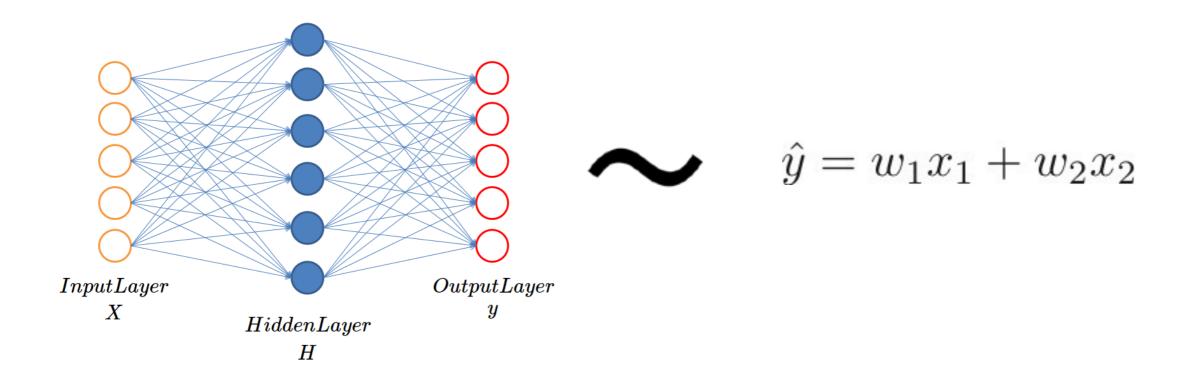
## ocal nterpretable odel-agnostic explanations



## **Integrated Gradients**



## **Sensitivity**

is the attribution sensitive to **relevant features**: inputs differ by a single feature

Axioms

## **Implementation Invariance**

the attributions are always identical of **functionally equivalent** networks

outputs are identical for all inputs