Stat 992 - Motivation + Vocabulary
Wednesday, January 22, 2025

General Motivation

1. ML models are changing the world, but we know surprisingly little about why they work. This has some concrete risks:
   - Fairness
   - Safety
   - Misinformation

2. In the past, people were directly responsible for all their decisions, discoveries, and creative work. You could ask them Q's.

Specific Motivation

Interpretability can help address formulation gaps.

Formulation Gaps
(Things that matter, but which are difficult to include in an objective function)

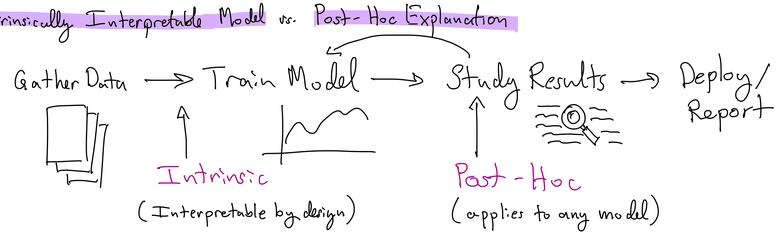
- Ethics
  - Fairness
  - Privacy
  - Copyright
- Reliability
  - Will the model generalize in the ways we want it to?
  - Is the model using the "right" features?
- Usefulness
  - Navigation
    e.g. refer a doctor to related cases.
  - Control
    e.g. use a model for image editing or translation

Types of Interpretability

1. Considering all the different motivations for interpretability, it's not surprising that many approaches have been proposed.

2. Local vs. Global Explanations
   - **Local:** Give information about the prediction for a specific example
   - **Global:** Give information about the entire model

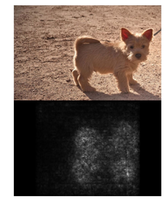
3. **Intrinsically Interpretable Model** vs. **Post-Hoc Explanation**

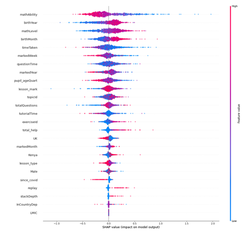


4. What makes some models intrinsically interpretable?
   - **Sparsity:** Parts (e.g. coefs) ↓ Interpretability ↑
   - **Simulatability:** Can you make the prediction "by hand"
   - **Modularity:** Can it be broken down?
     e.g. $f(x) = \sum f_j(x_j)$

Q: Is linear regression intrinsically interpretable?

5. Post-Hoc Explanations give one of two things:
   - Important features (can be local or global, isolated or within interactions)





   - Additional Context (e.g. natural language or related examples)

Laysan Albatross

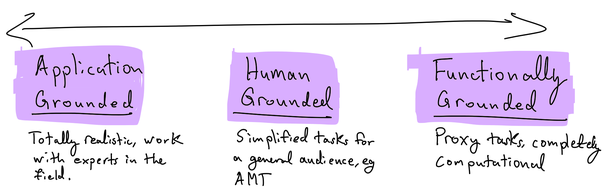
Description: This is a large flying bird with black wings and a white belly.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a large wingspan, hooked yellow beak, and white belly.

Q: Have you ever seen a trade-off between prediction accuracy and interpretability? Have you ever improved both at once?

Evaluating Interpretability

1. Murdoch et al. suggest the PDR approach
   - [P]redictive Accuracy - No point interpreting a bad model
   - [M]odel Accuracy - Is the explanation faithful to the model?
   - [R]elevance - Are the outputs useful to "domain experts"?

2. Doshi-Velez and Kim recognize a spectrum of experimental evaluations:



3. To build on each others work, we should clearly report (explain?!) the type of interpretability and evaluation criteria we are targeting.