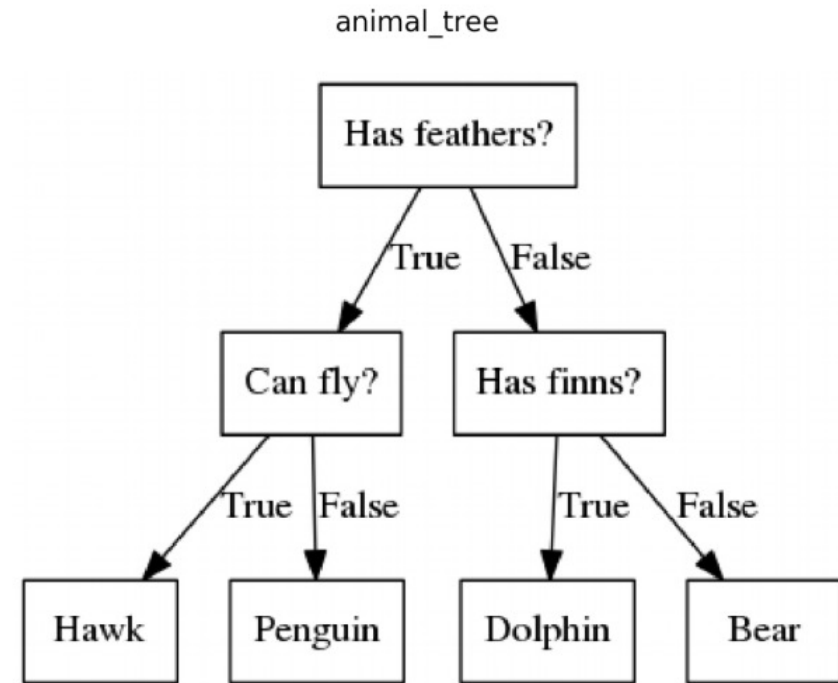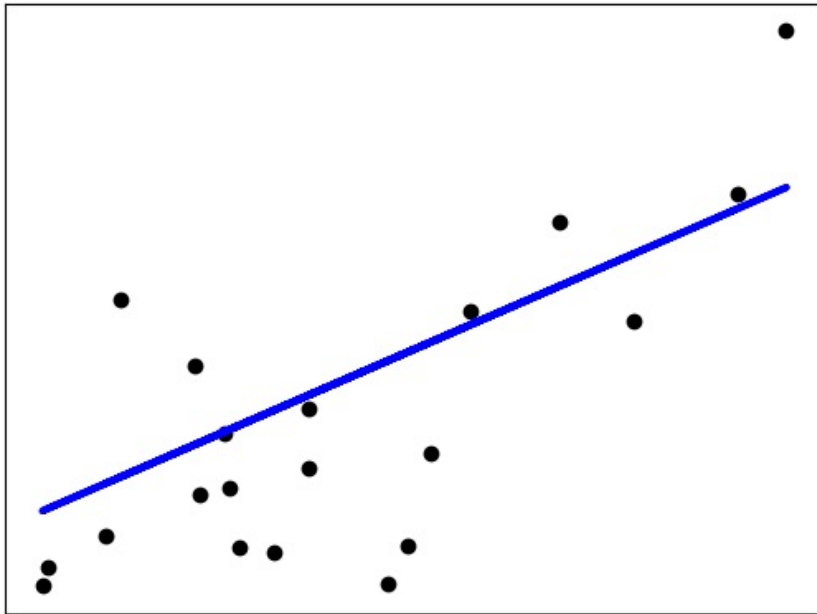
Develop an "explainable" ML model

- Black-box approaches are hard to explain

Deep neural network

input layer   hidden layer 1   hidden layer 2   hidden layer 3

output layer

Develop an "Explainable" ML Model

- Black-box approaches are hard to explain
- Classical white-box approaches: linear models, decision trees,…
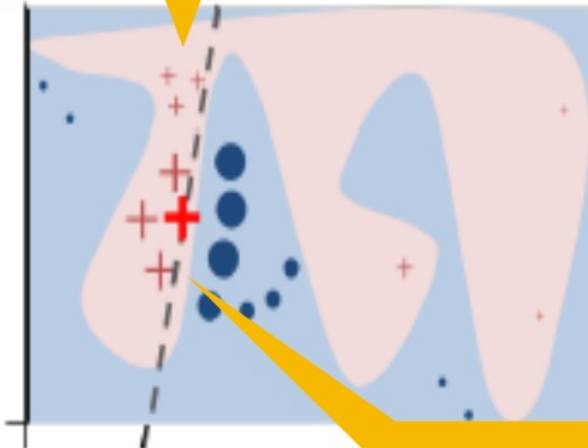

animal_tree

Develop an "Explainable" ML Model

- Black-box approaches are hard to explain
- Classical white-box approaches: linear models, decision trees,...
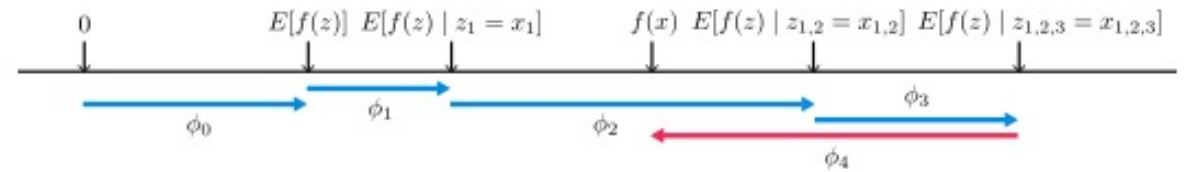- The difference is not always clear
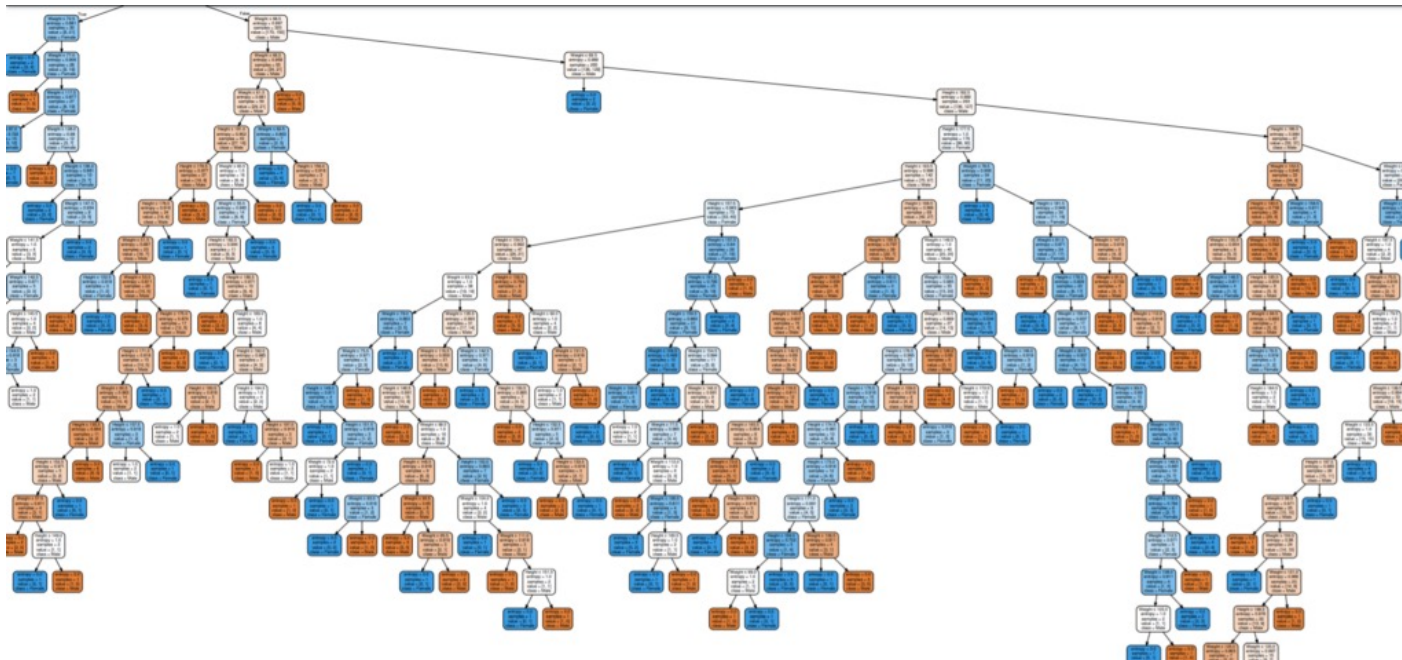
Post-hoc Process

- LIME [Ribeiro et al. 2016]
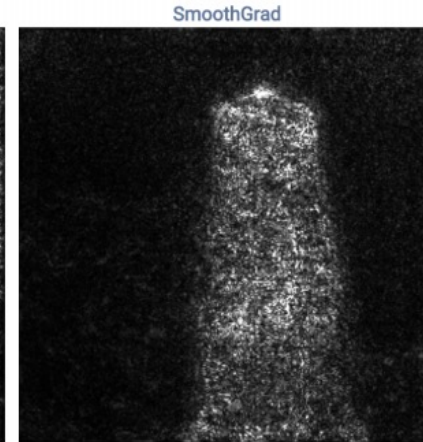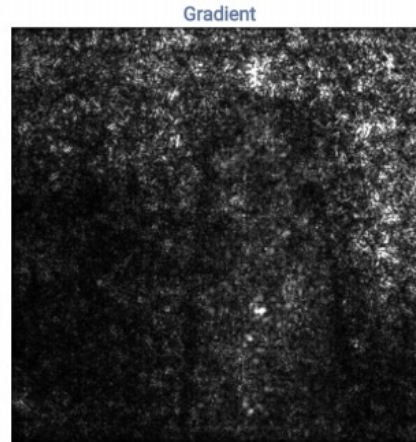
- SHAP [Lundberg and Lee. 2017]

Post-hoc Process

- Saliency map
  - Understand how changing the feature impacts the predictions

  - Take the gradient $\dfrac{\partial y}{\partial x_{i,j}}$
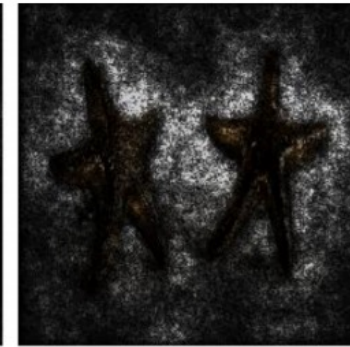
Post-hoc Process

- Saliency map

SmoothGrad [Smilkov et al. 17]

Gradient SmoothGrad

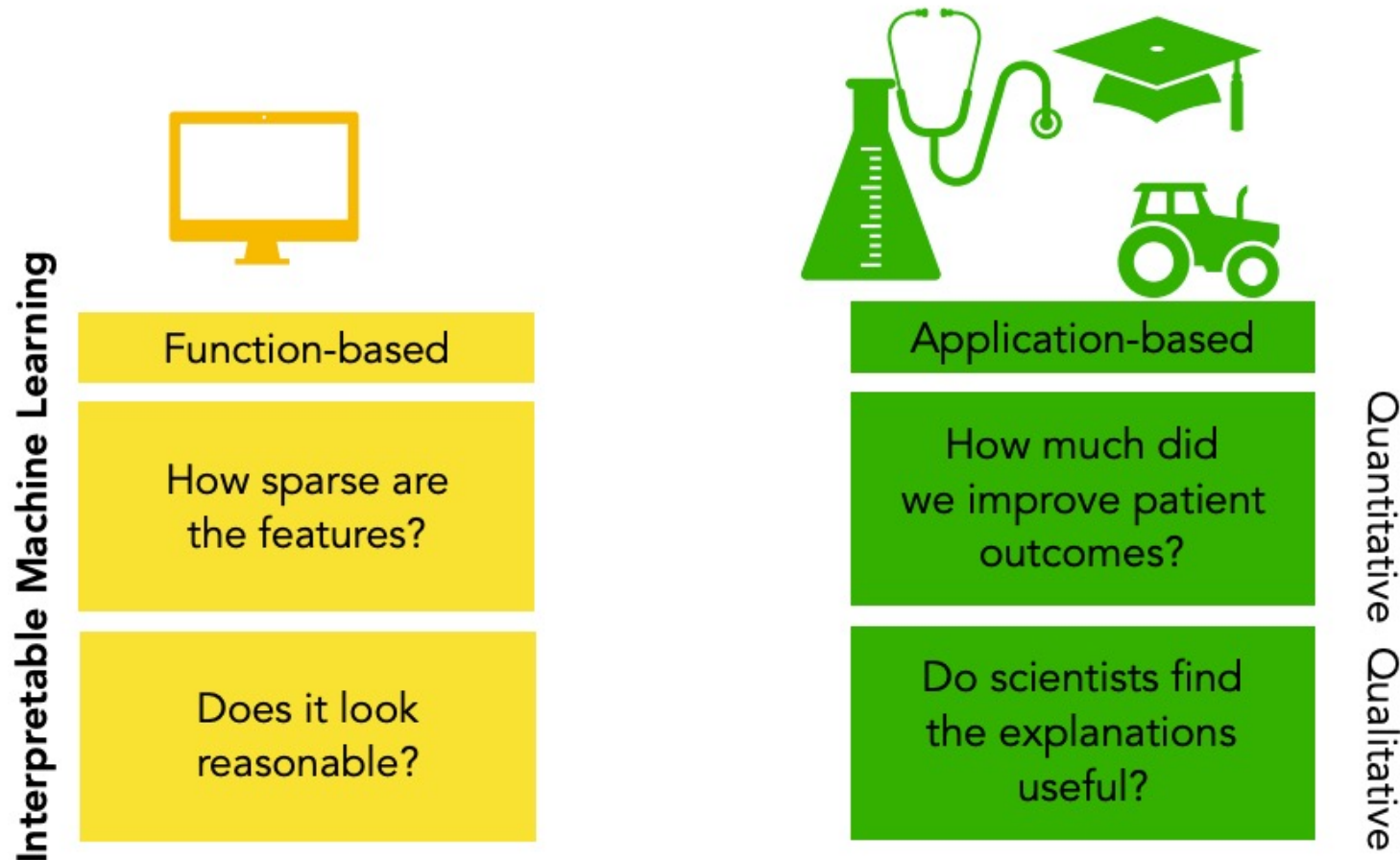**Integrated gradients [Sundararajan et al. 17]**

Top label: starfish
Score: 0.999992

Who Decides Whether Explanations Make Sense?

- Humans make the call

Who Decides Whether Explanations Make Sense?

- Humans make the call

- Humans' mental models of AI
  - We have talked about modeling human behavior in this course
  - Maybe we want to model how human models AI's behavior

- We'll talk a little bit about these in the lecture on Dec 10.

Concerns of Interpretability?

- Being interpretable
    => humans understand how AI operate
    => humans might want to alter the data to algin with their needs
    => "Gaming", or manipulation issue