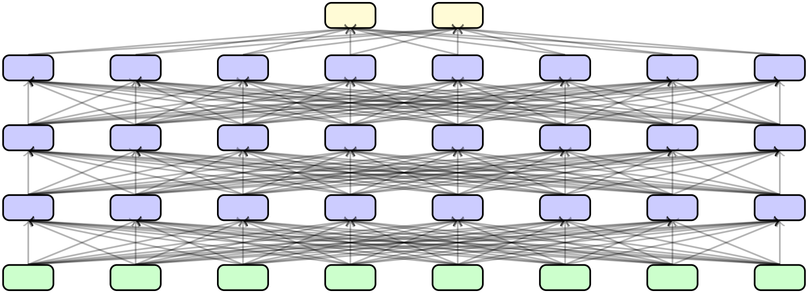
Tree-Structured Algorithms as Causal Abstractions of Neural Networks

Atticus Geiger

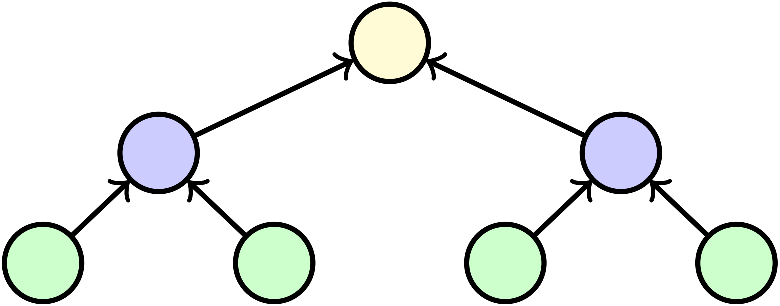
Stanford Linguistics and the Stanford NLP Group

Introduction

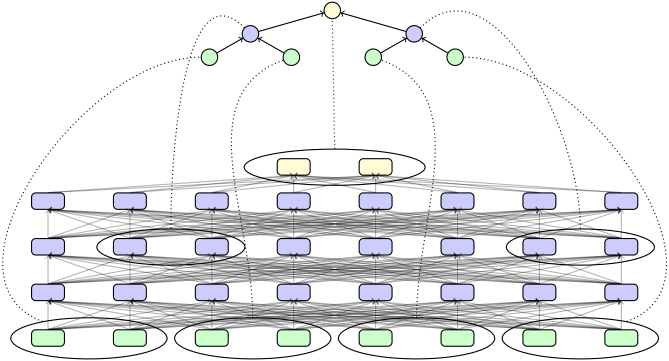
Deep Learning Model



Tree Structured Algorithm

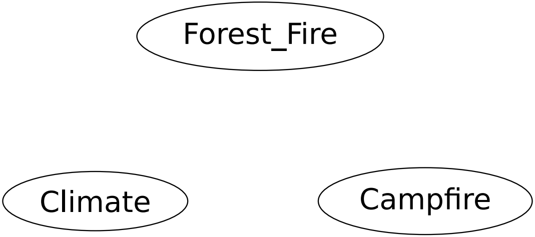


Constructive Causal Abstraction

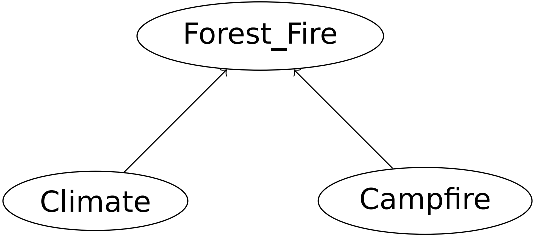


Acyclic Causal Models

Dynamics with an Intervention

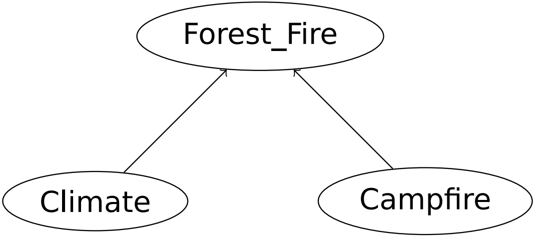


Dynamics with an Intervention



Dynamics with an Intervention

$$f_{\text{Forest\_Fire}}(\text{dry\_climate}, \text{campfire}) = \text{dry\_climate} \wedge \text{camp fire}$$

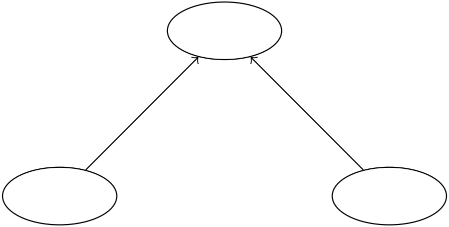


$f_{\text{Dry\_Climate}} = \mathit{True}$

$f_{\text{Campfire}} = \mathit{True}$

Dynamics with an Intervention

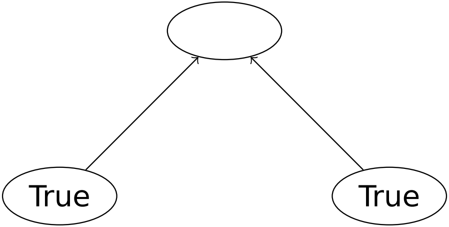
$$f_{\text{Forest\_Fire}}(\text{dry\_climate}, \text{campfire}) = \text{dry\_climate} \land \text{camp fire}$$



$$f_{\text{Dry\_Climate}} = \mathit{True} \qquad f_{\text{Campfire}} = \mathit{True}$$

Dynamics with an Intervention

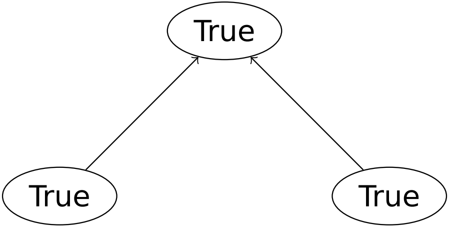
$$f_{\text{Forest\_Fire}}(\text{dry\_climate}, \text{campfire}) = \text{dry\_climate} \wedge \text{camp fire}$$



$f_{\text{Dry\_Climate}} = \mathit{True}$ $f_{\text{Campfire}} = \mathit{True}$

Dynamics with an Intervention

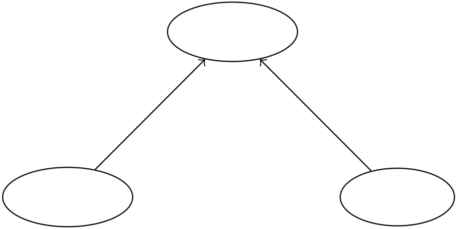
$$f_{\text{Forest\_Fire}}(\text{dry\_climate}, \text{campfire}) = \text{dry\_climate} \wedge \text{camp fire}$$



$f_{\text{Dry\_Climate}} = \mathit{True}$ $f_{\text{Campfire}} = \mathit{True}$

Dynamics with an Intervention

$$f_{\text{Forest\_Fire}}(\text{dry\_climate}, \text{campfire}) = \text{dry\_climate} \land \text{camp fire}$$

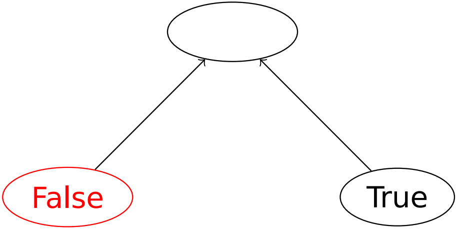


$f_{\text{Dry\_Climate}} = \mathit{True}$

$f_{\text{Campfire}} = \mathit{True}$

Dynamics with an Intervention

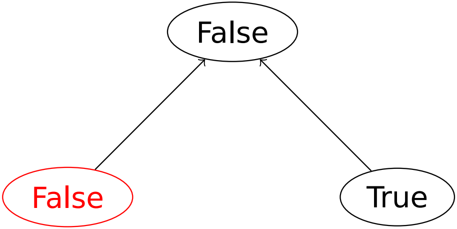
$$f_{\text{Forest\_Fire}}(\text{dry\_climate}, \text{campfire}) = \text{dry\_climate} \wedge \text{camp fire}$$



$f_{\text{Dry\_Climate}} = True$ $f_{\text{Campfire}} = True$

Dynamics with an Intervention

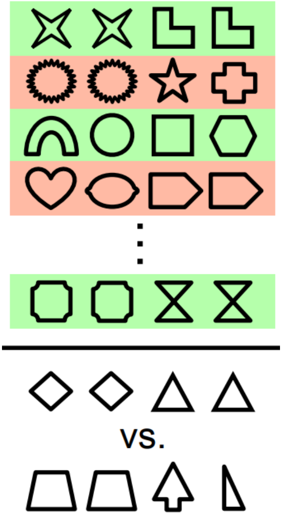
$$f_{\text{Forest\_Fire}}(\text{dry\_climate}, \text{campfire}) = \text{dry\_climate} \wedge \text{camp fire}$$



$f_{\text{Dry\_Climate}} = True$ $f_{\text{Campfire}} = True$

Hierarchical Equality Task

Hierarchical Equality Task



Algorithms as Acyclic Causal Models

Tree-Structured Algorithm

$\mathbf{Id_1}(\,\cdot\,,\,\cdot\,) =$

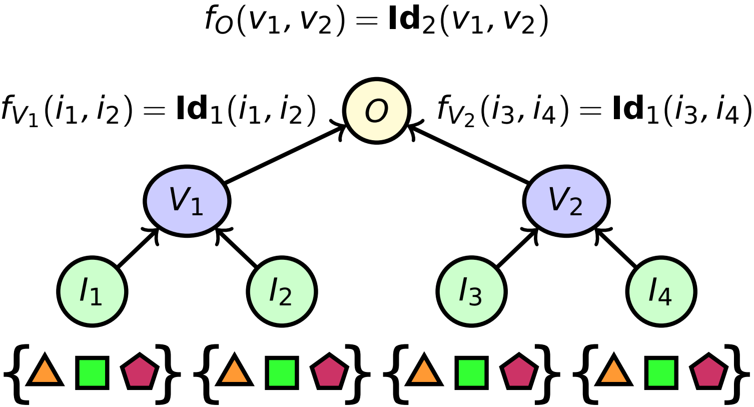
	△	🟩	⬟
△	T	F	F
🟩	F	T	F
⬟	F	F	T

$\mathbf{Id_2}(\,\cdot\,,\,\cdot\,) =$

	T	F
T	T	F
F	F	T

```
function EqualityTask($shape1, shape2, shape3, shape4$)
    $same1 \leftarrow \mathbf{Id_1}(shape1, shape2)$
    $same2 \leftarrow \mathbf{Id_1}(shape3, shape4)$
    $same3 \leftarrow \mathbf{Id_2}(same1, same2)$
    return $same3$
```

Causal Model of Algorithm



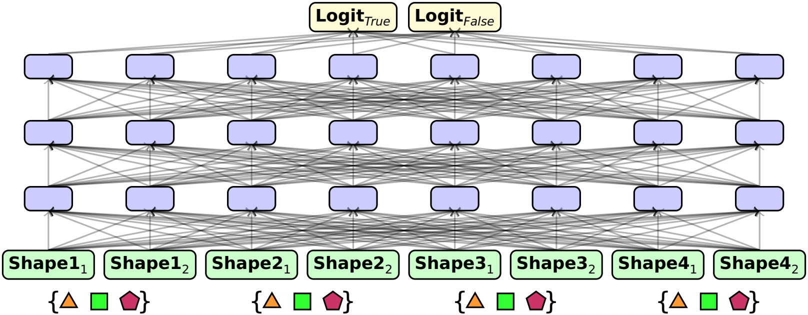
$$f_O(v_1, v_2) = \mathbf{Id}_2(v_1, v_2)$$

$$f_{V_1}(i_1, i_2) = \mathbf{Id}_1(i_1, i_2)$$

$$f_{V_2}(i_3, i_4) = \mathbf{Id}_1(i_3, i_4)$$

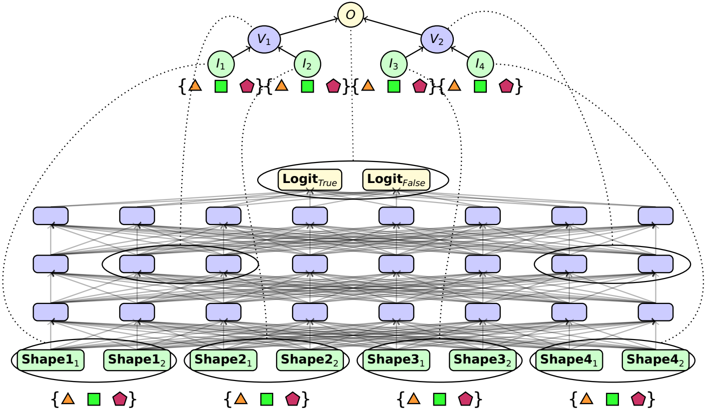
Deep Learning Models as Acyclic Causal Models

Deep Learning Models



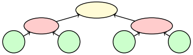
Constructive Causal Abstraction

Alignment

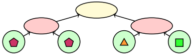


O
V_1
V_2
I_1
I_2
I_3
I_4
{▲ ■ ⬟}{▲ ■ ⬟}{▲ ■ ⬟}{▲ ■ ⬟}
Logit$_{True}$
Logit$_{False}$
Shape1$_1$
Shape1$_2$
Shape2$_1$
Shape2$_2$
Shape3$_1$
Shape3$_2$
Shape4$_1$
Shape4$_2$
{▲ ■ ⬟}
{▲ ■ ⬟}
{▲ ■ ⬟}
{▲ ■ ⬟}

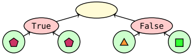
Interchange Interventions



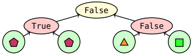
Interchange Interventions



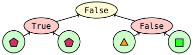
Interchange Interventions

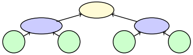


Interchange Interventions

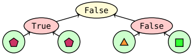


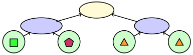
Interchange Interventions



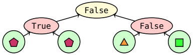


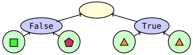
Interchange Interventions



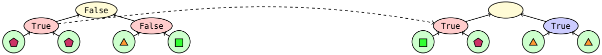


Interchange Interventions

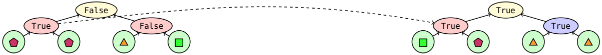




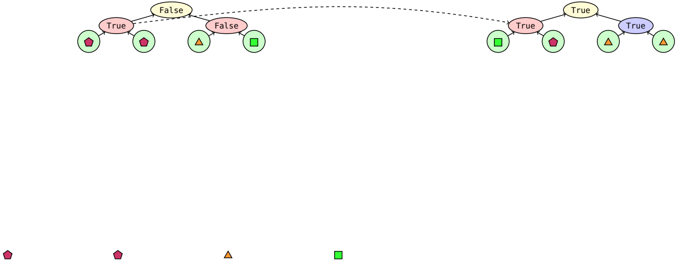
Interchange Interventions



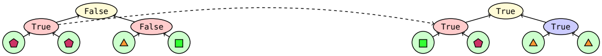
Interchange Interventions

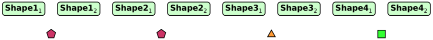


Interchange Interventions

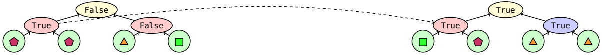


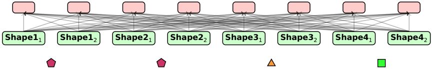
Interchange Interventions



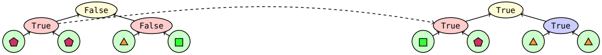


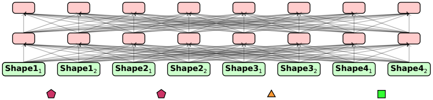
Interchange Interventions



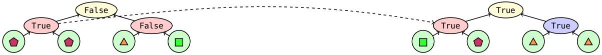


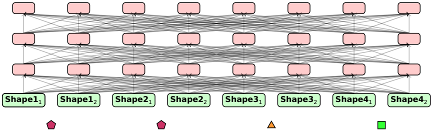
Interchange Interventions



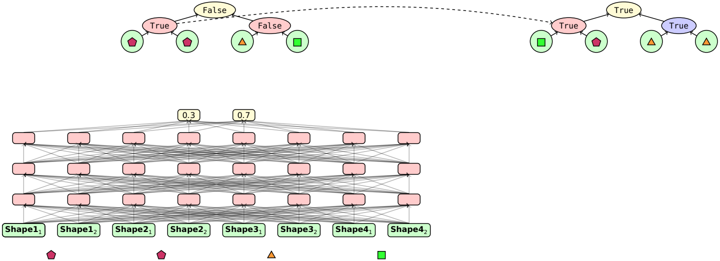


Interchange Interventions

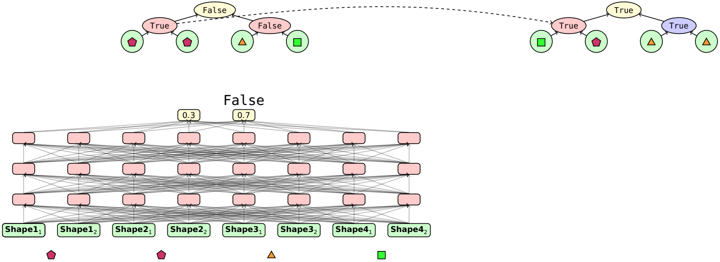




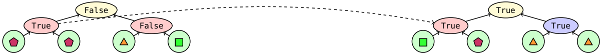
Interchange Interventions

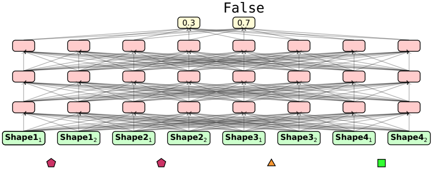


Interchange Interventions

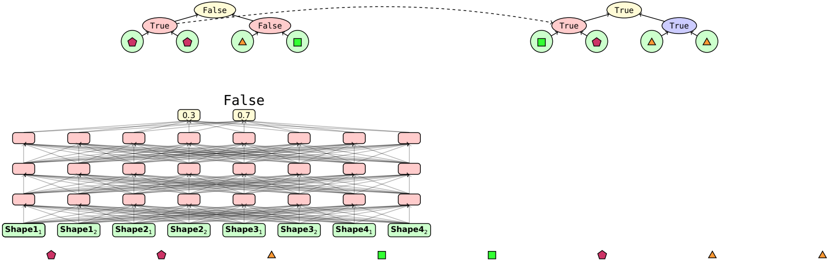


Interchange Interventions

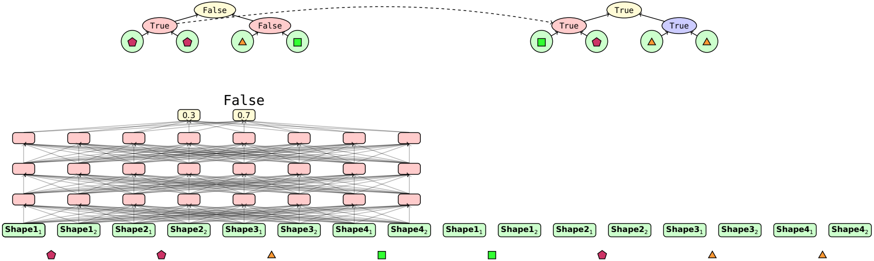




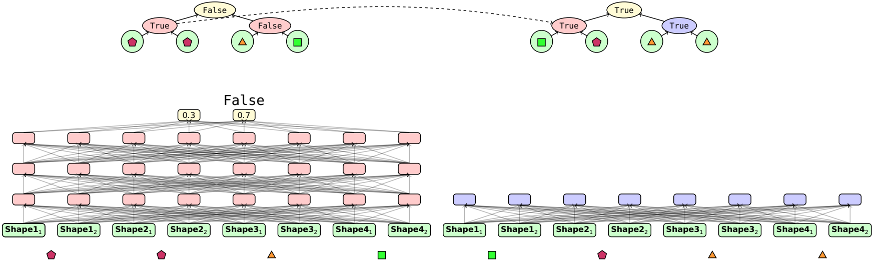
Interchange Interventions



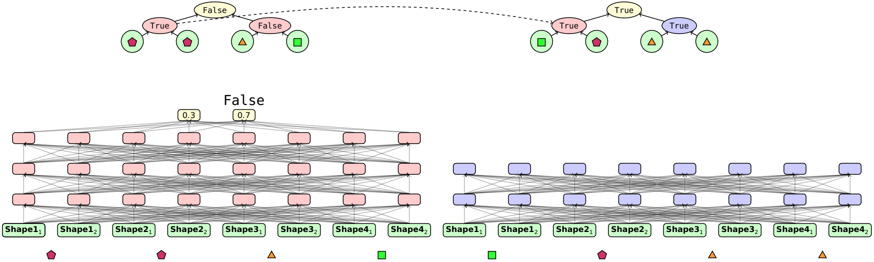
Interchange Interventions



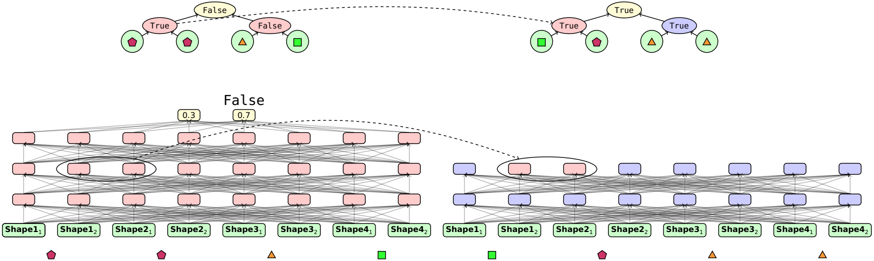
Interchange Interventions



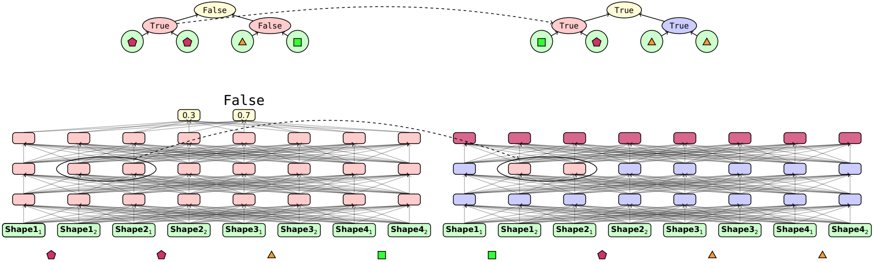
Interchange Interventions



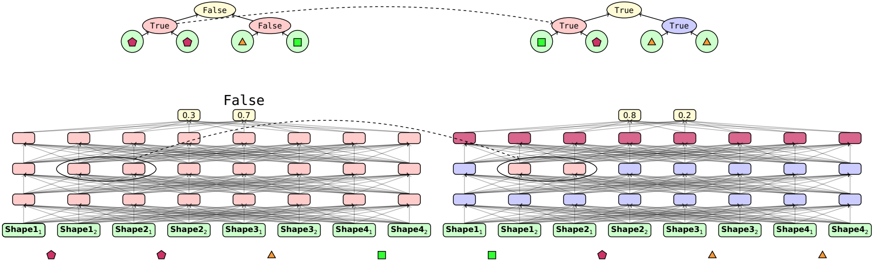
Interchange Interventions



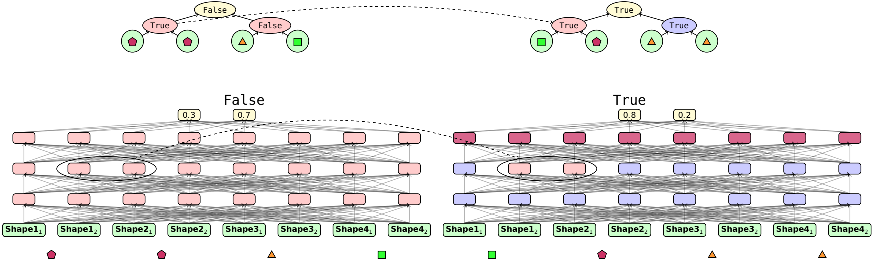
Interchange Interventions



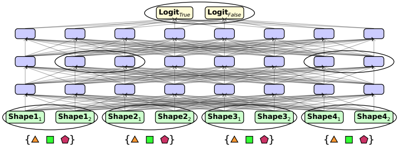
Interchange Interventions



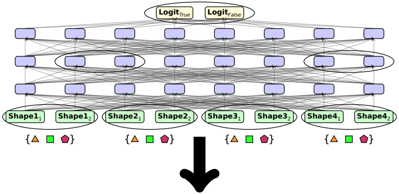
Interchange Interventions



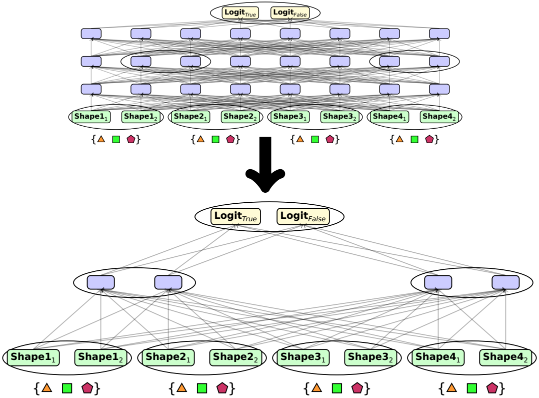
Marginalize



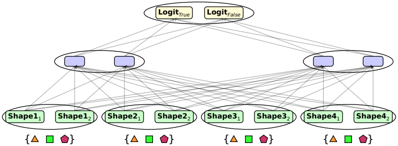
Marginalize



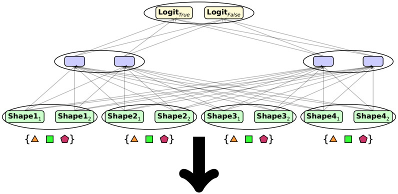
Marginalize



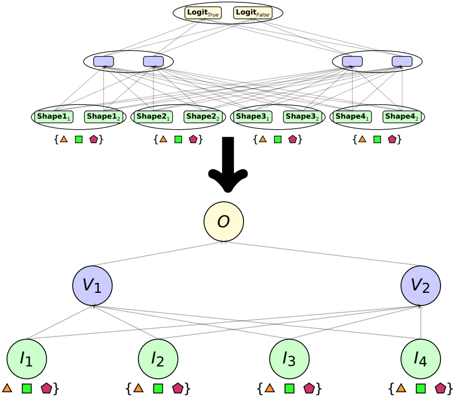
Variable Merge



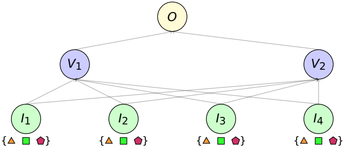
Variable Merge



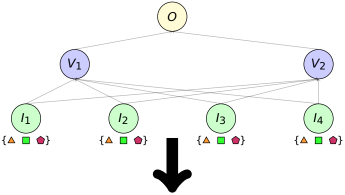
Variable Merge



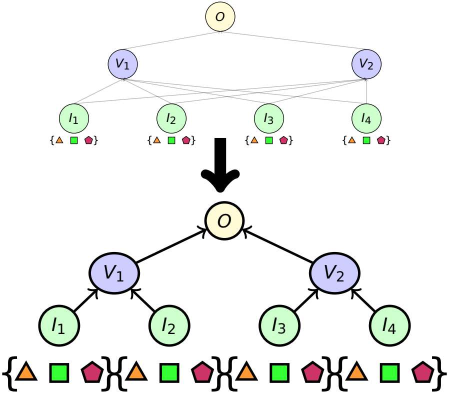
Value Change



Value Change



Value Change


O
V1
V2
I1
I2
I3
I4
O
V1
V2
I1
I2
I3
I4