Reduced Products of Abstract Domains for Fairness Certification of Neural Networks

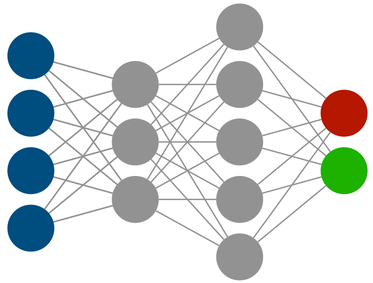
Denis Mazzucato and Caterina Urban

SAS 2021

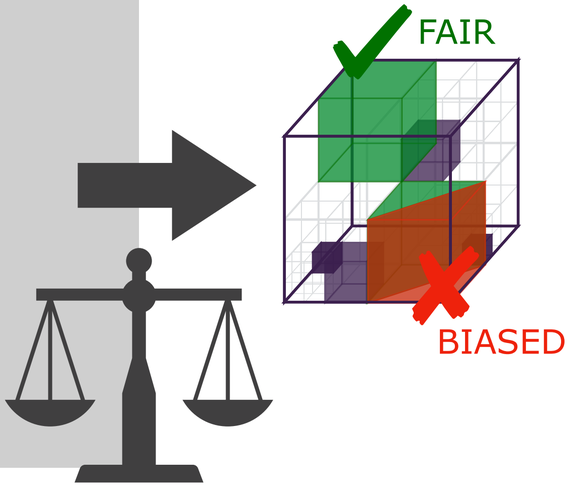
Social Impact of Machine Learning

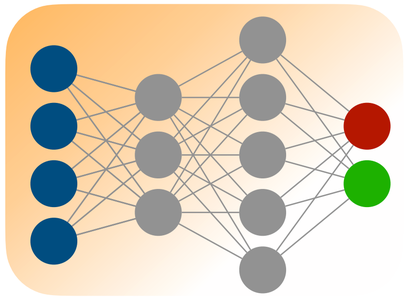
Artificial Intelligence Act

April 2021



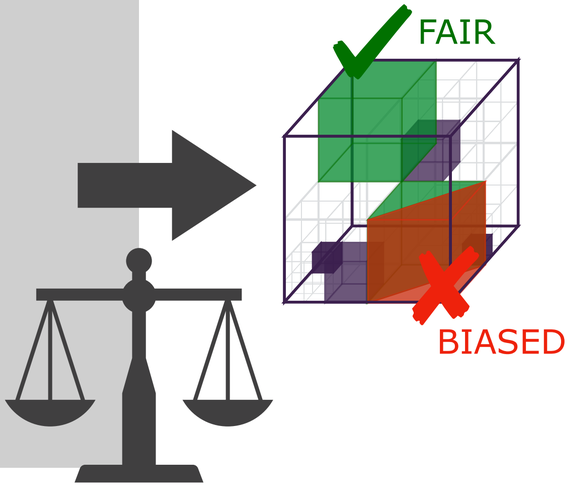
Libra



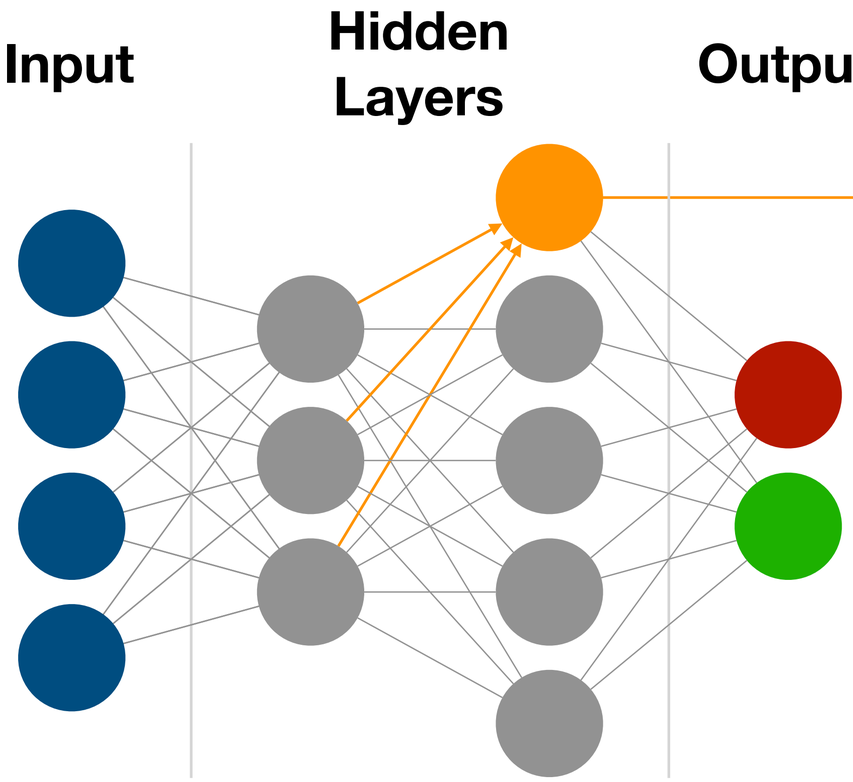


Neural Network

Libra



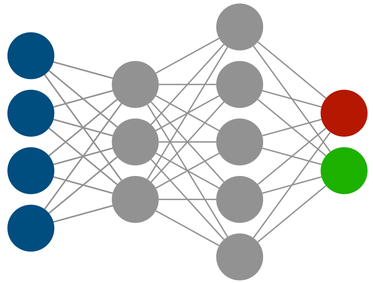
Feed-Forward **Neural Networks** with **ReLU Activations**



$$x_{i,j} = \max(\sum_{k} w_k \cdot x_{i-1,k} + b_{i,j}, 0)$$

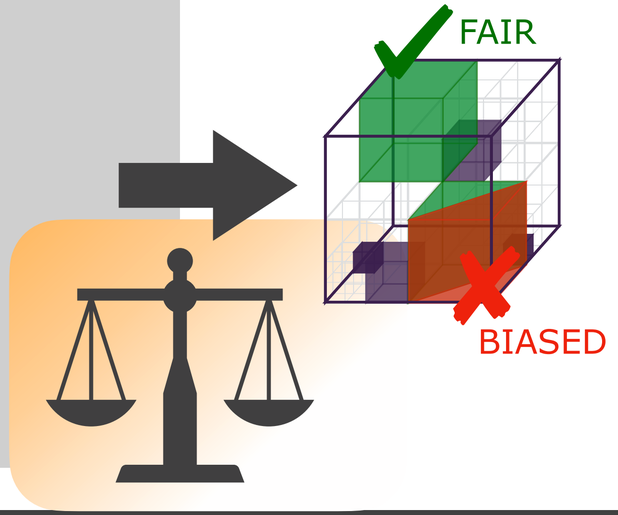
Classification Outcome

$$\underset{j}{\operatorname{argmax}}\ x_{i,j}$$



Neural Network

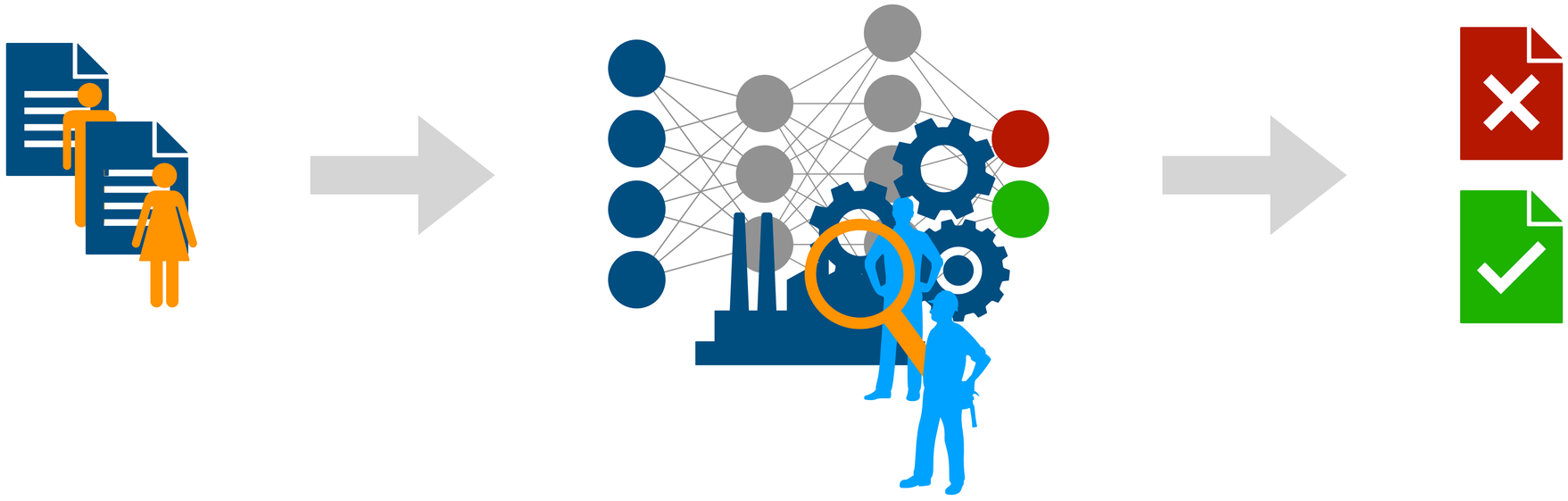
Libra



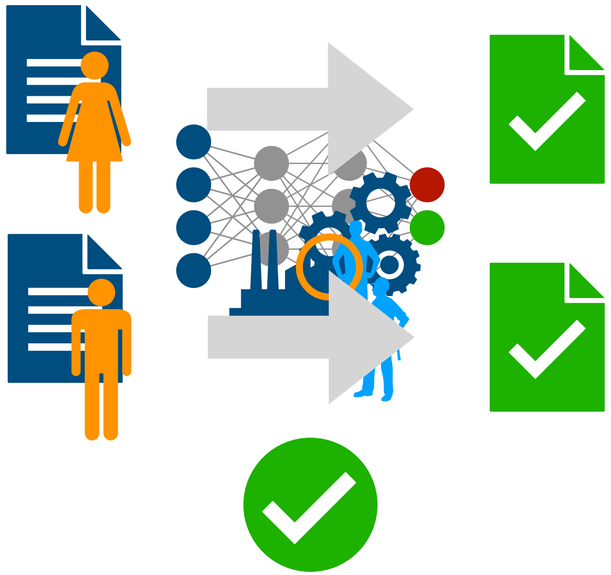
Dependency **Fairness**

The classification outcome is
Independent on the
Sensitive Features

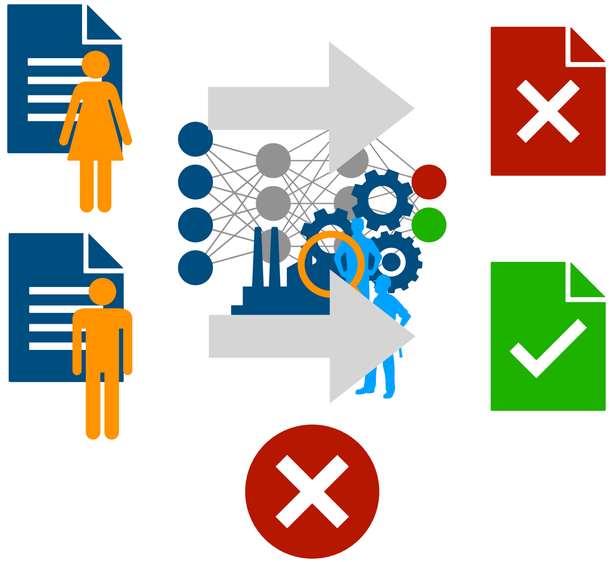
Recruiting Process



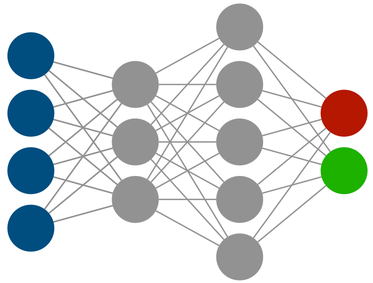
Recruiting Process



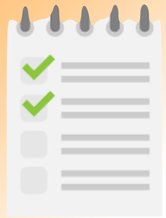
Fair



Unfair

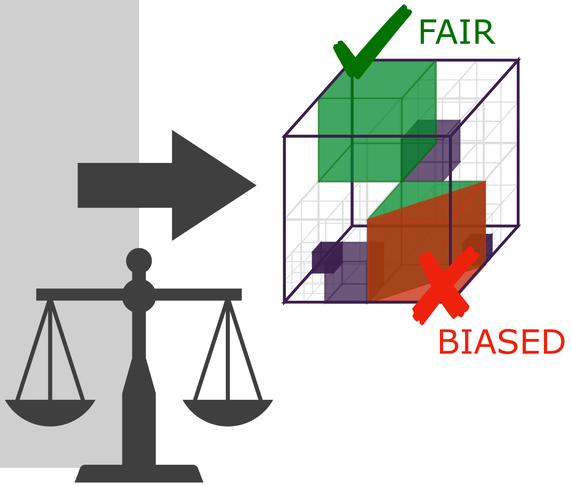


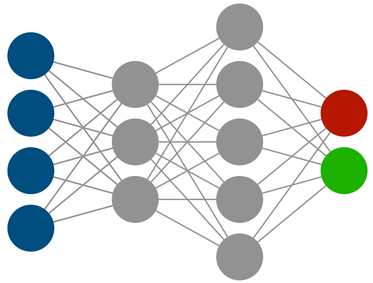
Neural Network



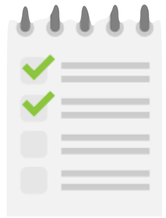
Specification

Libra

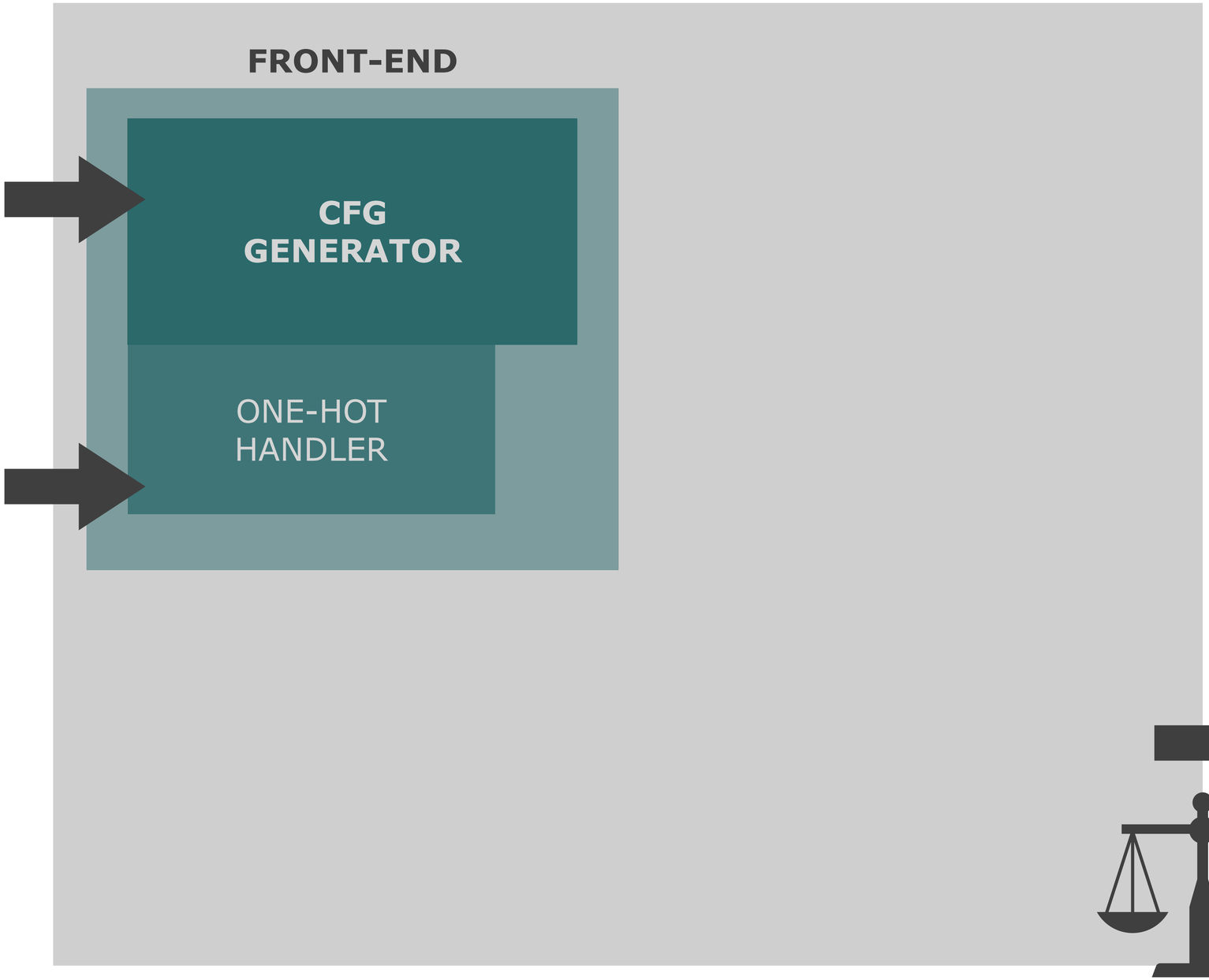


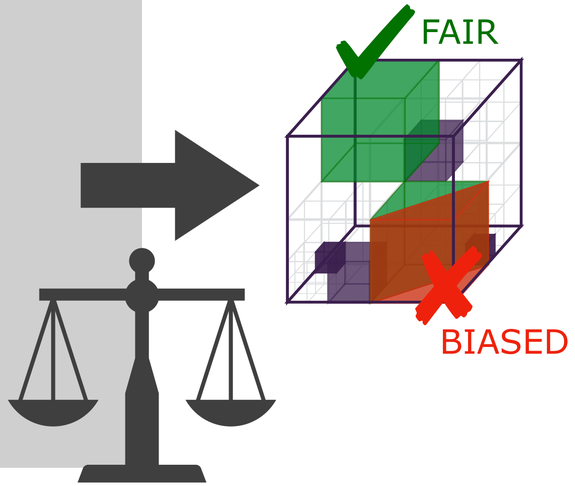


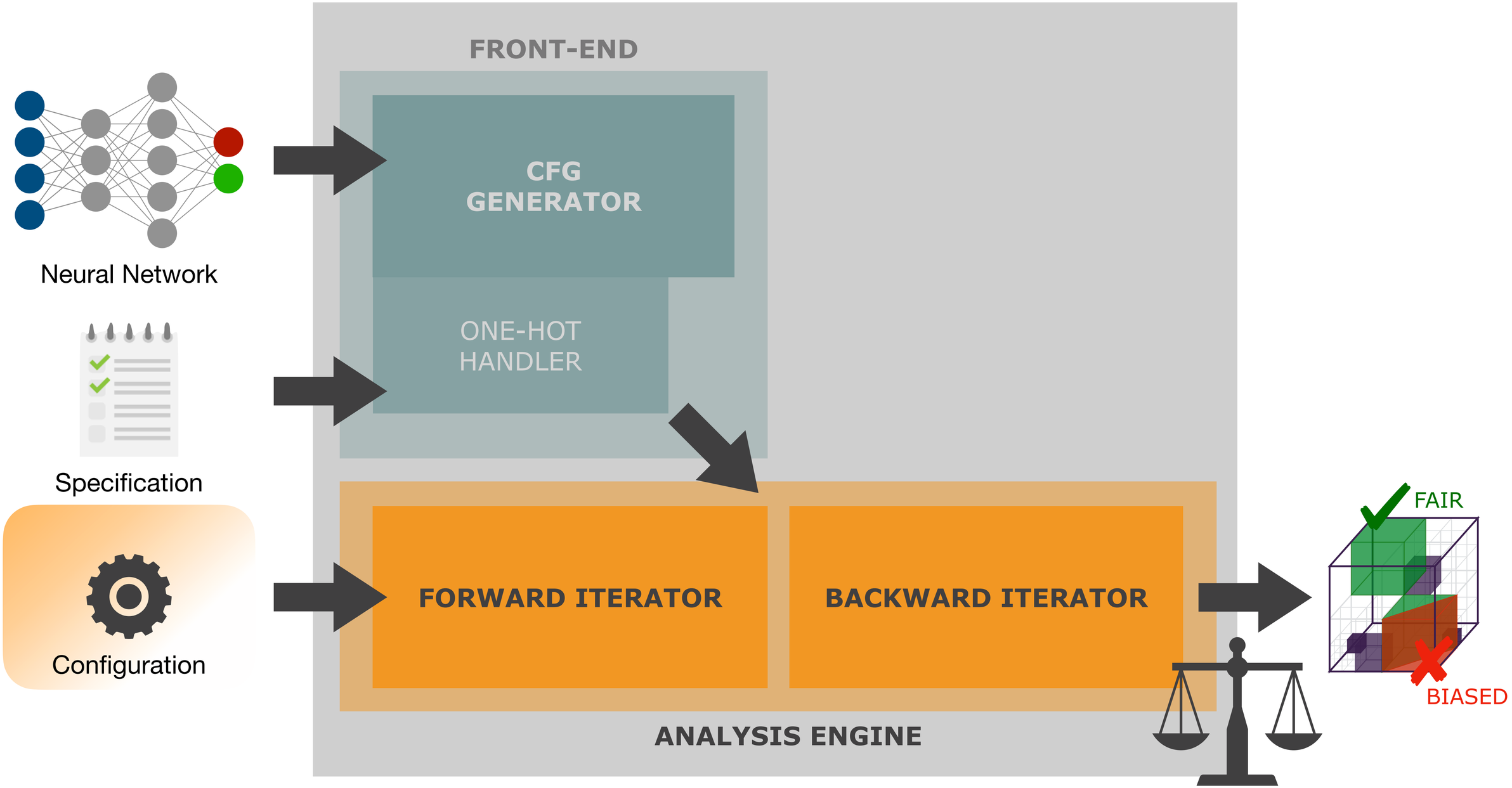
Neural Network



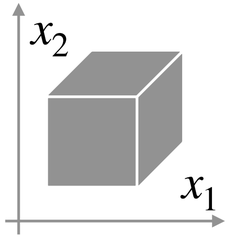
Specification



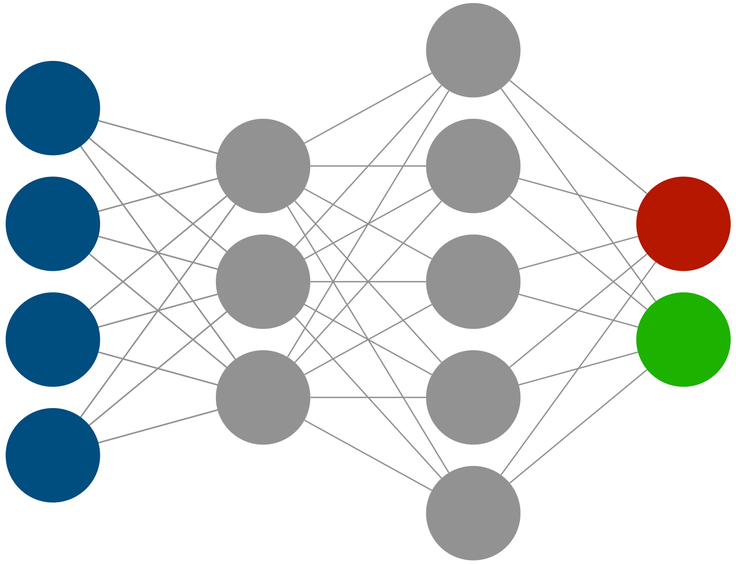


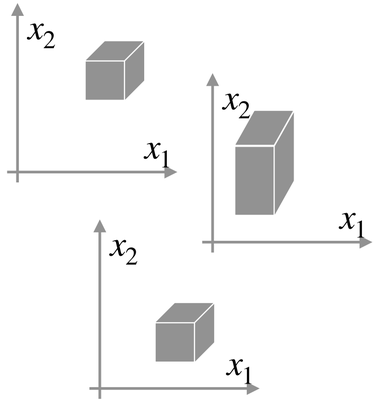

FRONT-END
CFG GENERATOR
ONE-HOT HANDLER
Neural Network
Specification
Configuration
FORWARD ITERATOR
BACKWARD ITERATOR
ANALYSIS ENGINE
FAIR
BIASED

Cheap **Forward Pre-Analysis**

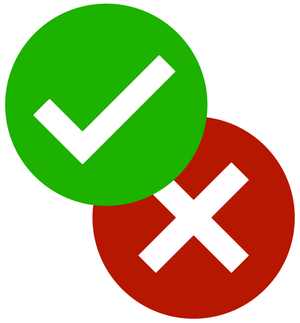


Input Space



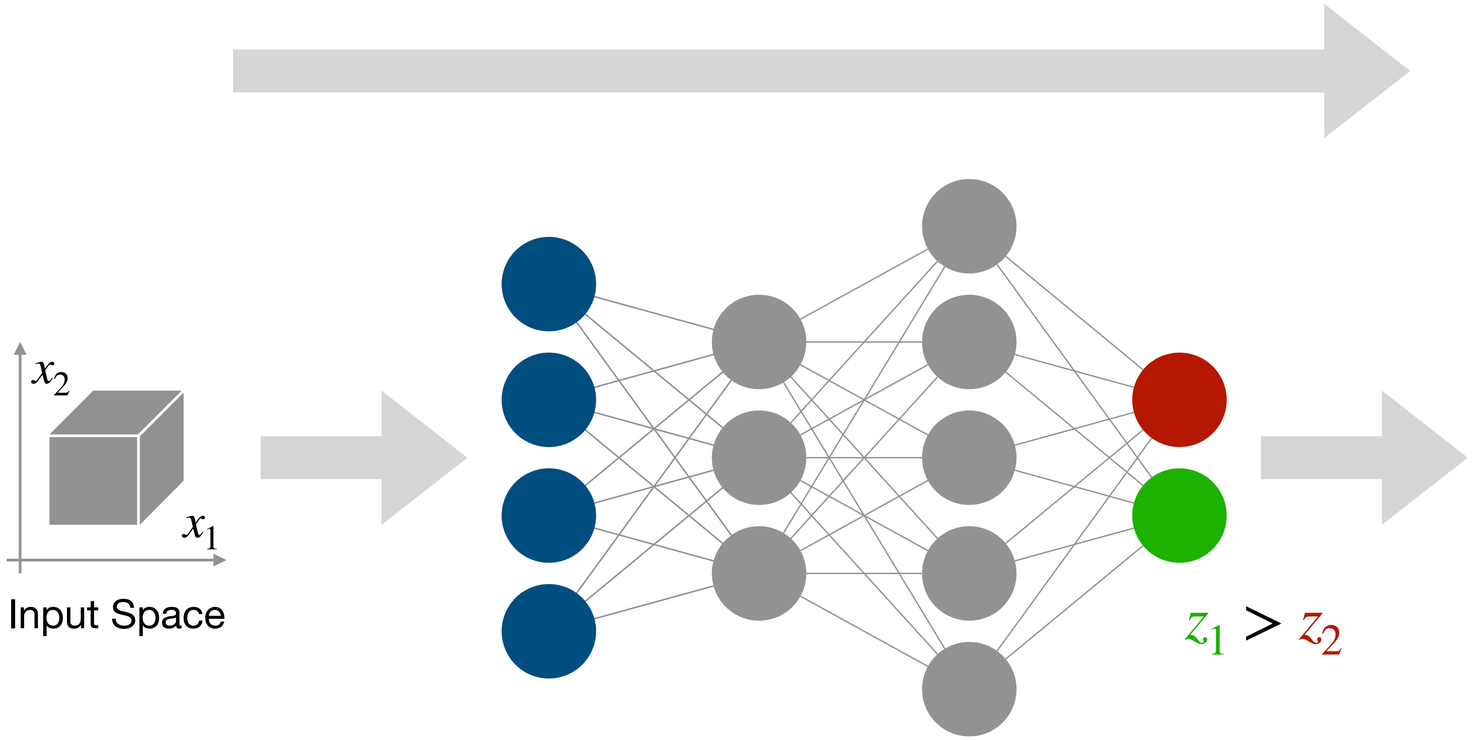


Partitions



Exact **Backward Analysis**
using **Polyhedra**

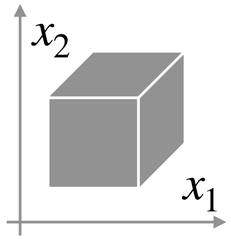
Cheap **Forward Pre-Analysis**



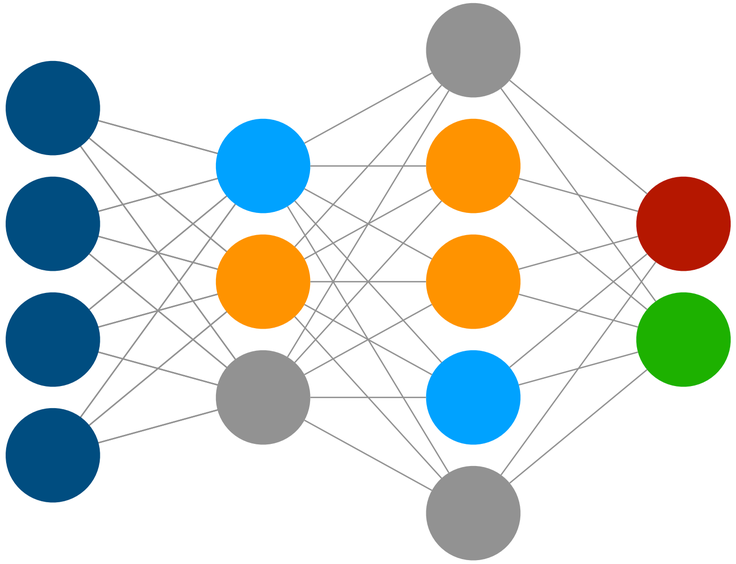
- **Fair**

Propagate the partition through the network via **abstract domains**

Cheap **Forward Pre-Analysis**



Input Space



- **Fair**
- Or, …

Activation Patterns

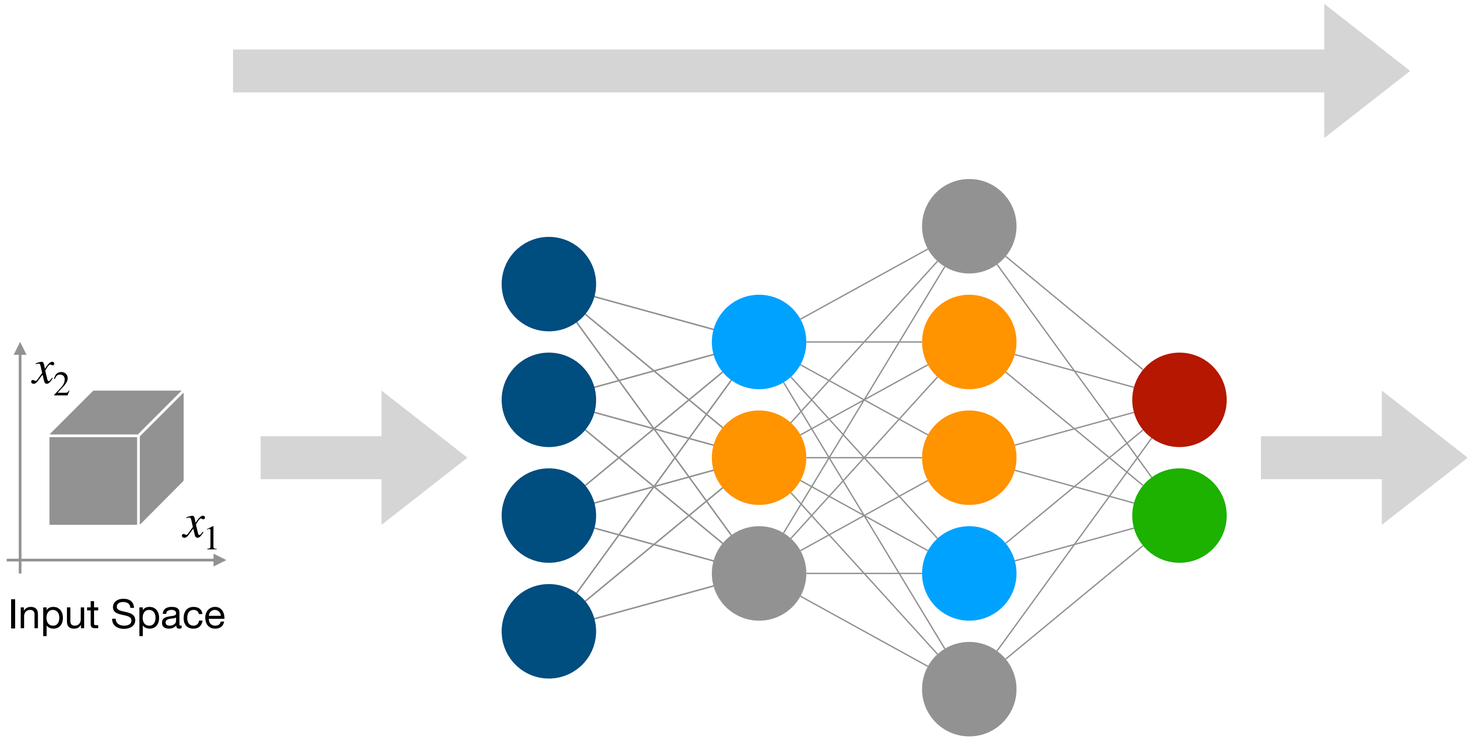
Always Negative $\max(x,0) = 0$

Always Positive $\max(x,0) = x$

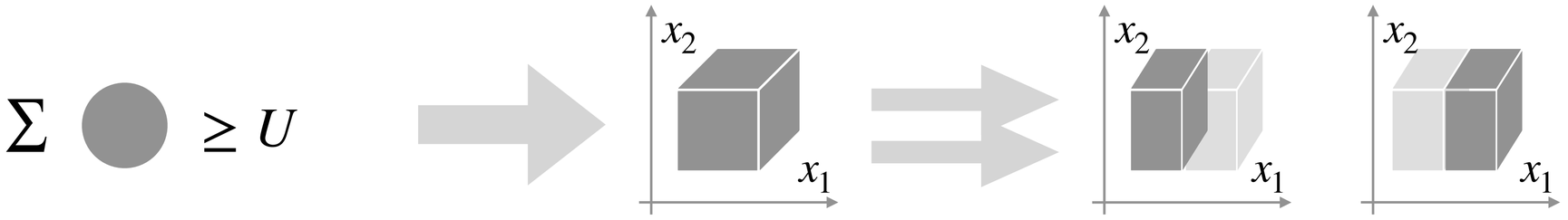
Fixed Activation Status

Uncertain $x \in [a, b]$ for $a < 0 < b$

Cheap **Forward Pre-Analysis**

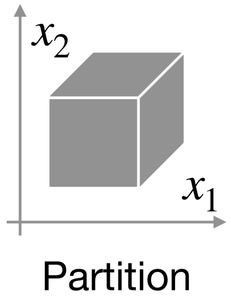


- **Fair**
  - **Partitioned**

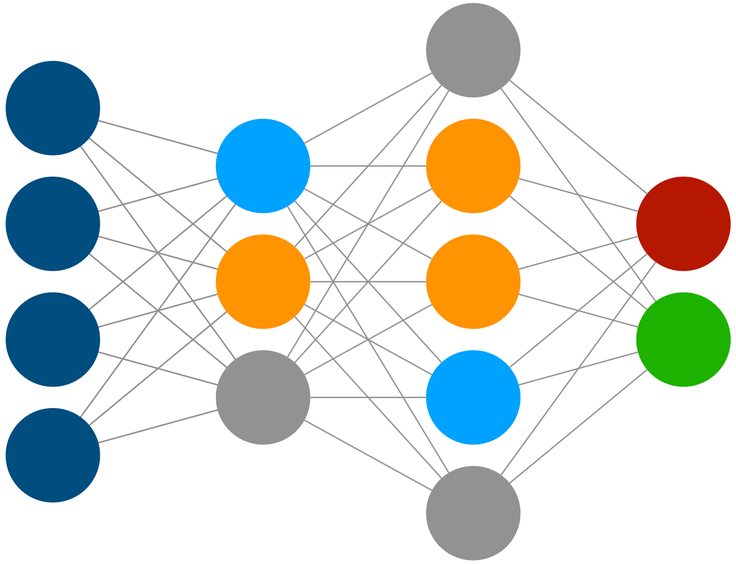


Along non-sensitive features only

Cheap **Forward Pre-Analysis**



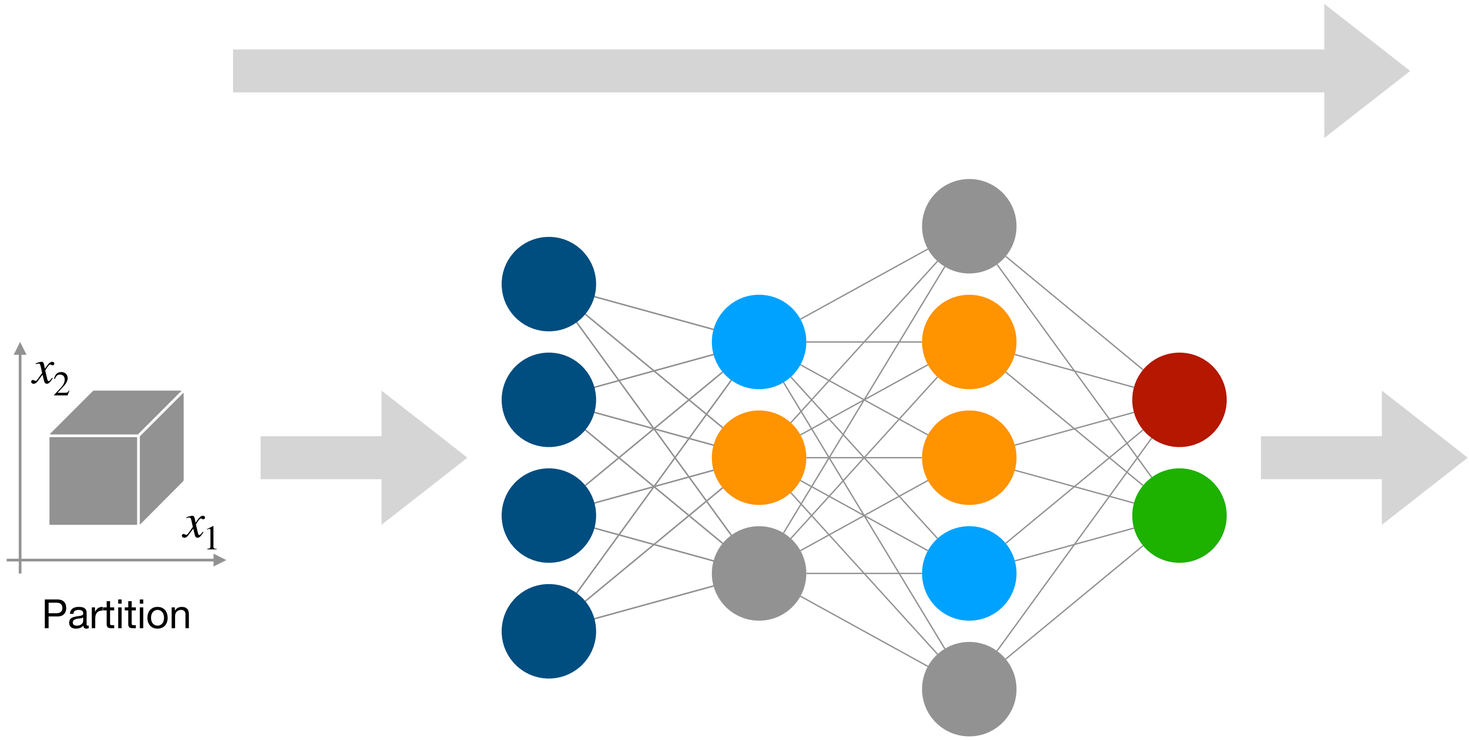
Partition



- **Fair**
- **Partitioned**
  - **Feasible**

$$\sum \quad \leq U$$

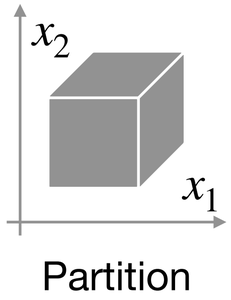
Cheap **Forward Pre-Analysis**

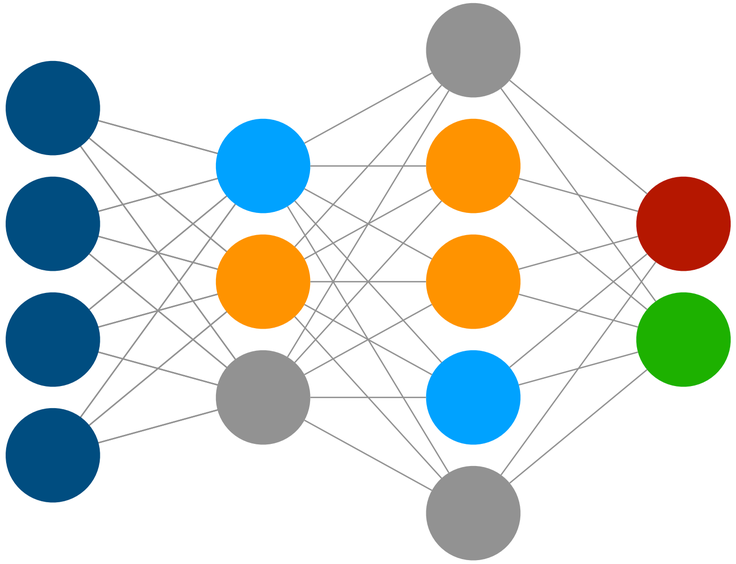


- **Fair**
- **Partitioned**
- **Feasible**
- ▸ **Excluded**

$\sum$ ● $\geq U$, and the partition becomes smaller than L

Cheap **Forward Pre-Analysis**





- **Fair**
- **Partitioned**
- **Feasible**
- **Excluded**

Budget constraints (L, U) can be
Automatically Configured to $(L_{min}, U_{\max})$

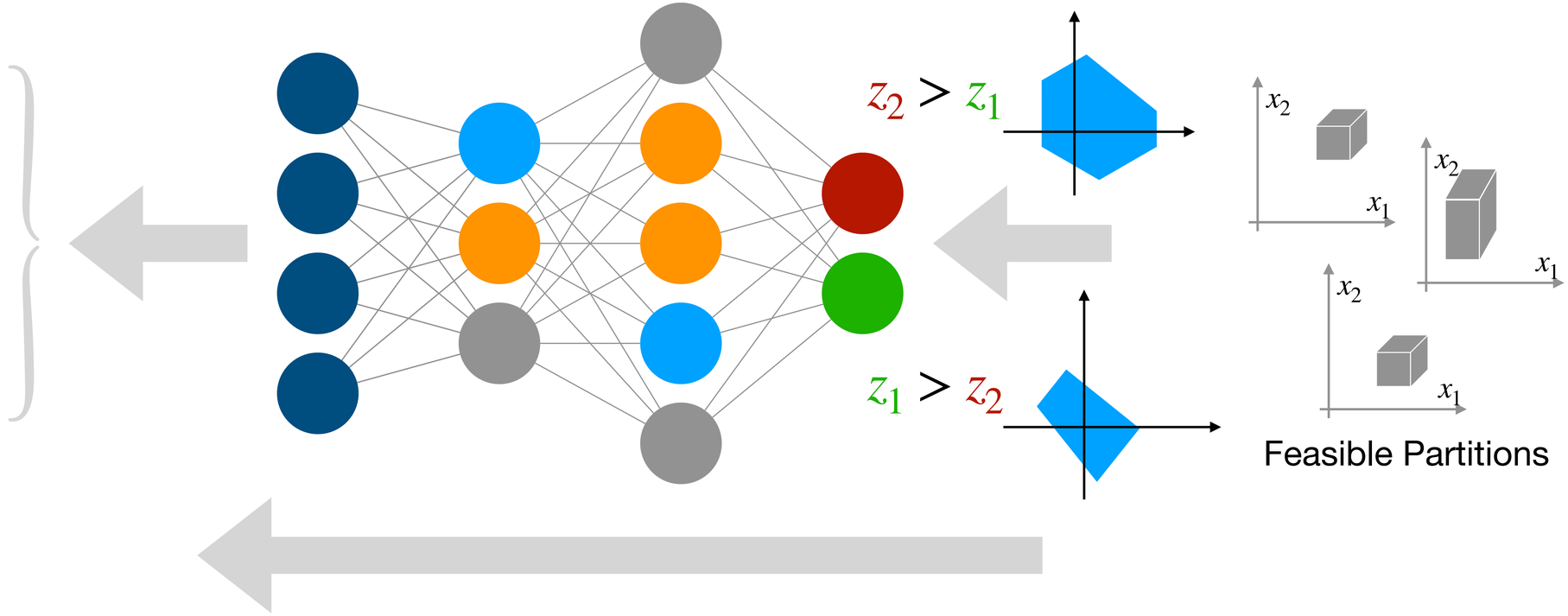
L down-to $L_{\min}$

U up-to $U_{\max}$

Proceed **Backwards**
for each **Feasible** partitions
for each possible **Outcome**

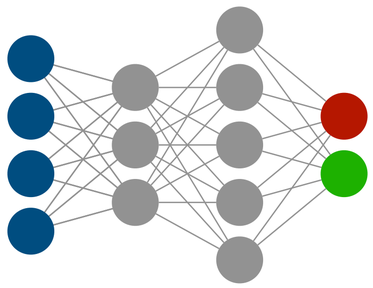
Empty
Intersection

Otherwise

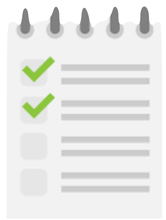


Feasible Partitions

Exact **Backward Analysis**
using **Polyhedra**



Neural Network



Specification

Configuration

FRONT-END

CFG GENERATOR

ONE-HOT HANDLER

ABSTRACT DOMAINS

BOXES

SYMBOLIC

DEEPPOLY

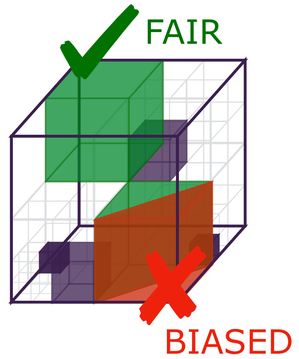
NEURIFY

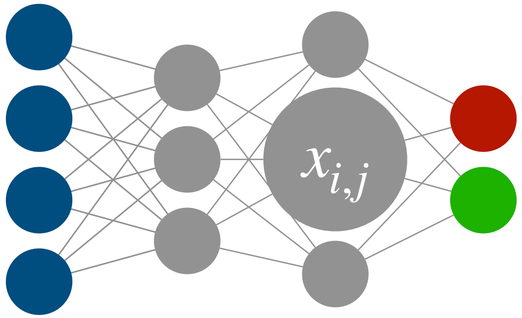
PRODUCT BUILDER

FORWARD ITERATOR

BACKWARD ITERATOR

ANALYSIS ENGINE



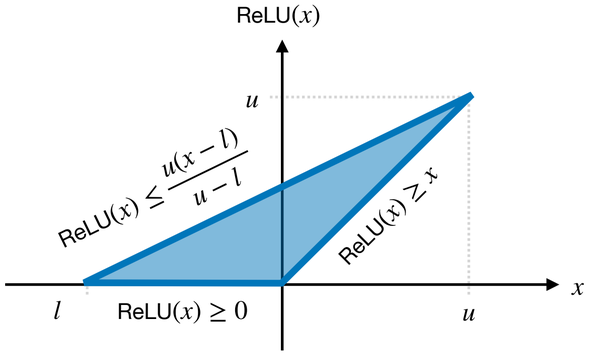


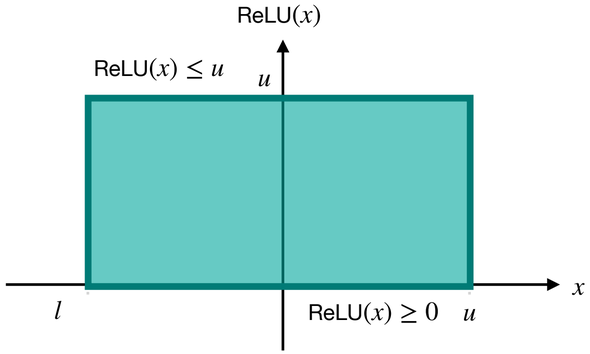
Symbolic

Li et al. @ SAS 2019

$[l, u]$

$$\sum_k m_k \cdot x_k + q$$



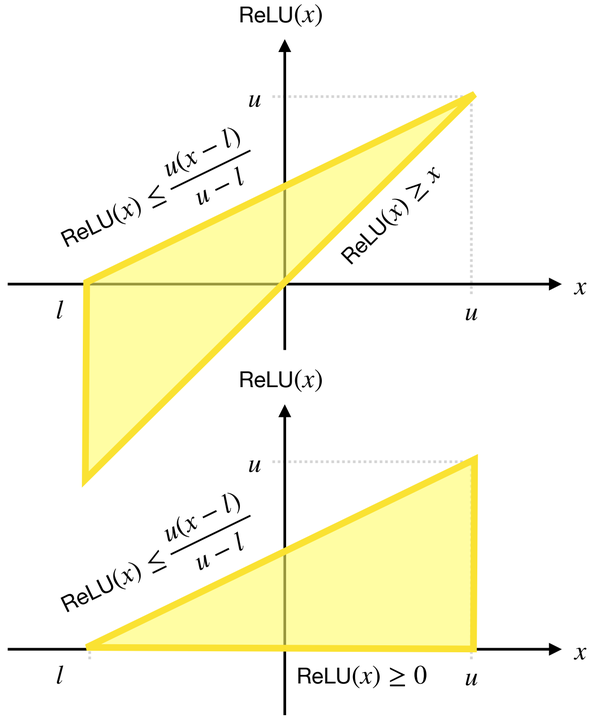


DeepPoly

Singh et al. @ POPL 2019

$[l, u]$

$[\text{eq}_{\text{low}}, \text{eq}_{\text{up}}]$

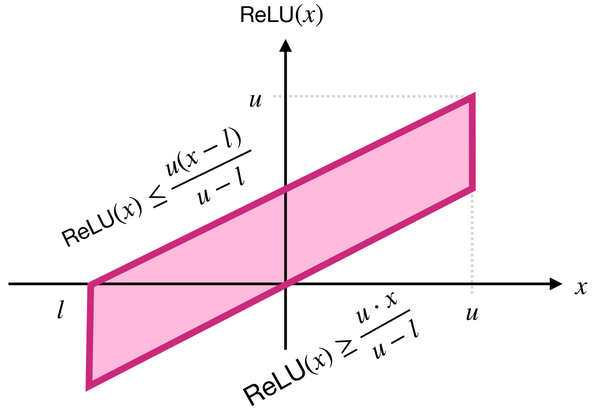


Neurify

Wang et al. @ NeurIPS 2018

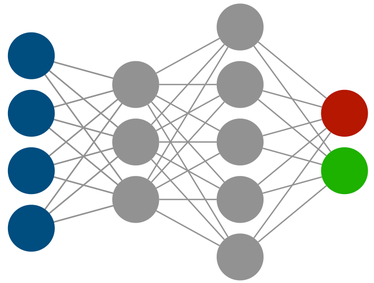
$[l_{\text{low}}, l_{\text{up}}, u_{\text{low}}, u_{\text{up}}]$

$[\text{eq}_{\text{low}}, \text{eq}_{\text{up}}]$

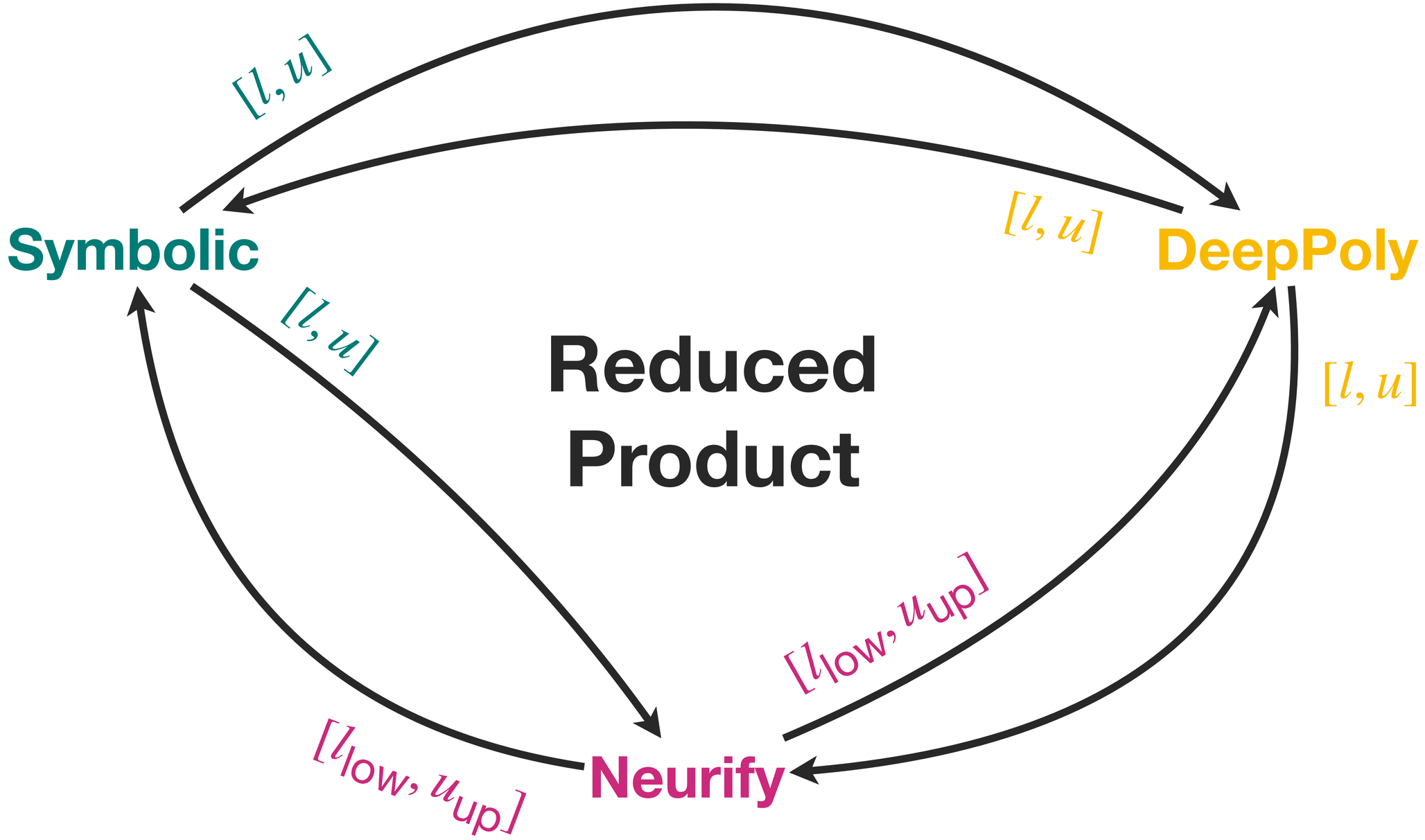


Precision-vs-Scalability

L	U	Symbolic	DeepPoly	Neurify
0.5	3	48.78%	49.01%	46.49%
	5	56.11%	56.15%	53.06%
0.25	3	83.63%	81.82%	81.40%
	5	91.67%%	91.58%	92.33%



- 4 Hidden Layers
- 5 Neuron per Layer
- 23 inputs $\in [0,1]$
- 2 Output classes


Symbolic
DeepPoly
Neurify
Reduced
Product
$[l, u]$
$[l, u]$
$[l, u]$
$[l, u]$
$[l_{low}, u_{up}]$
$[l_{low}, u_{up}]$

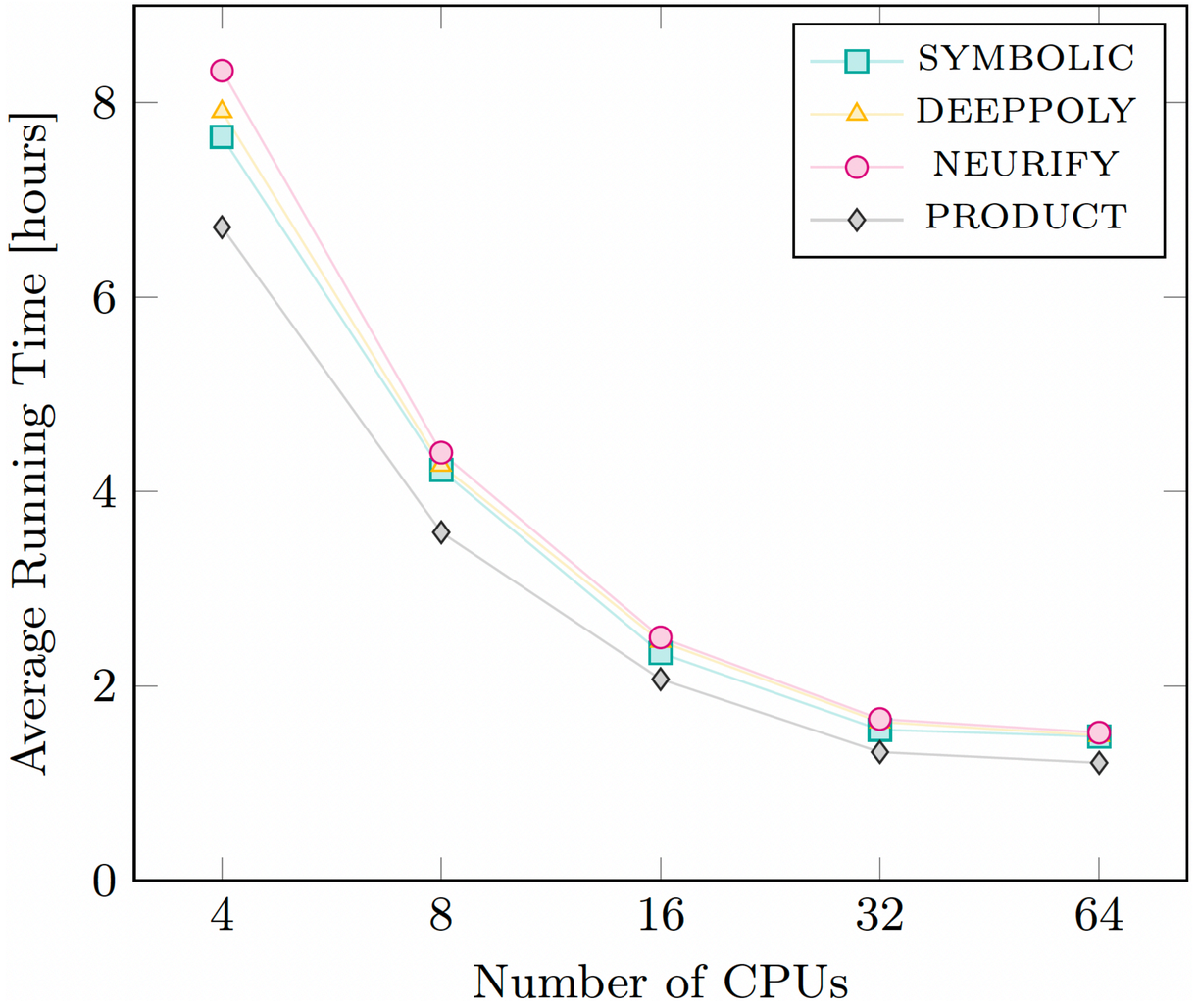
Precision-vs-Scalability

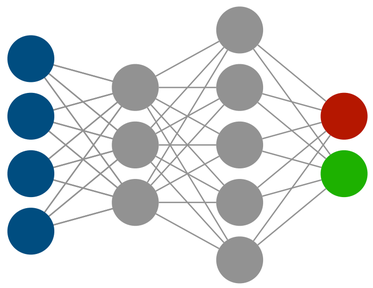
L	U	Symbolic	DeepPoly	Neurify	Product	
0.5	3	48.78%	49.01%	46.49%	59.20%	+10.3%
	5	56.11%	56.15%	53.06%	68.23%	+11.9%
0.25	3	83.63%	81.82%	81.40%	87.04%	+3.4%
	5	91.67%%	91.58%	92.33%	95.48%	+3.2%

Effect of Neural Network Structure

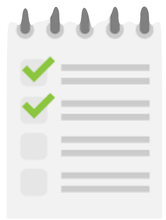
Size	Symbolic	DeepPoly	Neurify	Product	
10	98.72%	98.37%	98.51%	99.44%	**+0.7%**
12	76.70%	66.39%	64.58%	77.29%	**+0.6%**
20	56.11%	56.10%	53.06%	68.23%	**+12.1%**
40	34.72%	38.69%	41.22%	51.18%	**+10%**
45	43.78%	51.21%	50.59%	55.53%	**+4.3%**

Leveraging Multiple CPUs





Neural Network



Specification

Configuration

FRONT-END

ABSTRACT DOMAINS

CFG GENERATOR

ONE-HOT HANDLER

BOXES

SYMBOLIC

DEEPPOLY

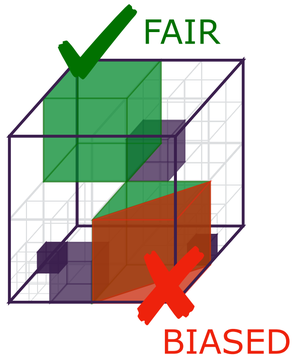
NEURIFY

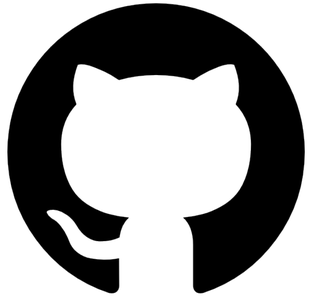
PRODUCT BUILDER

FORWARD ITERATOR

BACKWARD ITERATOR

ANALYSIS ENGINE





Check it out on **GitHub**!
https://github.com/caterinaurban/libra

Ready-to-go **Docker image*** at
https://doi.org/10.5281/zenodo.4737450

* no installation needed!