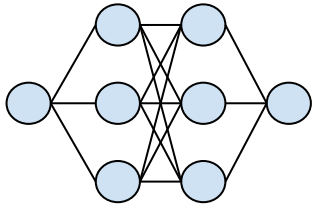
Interpretable Local Tree Surrogate Policies

AAAI SafeAI Workshop 2022

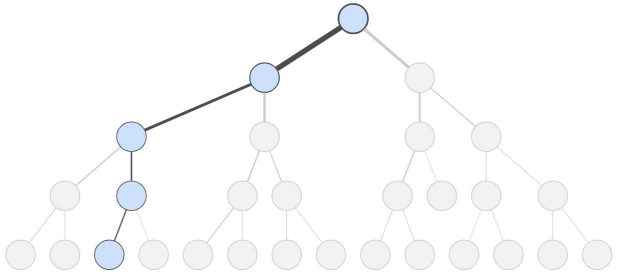
John Mern, Sidhart Krishnan, Anil Yildiz, Kyle Hatch,
Mykel J Kochenderfer

SISL
Stanford Intelligent
Systems Laboratory

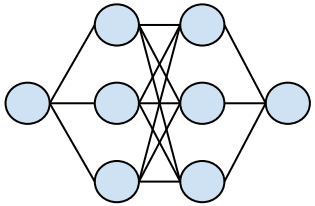
Trees as Interpretable Surrogates



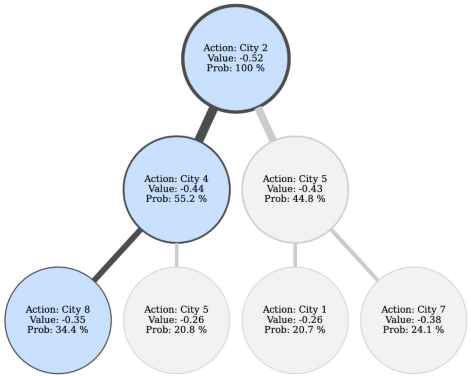
s_0



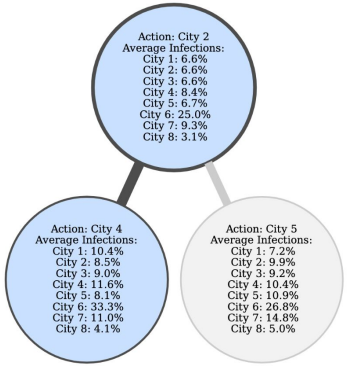
Trees as Interpretable Surrogates



s_0

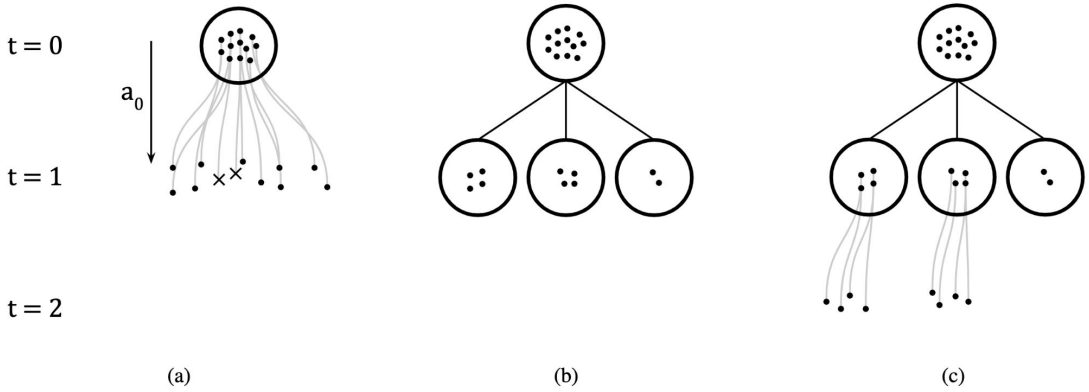


(a)



(b)

Monte Carlo Clustering



$$\delta = \|\hat{Q}(s,a) - \max_{a' \in \mathcal{A}} \hat{Q}(s,a')\|$$

SISL

Stanford Intelligent
Systems Laboratory