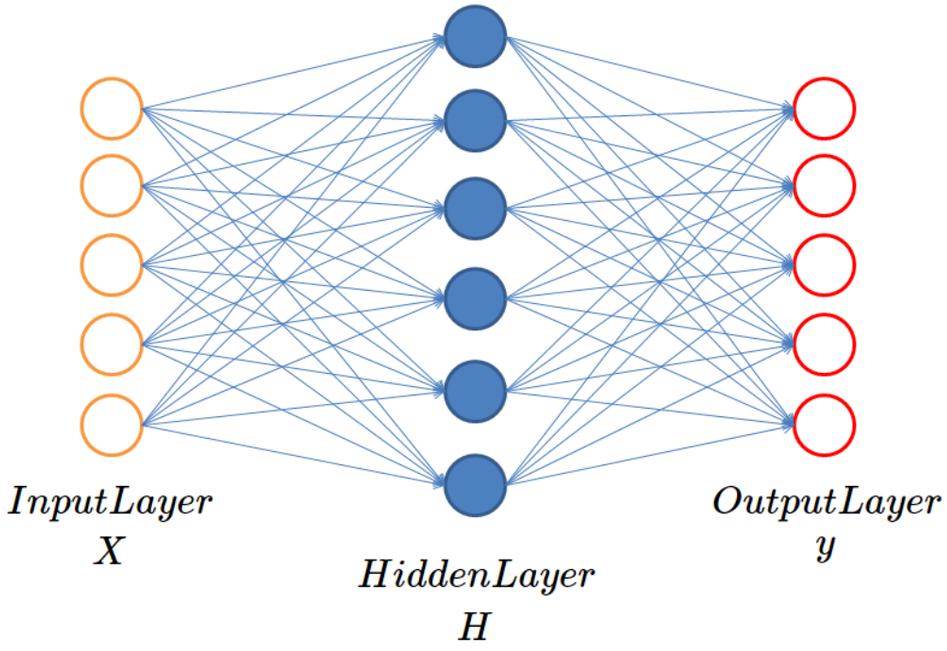
Local **I**nterpretable **M**odel-agnostic **E**xplanations

“Why Should I Trust You?": Explaining the Predictions of Any Classifier
(Aug 2016) by Ribeiro et al

LIME

Local **I**nterpretable **M**odel-agnostic **E**xplanations



$$\sim \quad \hat{y} = w_1 x_1 + w_2 x_2$$