Deep Neural Cryptography

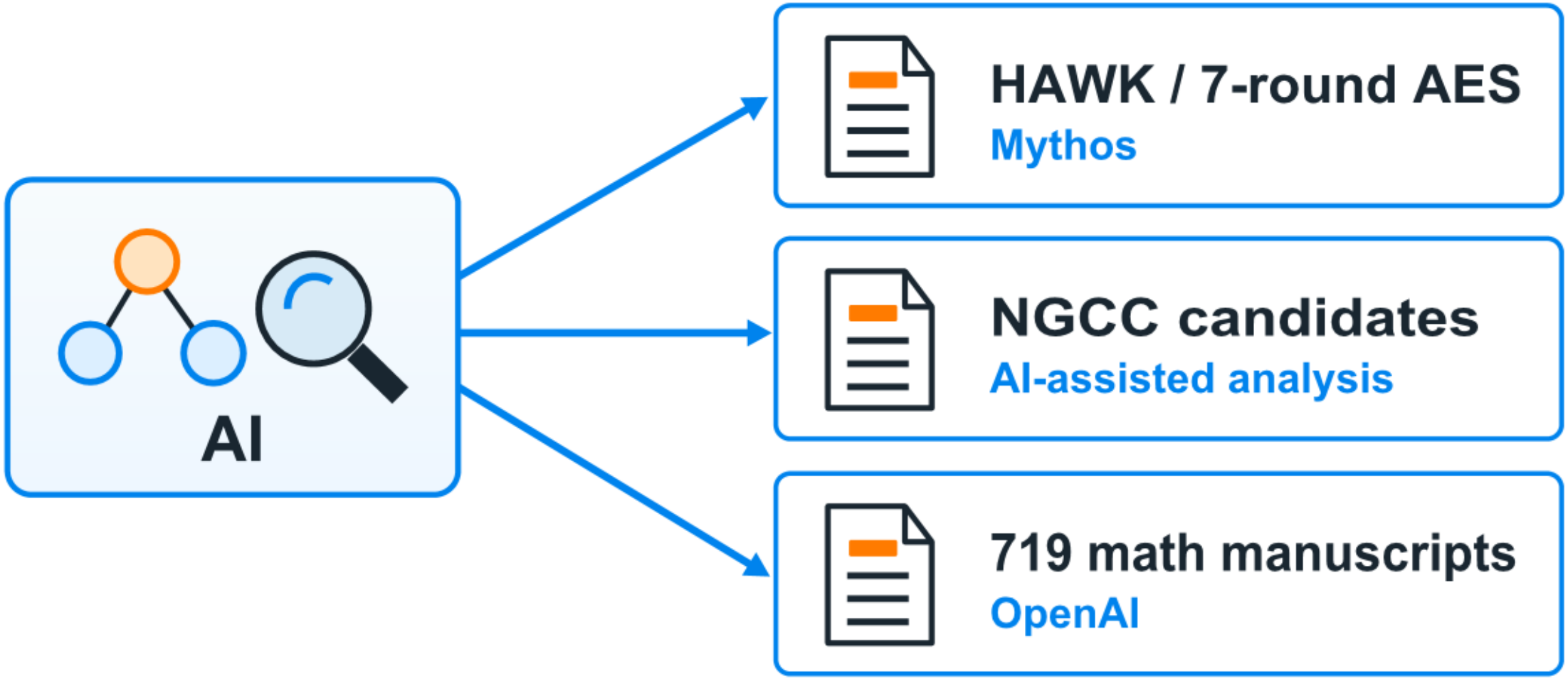
Cryptography Inside Neural Networks

David Gerault

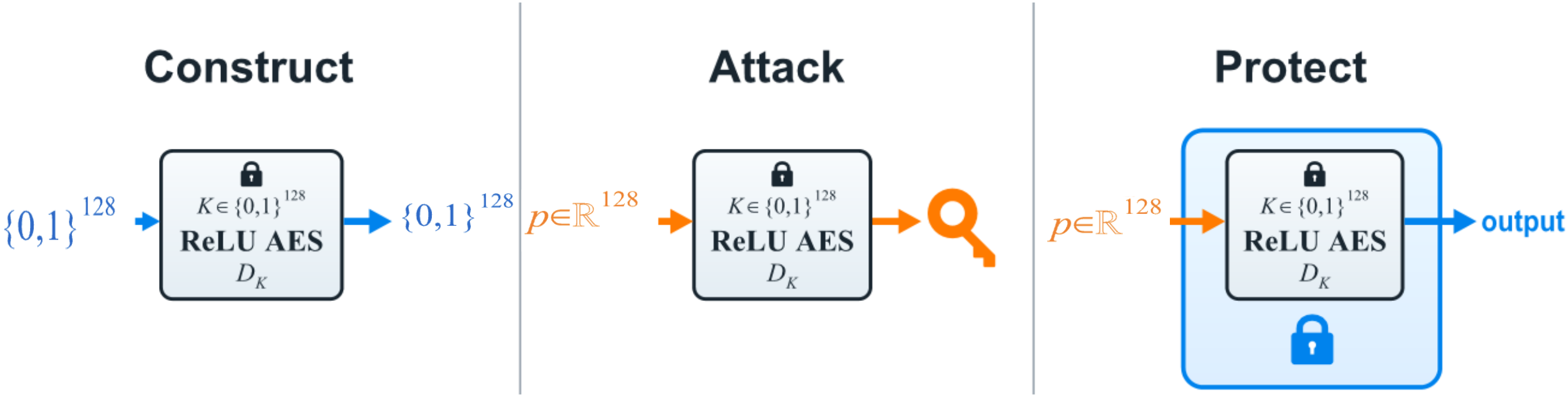
Joint work with Anna Hambitzer, Eyal Ronen and Adi Shamir

LUXEMBOURG INSTITUTE OF SCIENCE AND TECHNOLOGY | LIST

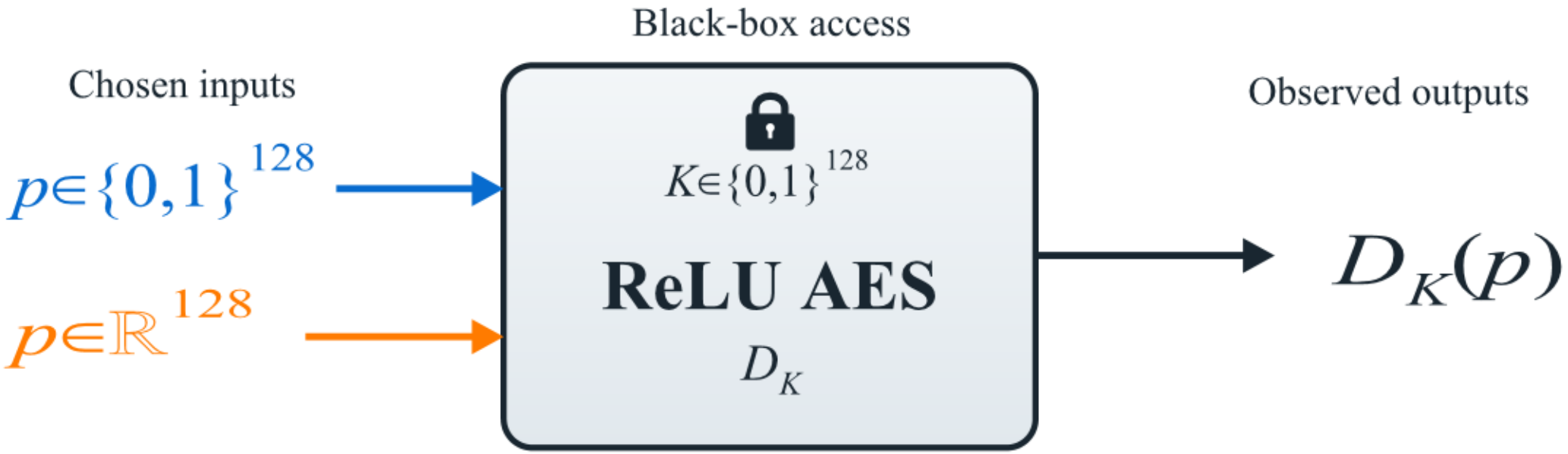
2021: So close yet so far



Construct, attack, protect

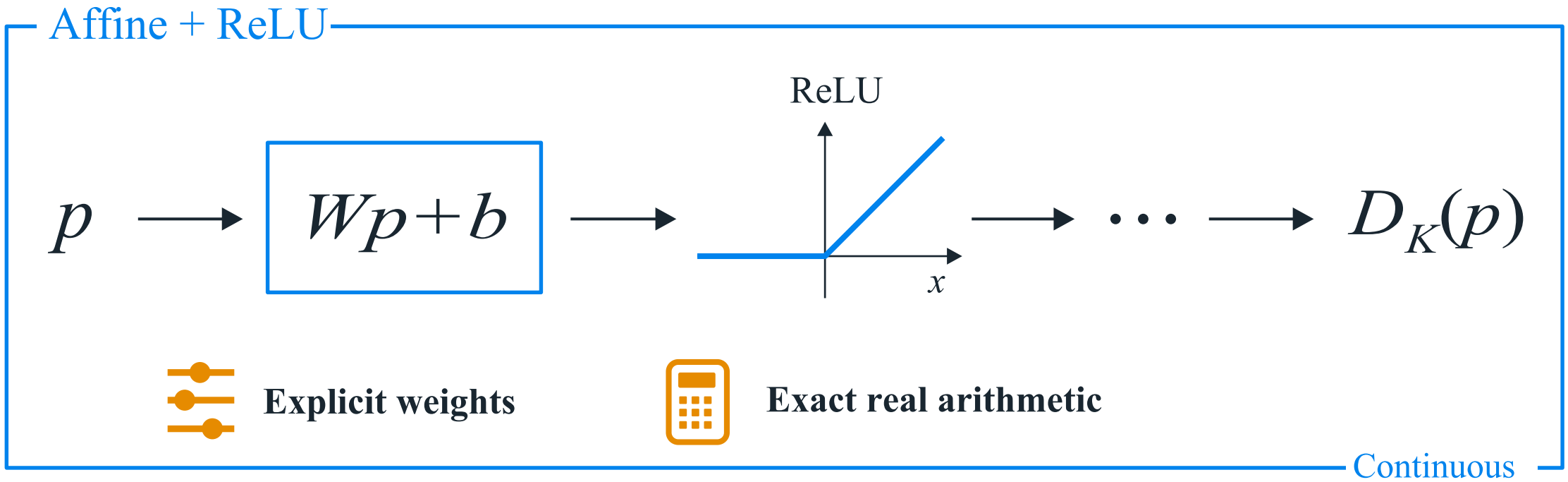


AES through a neural oracle

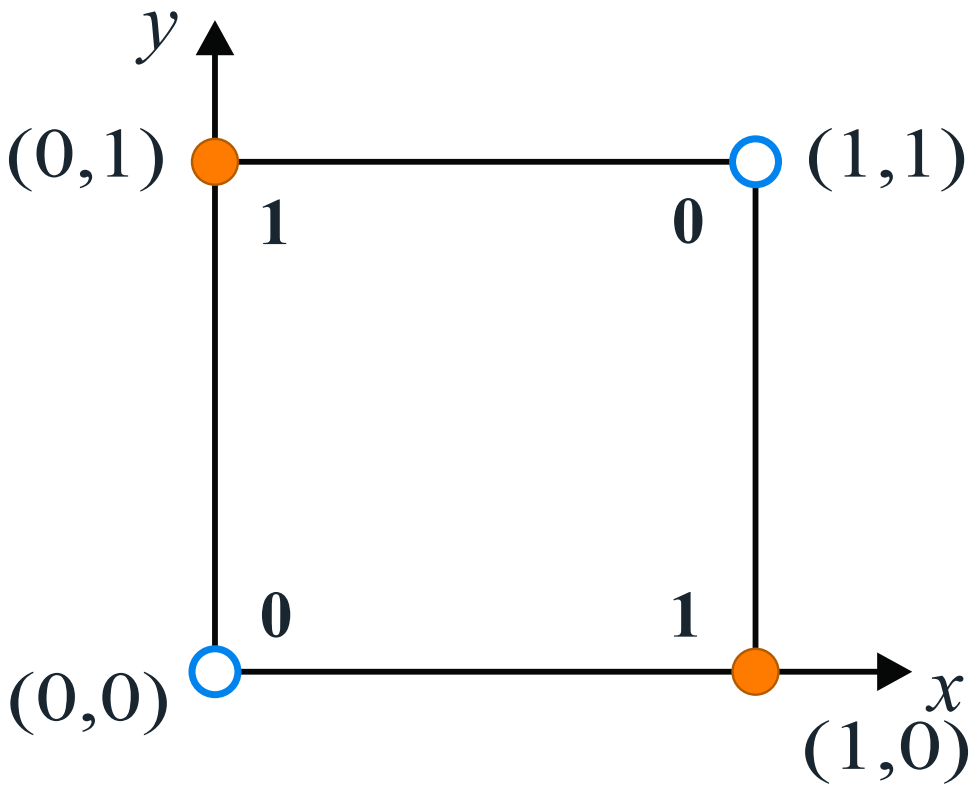


$$D_K(p) = \mathrm{AES}_K(p), \quad p \in \{0,1\}^{128}$$

The computational model



XOR from corners



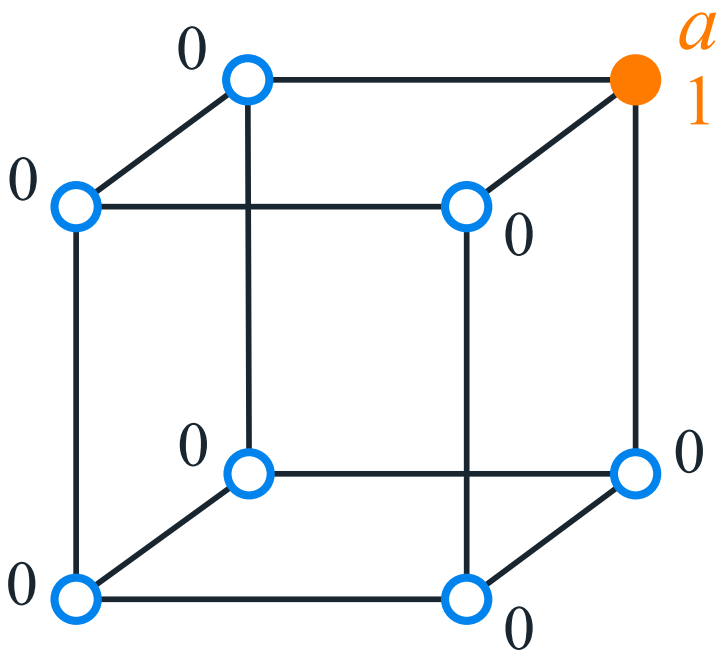
$$C_{01}(x,y)=\mathrm{ReLU}(y-x)$$

$$C_{10}(x,y)=\mathrm{ReLU}(x-y)$$

$$C_{01}+C_{10}=|x-y|$$

On $\{0,1\}^2$: exact XOR

General corner functions

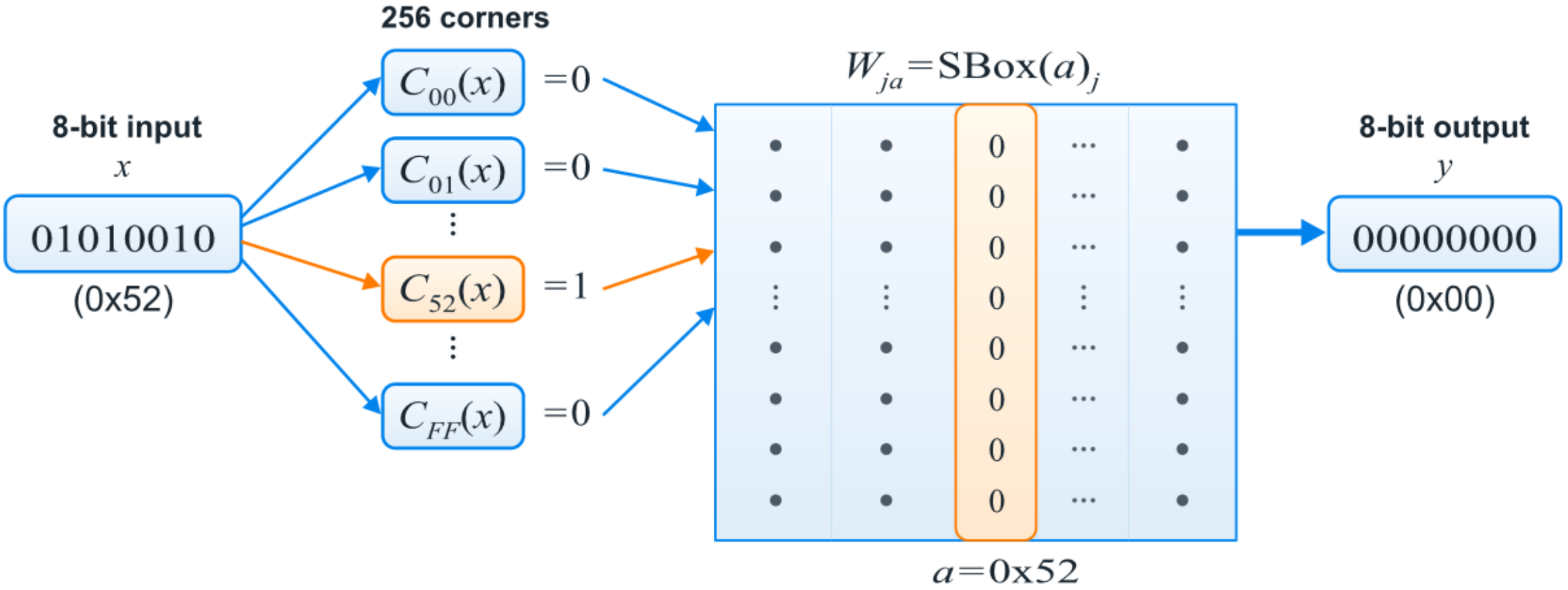


$$a \in \{0,1\}^n$$

$$F(x) = \sum_{a : f(a)=1} C_a(x)$$

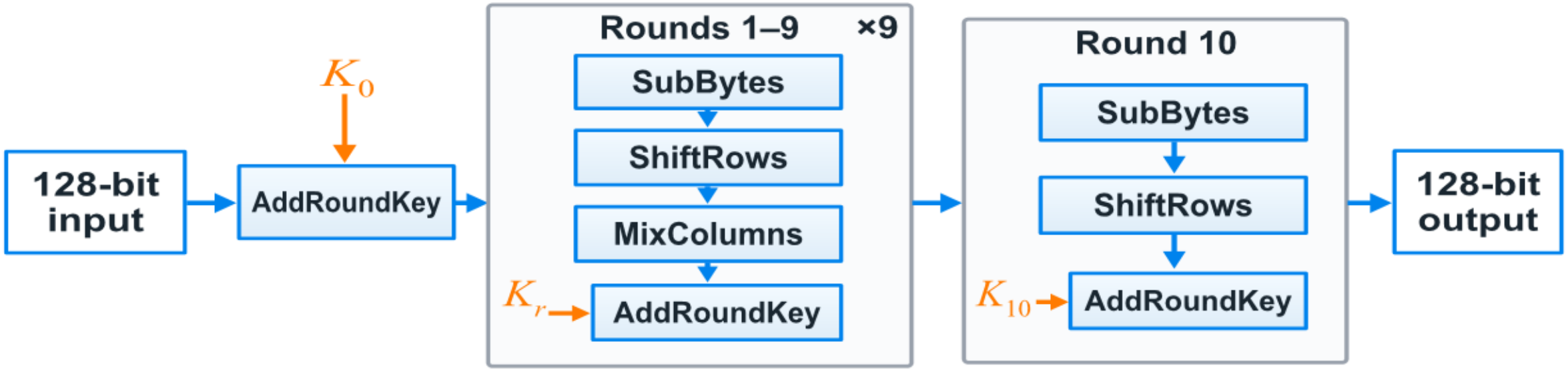
$$C_a(x) = \mathrm{ReLU}\left(1 - \underbrace{\sum_i [a_i(1-x_i) + (1-a_i)x_i]}_{\text{mismatches on bits}}\right)$$

A neural S-box

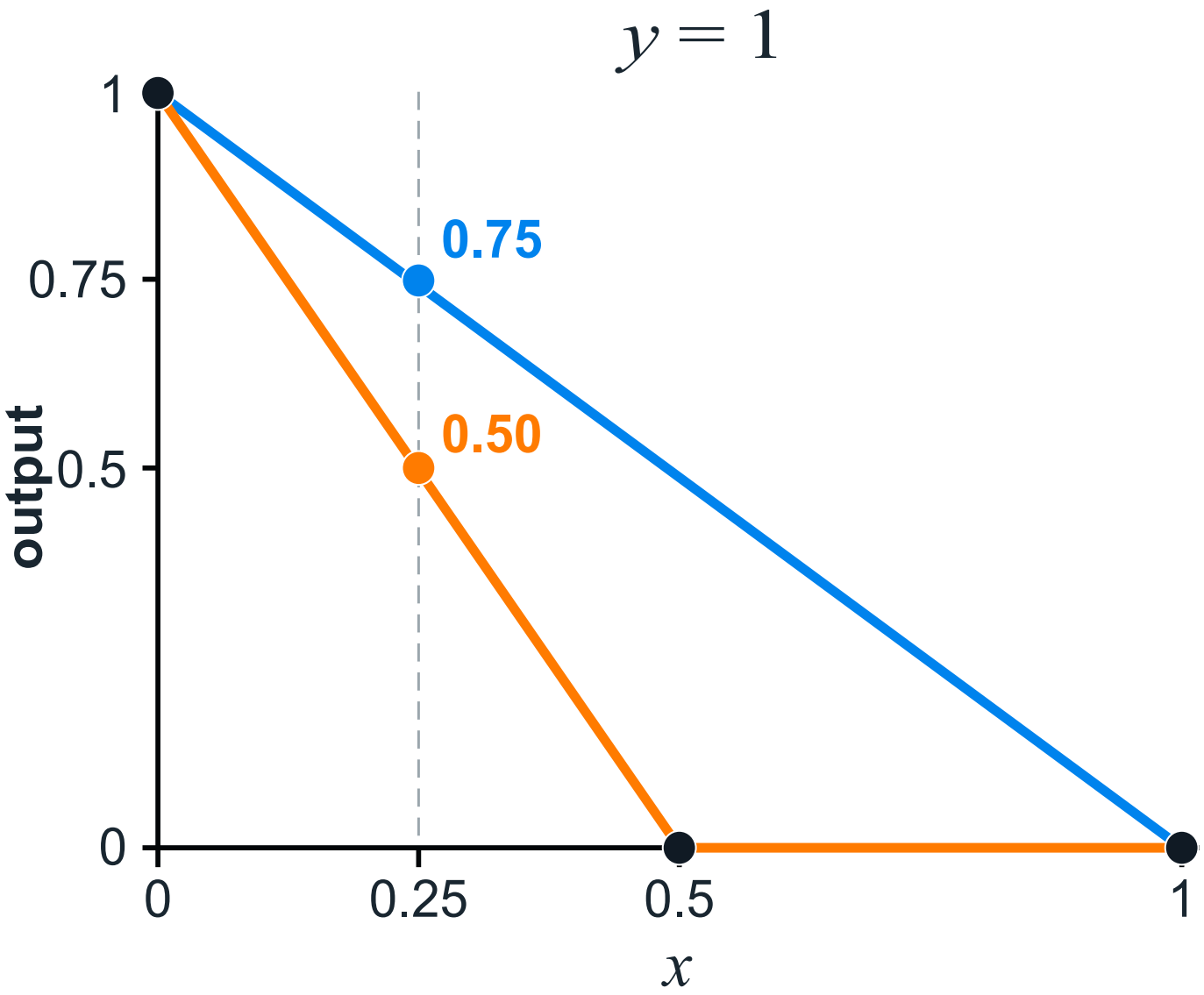


$$y_j=\sum_a \mathrm{SBox}(a)_j C_a(x)$$

Composing AES



Many continuous extensions



$$f_1(x,y)=|x-y|$$

$$f_2(x,y)=\mathrm{ReLU}(2|x-y|-1)$$

(x,y)	f_1	f_2
$(0,0)$	0	0
$(0,1)$	1	1
$(1,0)$	1	1
$(1,1)$	0	0

Conditional collisions

$$g(x, k_i) = |x - k_i|$$

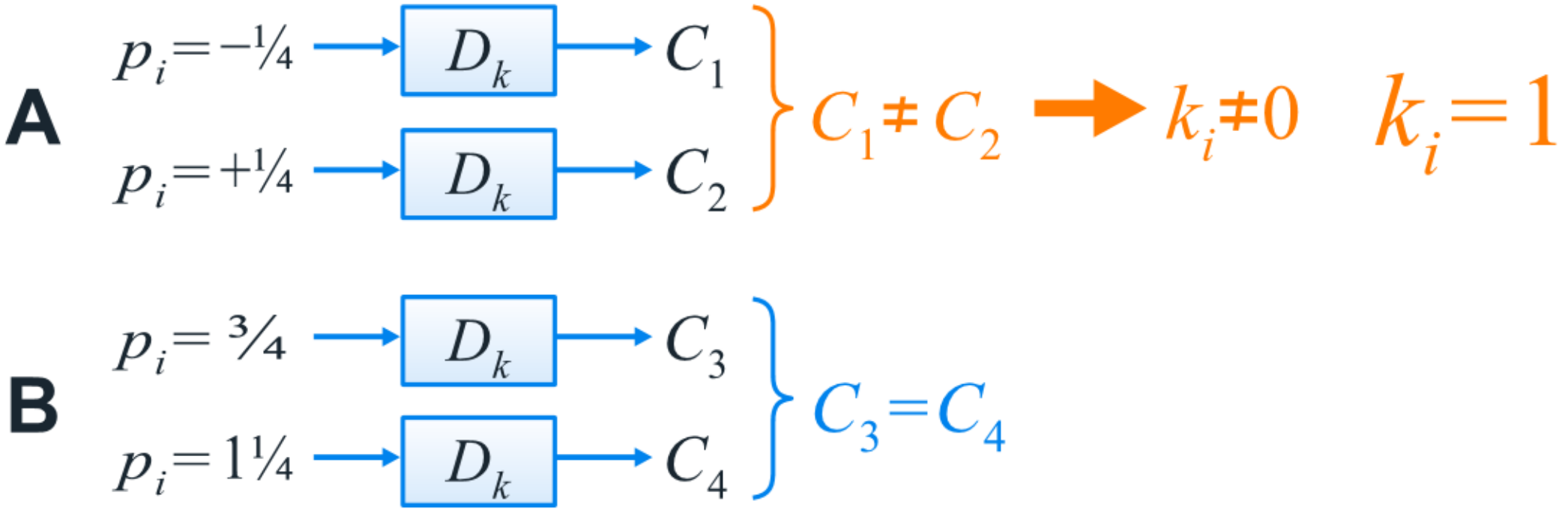
after XOR

Input pair	$k_i = 0$	$k_i = 1$
$-\frac{1}{4}, +\frac{1}{4}$	$\frac{1}{4} = \frac{1}{4}$	$1\frac{1}{4} \neq \frac{3}{4}$
$\frac{3}{4}, 1\frac{1}{4}$	$\frac{3}{4} \neq 1\frac{1}{4}$	$\frac{1}{4} = \frac{1}{4}$

Recovering a key bit

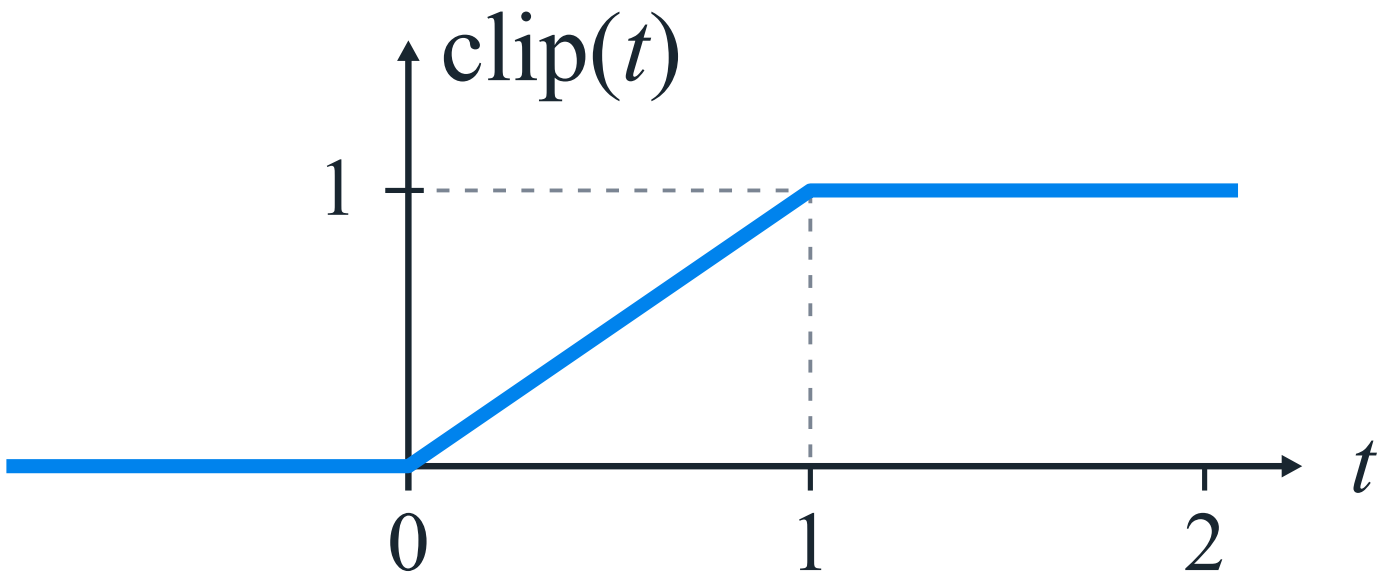
Fixed background

$$g(x,k_i)=|x-k_i|$$



Both equal: retry

Clipping the inputs



clip

$-\frac{1}{4} \rightarrow 0$

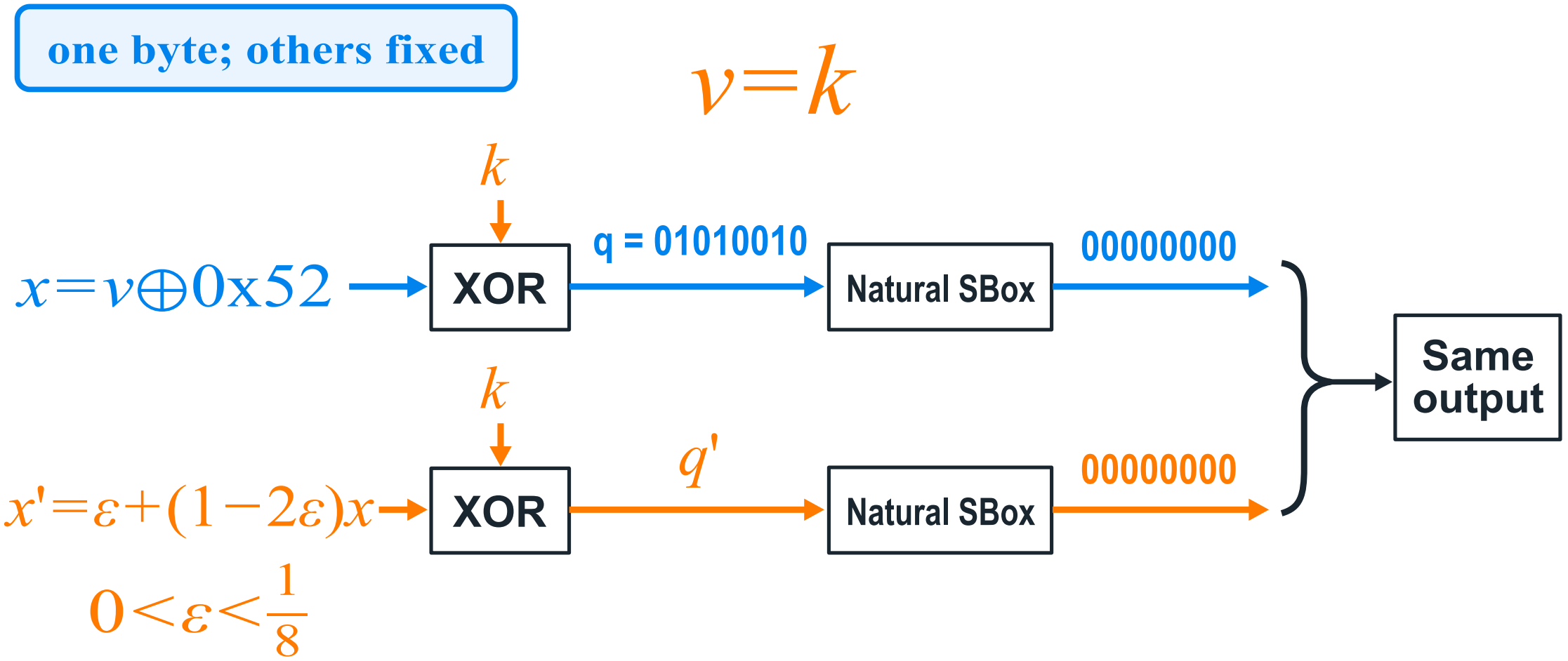
$1\frac{1}{4} \rightarrow 1$

$$\text{clip}(t) = \text{ReLU}(t) - \text{ReLU}(t-1)$$

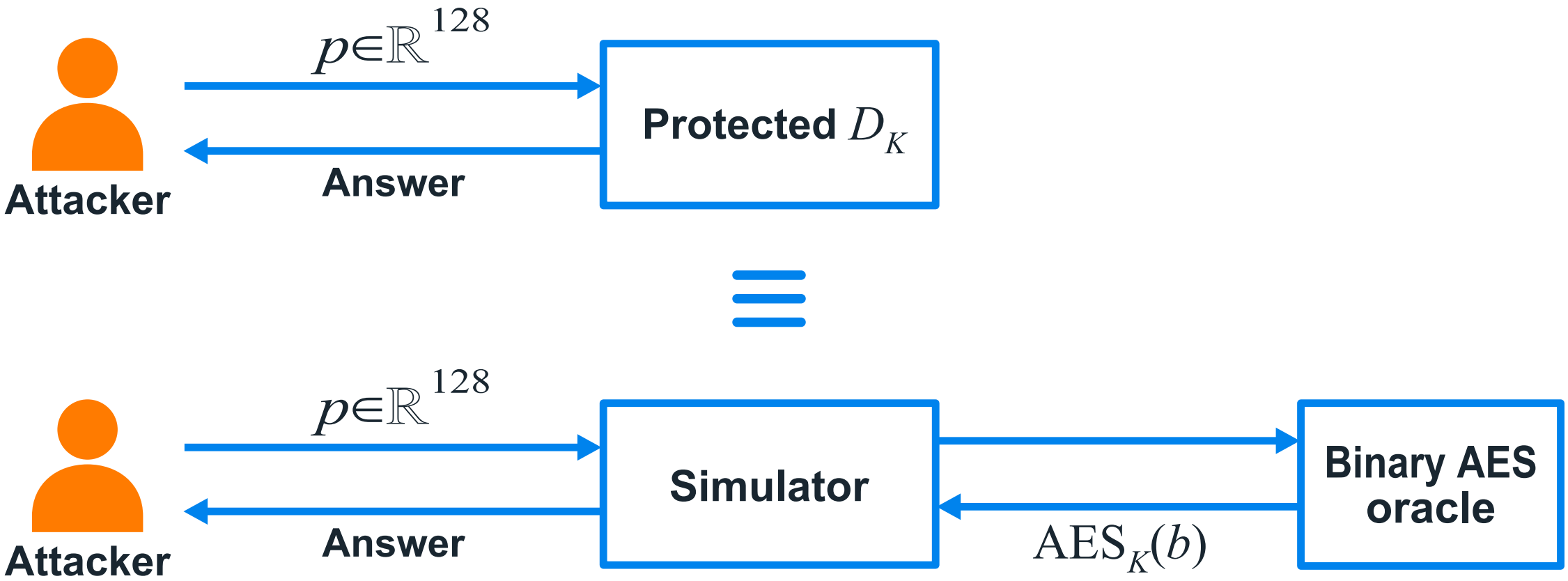
$$g(x,0) = x \qquad g(x,1) = 1 - x$$

$$x \in [0,1]$$

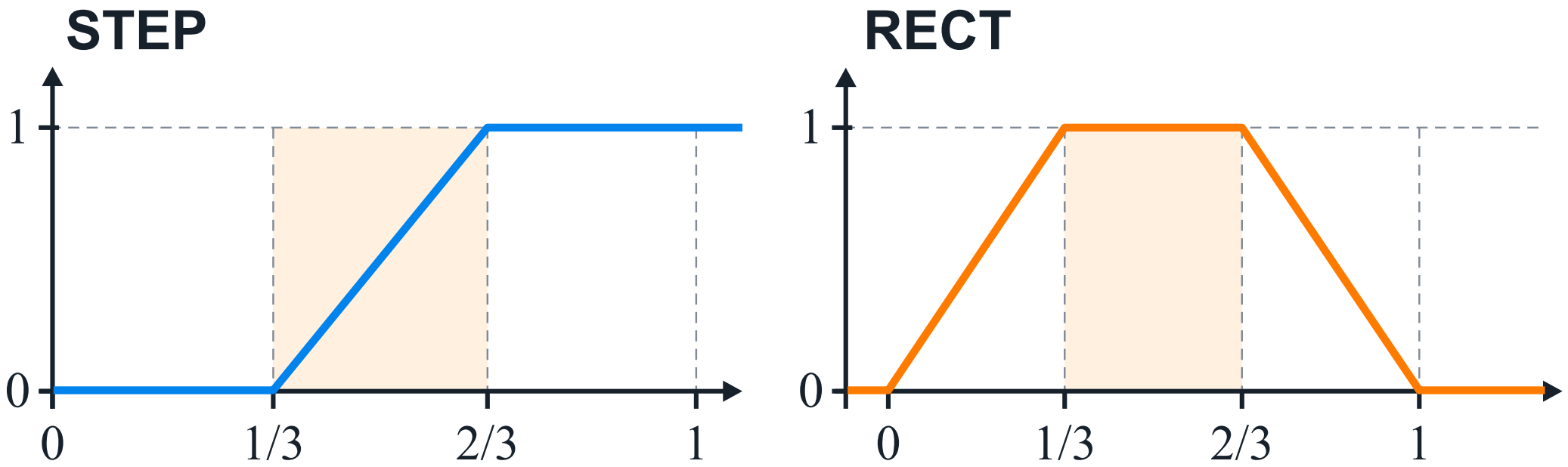
An attack inside the cube



The security goal



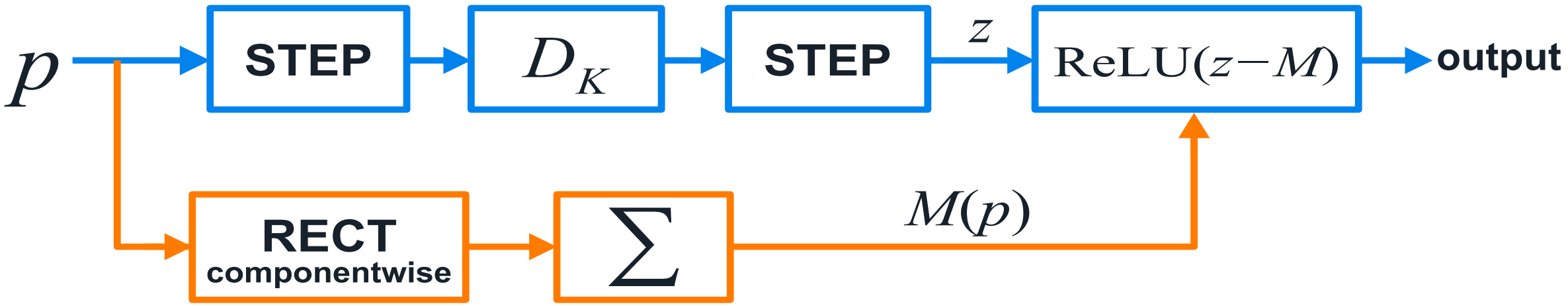
STEP and RECT



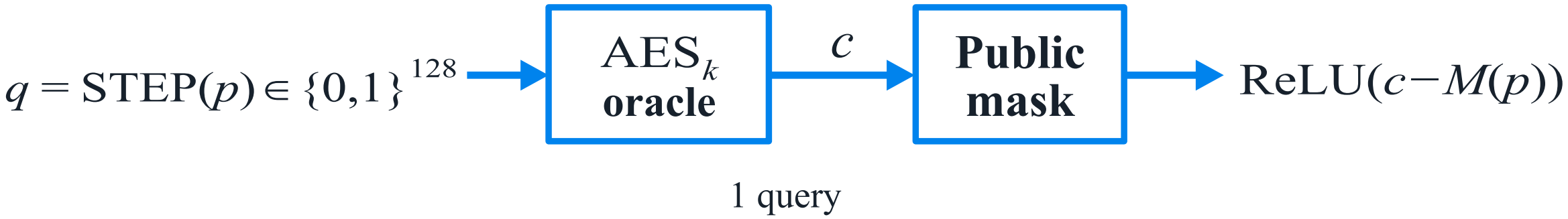
$$M(p) = \sum_i \mathrm{RECT}(p_i)$$

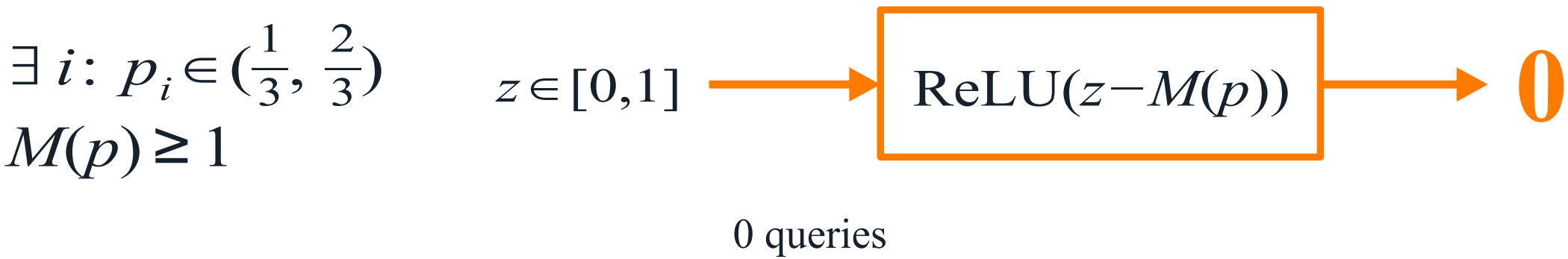
$$p_i \in (1/3, 2/3) \Rightarrow M(p) \geq 1$$

The protection wrapper

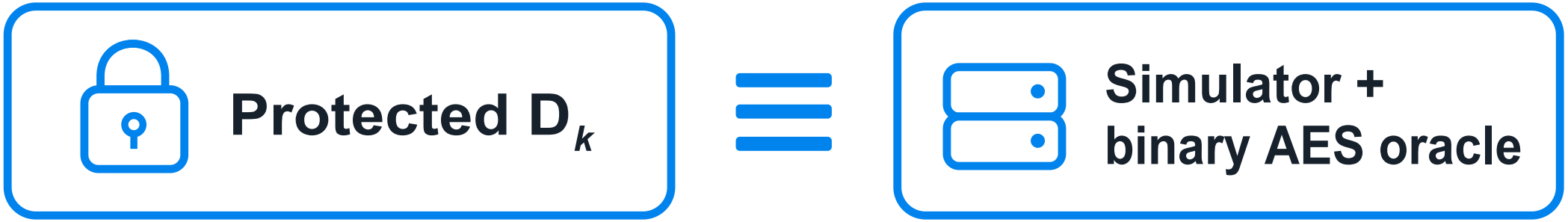


Simulating real-valued queries

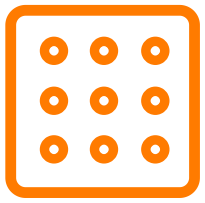




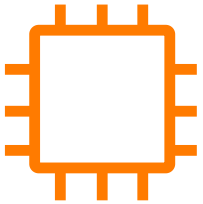
Scope of the guarantee



Outside this model



Weight access



Floating-point execution



Timing / power

Measured throughput

Implementation	Encryption (Mbps)	Decryption (Mbps)
NeuralAES Base	82.7 ± 0.2	49.9 ± 0.1
NeuralAES Simplified	206.9 ± 0.8	114.4 ± 0.5
T-Tables NeuralAES	280.1 ± 1.2	114.4 ± 0.4
Protected T-Tables NeuralAES	279.7 ± 1.5	114.4 ± 0.3

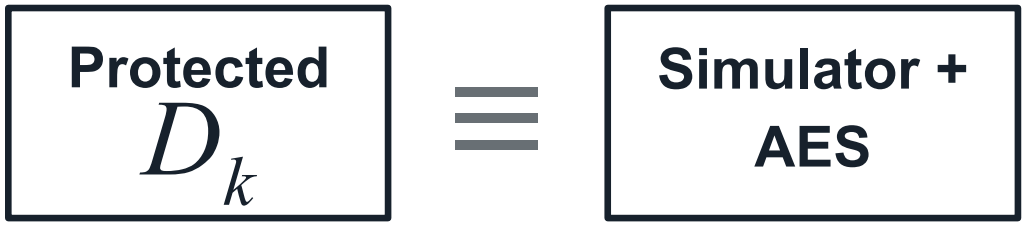
Takeaways



Exact on bits



leakage



Security-preserving wrapper

Questions?

