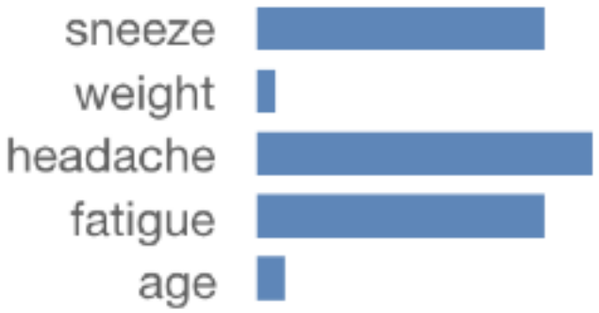
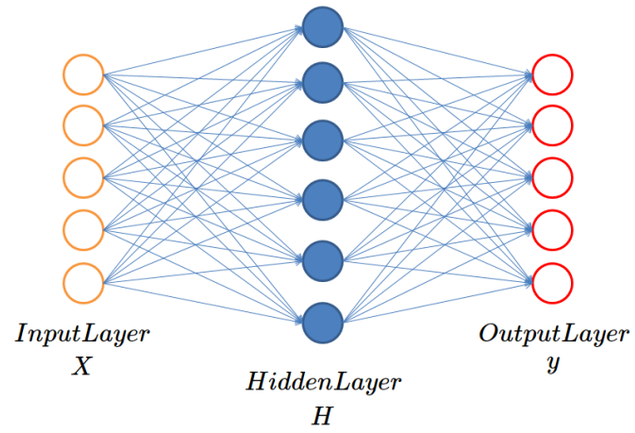


LIMITATIONS

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

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

LIME

Local Interpretable Model-agnostic explanations