**Robust and Explainable Machine Learning**

Interpretability: how easy it is to interpret

* Tradeoff between interpretability vs accuracy

Goal: have both interpretability and accuracy

* Additive Model
  + Each feature has a function
* Propose neural additive model
  + Give neural network for each feature
* Architecture
  + Concept network: maps from input space to concept space
  + Task network: maps from concept space to input space
* Advantages
  + Flexibility: reshape any network of standard NN
* SENN: Self-Explaining Neural Networks
  + Alignment is the biggest problem
* Use contrastive learning for better alignment
* Prototype-based
  + Extract prototype-based concept
  + Extract importance
  + Combine
  + Tested with cells
* Using LLMS to generate concepts

Current Project

* Stage 1: concept extraction
* Stage 2: symbol formation
* Stage 3: neurosymbolic evaluations
* Desirable Properties
  + Grounding
  + Hierarchy
  + Task adaptability
* Use LLM for scientific exploration
  + Unbiased/diverse thinking
  + Interdisciplinary reach
  + Efficiency
  + How to generate better hypotheses?
* Project: spurious correlation detection and mitigation
  + Spurious correlation => find correlation between classes and spurious attributes (e.g bird and background)
  + Robustness issue: reliance on spurious correlation