Lecture I: Background & review

CSEP 590B: Explainable AI

Ian Covert & Su-In Lee
University of Washington

Today

- Section 1
  - Motivation & aims
  - Course logistics
  - Examples in the healthcare space
- Section 2
  - Discussion: “Statistical Modeling: The Two Cultures”
- Section 3
  - Example scenario and ML review 

Example scenario

- You work at a bank, and your boss asks you to automate **credit risk evaluation**

- **Step 1:** obtain historical data
  - Input features $x = [\text{age}, \text{income}, \text{job}, \text{savings}, \ldots]$
  - Label $y \in \{0, 1\}$ (repaid loan or not)
- **Step 2:** train model
  - What model type? Let's review some options

Linear regression

- Model continuous response as linear function:

$$f(x) = \beta_0 + \sum_{i=1}^{d} \beta_i \cdot x_i$$

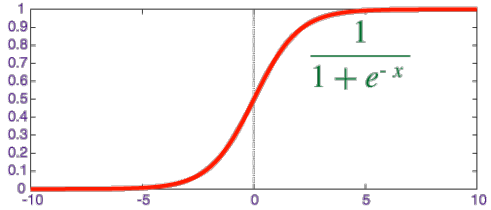
- **Parameters:** one coefficient per feature, plus bias term
- **Training:** least squares loss, closed-form solution

Logistic regression

- Model discrete response log-probability as linear function:

$$f(x) = \sigma\left(\beta_0 + \sum_{i=1}^{d} \beta_i \cdot x_i\right)$$

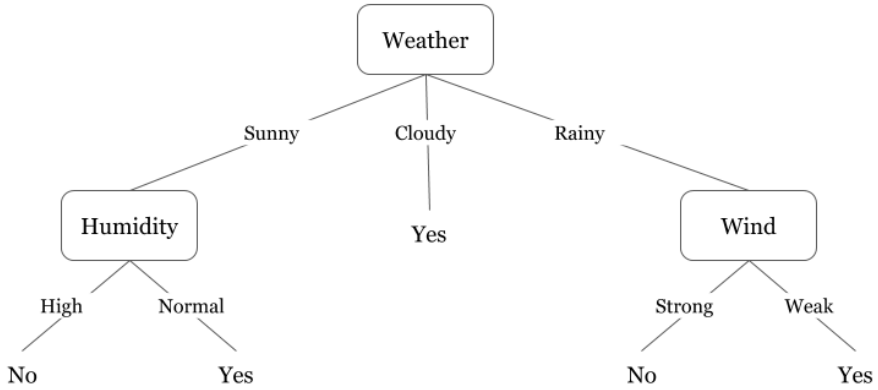
where $\sigma(z) = \frac{1}{1+e^{-z}}$



- **Parameters:** one coefficient per feature, plus bias term
- **Training:** log-loss; no closed-form solution, optimize with gradient descent

Decision tree

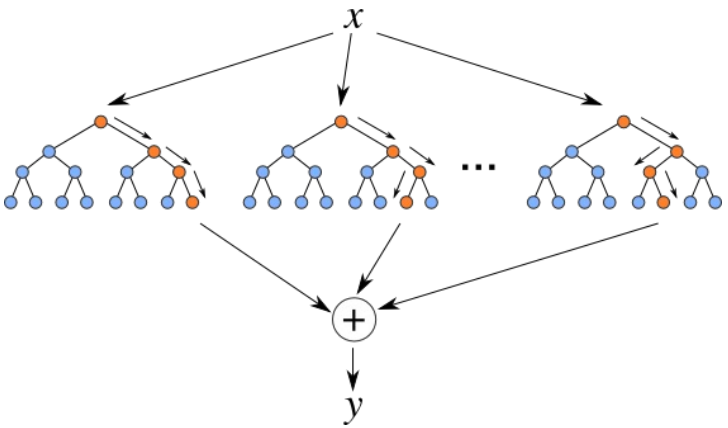
- Model discrete or continuous outcomes using tree structure



- **Parameters:** one feature/threshold per internal node, one prediction per leaf node
- **Training:** greedily minimize entropy or variance

Tree ensembles

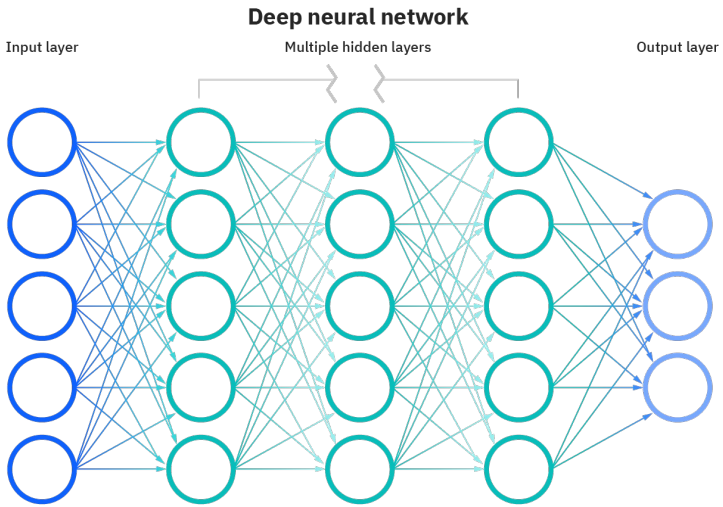
- Sum predictions across multiple decision trees



- **Random forests:** bootstrap training data and aggregate trees (bagging), ensemble independent strong learners
- **Gradient boosting machines:** combine weak learners, let new trees improve on previous ones (gradient boosting)

Neural networks

- Model continuous or discrete outcomes using hierarchical, nonlinear, differentiable functions



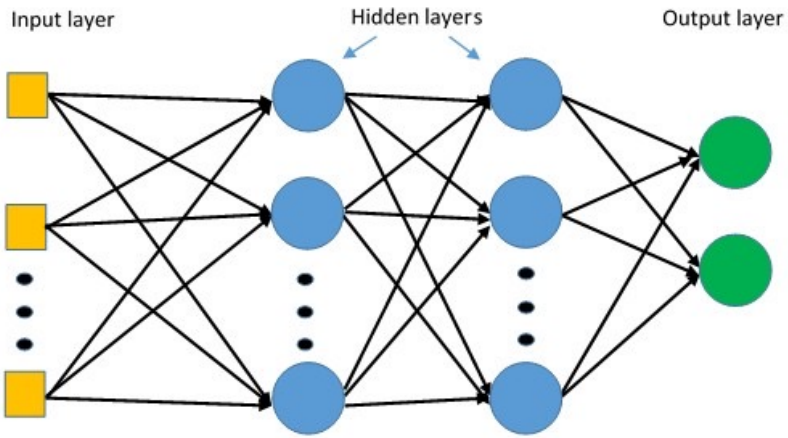
- **Parameters:** many parameters per layer
- **Training:** use stochastic gradient descent (SGD) to find local minimum of loss function

Multilayer perceptron (MLP)

- For vector input ("tabular" data)
- Each layer applies linear operation and nonlinear activation:

$$a^{(l+1)} = \sigma\left(W^{(l)}a^{(l)} + b^{(l)}\right)$$

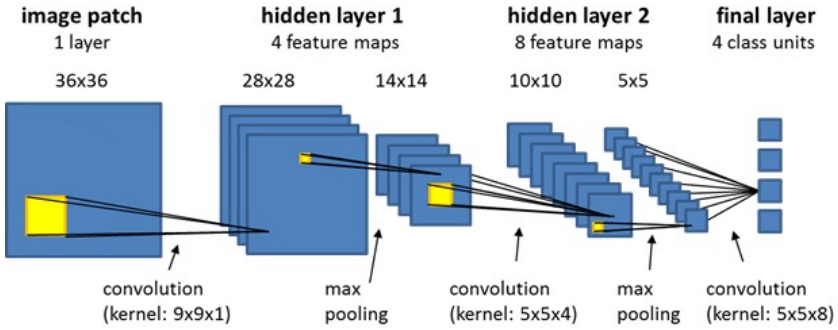
where $\sigma(z) = \max(z, 0)$ (for ReLU activation)



Convolutional neural network (CNN)

- For grid-shaped data (images)
- Alternate convolutional layers, nonlinear activations, pooling:

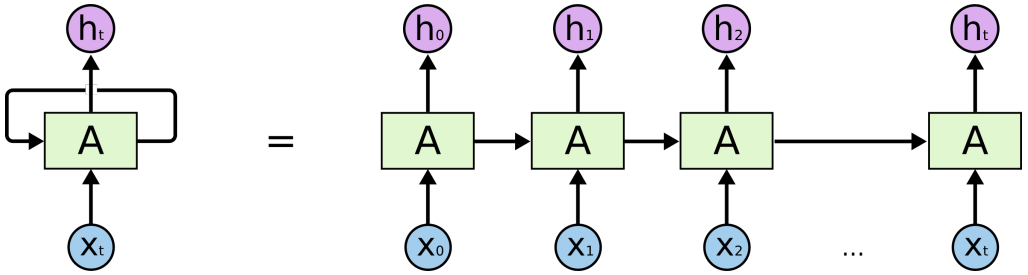
$$a^{(l+1)} = \sigma\left(W^{(l+1)} * a^{(l)} + b^{(l+1)}\right)$$



- Variations involve residual connections, normalization layers, dropout, etc.

LSTMs, Transformers

- Deep learning architectures designed for sequential data
- **LSTMs:** a version of recurrent neural networks (RNNs), maintains internal state for each time point in a sequence



- **Transformers:** use self-attention mechanism to generate predictions that depend on all previous time points

Deep learning hyperparameters

- Activation function
  - Sigmoid, ReLU, ELU, GELU, etc.
- Depth
  - How many layers?
- Width
  - How many hidden units/channels per layer?
- Optimizer
  - SGD, Adam, RMSProp, etc.
- Regularization
  - Dropout, batch norm, etc.

Math background

- ML is built on tools from calculus, optimization, linear algebra, and probability

- If your memory of these topics isn't 100% fresh, it's probably okay
  - Can recall important concepts as needed
  - See HW0 for a refresher. If you find it easy, you'll be fine

How to choose your model?

- What models have you used before?
- What factors influence your choice?

Back to the bank example…

- You examine the various model options
- Select a complex model because it gets the best performance: a gradient boosted tree
  - XGBoost, LightGBM

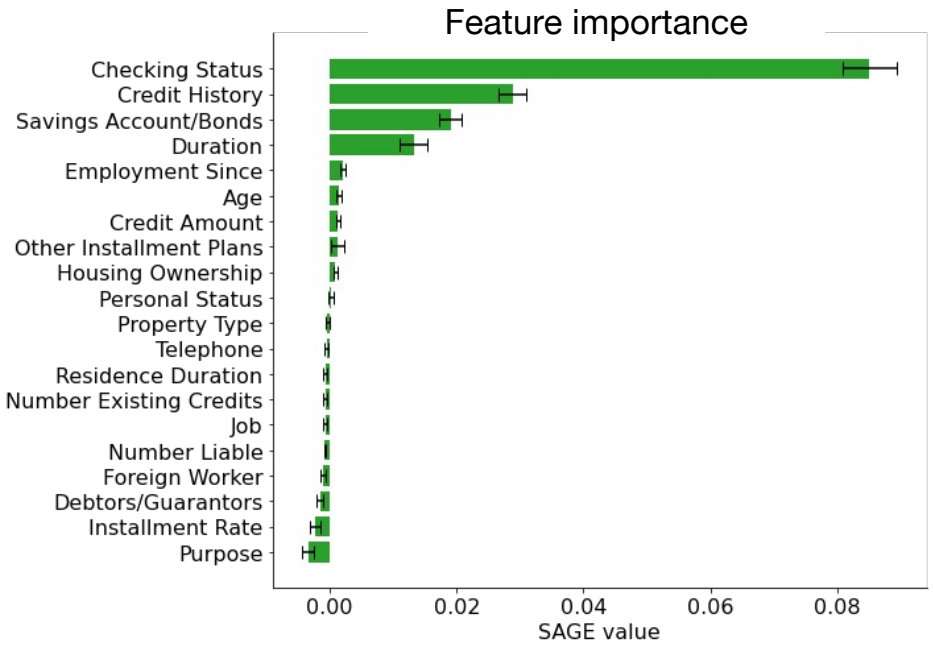
- **Boss:** Nice job, can I ask some questions about how the model works?
- **You:** Sure!

Q: What did the model learn, and how does it make decisions?

- We can't easily summarize the patterns, rules, concepts learned by a complex model
- Gradient boosted trees have too many parameters to examine
- Coarse summary: count number of splits on each feature

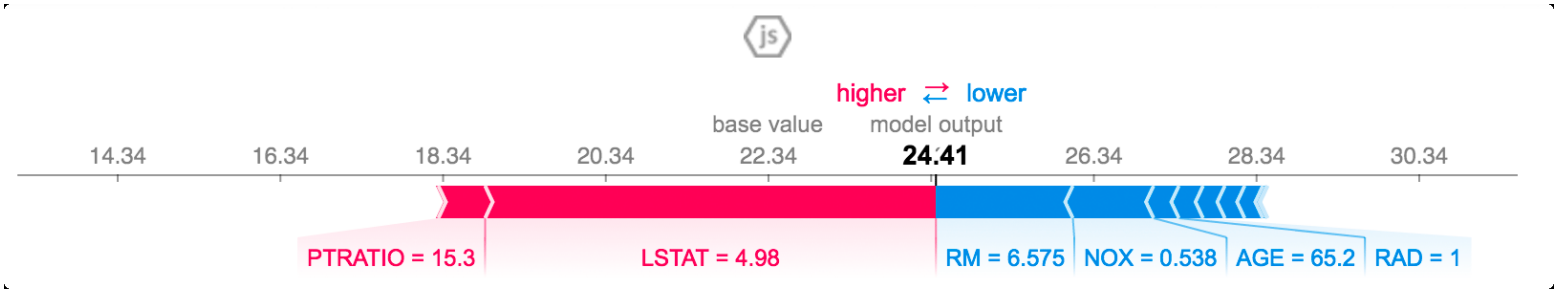
Q: Which features are most important overall?

- Can be answered by permutation tests (Breiman, 2001)
- There are in fact many methods for analyzing **global feature importance**



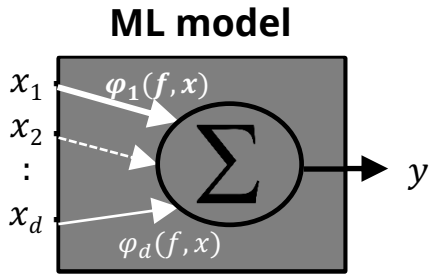
Q: For the customers whose loans are denied, can we tell which features led to the decision?

- In this case, must analyze individual predictions (not overall behavior)
- Many methods designed to assess **local feature importance**
  - E.g., shap, LIME

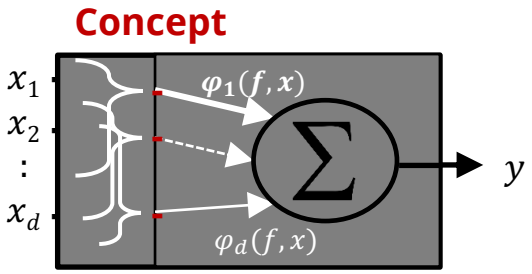


Course overview

- **Course introduction** (1 lecture)
- **Feature importance explanations** (3 lectures)
  - Removal-based explanations
  - Shapley values
  - Propagation-based explanations



- **Evaluating explanations** (1 lecture)
- **Inherently interpretable models** (1 lecture)
- **Other approaches** (2 lectures)
  - Concept-based explanations, neuron interpretation
  - Counterfactual explanations, instance explanations
- **Enhancing human-AI collaboration** (1 lecture)
- **XAI in industry, model improvement** (1 lecture)

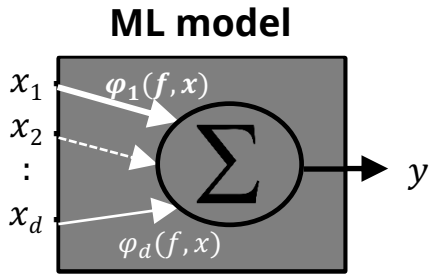


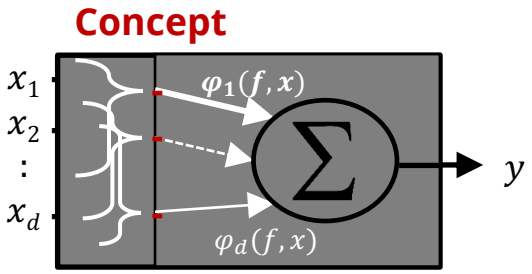
Q: How can we be sure the model explanations are correct?

- Our goal is to evaluate model explanations generated by XAI methods
- Feature importance can be evaluated based on qualitative or quantitative criteria

Course overview

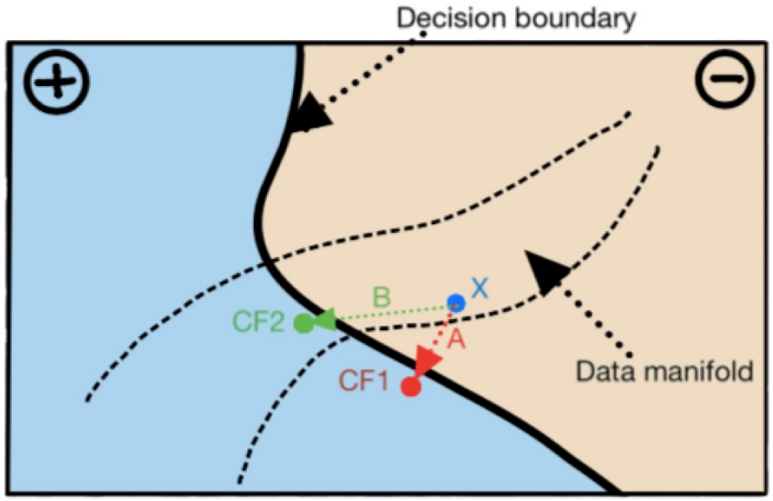
- **Course introduction** (1 lecture)
- **Feature importance explanations** (3 lectures)
  - Removal-based explanations
  - Shapley values
  - Propagation-based explanations
- **Evaluating explanations** (1 lecture)
- **Inherently interpretable models** (1 lecture)
- **Other approaches** (2 lectures)
  - Concept-based explanations, neuron interpretation
  - Counterfactual explanations, instance explanations
- **Enhancing human-AI collaboration** (1 lecture)
- **XAI in industry, model improvement** (1 lecture)





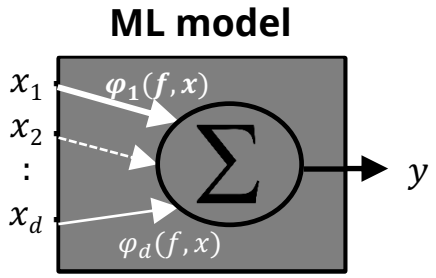
Q: Can we tell customers what to change to get approved next time?

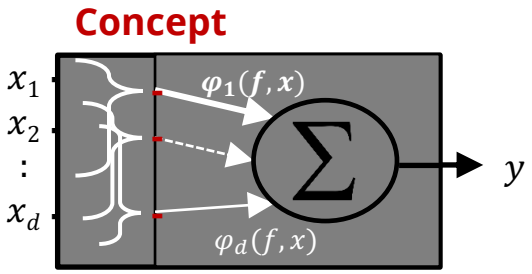
- Here, our goal is to identify small changes that can alter the model output
- These are called **counterfactual explanations**



Course overview

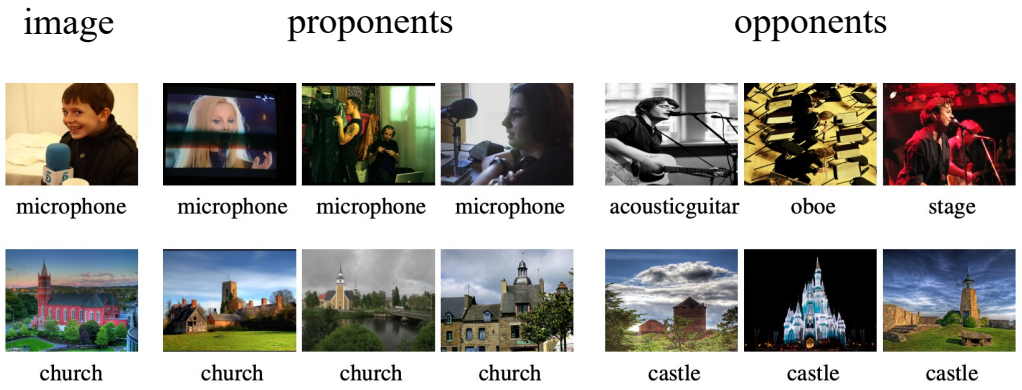
- **Course introduction** (1 lecture)
- **Feature importance explanations** (3 lectures)
  - Removal-based explanations
  - Shapley values
  - Propagation-based explanations
- **Evaluating explanations** (1 lecture)
- **Inherently interpretable models** (1 lecture)
- **Other approaches** (2 lectures)
  - Concept-based explanations, neuron interpretation
  - Counterfactual explanations, instance explanations
- **Enhancing human-AI collaboration** (1 lecture)
- **XAI in industry, model improvement** (1 lecture)





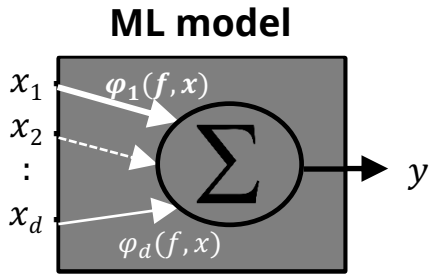
Q: Are there potentially misleading examples in our historical data, and can you get rid of them?

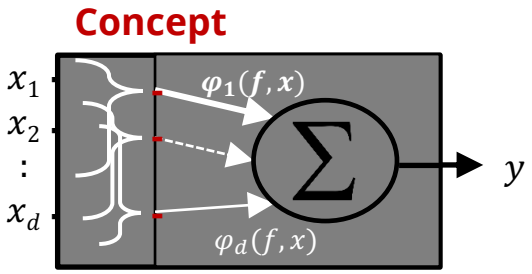
- No longer asking about role of features, now asking about *data examples*
- We can analyze the influence of dataset examples using **instance explanations**



Course overview

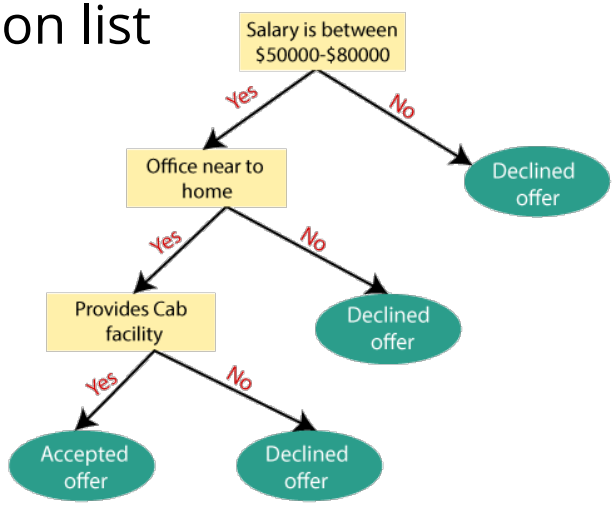
- **Course introduction** (1 lecture)
- **Feature importance explanations** (3 lectures)
  - Removal-based explanations
  - Shapley values
  - Propagation-based explanations
- **Evaluating explanations** (1 lecture)
- **Inherently interpretable models** (1 lecture)
- **Other approaches** (2 lectures)
  - Concept-based explanations, neuron interpretation
  - Counterfactual explanations, instance explanations
- **Enhancing human-AI collaboration** (1 lecture)
- **XAI in industry, model improvement** (1 lecture)





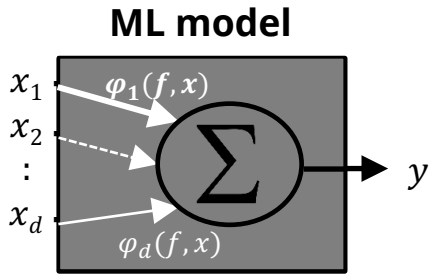
Q: No human can internalize your model, can you use something simpler instead?

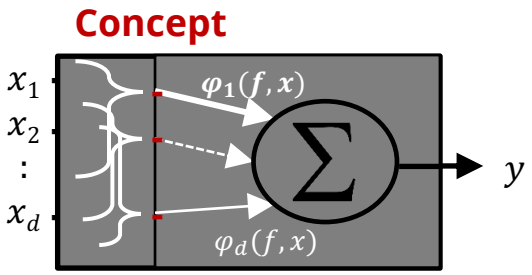
- Instead of using a black-box model, maybe an **inherently interpretable model** gets sufficiently high accuracy
- These can make a mental model manageable
  - E.g., single decision tree, or decision list



Course overview

- **Course introduction** (1 lecture)
- **Feature importance explanations** (3 lectures)
  - Removal-based explanations
  - Shapley values
  - Propagation-based explanations
- **Evaluating explanations** (1 lecture)
- **Inherently interpretable models** (1 lecture)
- **Other approaches** (2 lectures)
  - Concept-based explanations, neuron interpretation
  - Counterfactual explanations, instance explanations
- **Enhancing human-AI collaboration** (1 lecture)
- **XAI in industry, model improvement** (1 lecture)





We'll cover these topics, and more!

Reminder: for next time

- Office hours poll (see your email)
- Discussion leader volunteers (see your email)
- First discussion post
  - Petsiuk et al., “RISE: Randomized Input Sampling for Explanation of Black-box Models” (2018)
- HW0 due