

# **Reduced Products of Abstract Domains for Fairness Certification of Neural Networks**

Denis Mazzucato and Caterina Urban

**SAS 2021**

NEWS OCTOBER 10, 2018 / 5:12 AM / A YEAR AGO

WIRED

In 2019, predictive algorithms will start to make banking fair for all

WIRED

The AI Doctor Will See You Now

Advances in neural networks and other techniques promise to transform health care while raising profound questions about our bodies and society.

Google Translate

DETECT LANGUAGE ENGLISH FRENCH SPANISH

A nurse  
A doctor

16/5000

Une infirmière  
Un médecin

100%

Sign in

nature

SEARCH

NEWS · 24 OCTOBER 2019

UPDATE 26 OCTOBER 2019

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

WIRED

Amazon scraps secret AI recruiting tool that showed bias against women

ERIC NIILER BUSINESS 03.25.2019 07:00 AM

Can AI Be a Fair Judge in Court? Estonia Thinks So

Estonia plans to use an artificial intelligence program to decide some small-claims cases, part of a push to make government services smarter.

WIRED

Business 12.21.2019 08:00 AM

The Telegraph

AI used for first time in job interviews in UK

By Charles Hymas

27 SEPTEMBER 2019 · 10:00 PM

AUTOMATED BACKGROUND CHECKS ARE DECIDING WHO'S FIT FOR A HOME

By Colin Lecher | @colinlecher | Feb 1, 2019, 8:00am EST

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

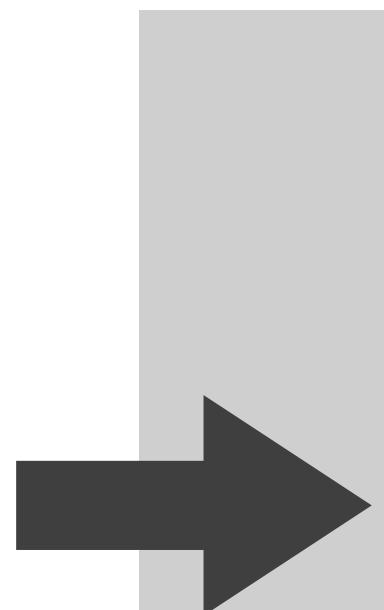
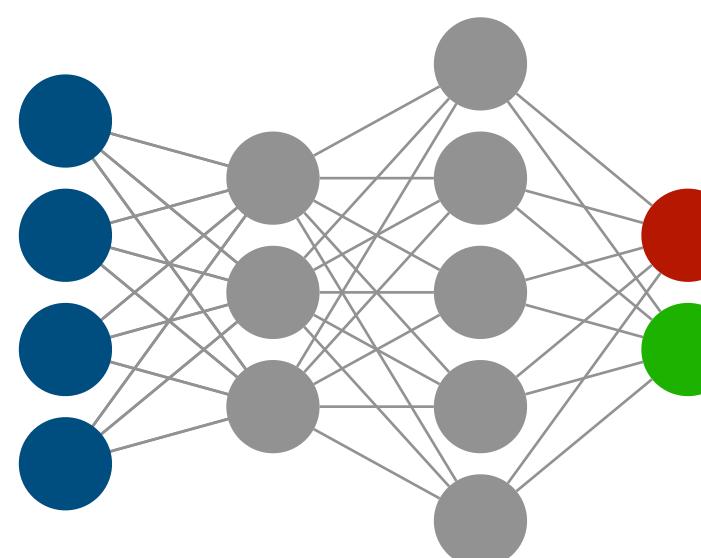
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

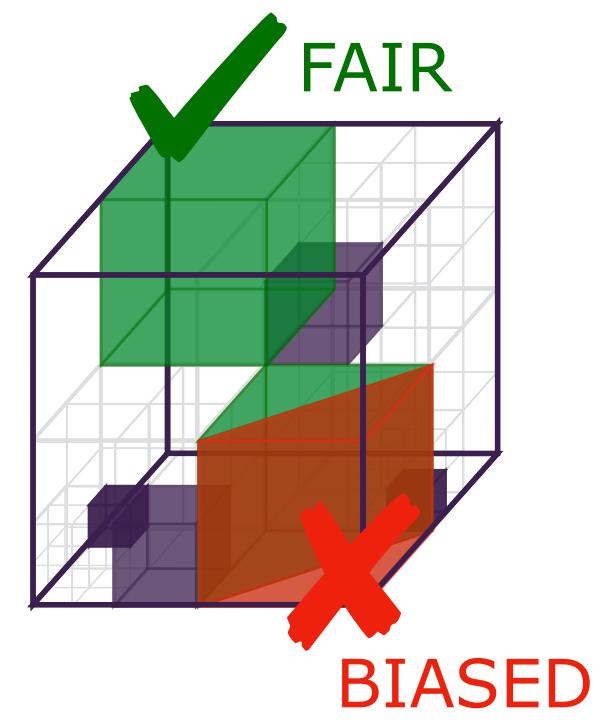


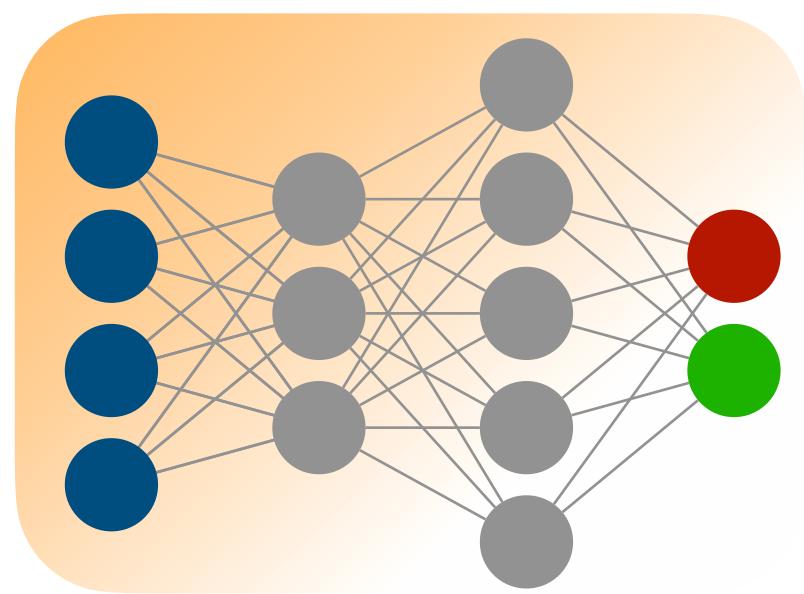
# Artificial Intelligence Act

April 2021



# Libra

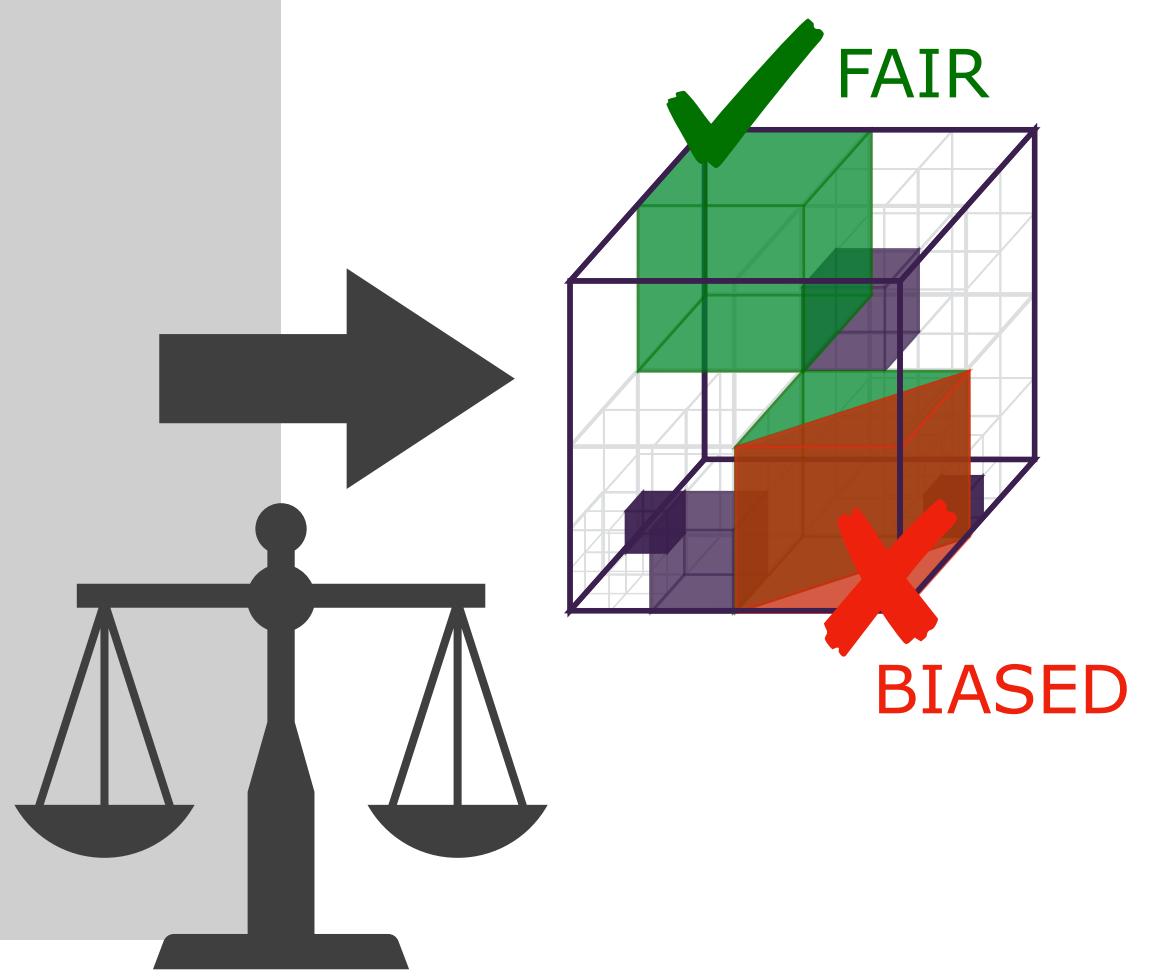




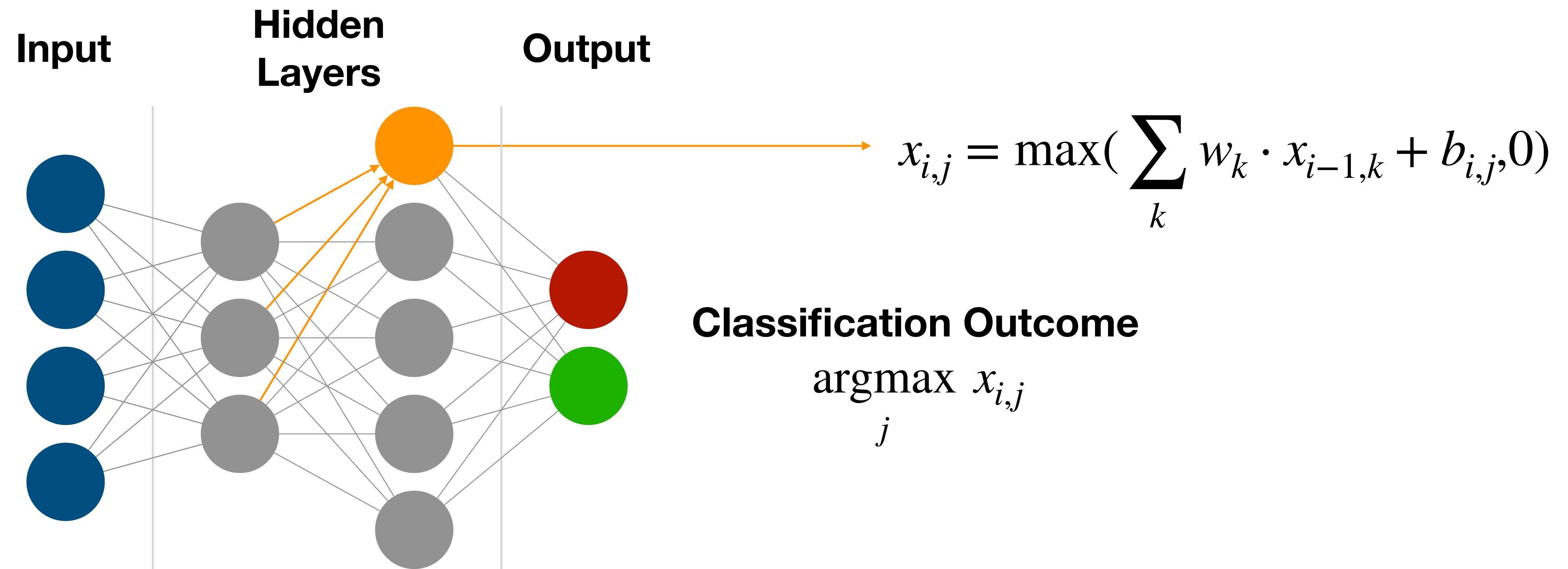
Neural Network

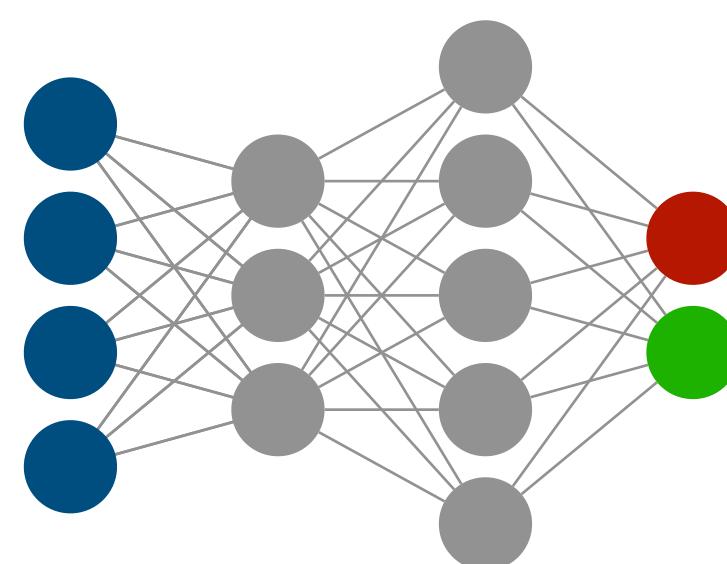


# Libra

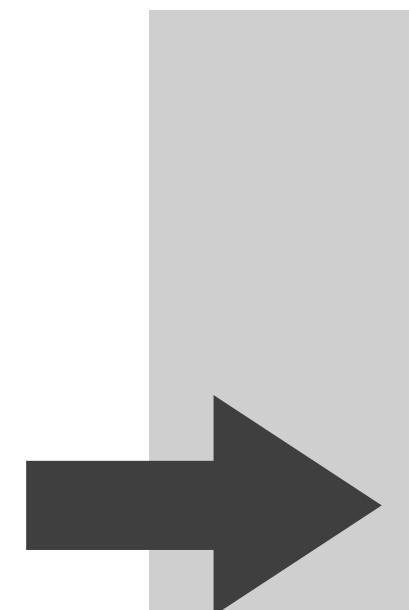


# Feed-Forward Neural Networks with ReLU Activations

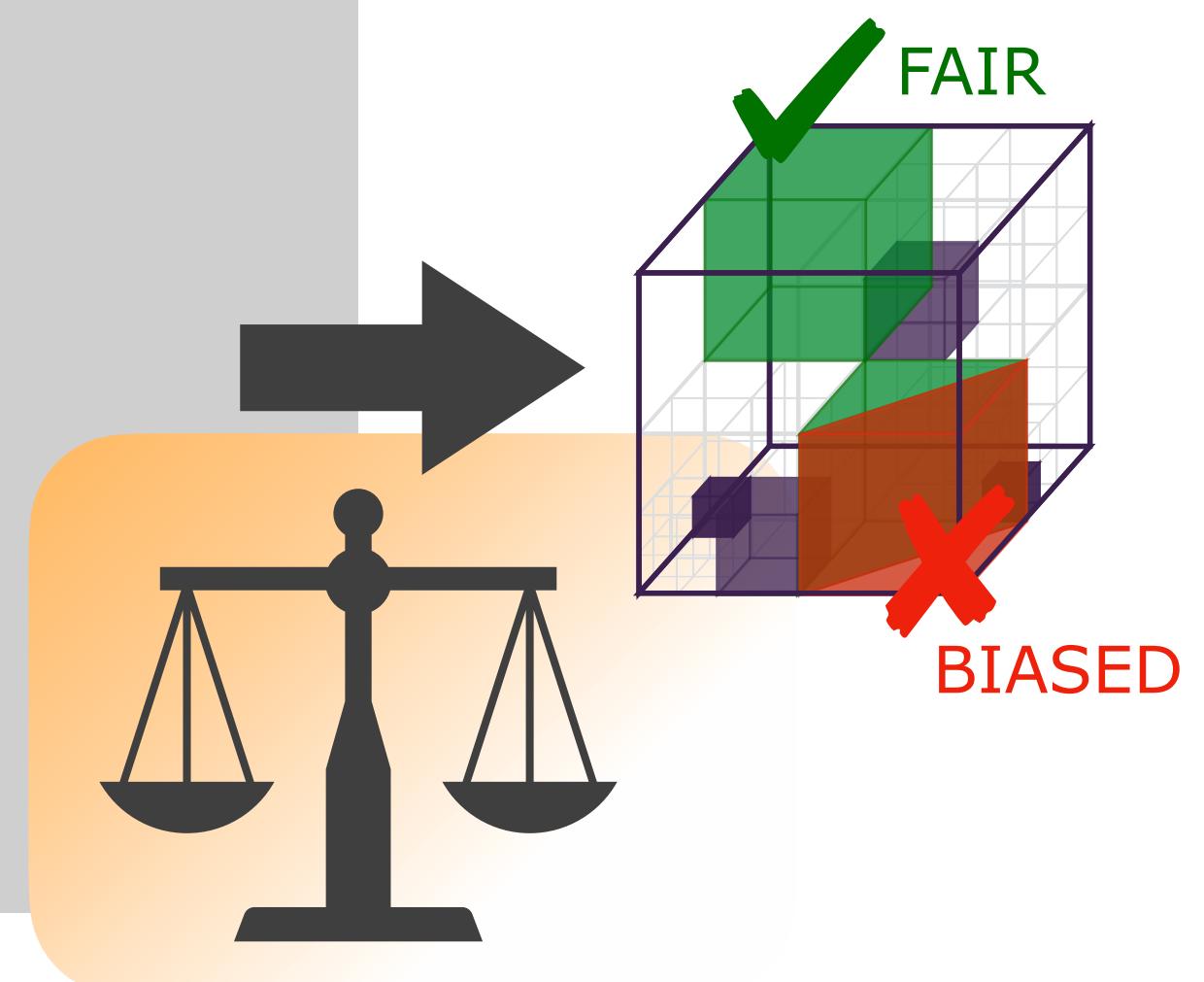




Neural Network



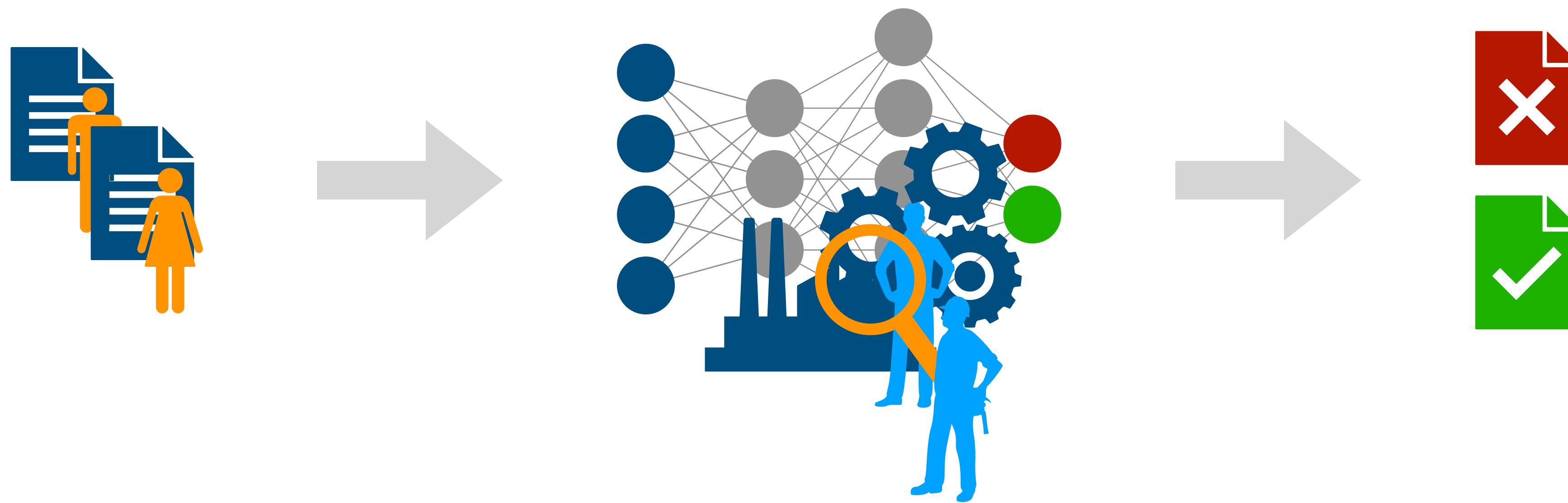
# Libra



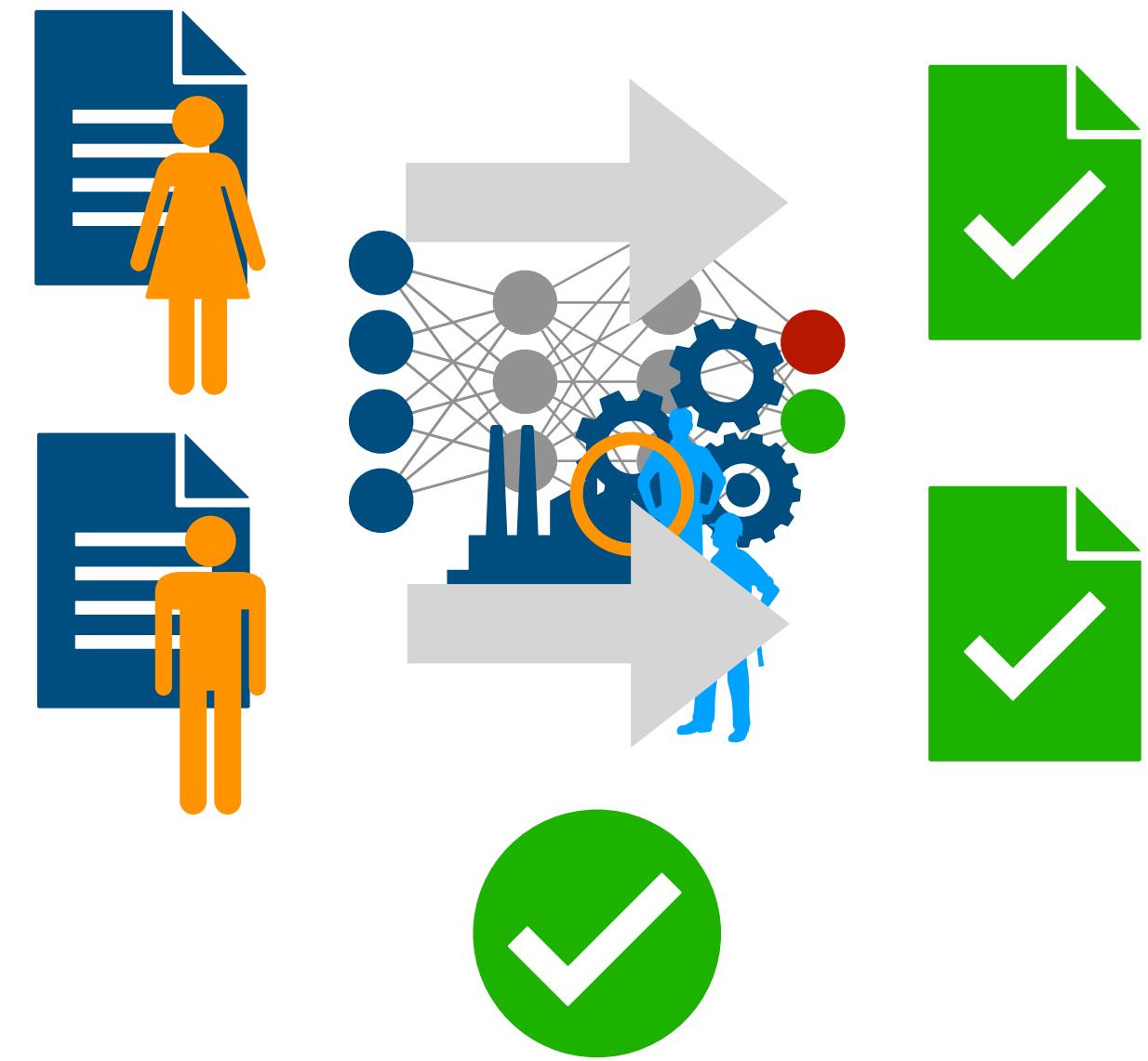
# Dependency **Fairness**

The classification outcome is  
**Independent** on the  
**Sensitive Features**

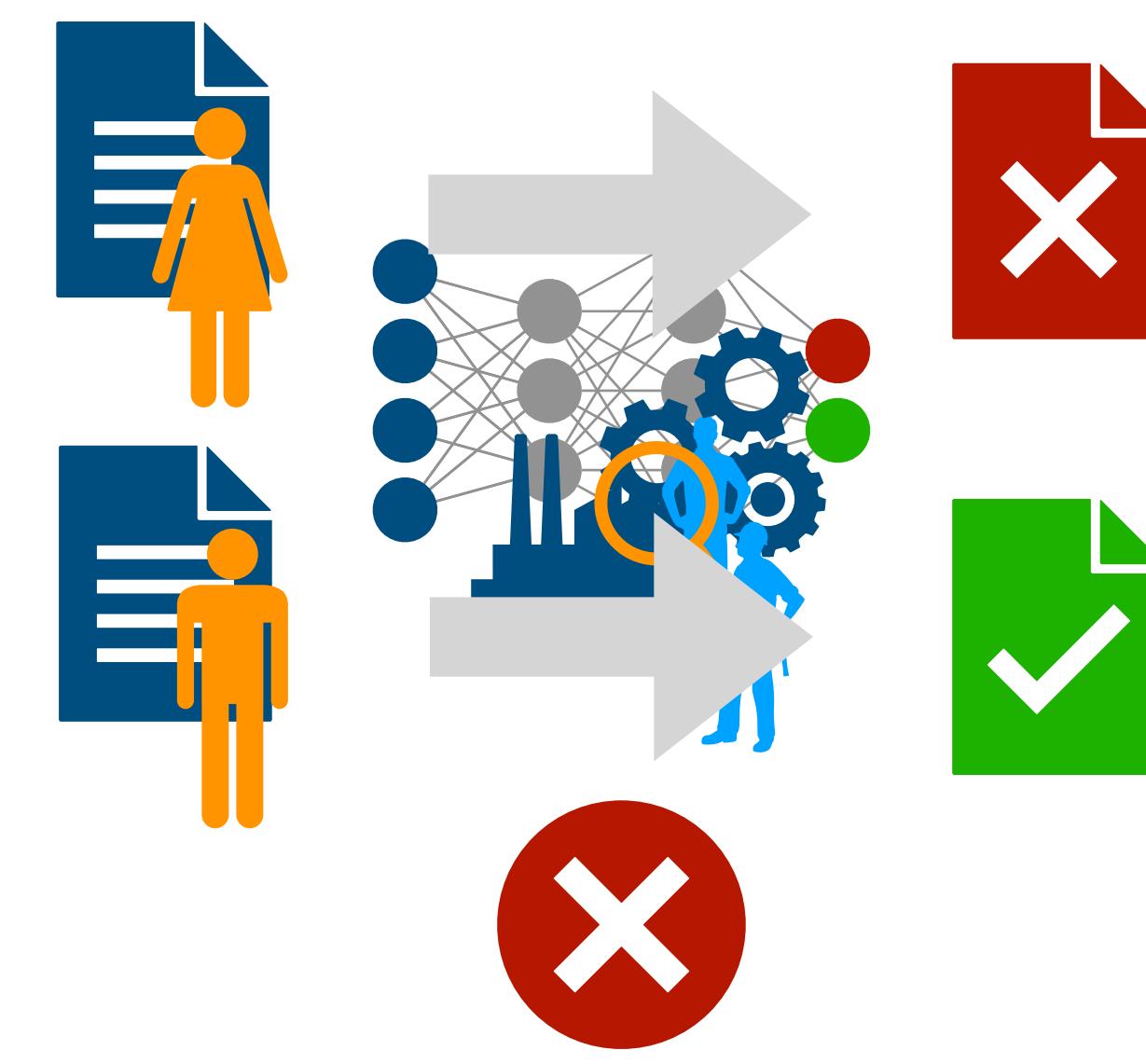
# Recruiting Process



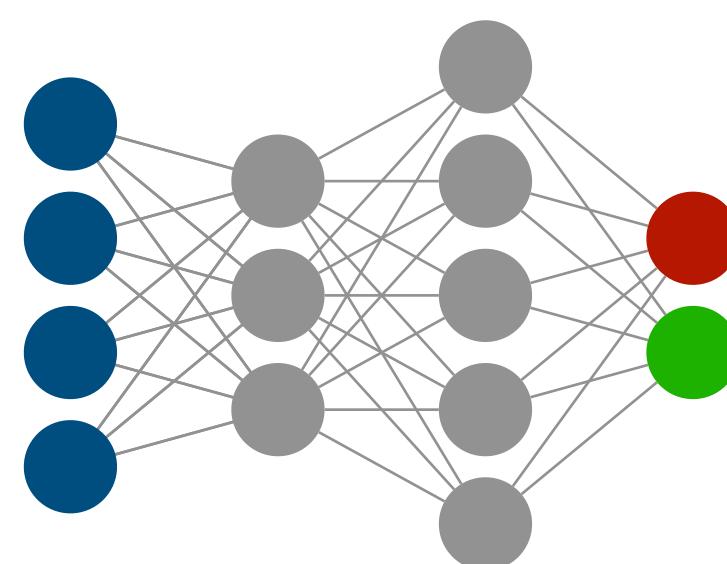
# Recruiting Process



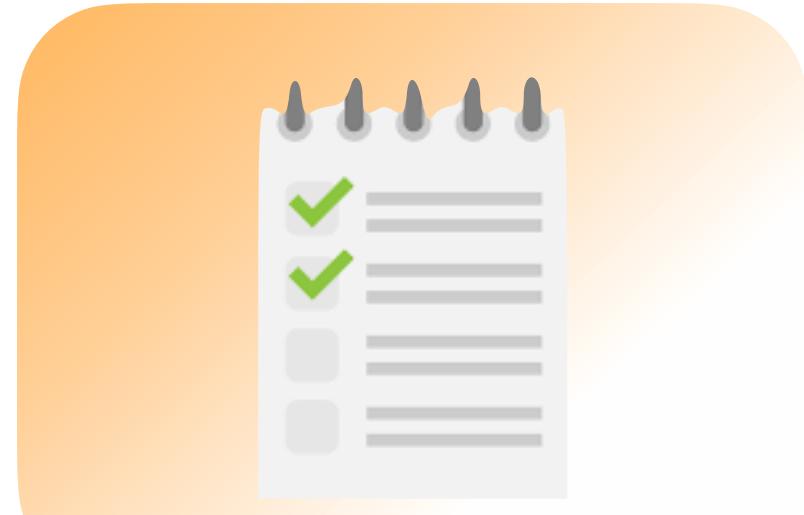
Fair



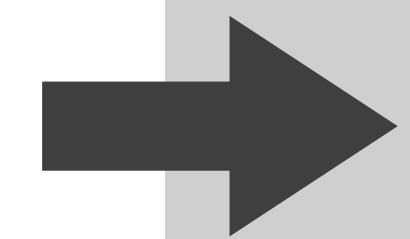
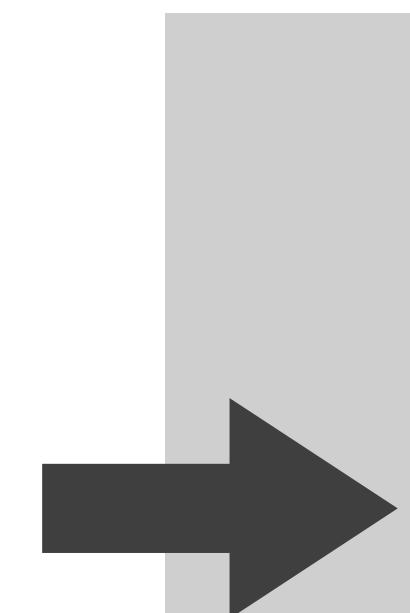
Unfair



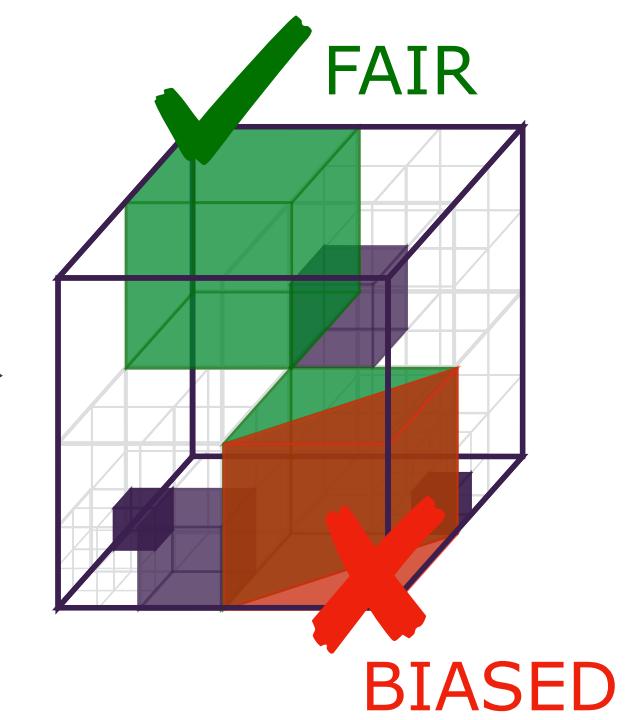
Neural Network

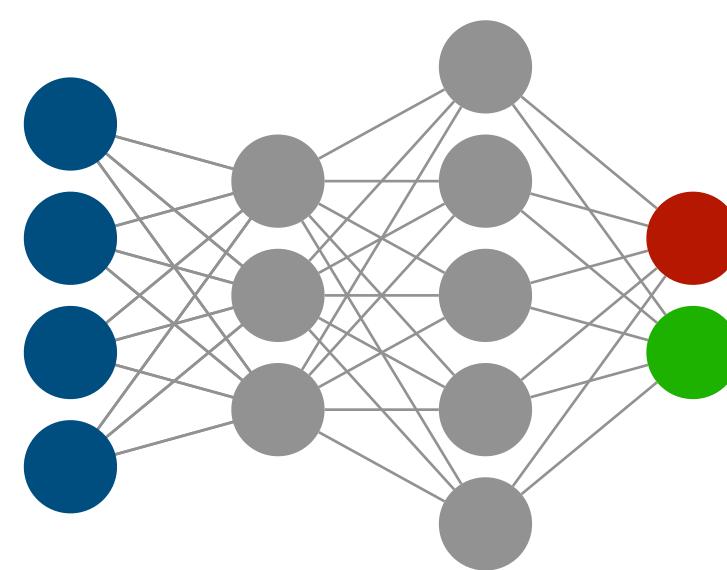


Specification

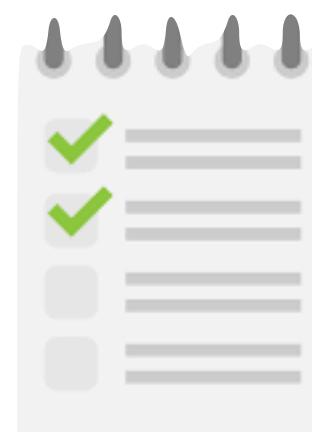


# Libra

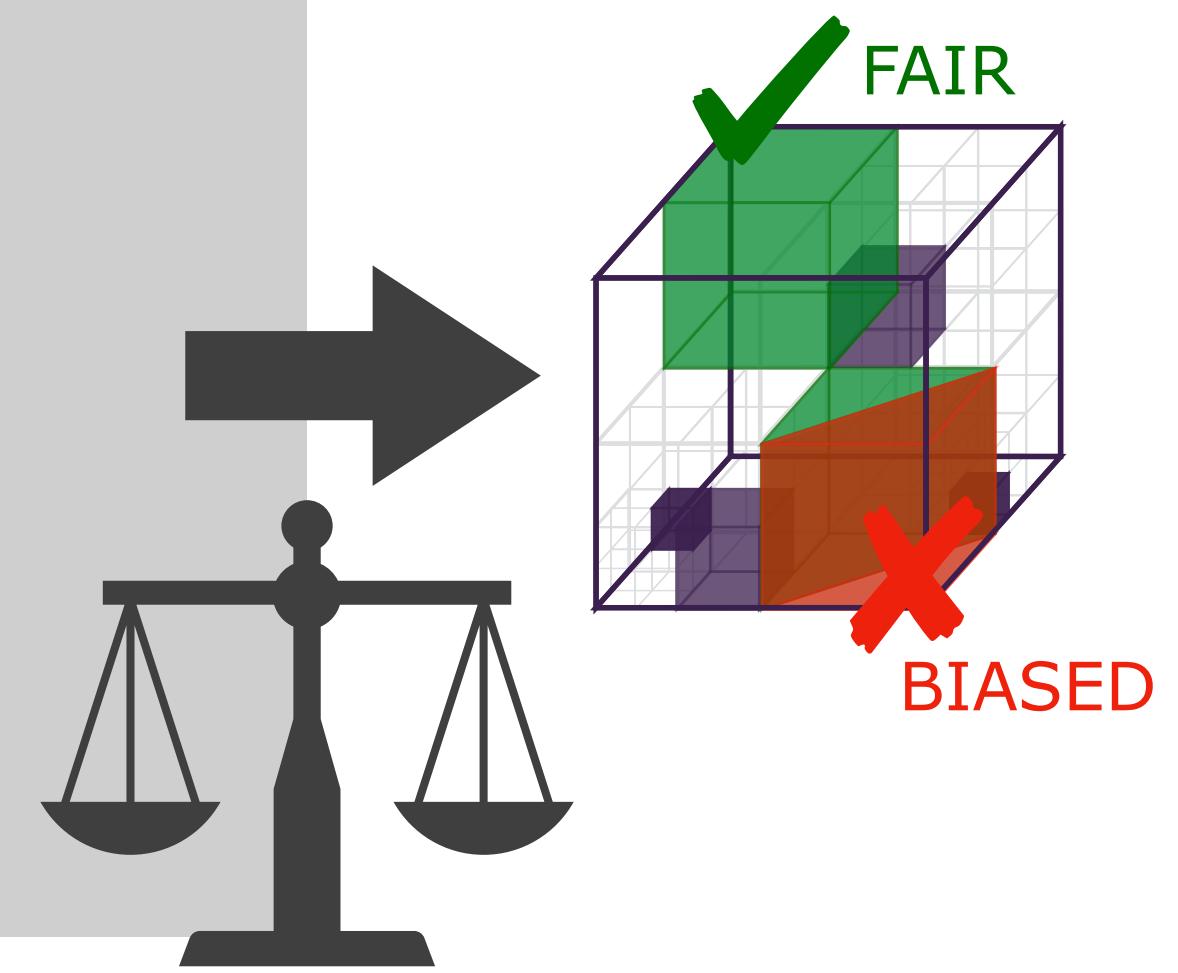
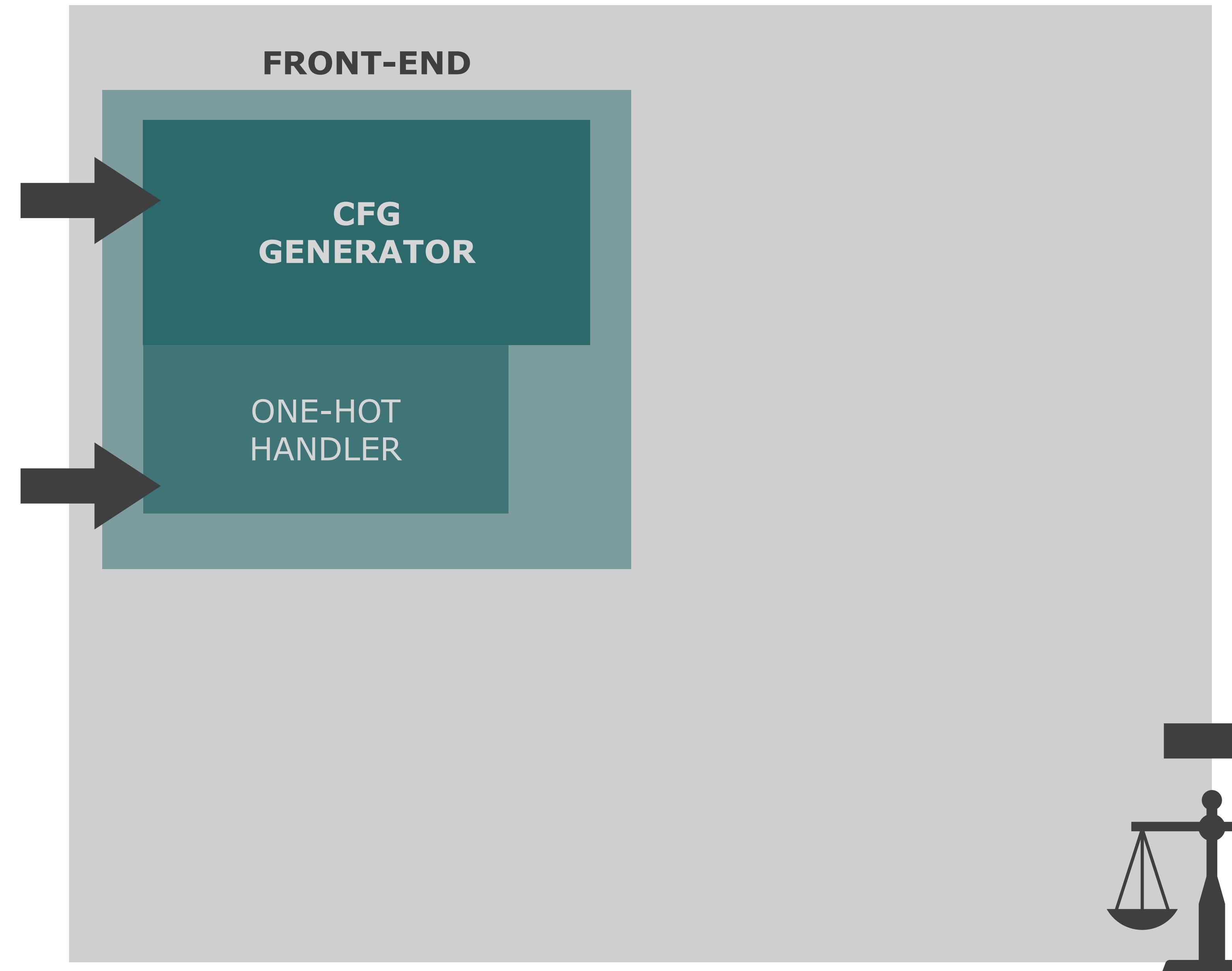


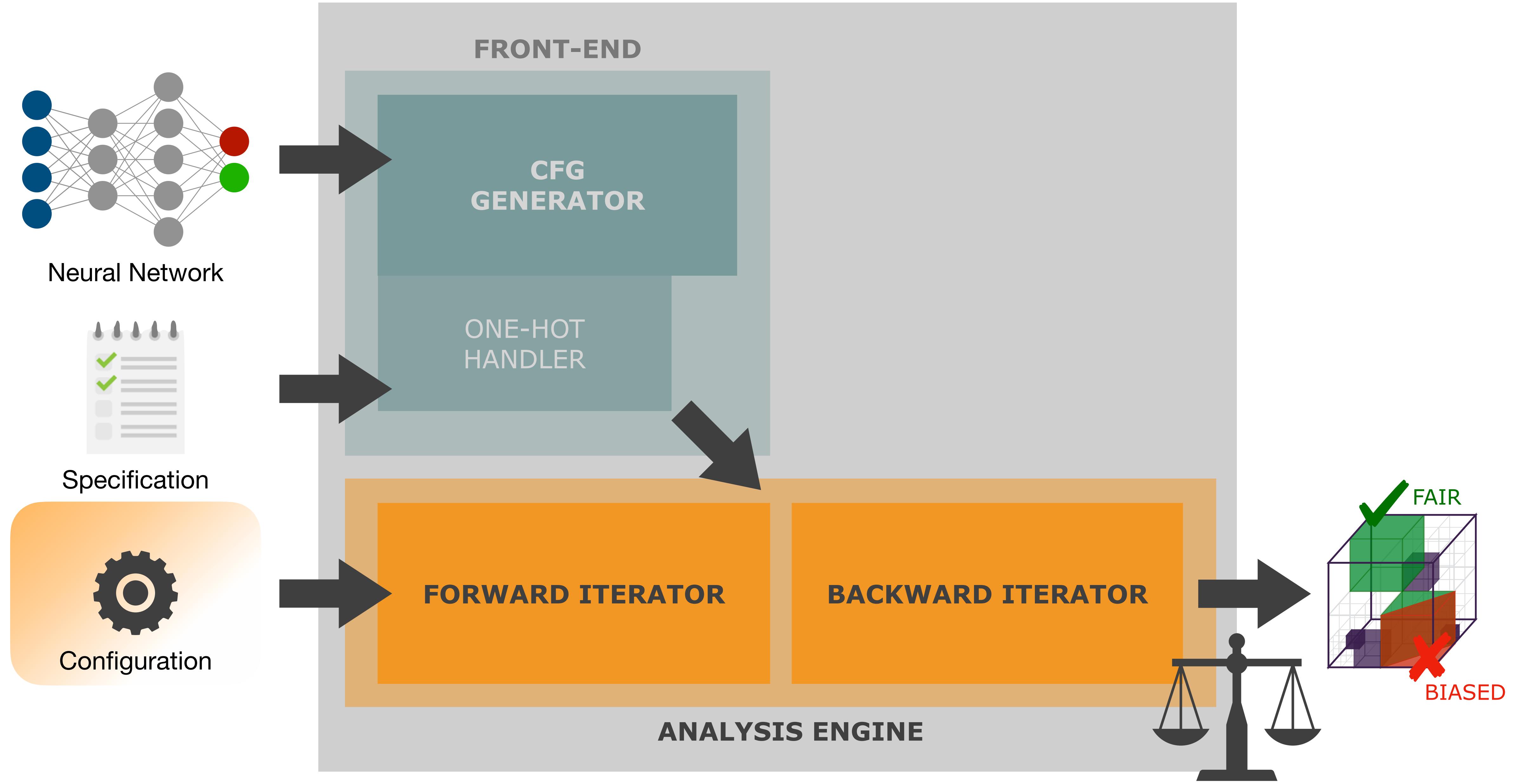


Neural Network

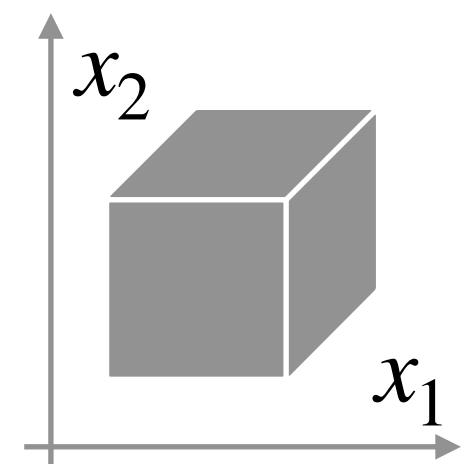


Specification

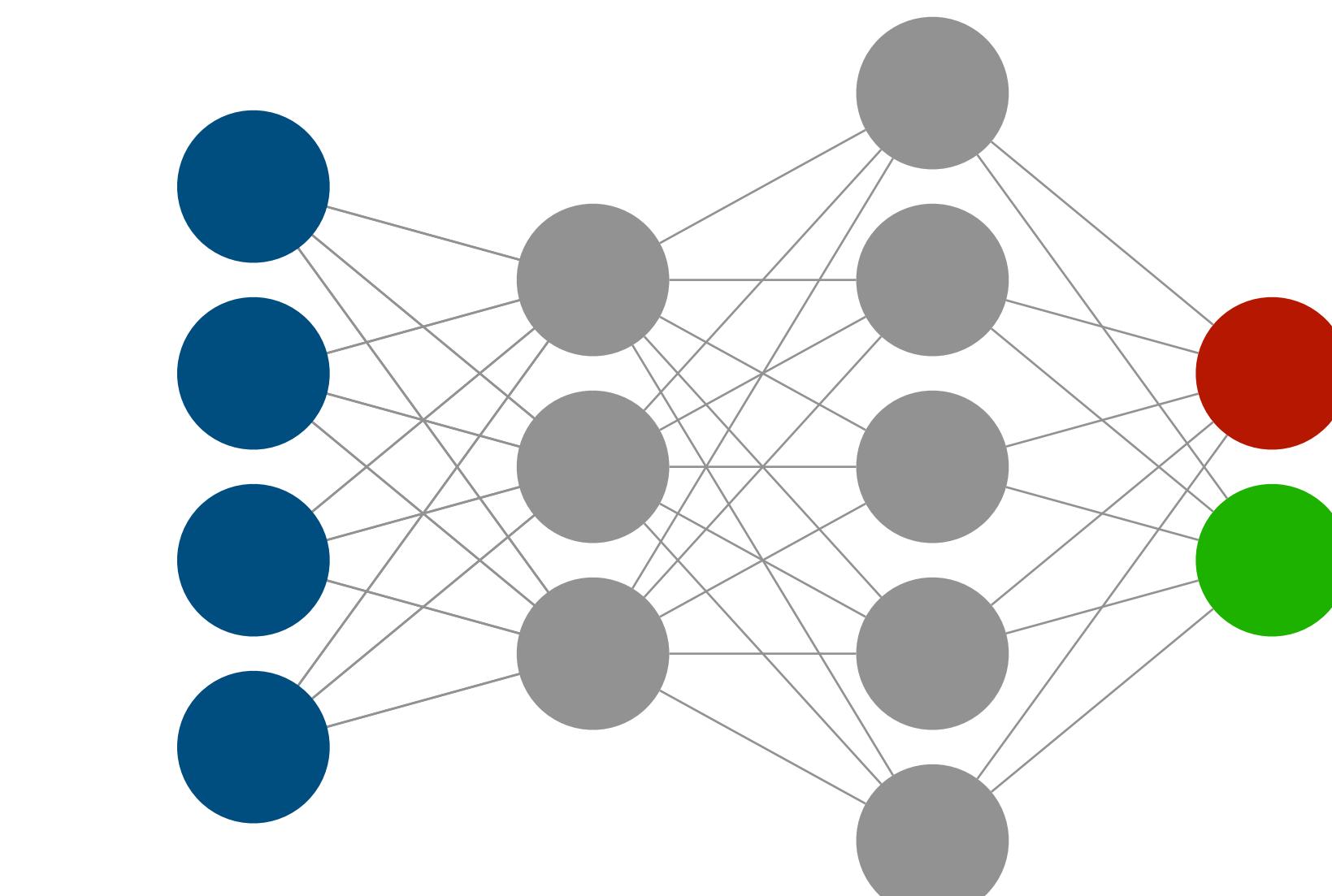
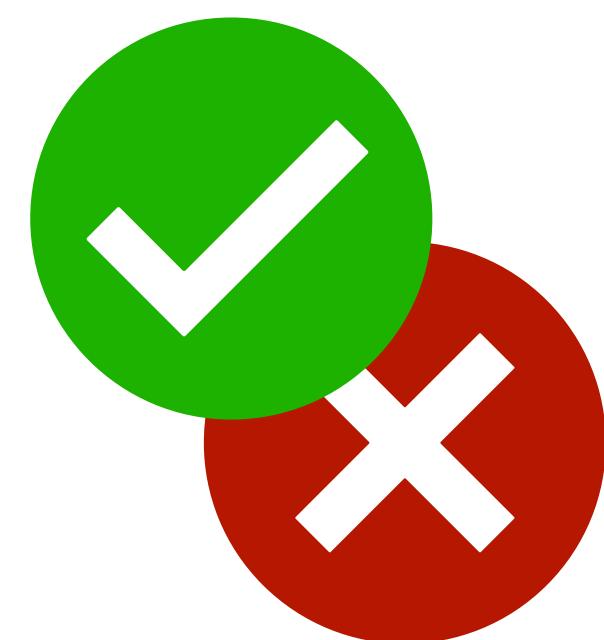




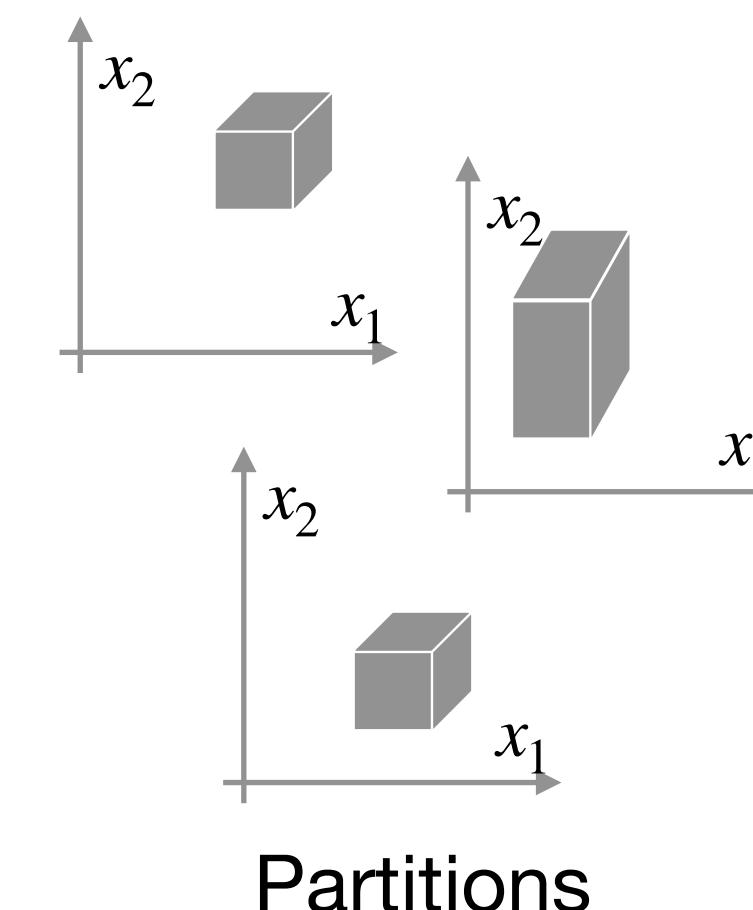
# Cheap Forward Pre-Analysis



Input Space

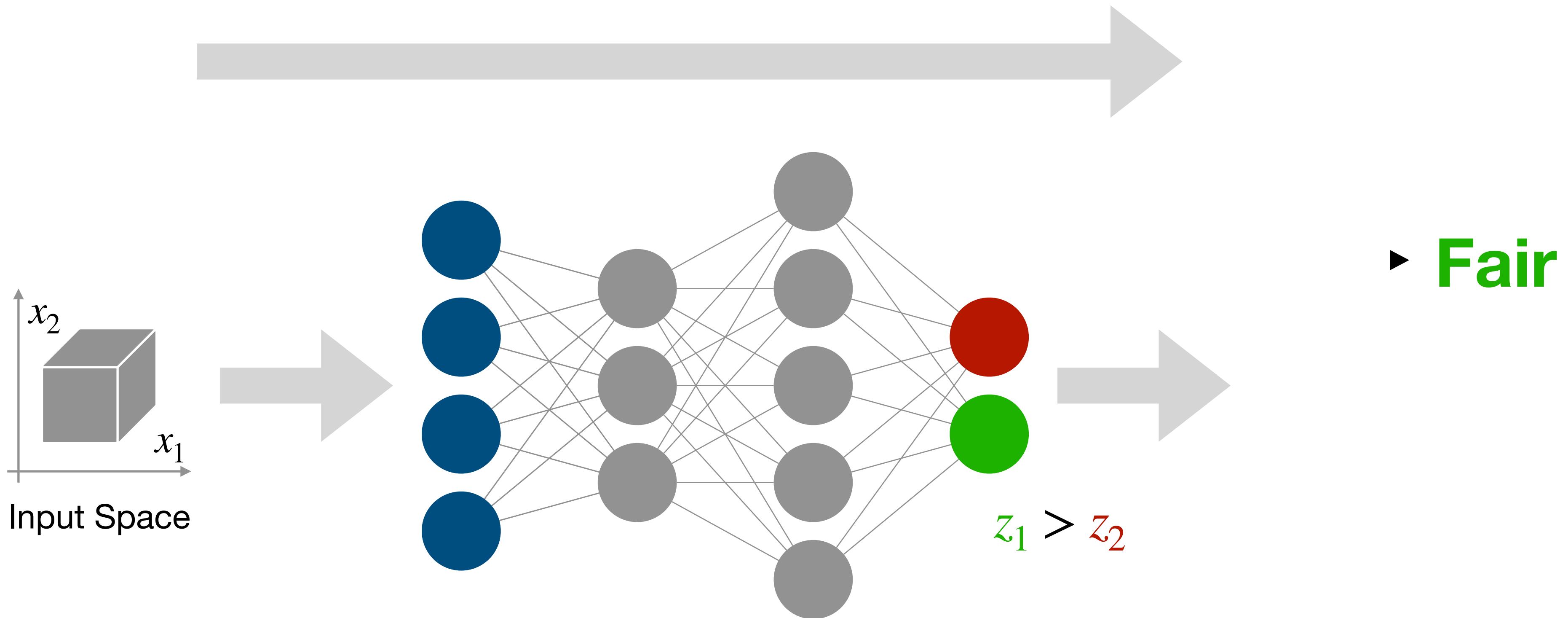


**Exact Backward Analysis  
using Polyhedra**



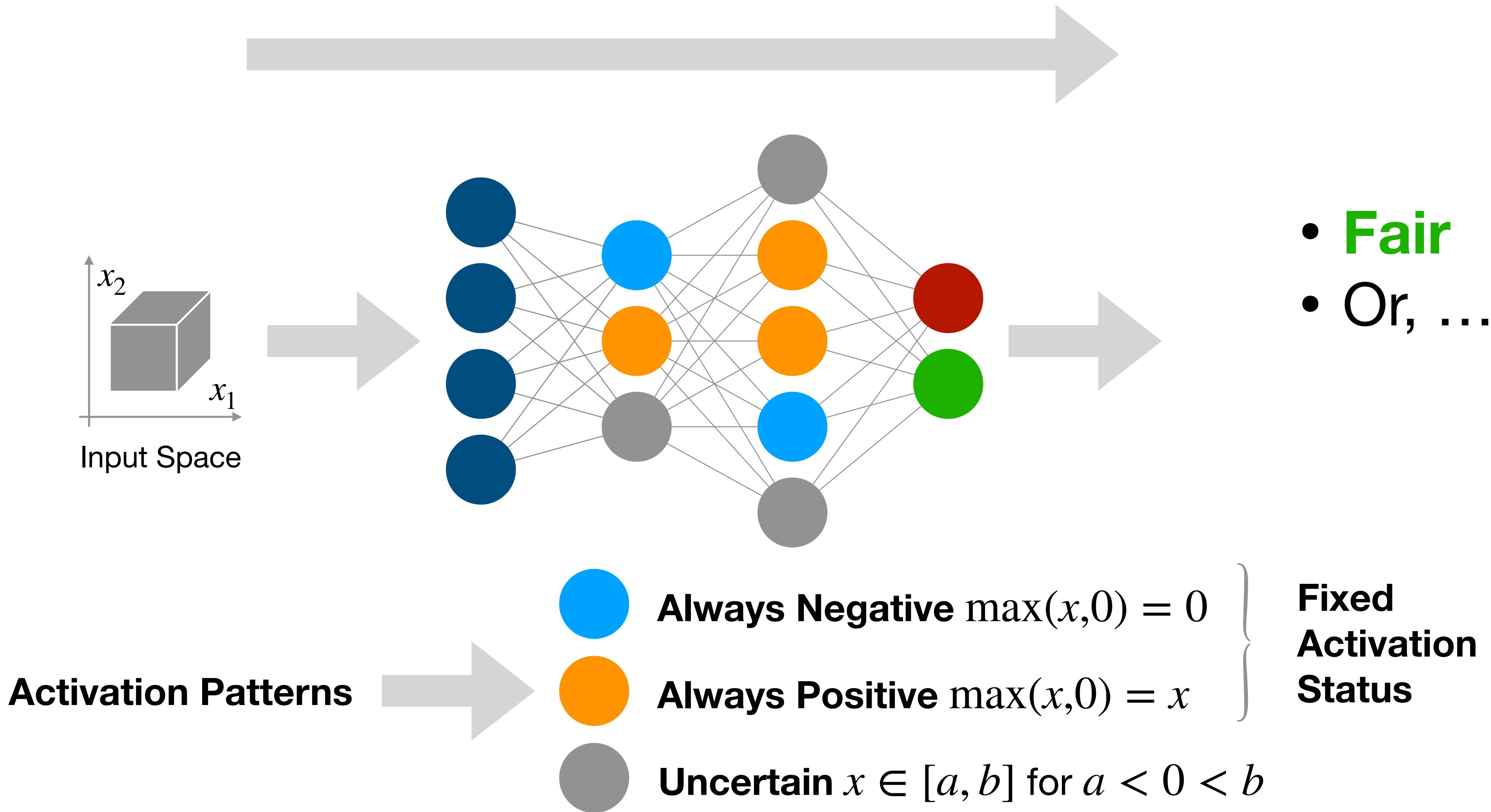
Partitions

# Cheap Forward Pre-Analysis

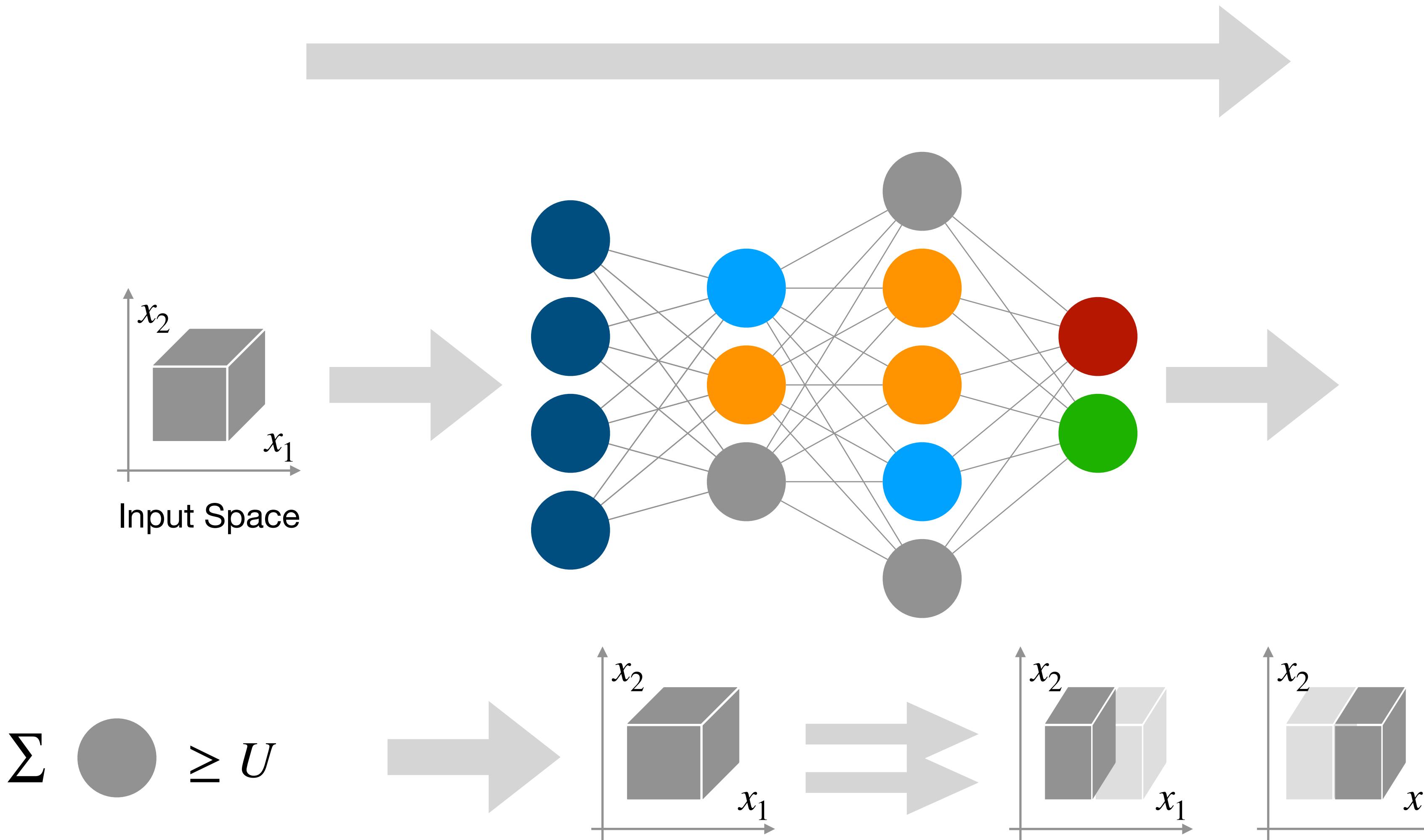


Propagate the partition through the network via **abstract domains**

# Cheap Forward Pre-Analysis



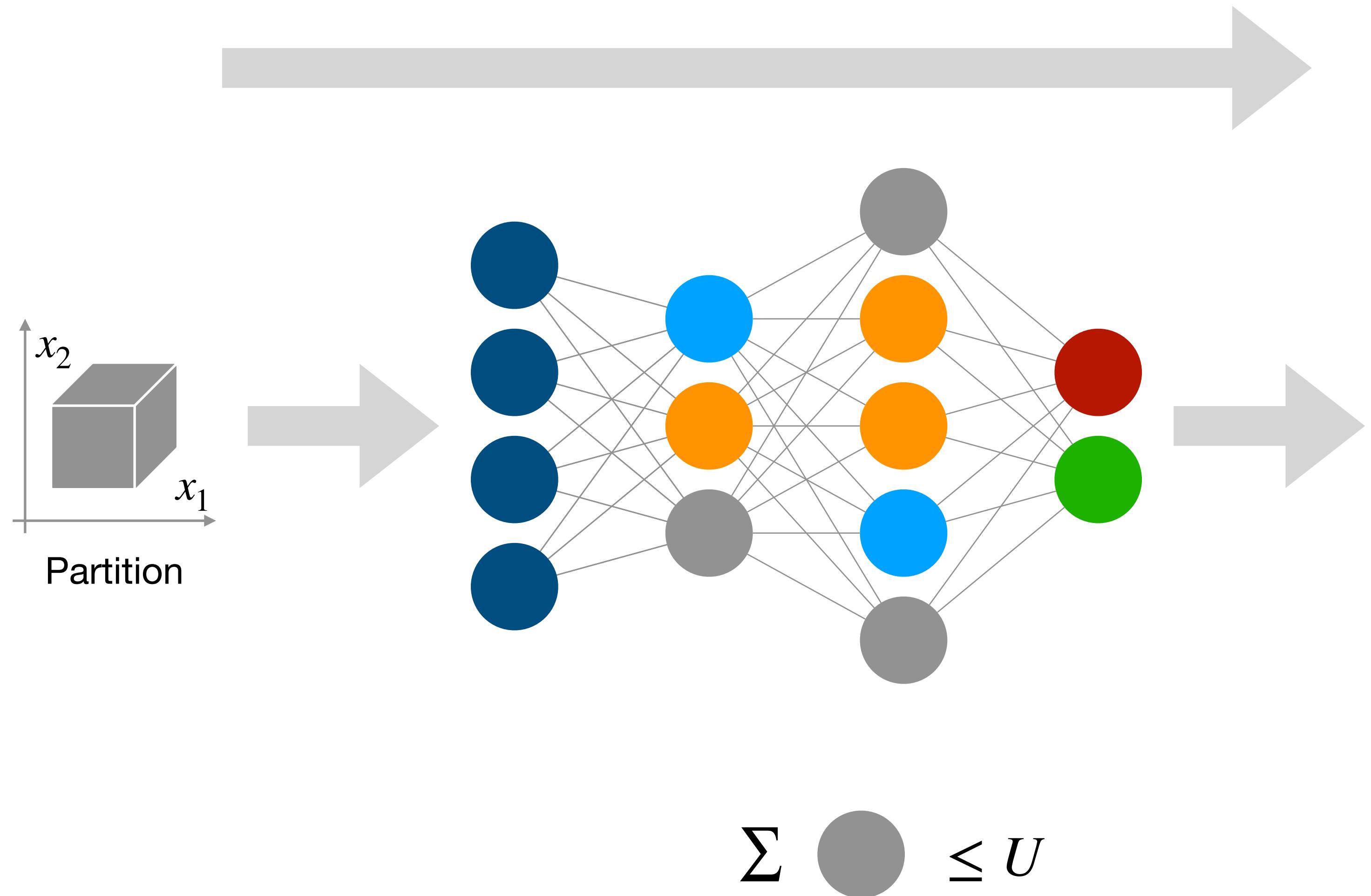
# Cheap Forward Pre-Analysis



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

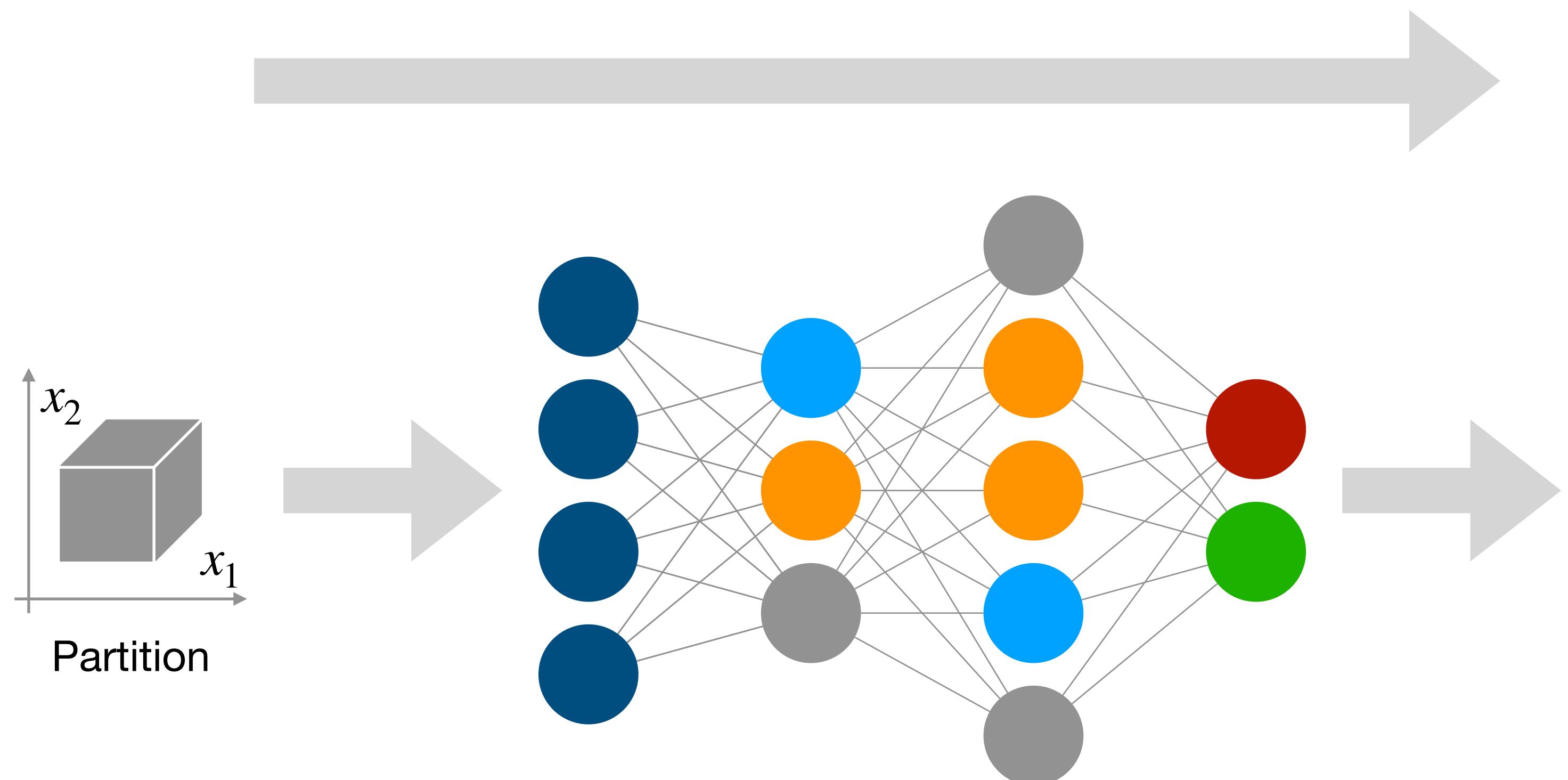
Along non-sensitive  
features only

# Cheap Forward Pre-Analysis



- Fair
- Partitioned
- Feasible

# Cheap Forward Pre-Analysis

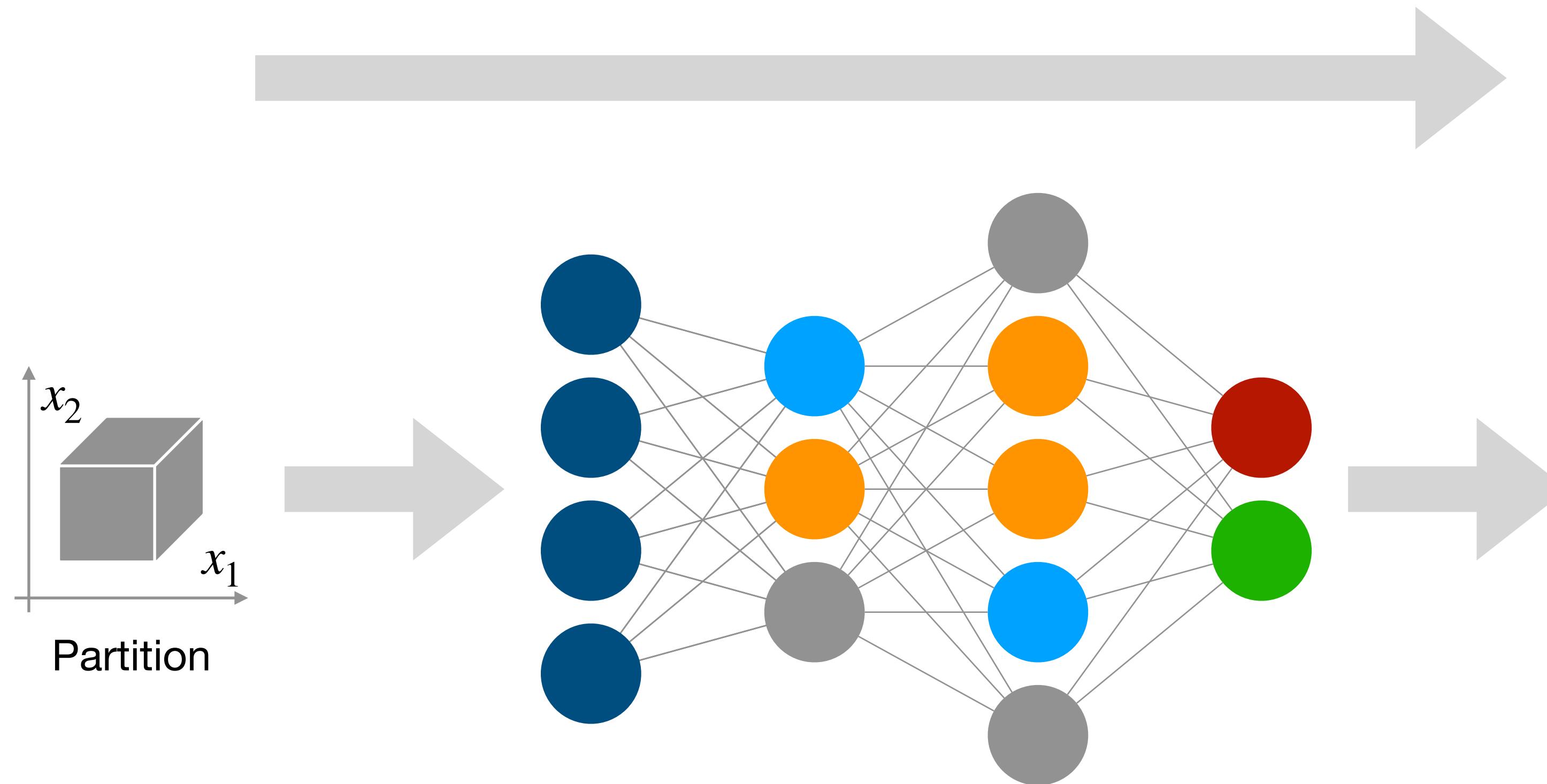


- Fair
  - Partitioned
  - Feasible

► Excluded

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

# Cheap Forward Pre-Analysis

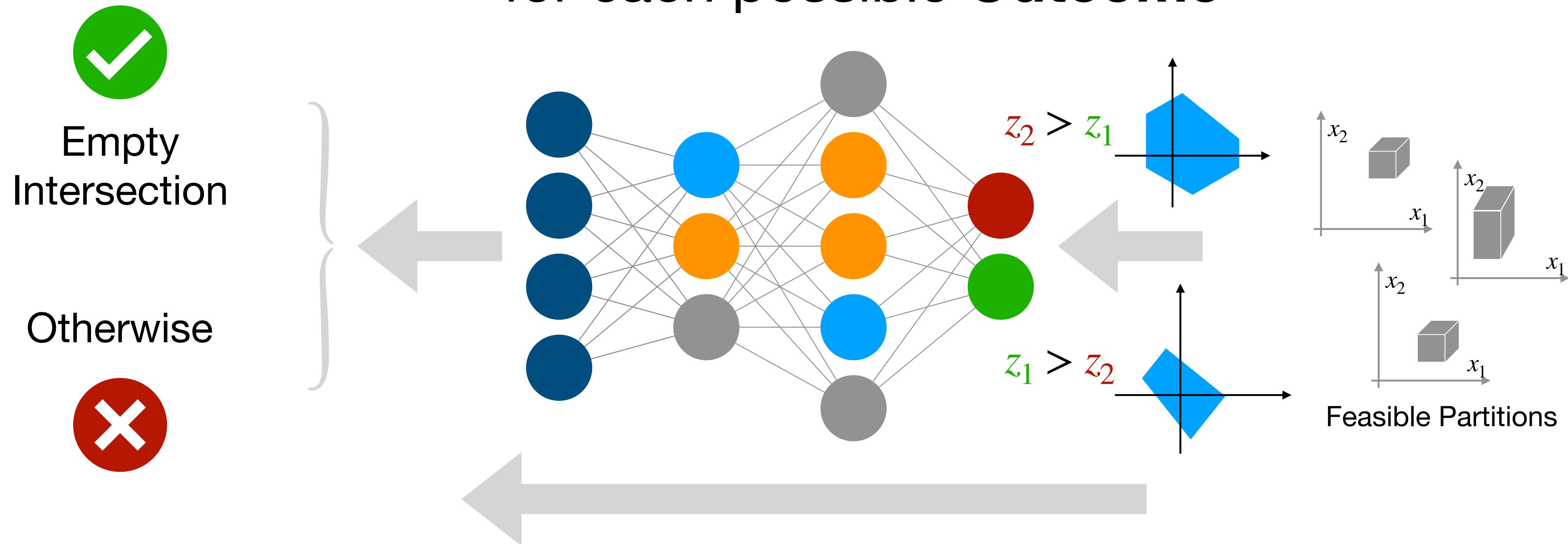


Budget constraints  $(L, U)$  can be  
**Automatically Configured** to  $(L_{min}, U_{max})$

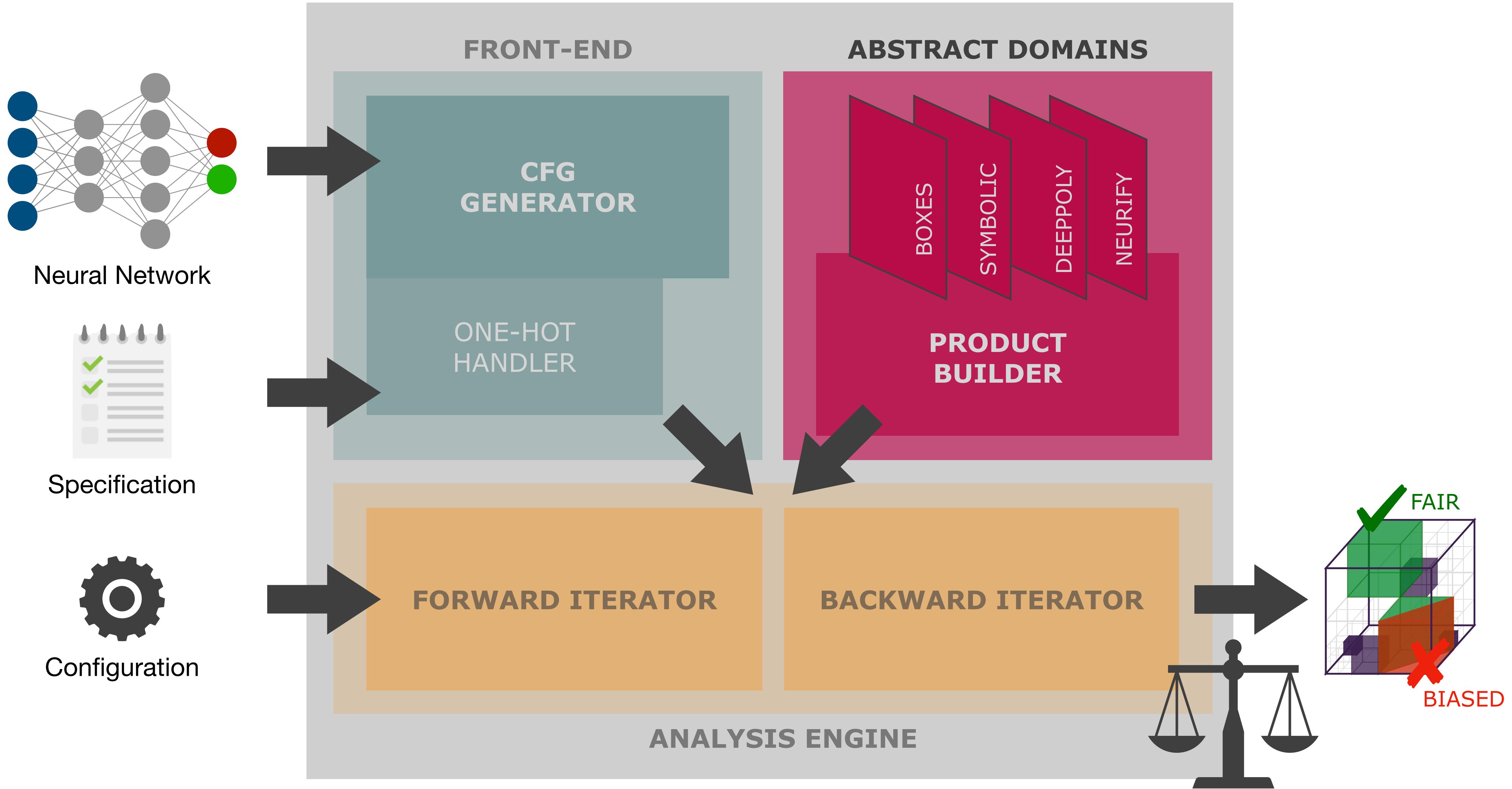
$$\left\{ \begin{array}{l} L \text{ down-to } L_{min} \\ U \text{ up-to } U_{max} \end{array} \right.$$

- Fair
- Partitioned
- Feasible
- Excluded

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

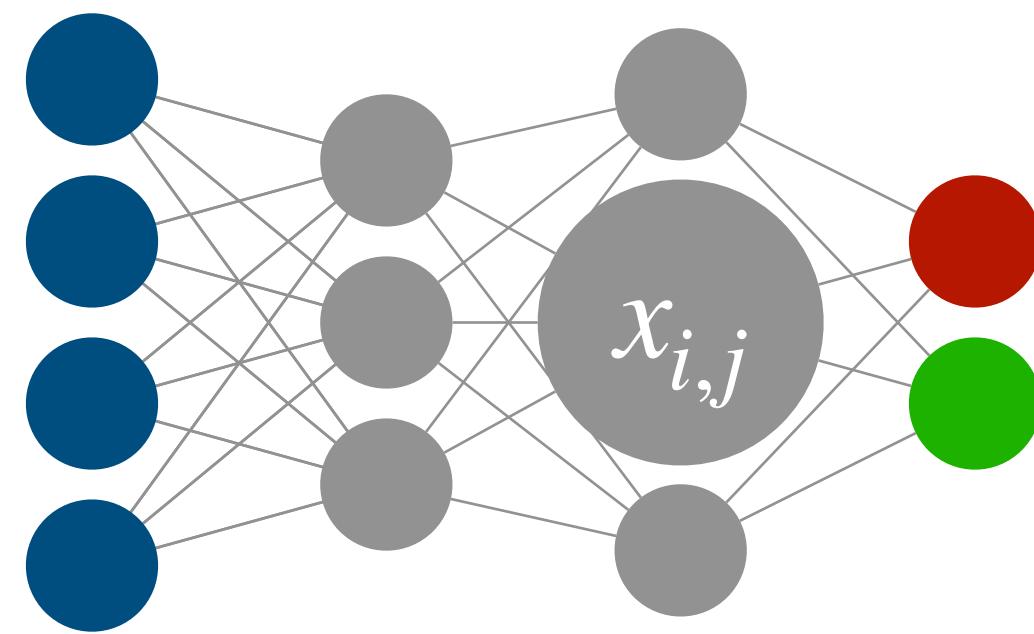


**Exact Backward Analysis**  
using **Polyhedra**



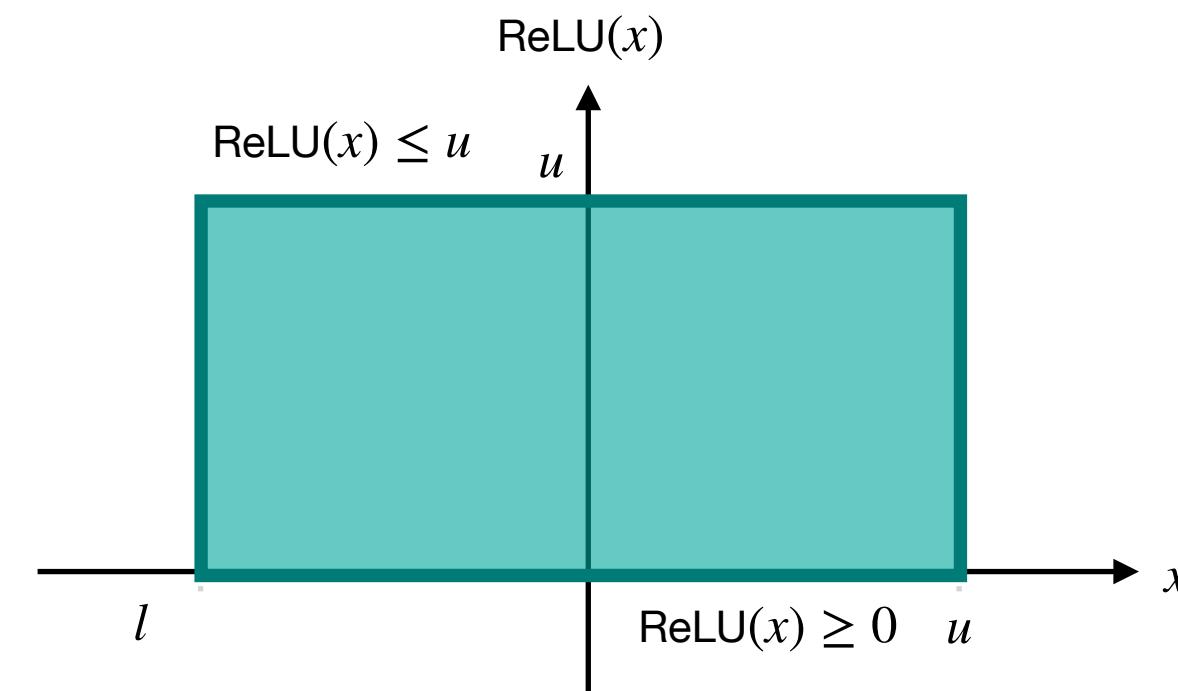
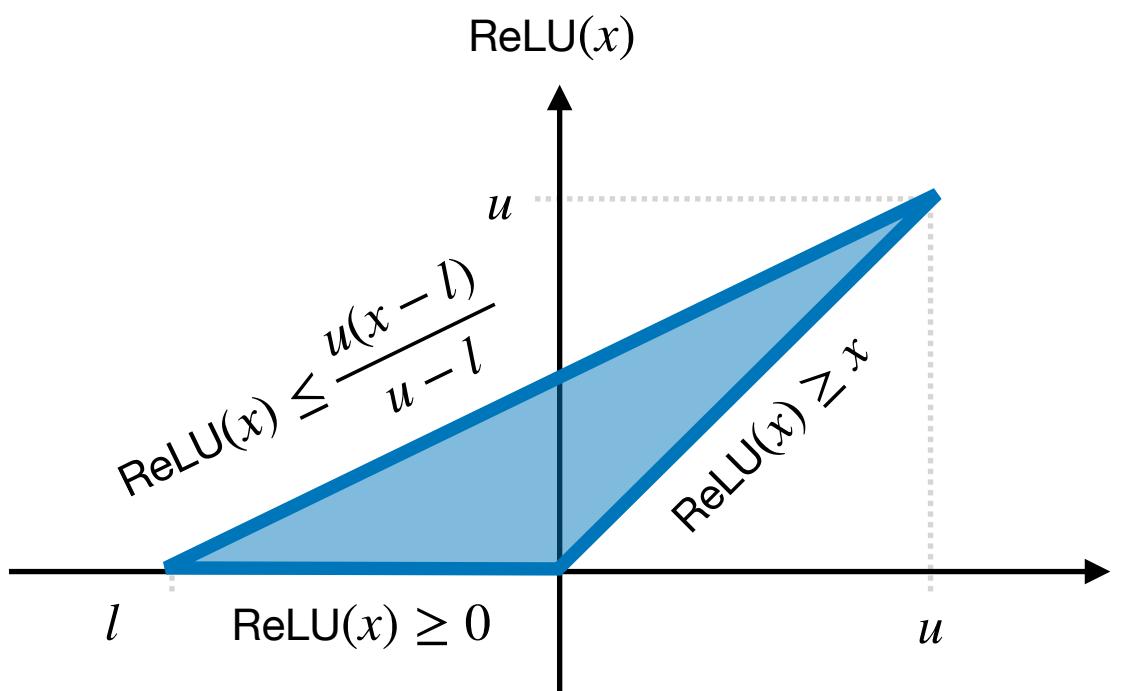
# Symbolic

Li et al. @ SAS 2019



$$[l, u]$$

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

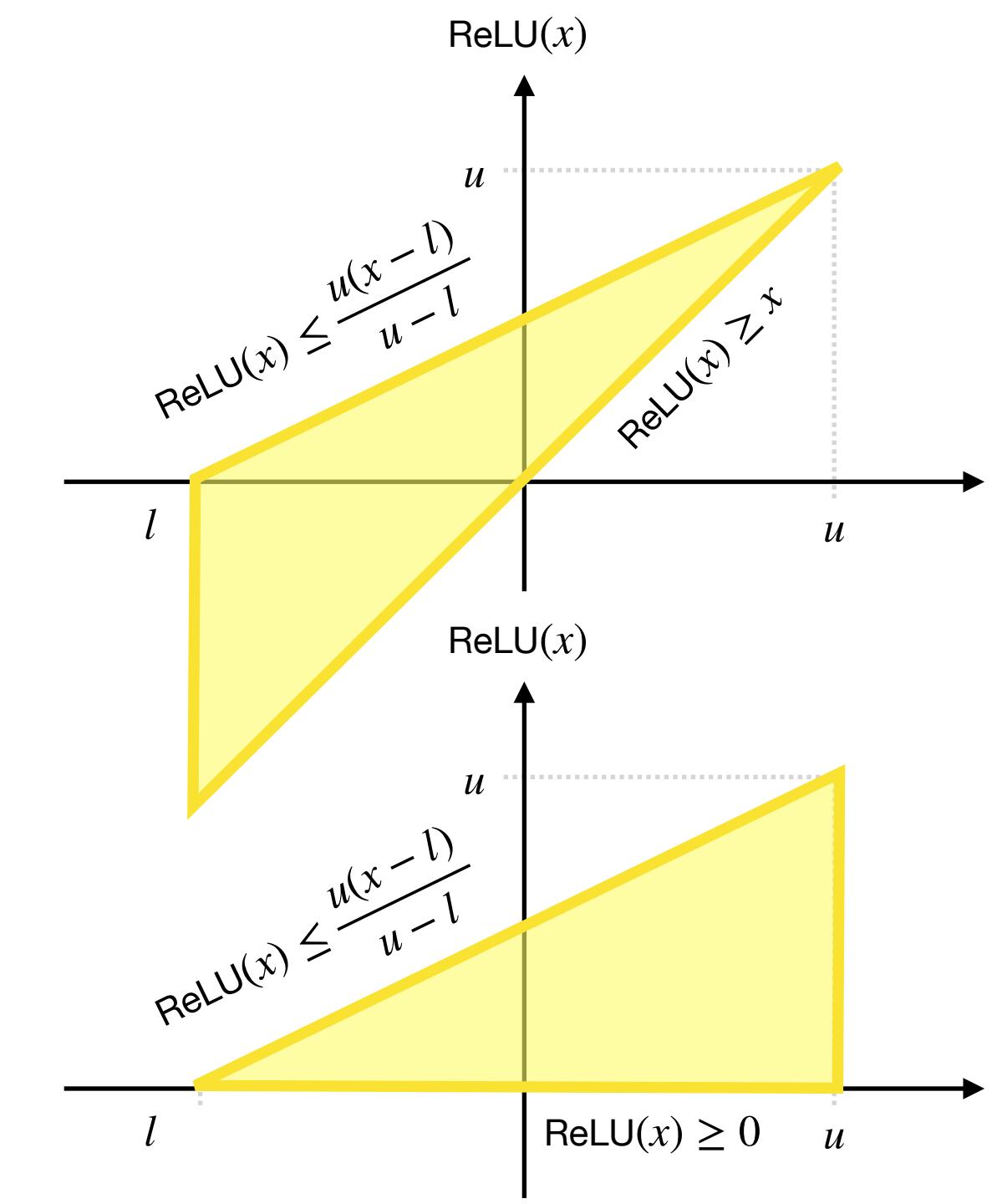


# DeepPoly

Singh et al. @ POPL 2019

$$[l, u]$$

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

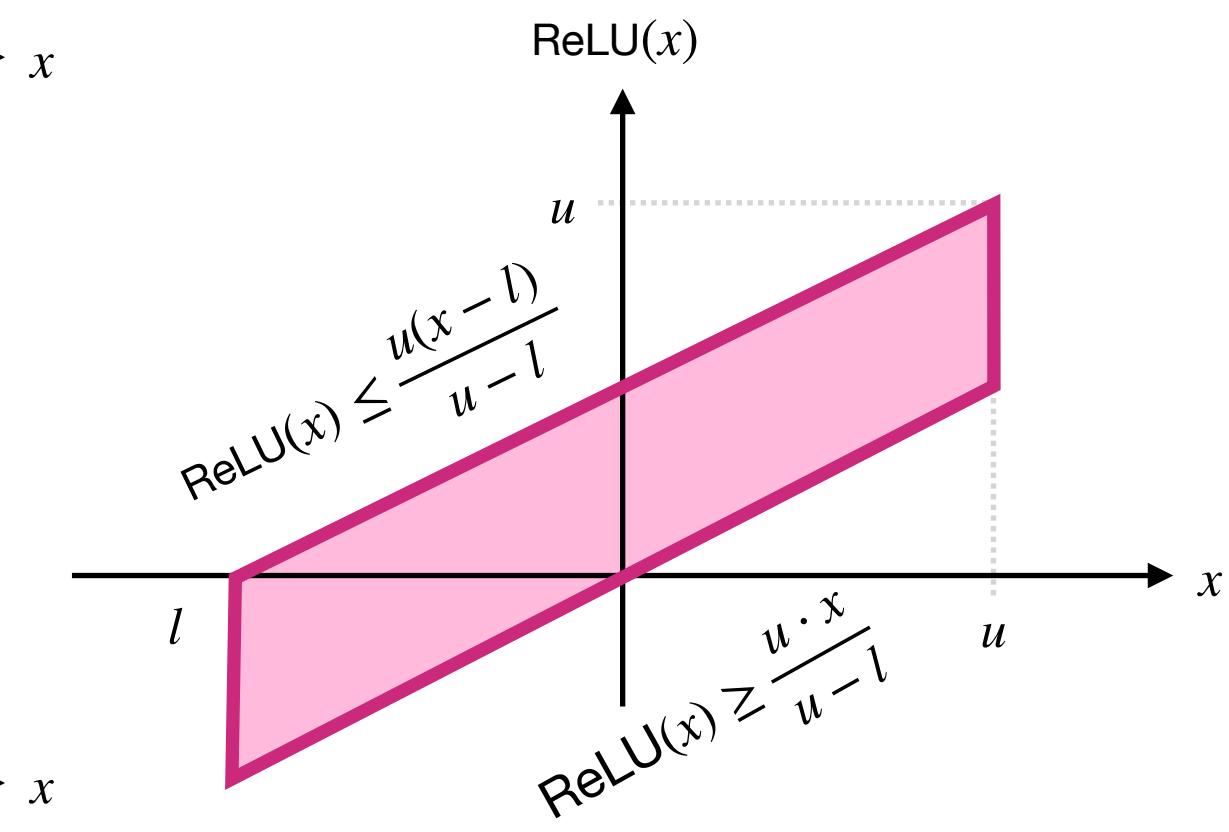


# Neurify

Wang et al. @ NeurIPS 2018

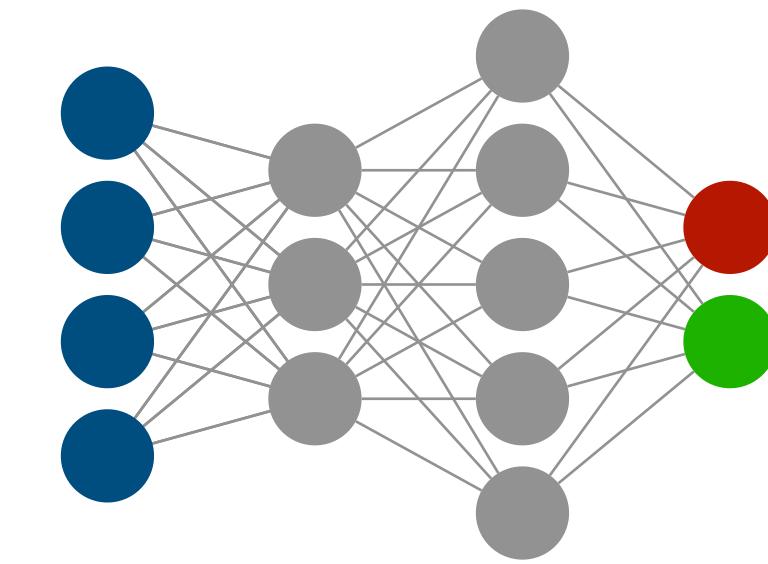
$$[l_{\text{low}}, l_{\text{up}}, u_{\text{low}}, u_{\text{up}}]$$

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

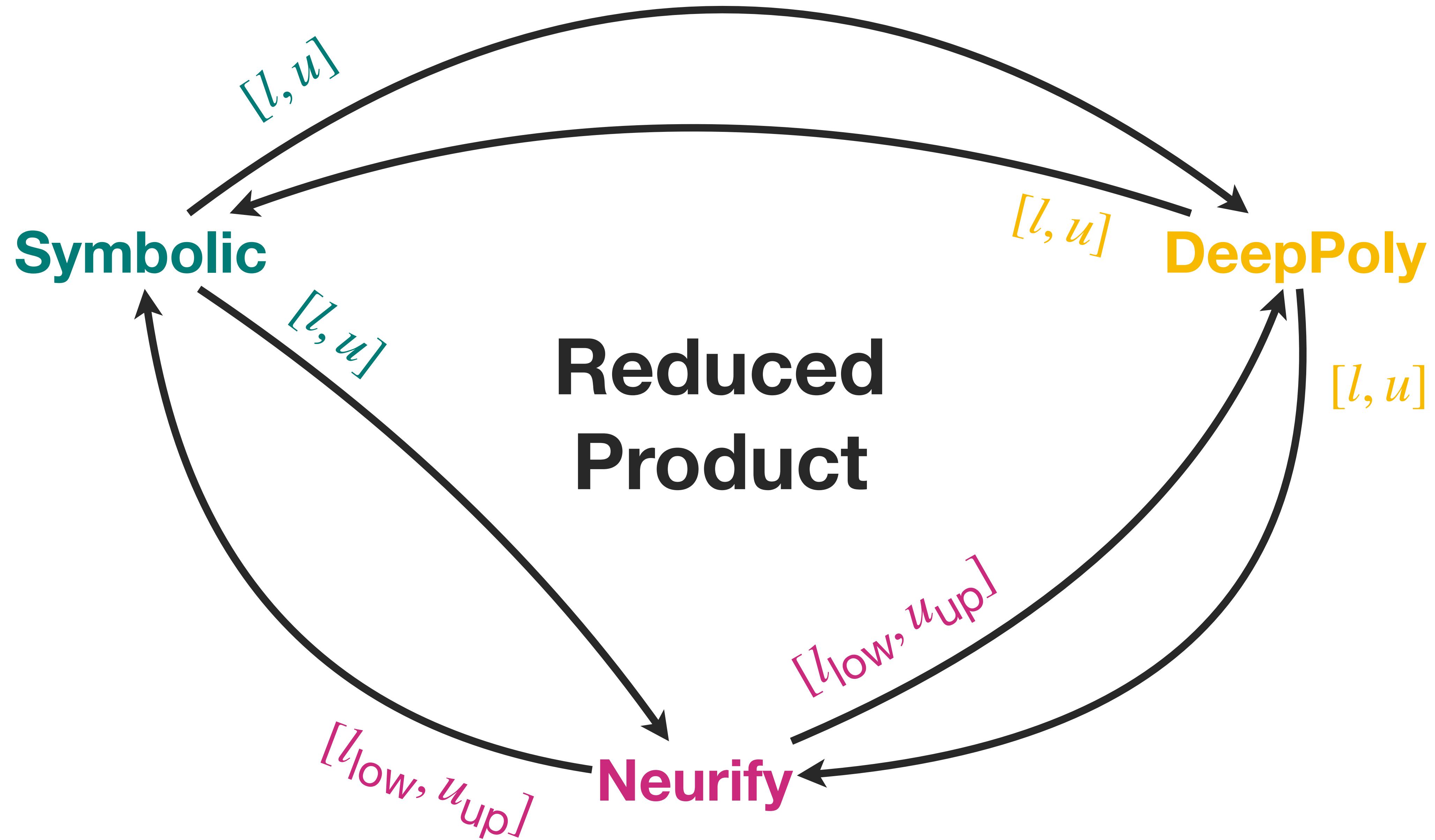


# Precision-vs-Scalability

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



# 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%

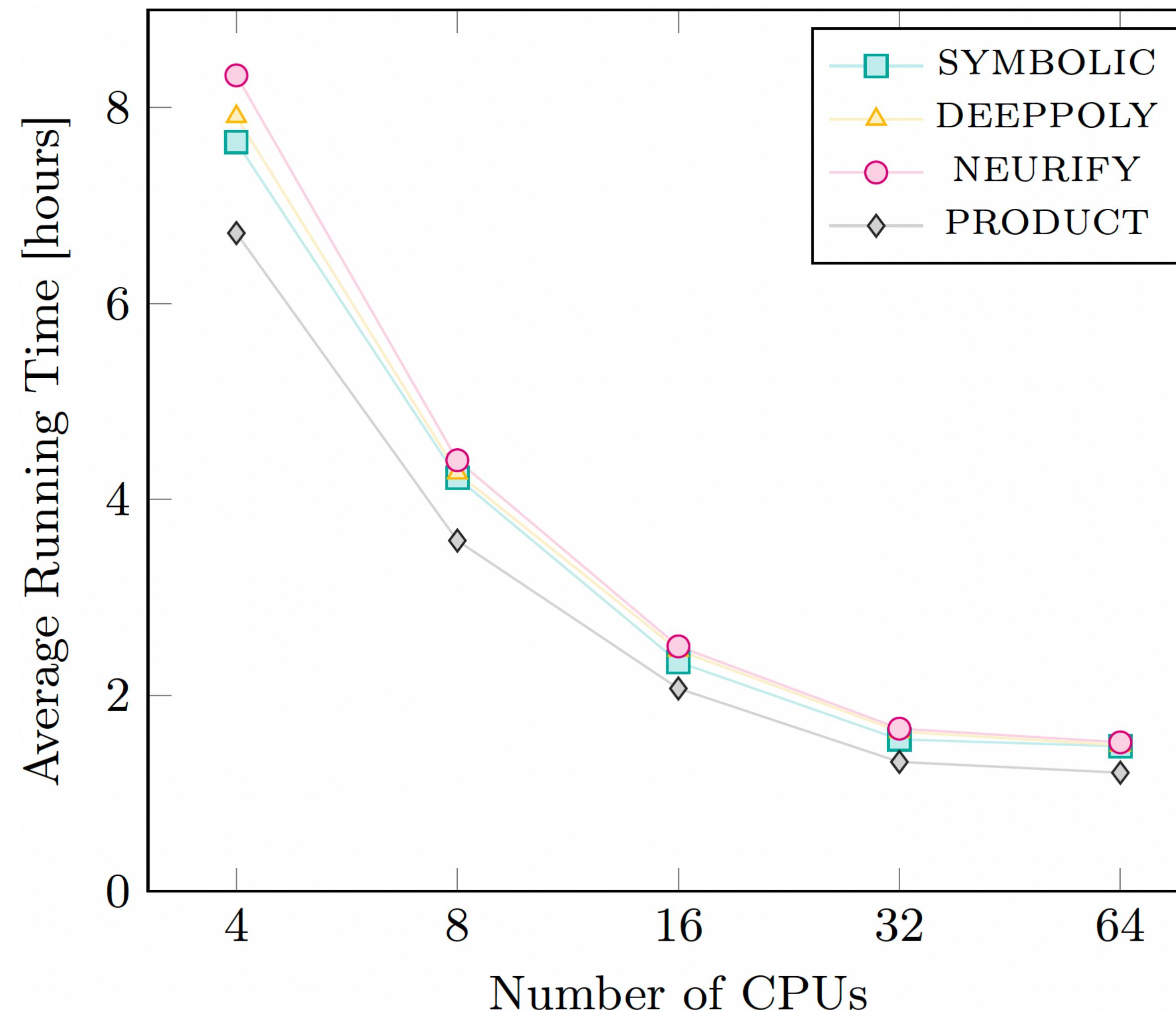


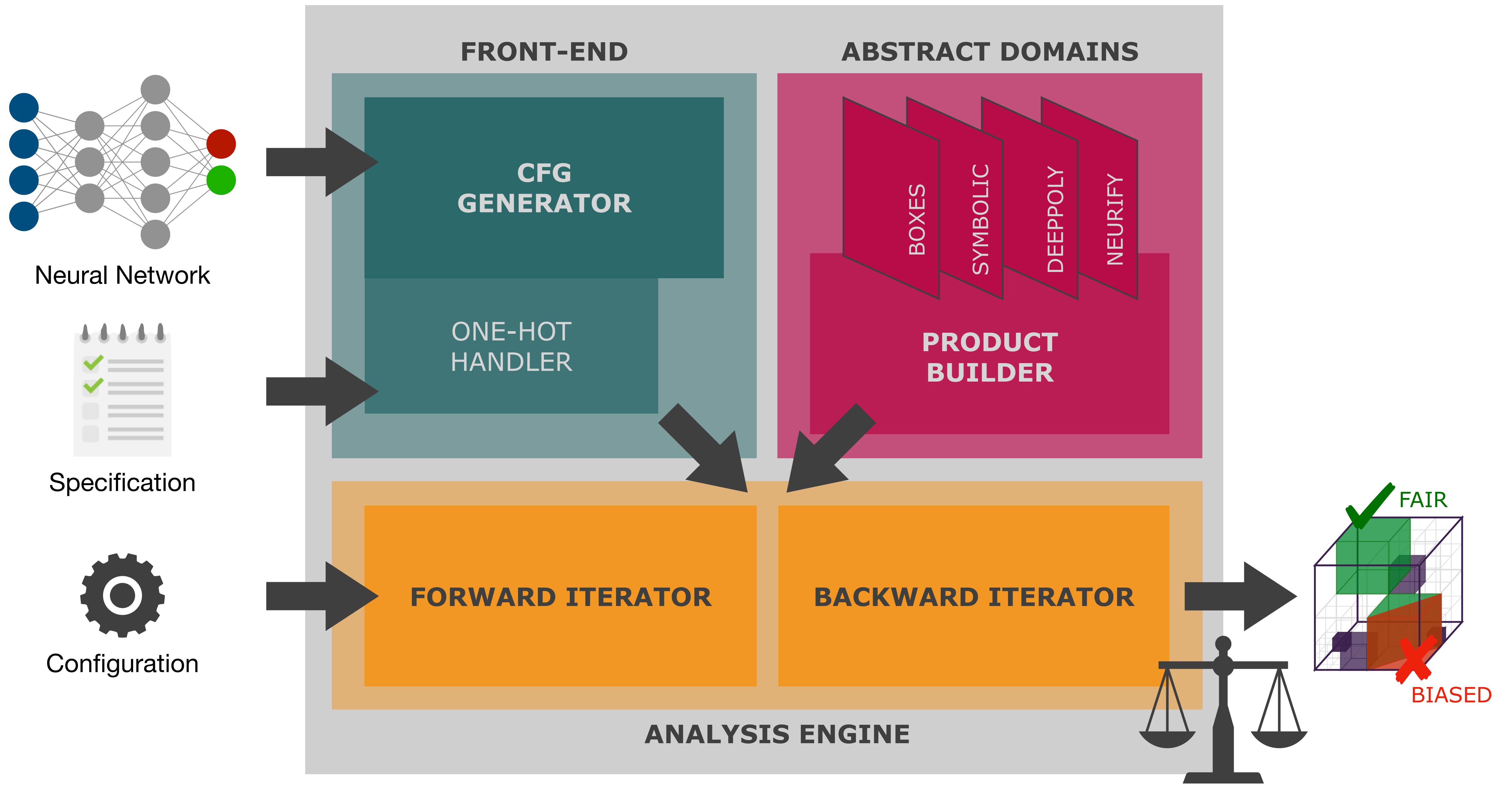
# Effect of Neural Network Structure

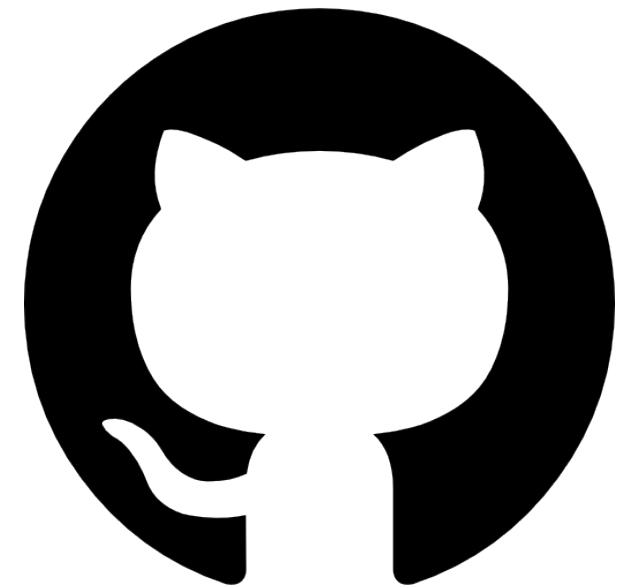
Size	Symbolic	DeepPoly	Neurify	Product	
10	98.72%	98.37%	98.51%	99.44%	+0.7%
12	76.70%	66.39%	64.58%	77.29%	+0.6%
20	56.11%	56.10%	53.06%	68.23%	+12.1%
40	34.72%	38.69%	41.22%	51.18%	+10%
45	43.78%	51.21%	50.59%	55.53%	+4.3%



# Leveraging Multiple CPUs







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!