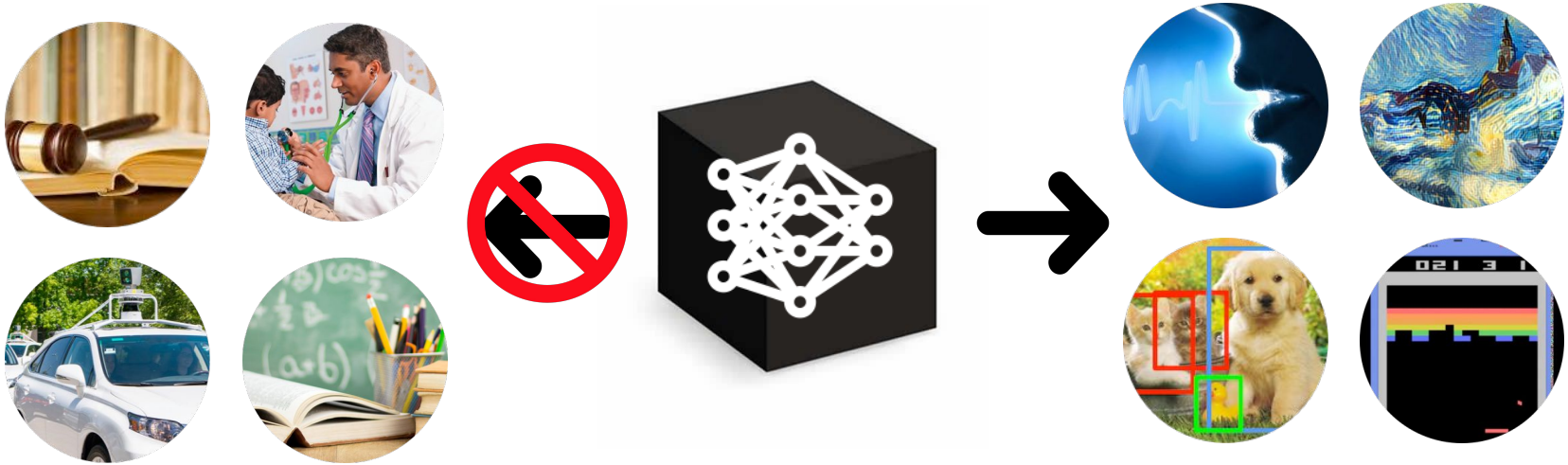
Beyond Sparsity: Tree Regularization of Deep Models for Interpretability

Mike Wu[1], Michael C. Hughes[2], Sonali Parbhoo[3], Maurizio Zazzi[4], Volker Ross[3], Finale Doshi-Velez[2]

[1] Stanford University, [2] Harvard University, [3] University of Basel, [4] University of Siena

NIPS TIML Workshop, Long Beach
December 9, 2017

Motivation: Deep Learning is not interpretable.



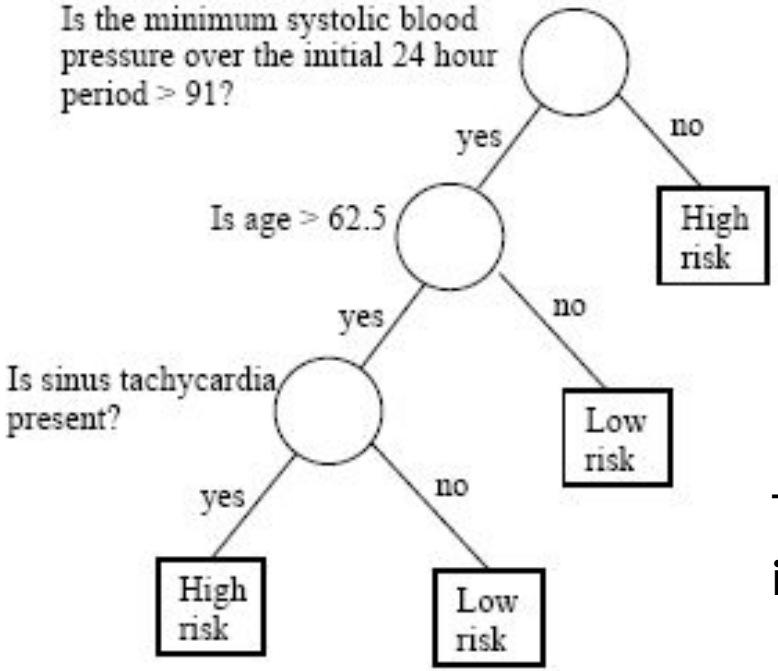
What is Interpretability?

Def: A model is **simulatable** if a human can "take in input data together with the parameters of the model and in *reasonable* time step through *every* calculation required to make a prediction." (Lipton 2016)

Advantages of simulation

- Check each step against expert knowledge
- Apply counterfactual reasoning i.e. *what if blood pressure was lower?*
- Identify systemic bias

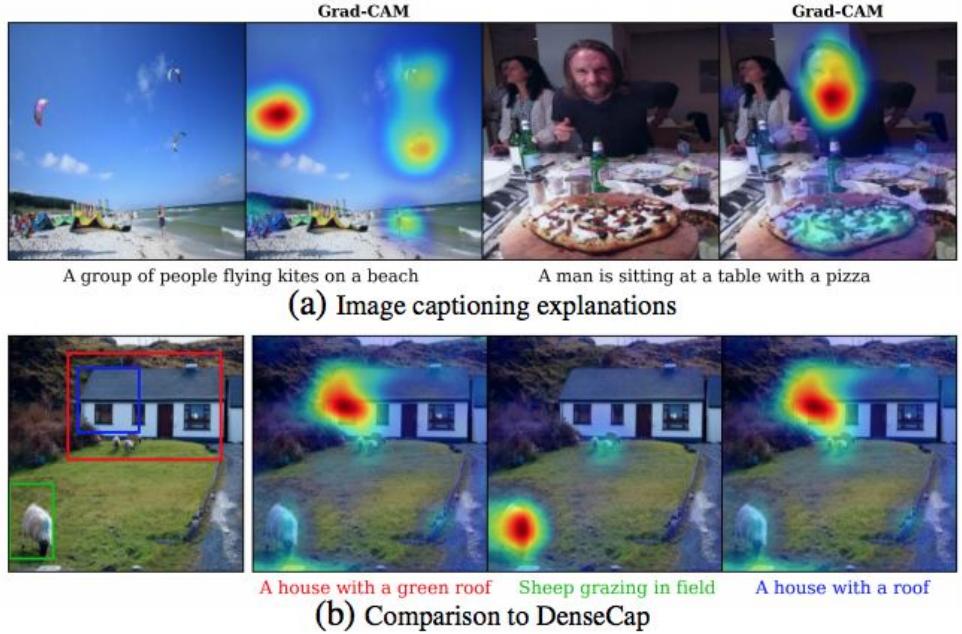
Decision Trees are Simulatable



Trees are pretty powerful but definitely **inferior** to modern deep models.

How to interpret a trained deep model?

Selvaraju et. al. 2017



Olah et. al. 2017

Dataset Examples show us what neurons respond to in practice

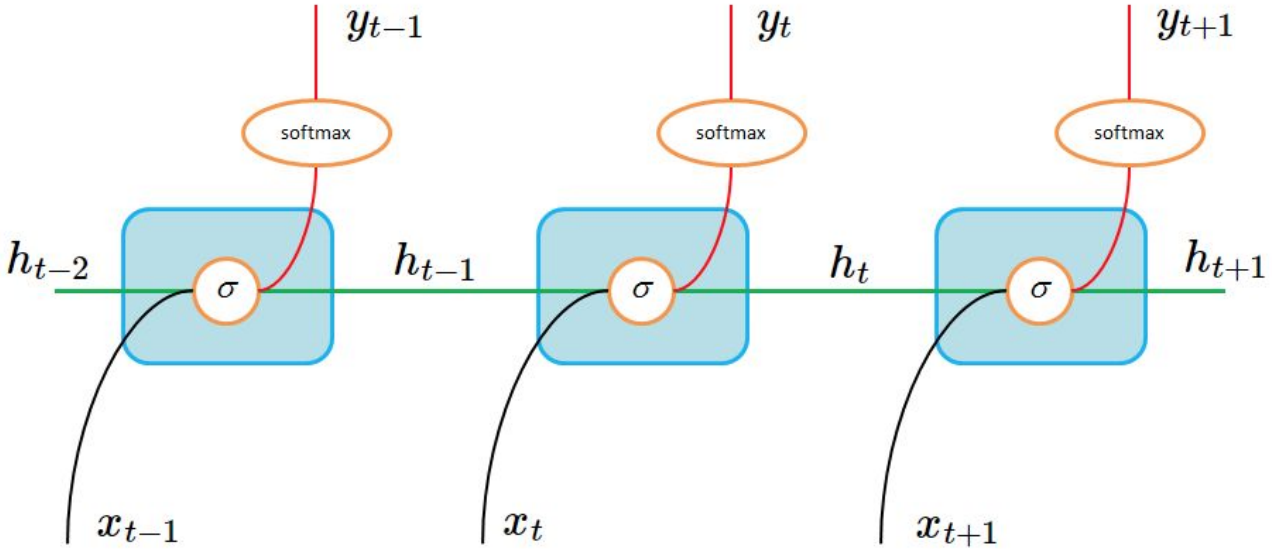
Optimization isolates the causes of behavior from mere correlations. A neuron may not be detecting what you initially thought.

Baseball—or stripes?
mixed4a, Unit 6

Animal faces—or snouts?
mixed4a, Unit 240

Can we directly **optimize** a deep model to be interpretable?

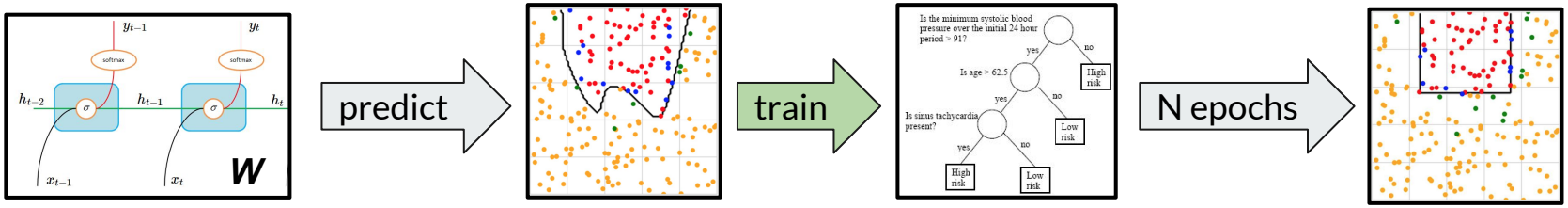
Model: Recurrent Neural Nets



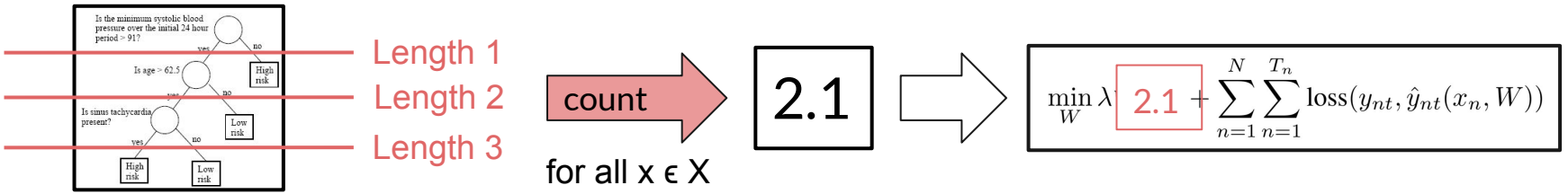
$$\min_W \lambda \Psi(W) + \sum_{n=1}^{N} \sum_{n=1}^{T_n} \text{loss}(y_{nt}, \hat{y}_{nt}(x_n, W))$$

Simulability Objective Function

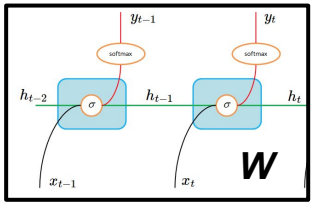
1) Train Decision Tree with similar predictions as deep model.



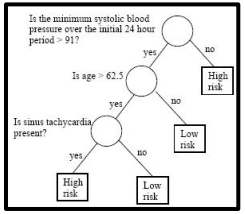
2) Count tree's average **path length** = cost of simulating the average example.



But Trees aren't Differentiable.



train



count

2.1

Surrogate MLP (approx. Tree)

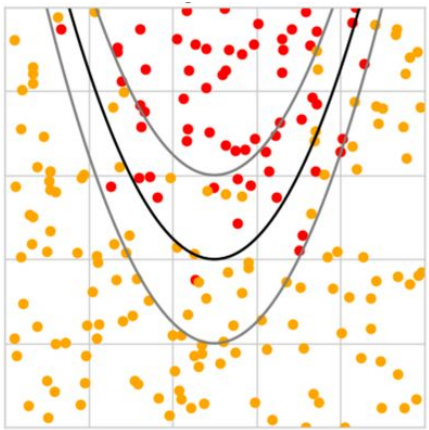
2.2

Given fixed Surrogate MLP, optimize ***W*** via gradient descent.

Given fixed ***W***, we can find the best Surrogate MLP .

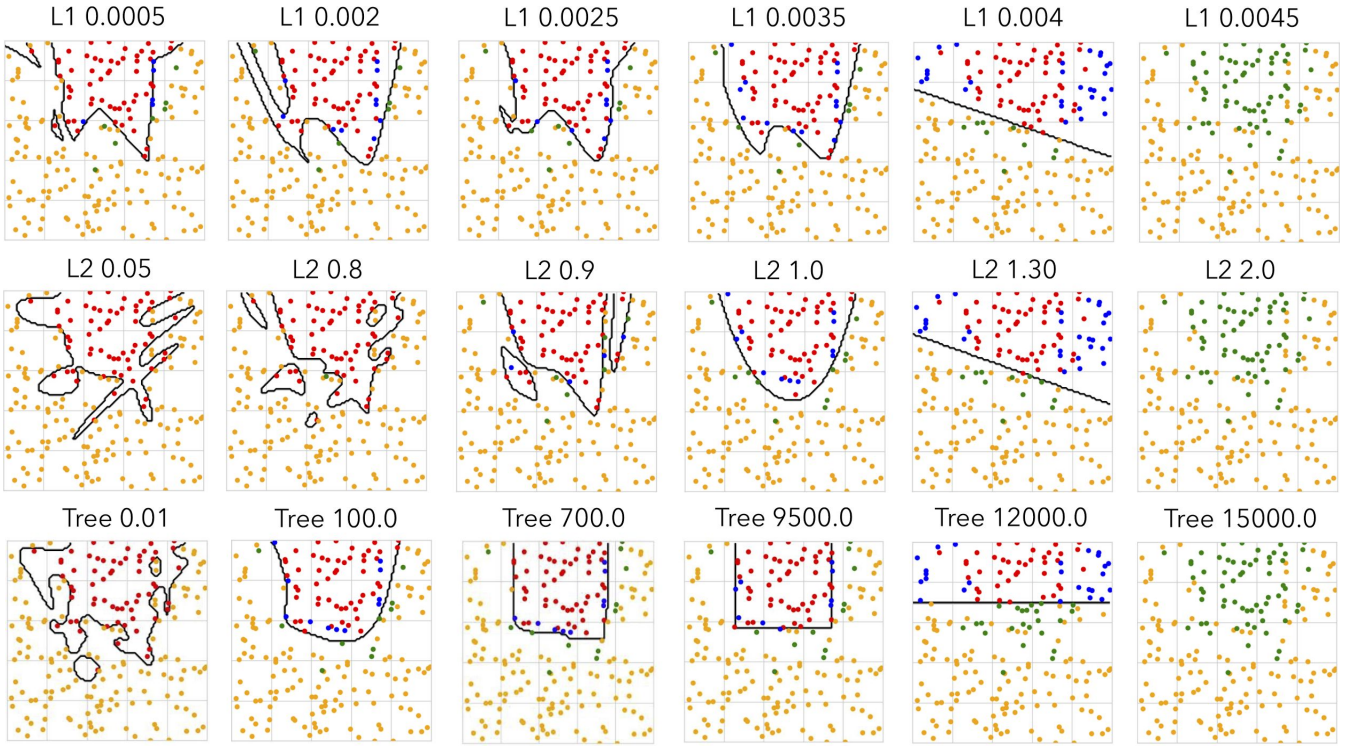
Tree-like Decision Boundaries for Deep Models.

Dataset

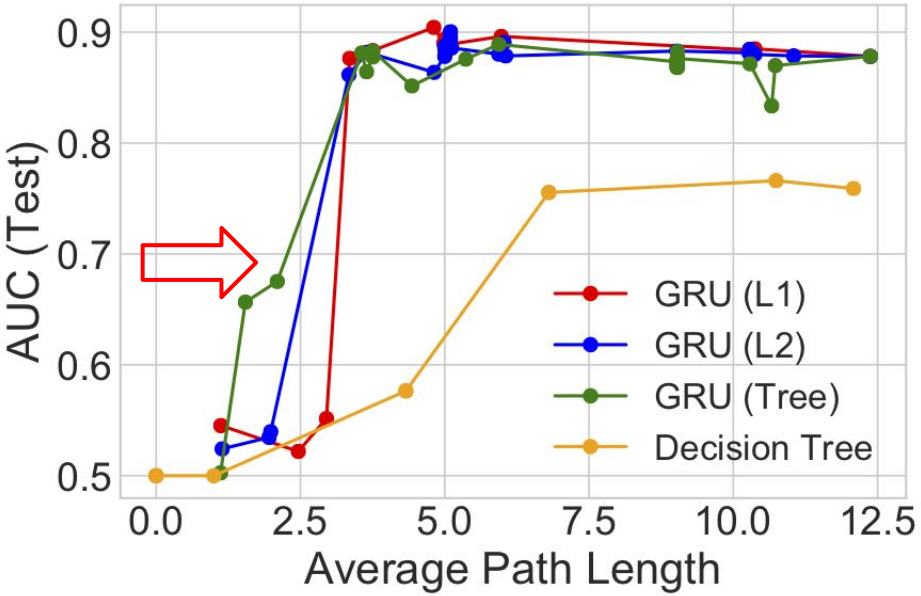


Red: Positive

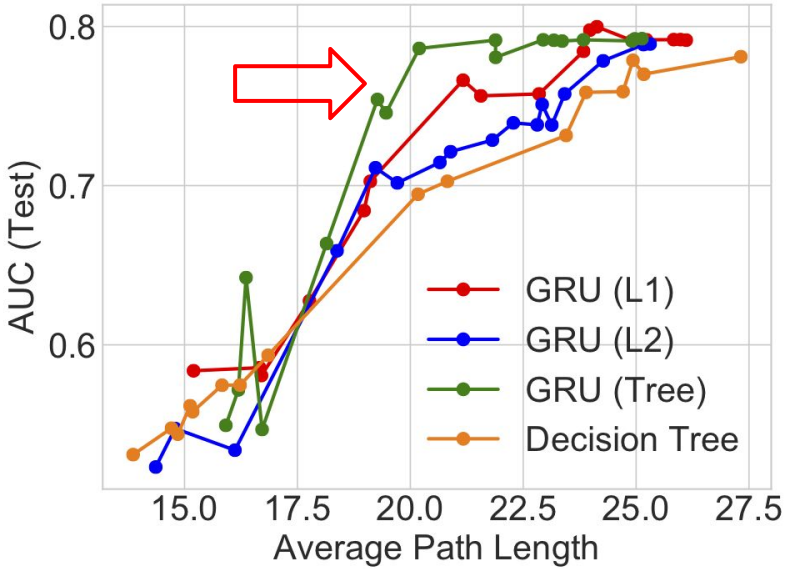
Yellow: Negative



Tree-reg finds sweet-spot that L2 and L1 can't.

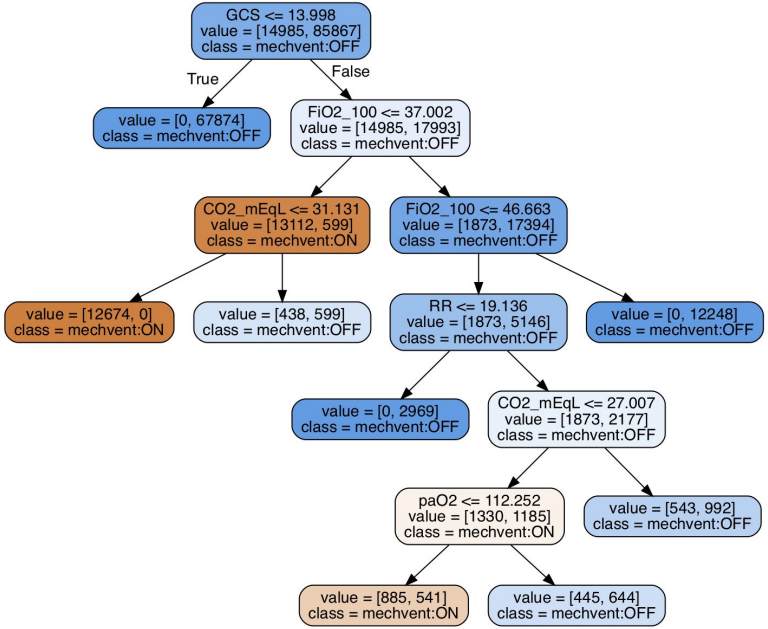


(a) Sepsis: Mechanical Ventilation

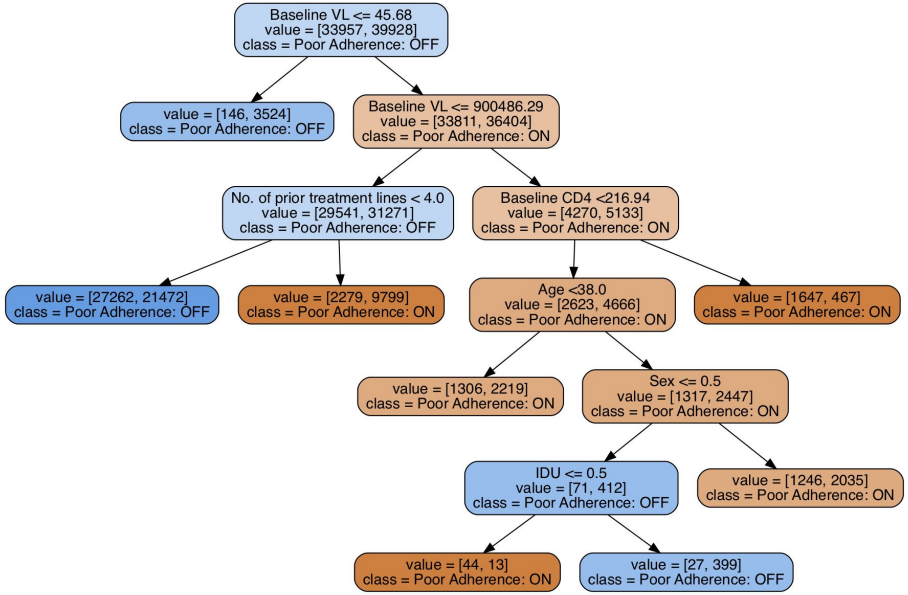


(b) HIV: Therapy Adherence

Tree-reg produces Trees that are interpretable



(a) Sepsis: Mechanical Ventilation



(b) HIV: Therapy Adherence

Future Work

- Current trees only use static features. What about **temporal** features?
- Can trees capture **local** interpretability?
- What if features are not *prima facie* interpretable (i.e. pixels)?

Come see our poster!