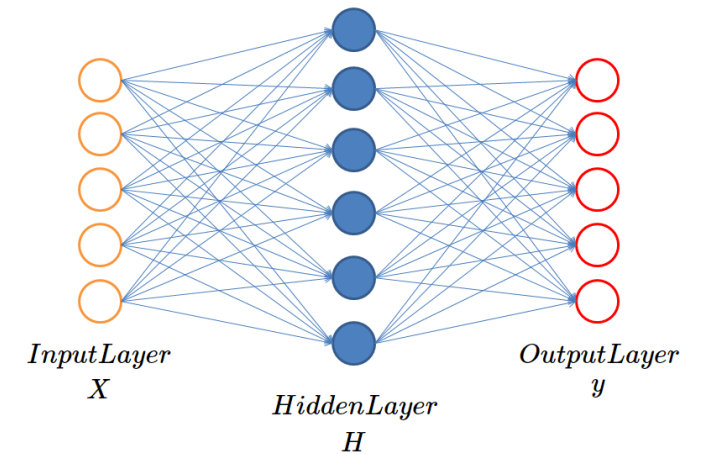


## LIMITATIONS

## no ground truth

- to assess an attribution
- no standard method for comparing attributions
- two equally good models may learn a different set of feature interactions



## single weights

attribution assigns a single weight to each feature, missing the non-linear combinations of inputs that yield predictions

$$\hat{y} = w_1x_1 + w_2x_2$$

**L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations