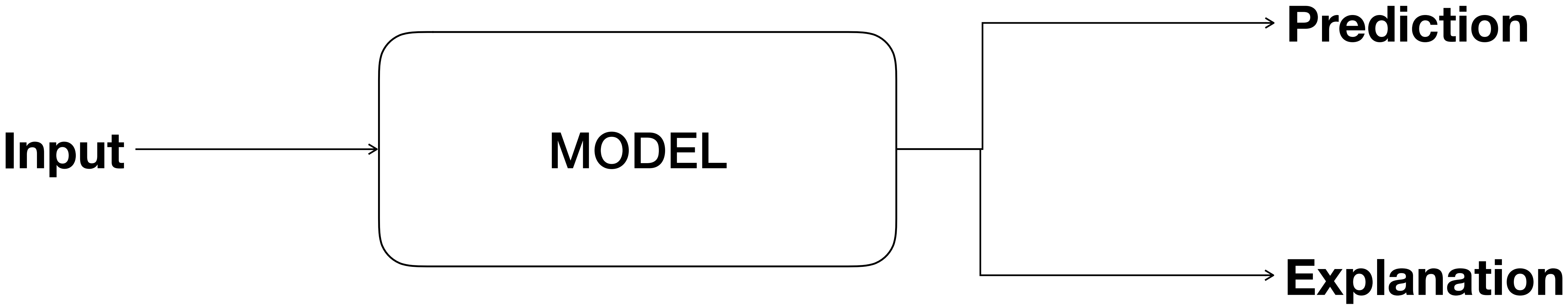
- Can we make our models explain their predictions as **intrinsic part** of their operation?

Beyond
Post-Hoc Explanations

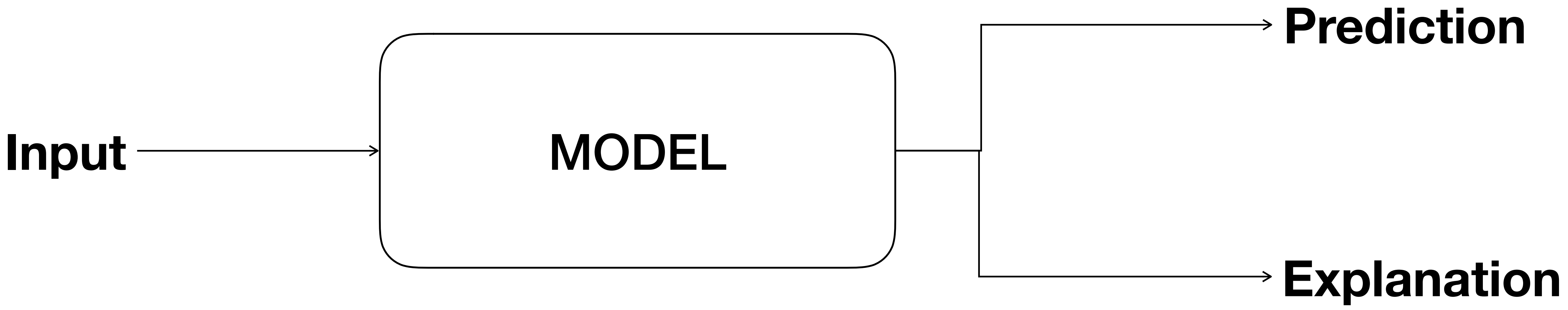

Input
MODEL
Prediction
Explanation

Goal:

Beyond Post-Hoc Explanations

- Can we make our models explain their predictions as **intrinsic part** of their operation?

Goal:



Self-Explaining Neural Networks (SENN)

AM + Jaakkola, *NeurIPS’18*