Od prostych sieci do GPT: Podróż przez wyzwania techniczne i etyczne w AI

Wojciech Szczepański

ChatGPT

30
11
2022

AI

Maszyna wykonująca ludzkie zadania poznawcze

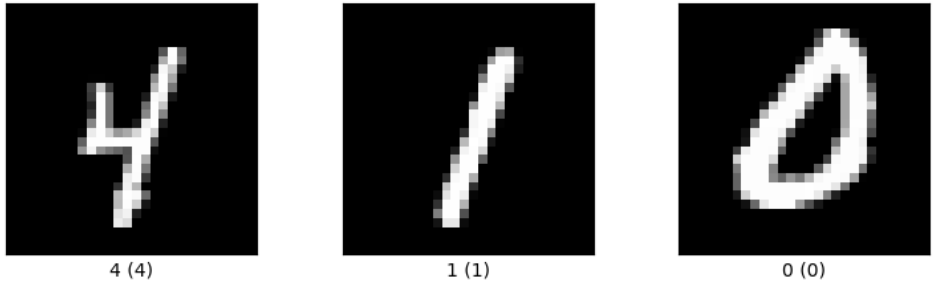
Sprawienie by maszyny wykonywały zadania poznawcze,
o których nie sądziliśmy, że są w stanie je wykonać

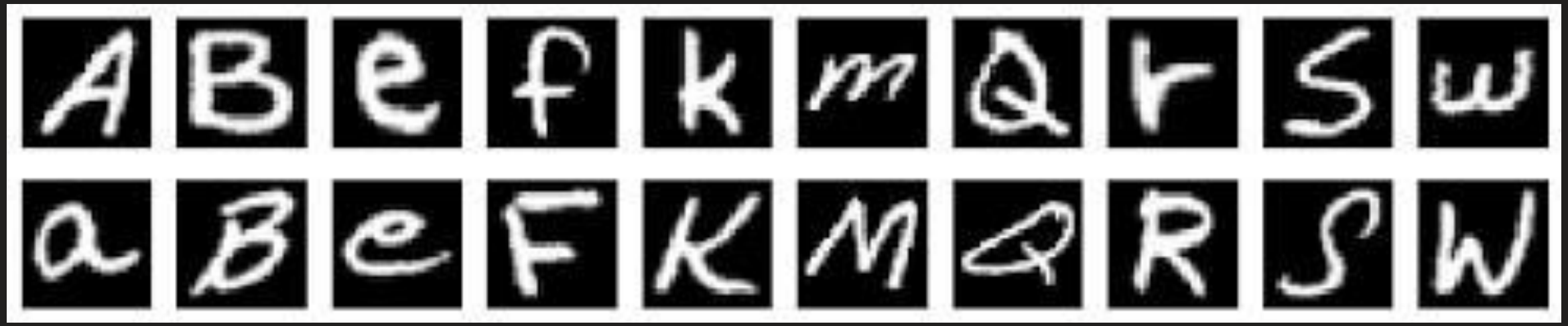
Formalizacja

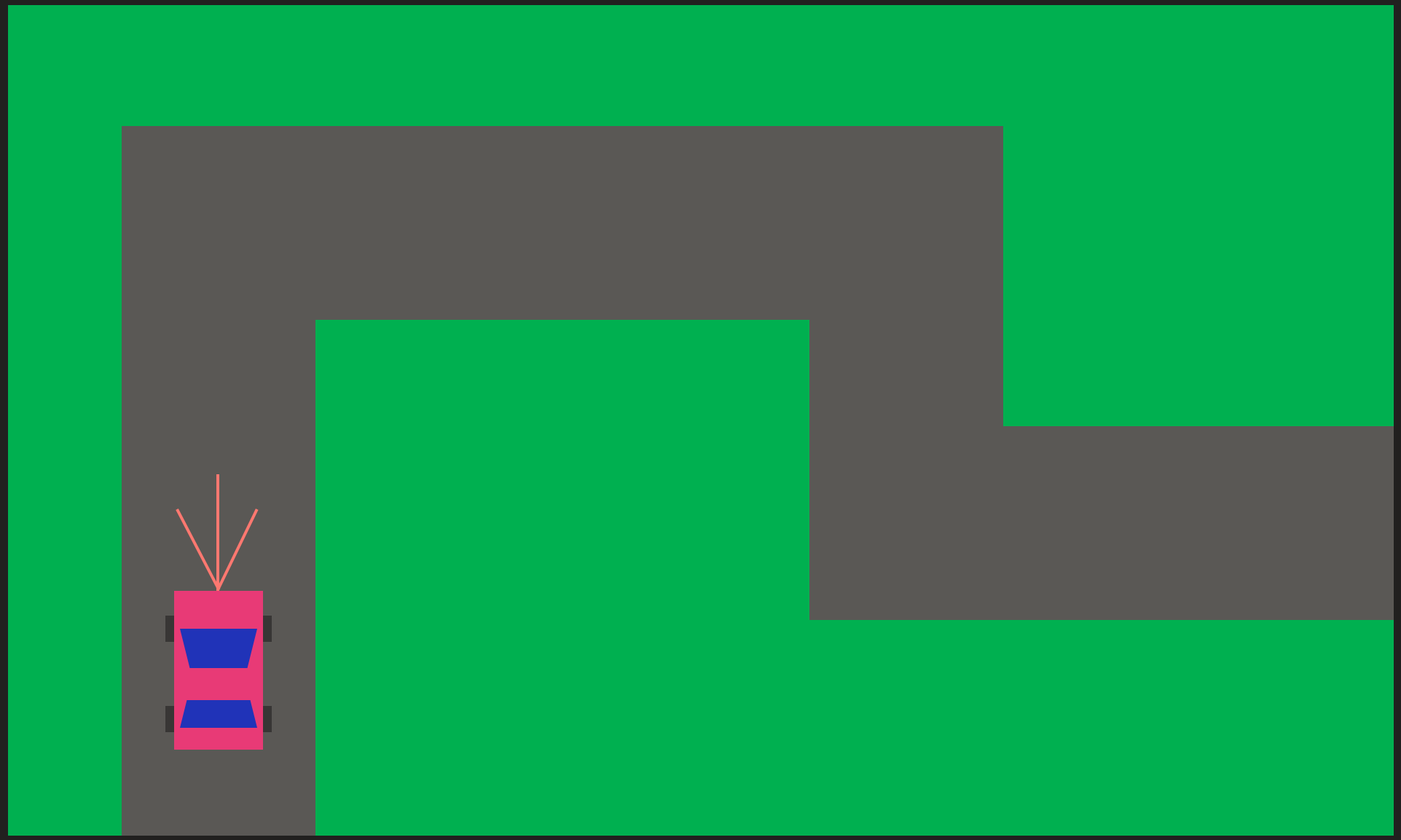
Napisanie programu w oparciu o wyprowadzone reguły

?

Napisanie programu w oparciu
o wyprowadzone reguły


4 (4)
1 (1)
0 (0)
7 (7)
8 (8)
1 (1)
2 (2)
7 (7)
1 (1)


A B e f k m Q r S w
a B e F K M Q R S W



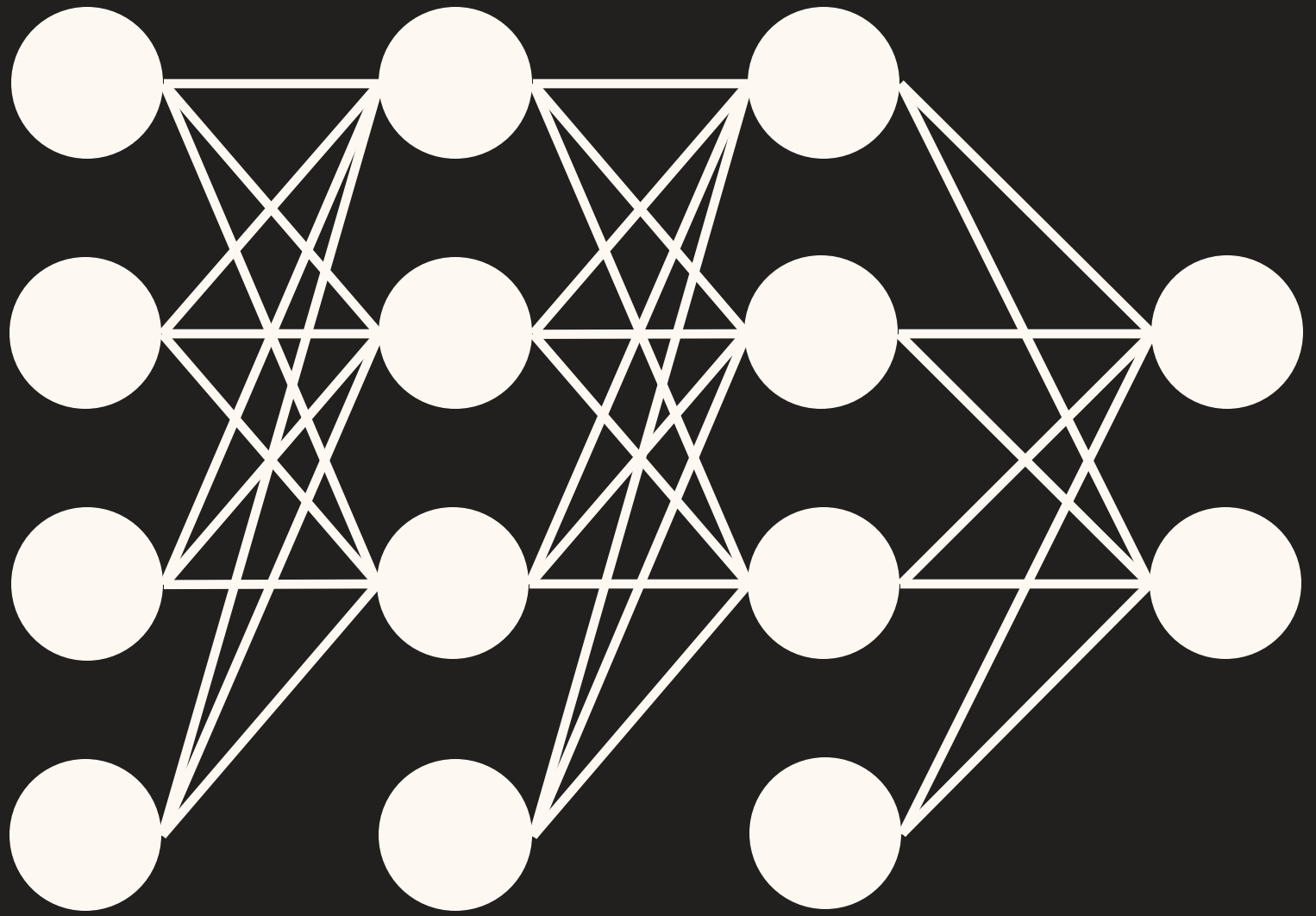
Znajdźmy dużo przykładów

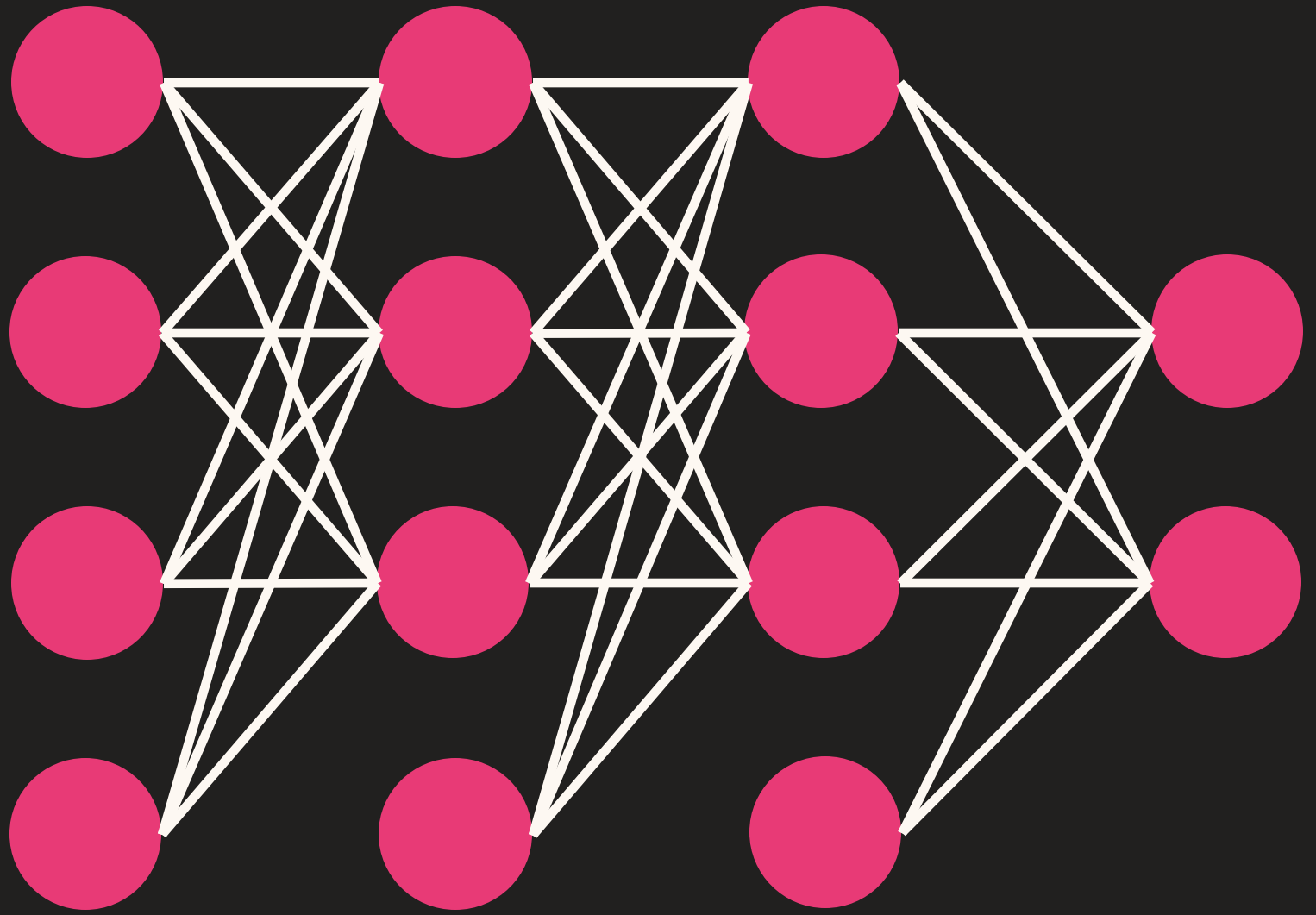
Nauczmy komputer
samemu dobrać reguły

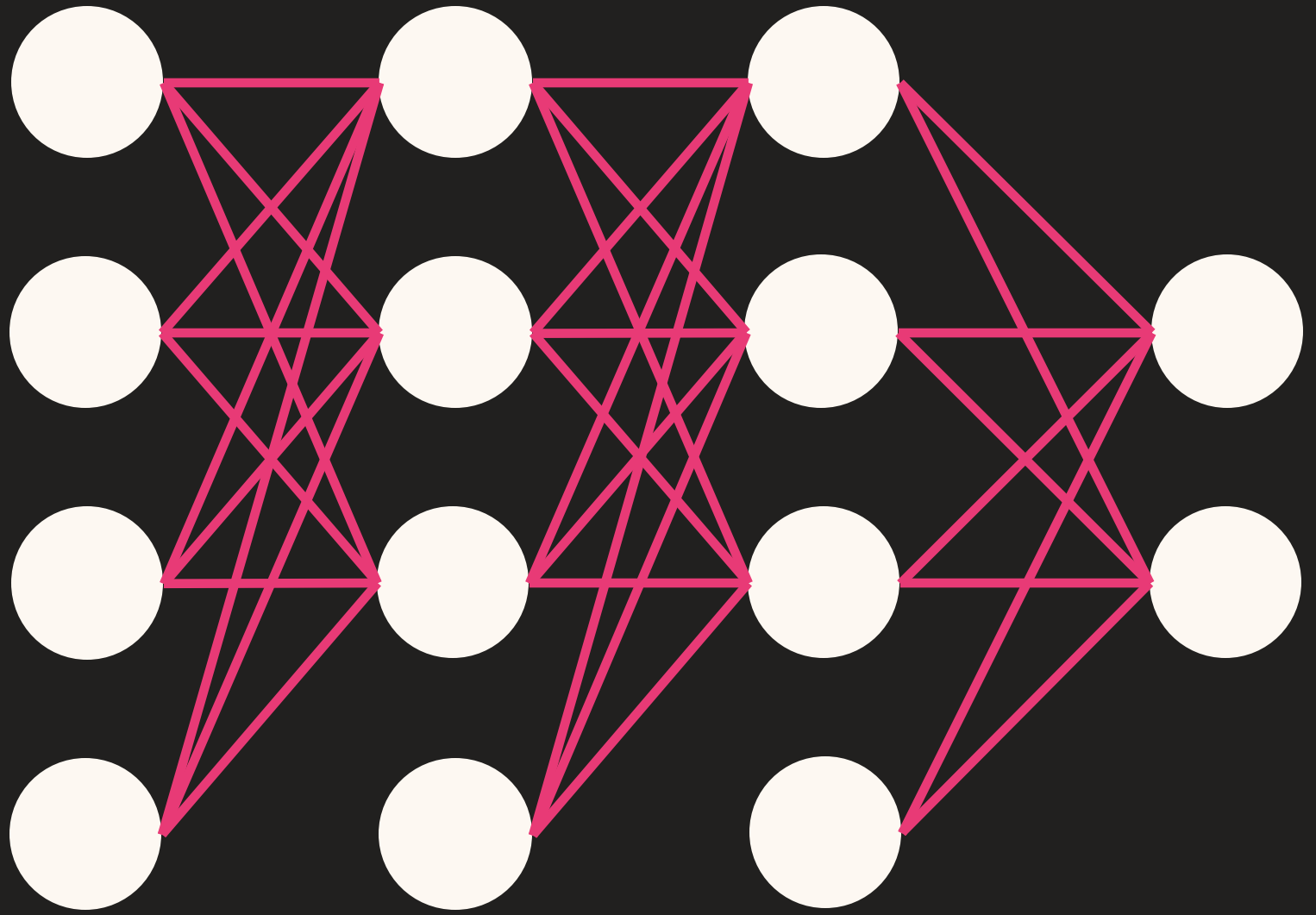
Sieć neuronowa

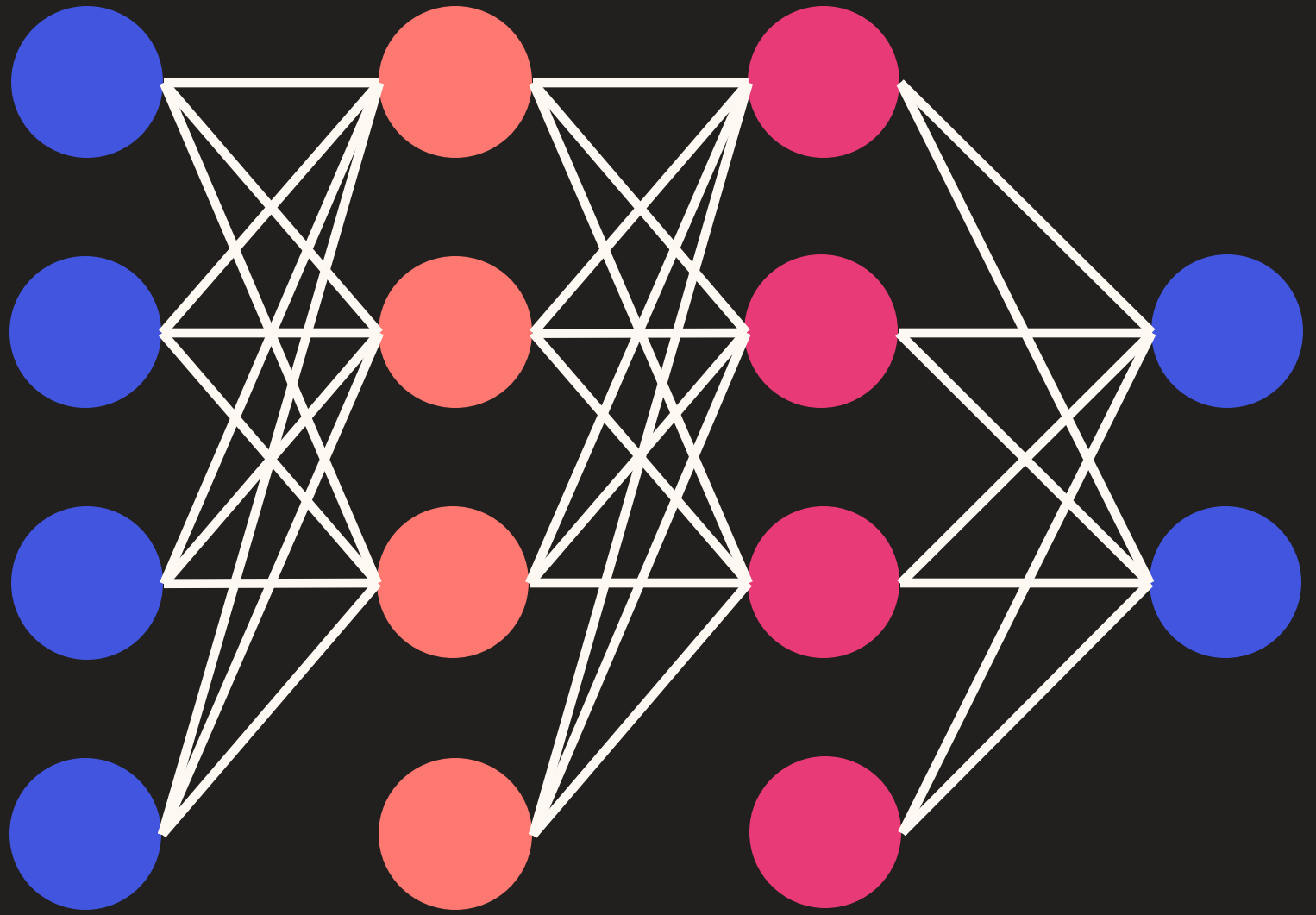
AKT I

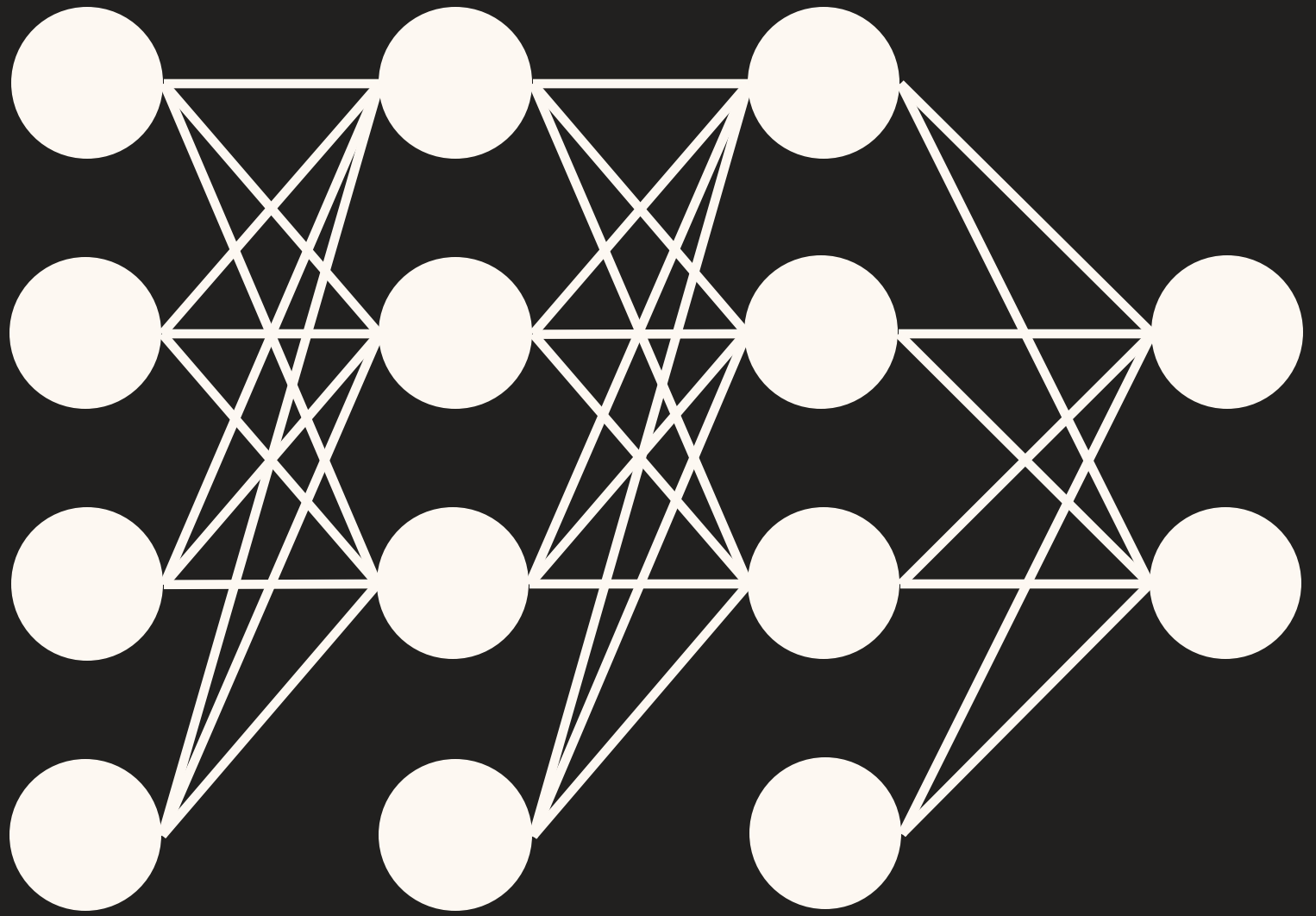
Na początku był neuron…

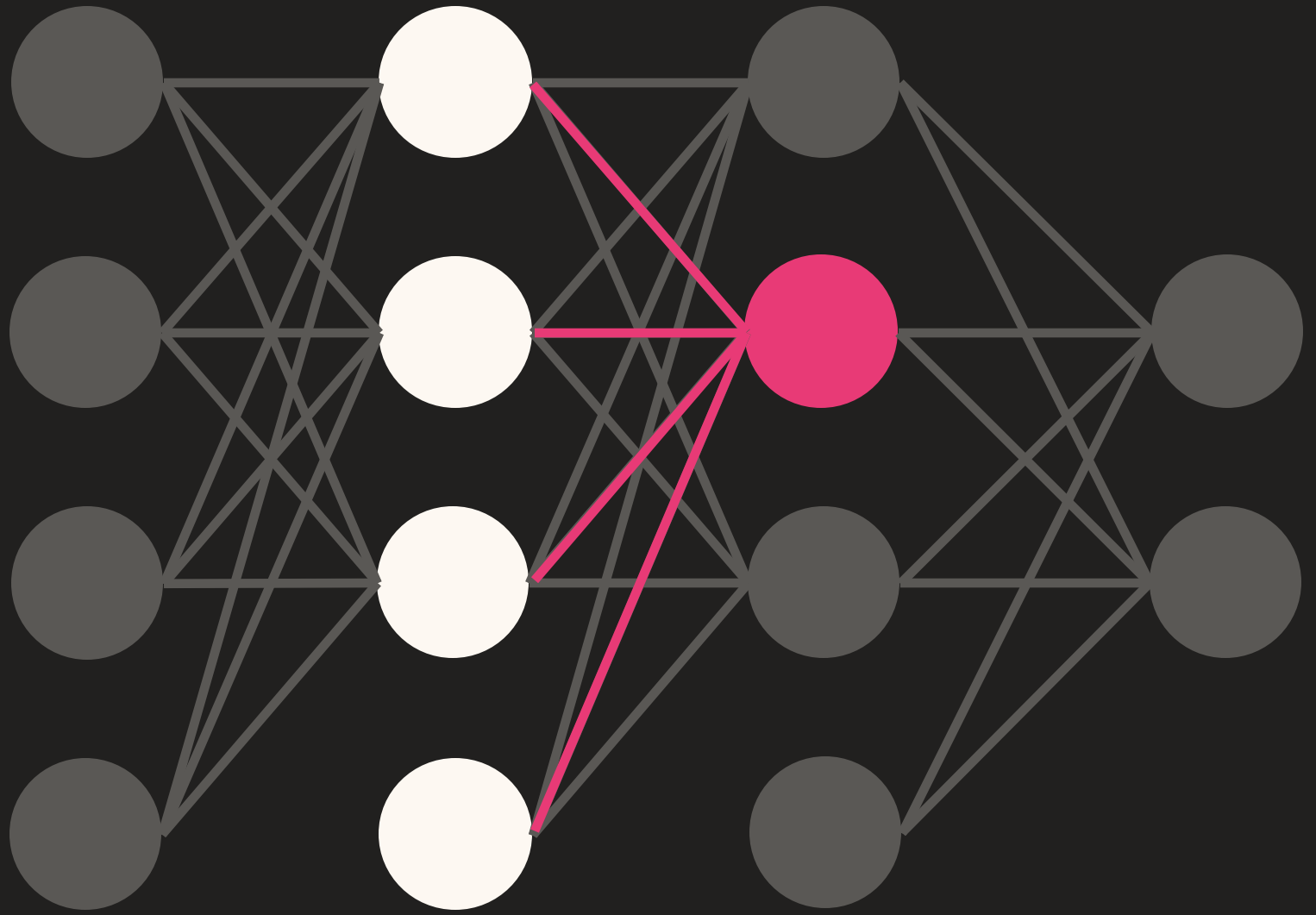


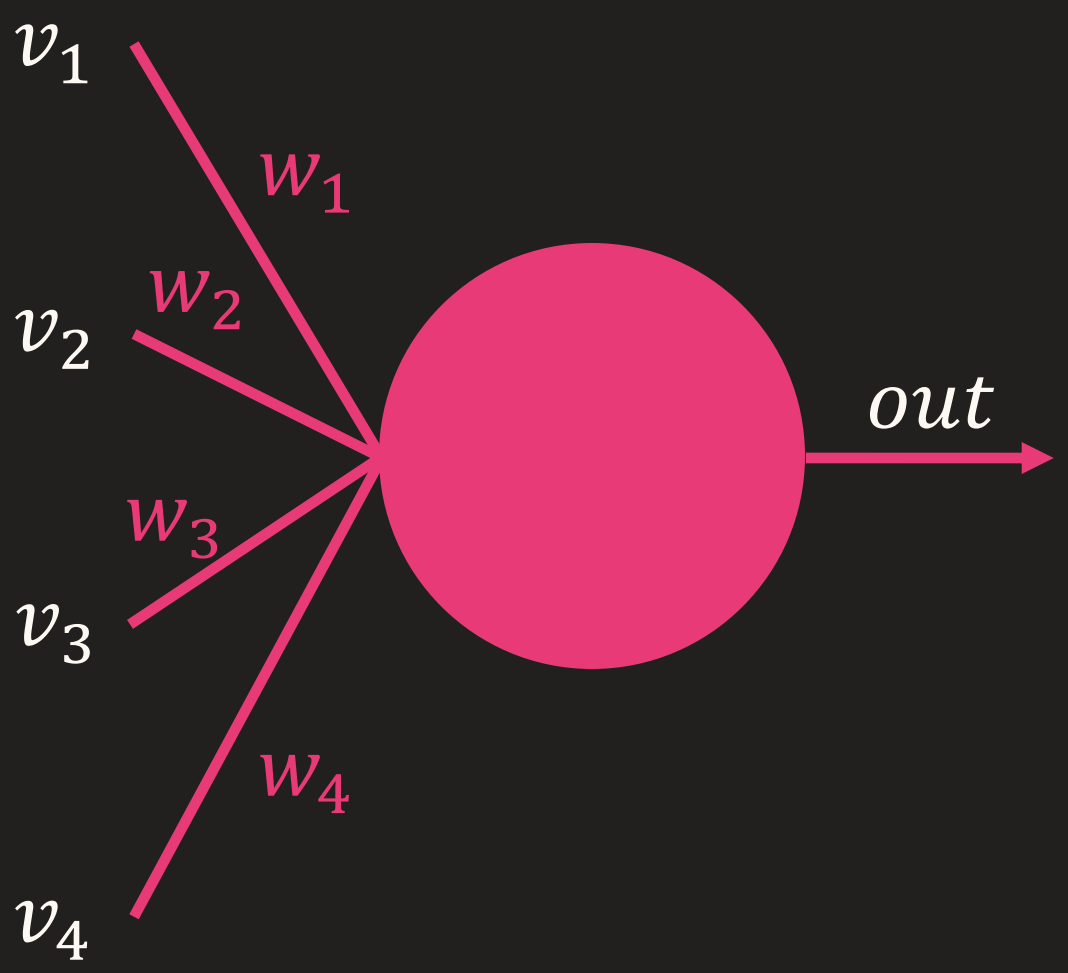


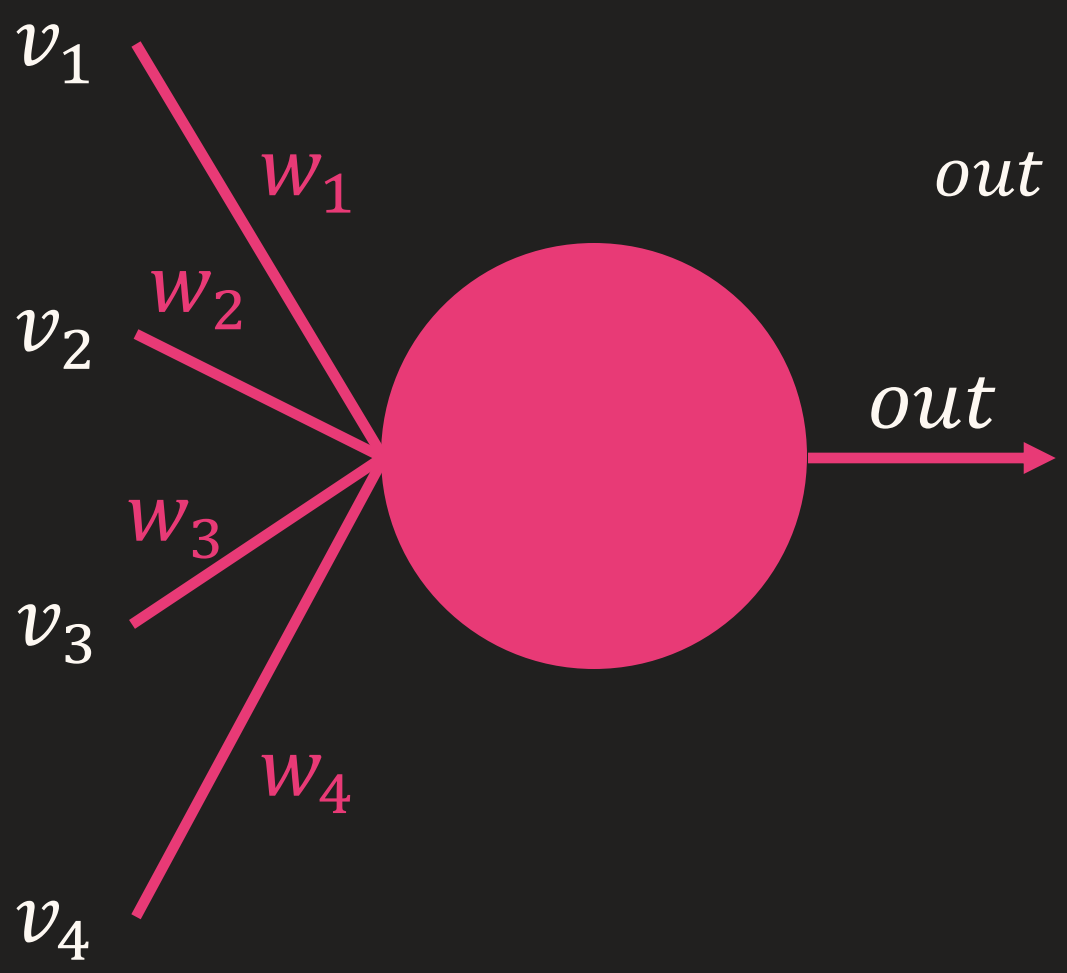




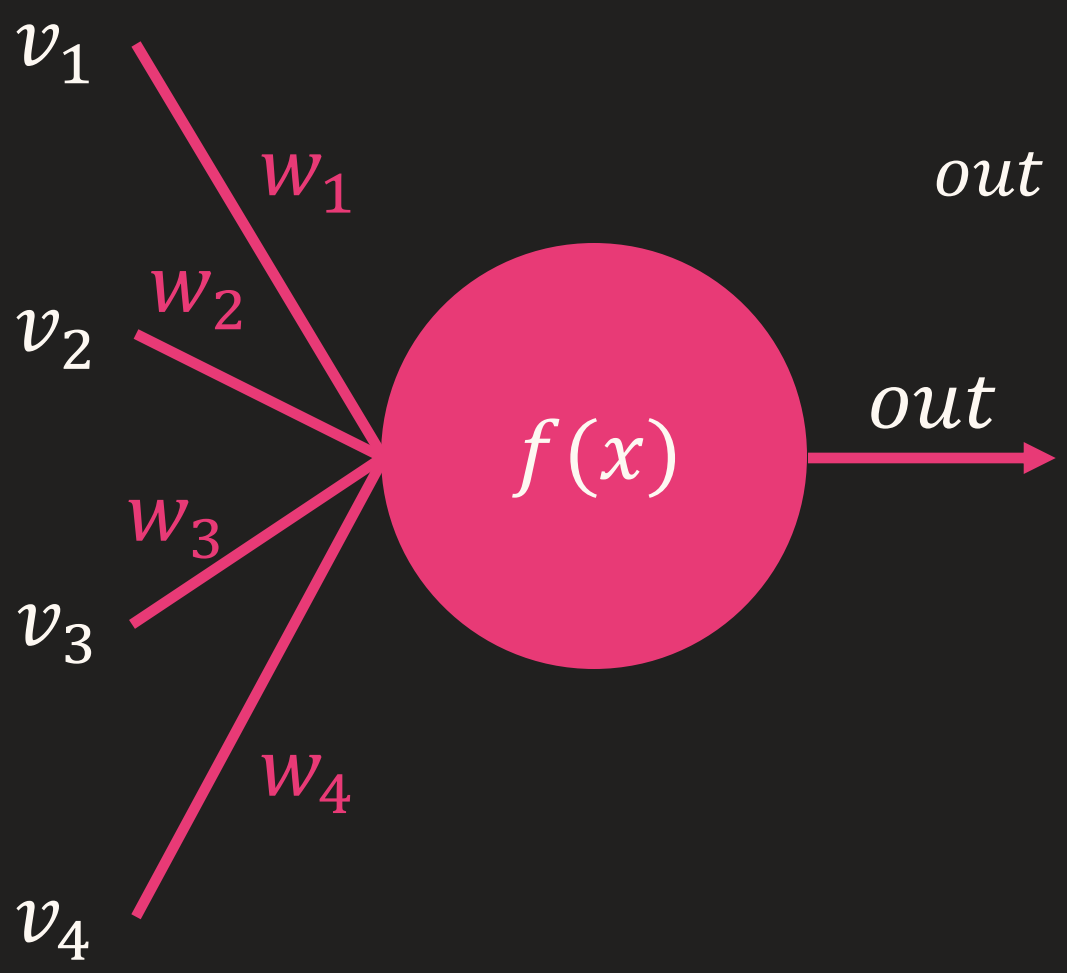





v_1
w_1
w_2
v_2
out
w_3
v_3
w_4
v_4



$$out = w_1 * v_1 + w_2 * v_2 + w_3 * v_3 + w_4 * v_4$$



$$out = w_1 * v_1 + w_2 * v_2 + w_3 * v_3 + w_4 * v_4$$

$$out = f(w_1 v_1 + w_2 v_2 + w_3 v_3 + w_4 v_4)$$

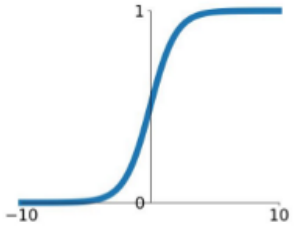
$$out = f(w_1 v_1 + \cdots + w_n v_n)$$

$$out = f\left(\sum_{i}^{n} w_i v_i\right)$$

Activation Functions

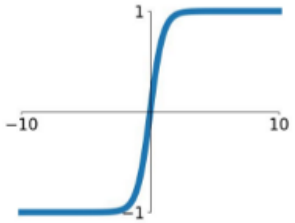
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



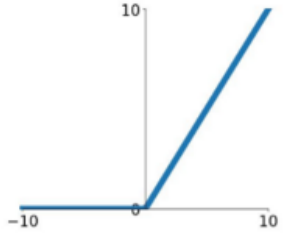
tanh

$$\tanh(x)$$



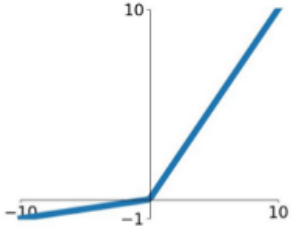
ReLU

$$\max(0, x)$$



Leaky ReLU

$$\max(0.1x, x)$$

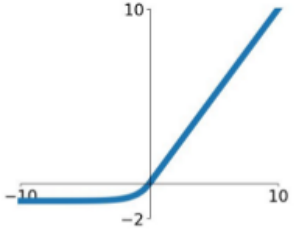


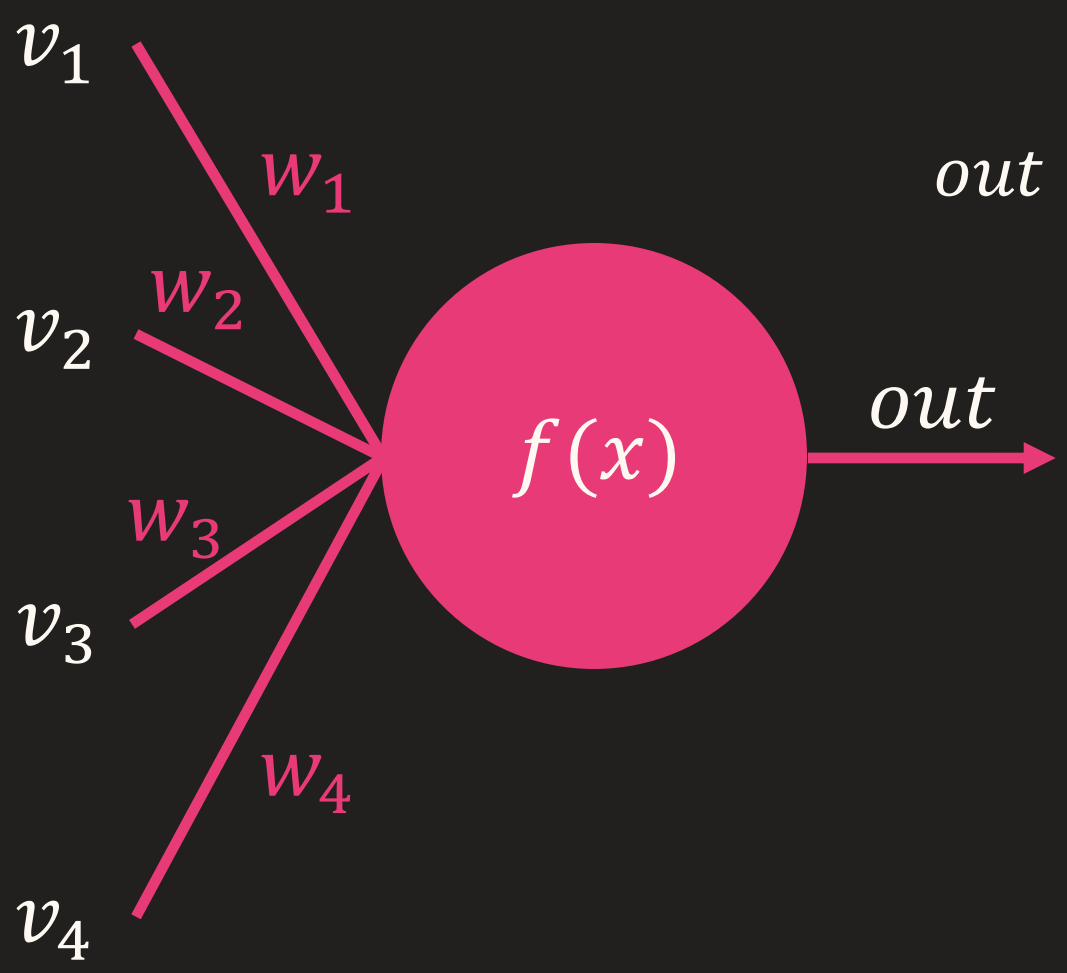
Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



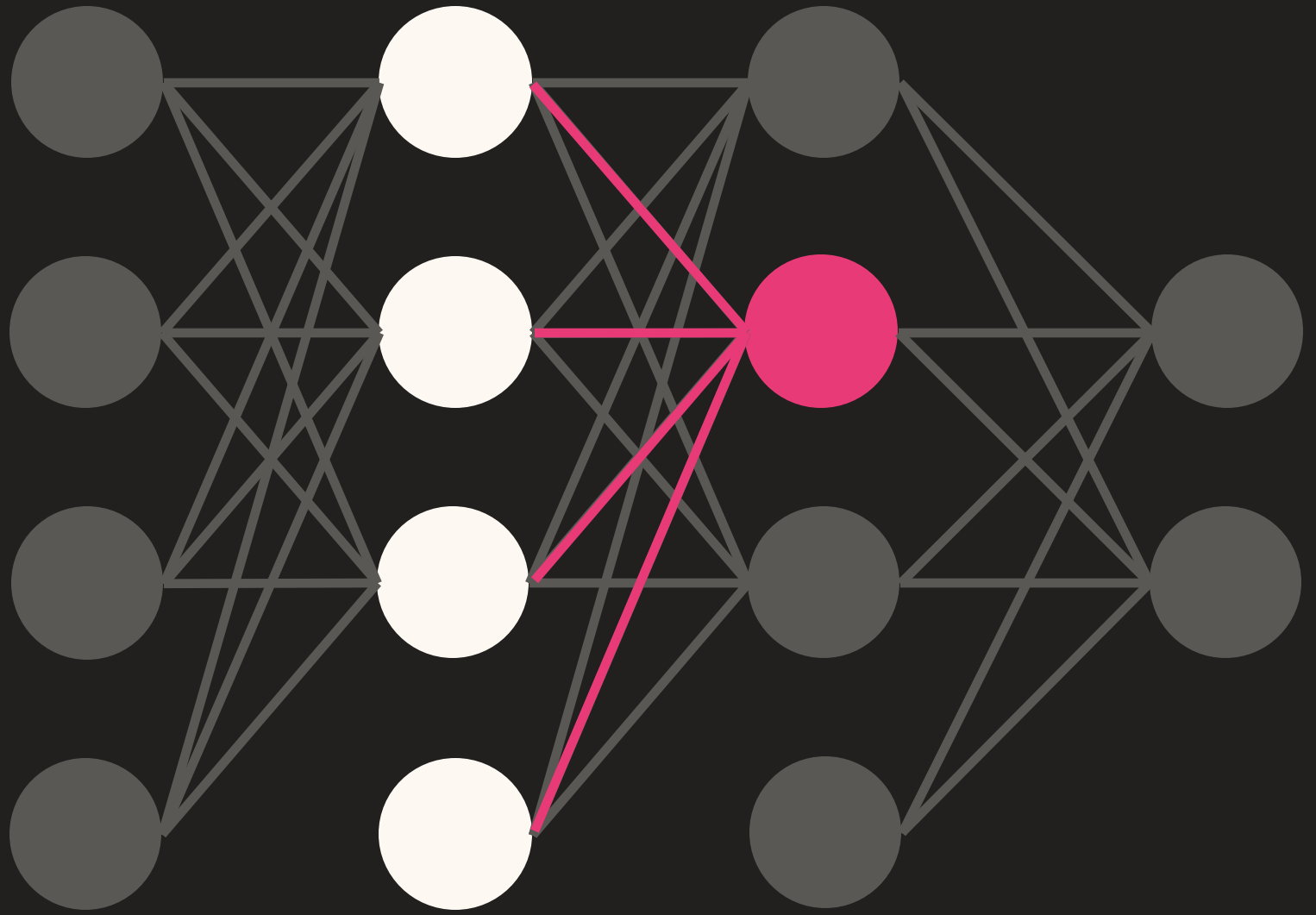


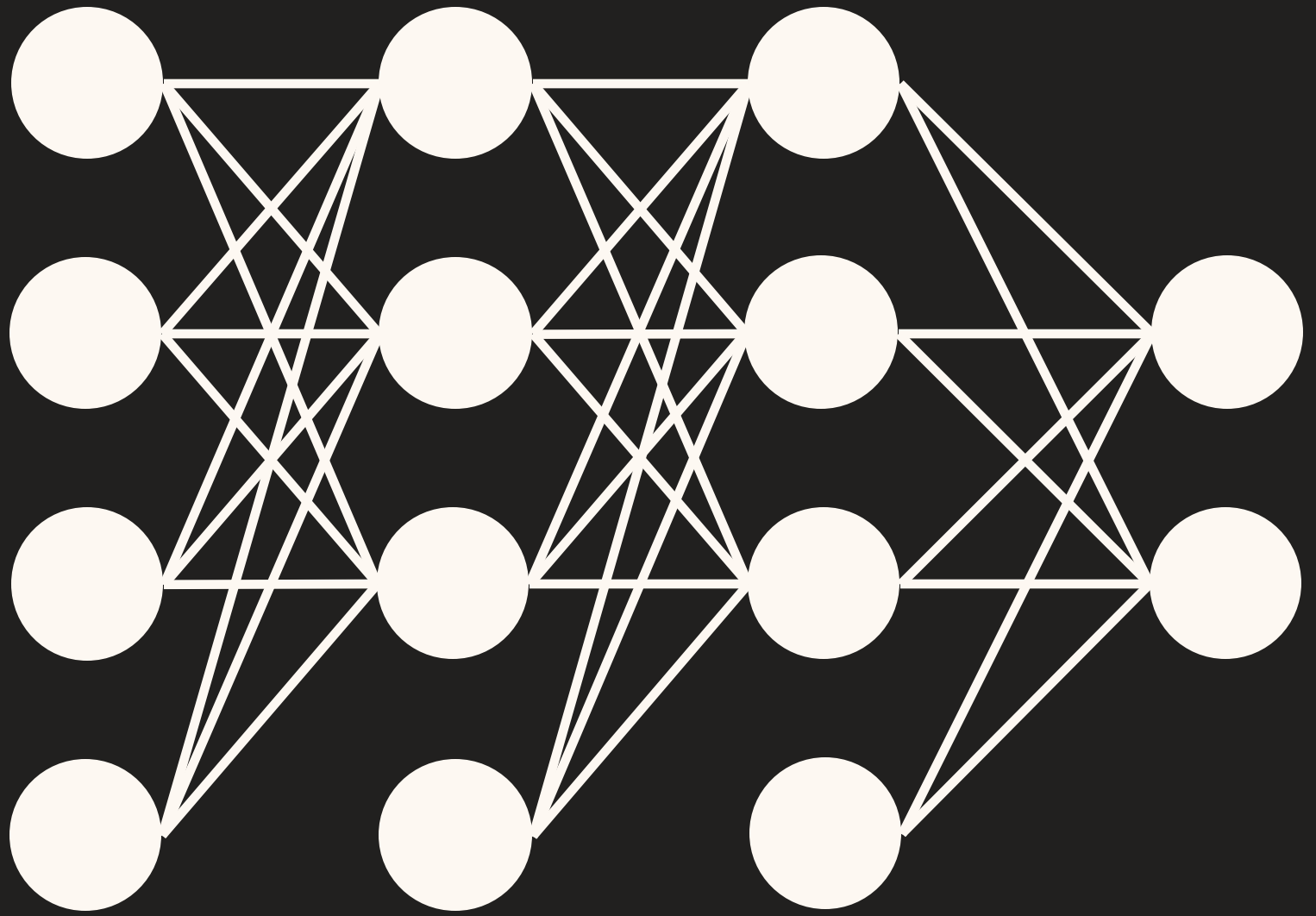
$$out = w_1 * v_1 + w_2 * v_2 + w_3 * v_3 + w_4 * v_4$$

$$out = f(w_1 v_1 + w_2 v_2 + w_3 v_3 + w_4 v_4)$$

$$out = f(w_1 v_1 + \cdots + w_n v_n)$$

$$out = f\left(\sum_{i}^{n} w_i v_i\right)$$





Jak sieć neuronowa się uczy?

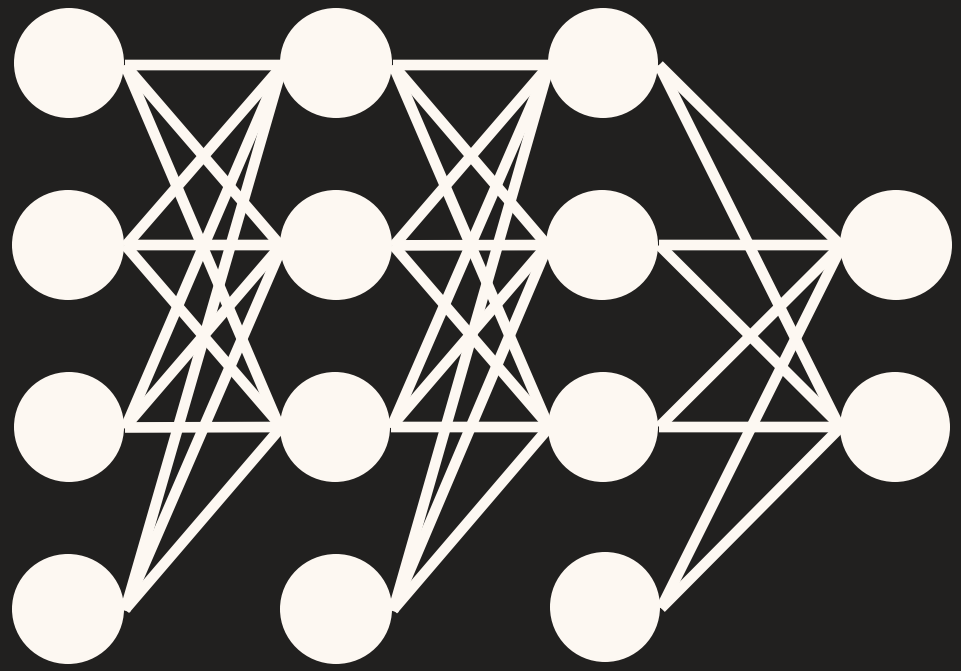
$$\begin{bmatrix} 132 & \cdots & 67 \\ \vdots & \ddots & \vdots \\ 15 & \cdots & 42 \end{bmatrix}$$

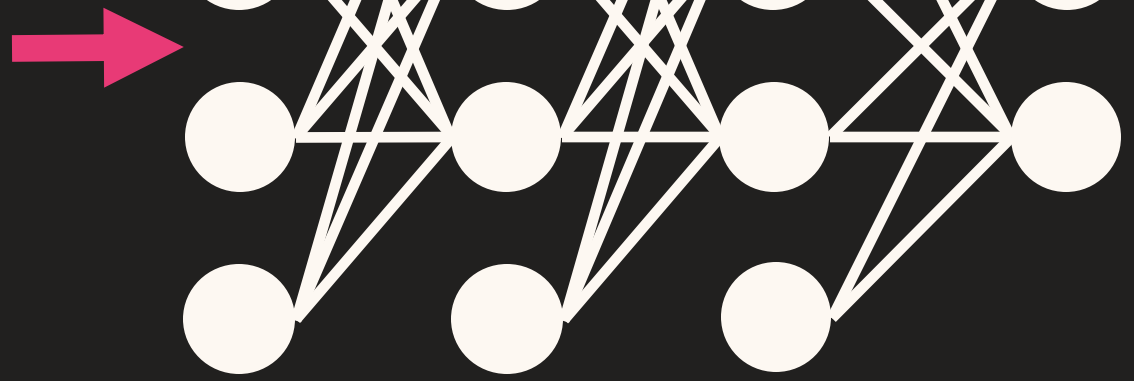
$$\begin{bmatrix} 0.52 & \cdots & 0.26 \\ \vdots & \ddots & \vdots \\ 0.06 & \cdots & 0.16 \end{bmatrix}$$

$\{Kot, Pies\}$ 

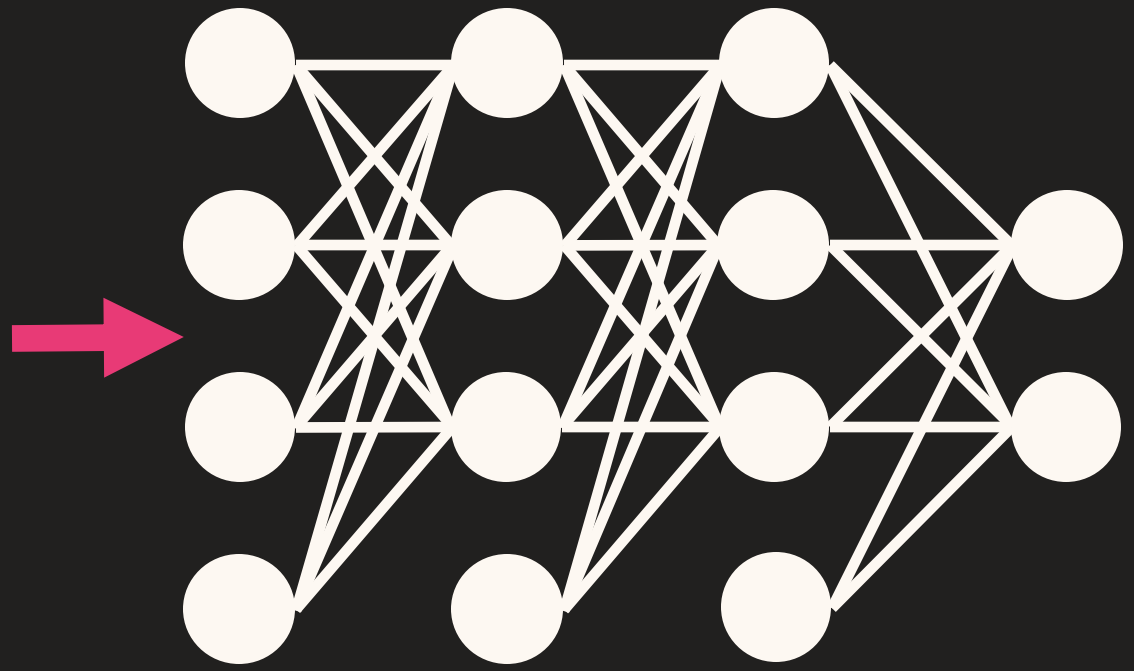
$\begin{bmatrix}1\\0\end{bmatrix}$ – Kot

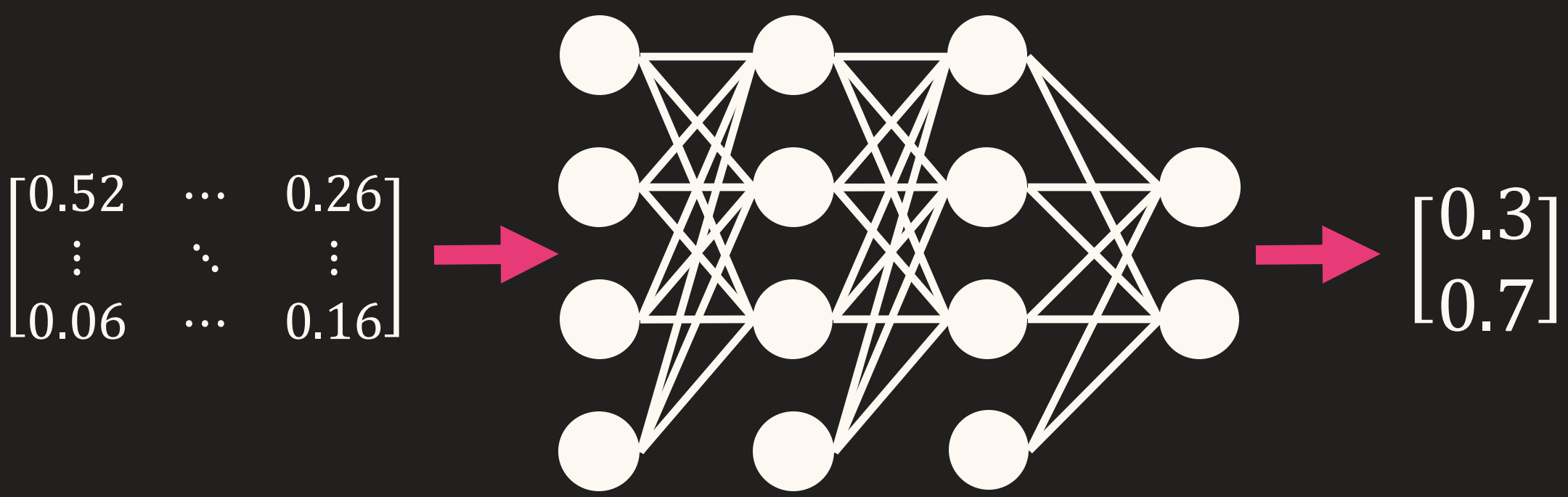
$\begin{bmatrix}0\\1\end{bmatrix}$ – Pies

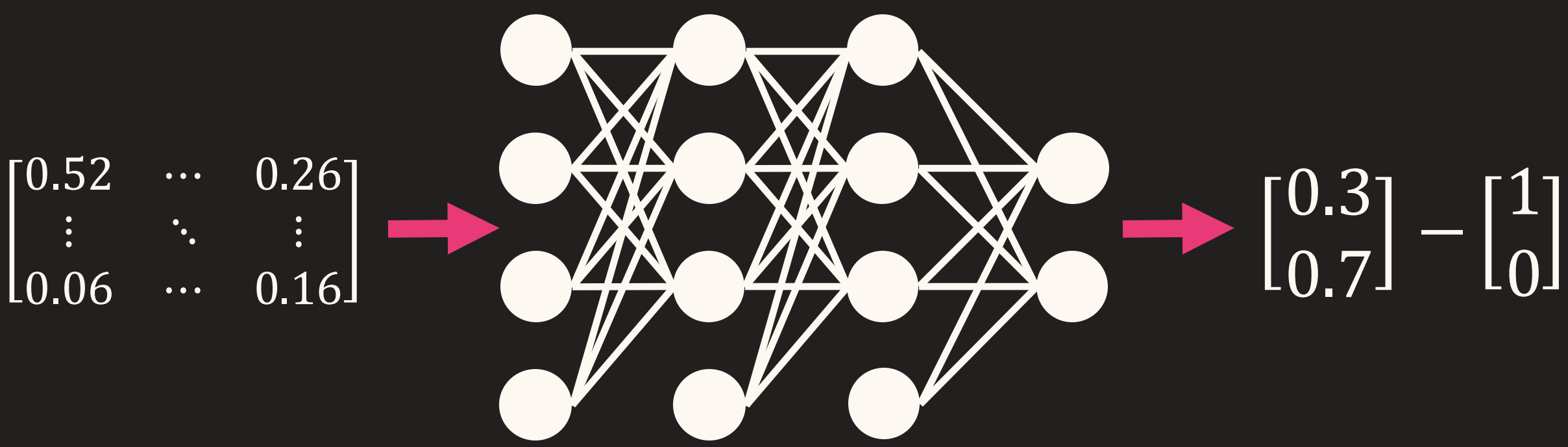




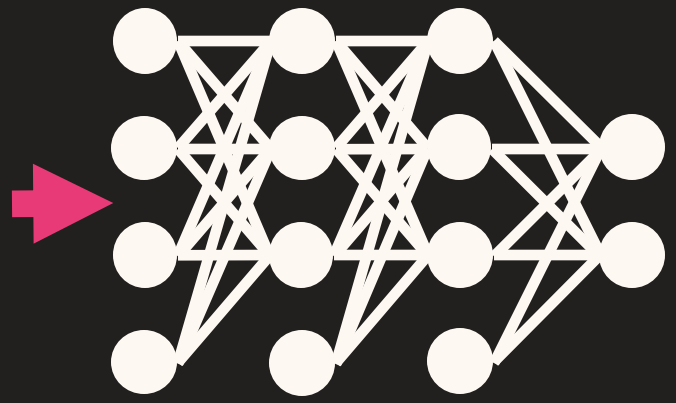
$$\begin{bmatrix} 0.52 & \cdots & 0.26 \\ \vdots & \ddots & \vdots \\ 0.06 & \cdots & 0.16 \end{bmatrix}$$




0.52 ⋯ 0.26
⋮ ⋱ ⋮
0.06 ⋯ 0.16
0.3
0.7

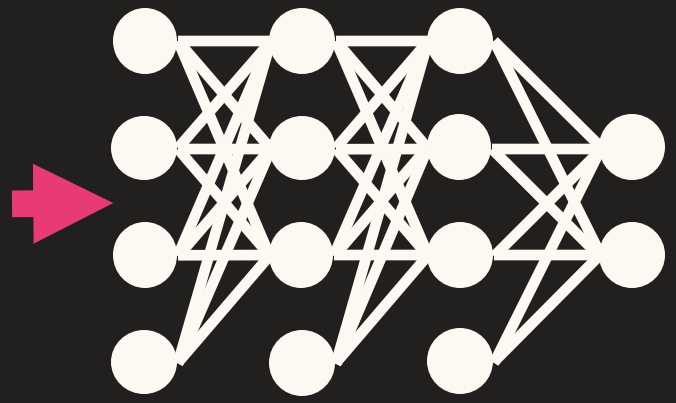

$\begin{bmatrix} 0.52 & \cdots & 0.26 \\ \vdots & \ddots & \vdots \\ 0.06 & \cdots & 0.16 \end{bmatrix}$
$\begin{bmatrix} 0.3 \\ 0.7 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$$\begin{bmatrix} 0.52 & \cdots & 0.26 \\ \vdots & \ddots & \vdots \\ 0.06 & \cdots & 0.16 \end{bmatrix}$$



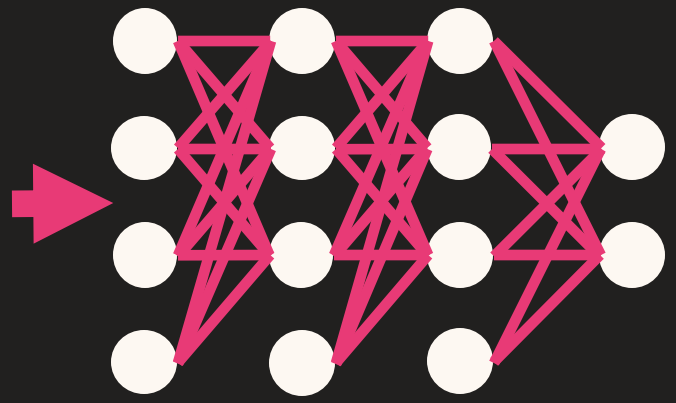
$$L(p, y) = \begin{bmatrix} 0.3 \\ 0.7 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0.52 & \cdots & 0.26 \\ \vdots & \ddots & \vdots \\ 0.06 & \cdots & 0.16 \end{bmatrix}$$



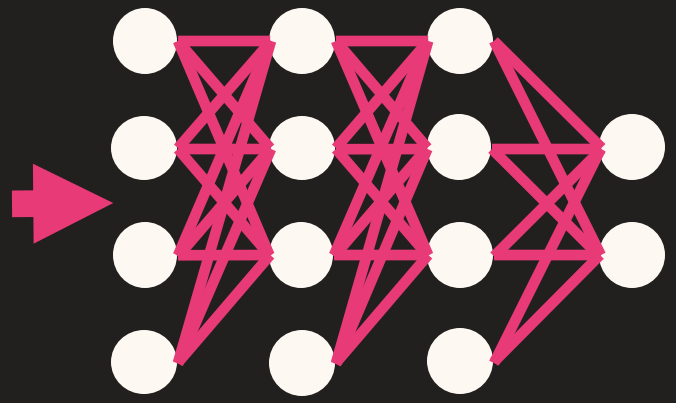
$$L(p, y) = p - y$$

$$\begin{bmatrix} 0.52 & \cdots & 0.26 \\ \vdots & \ddots & \vdots \\ 0.06 & \cdots & 0.16 \end{bmatrix}$$



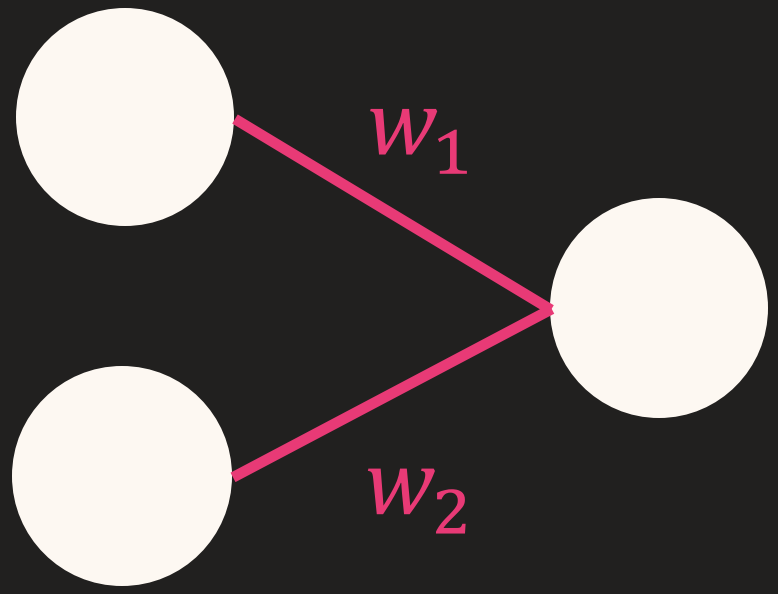
$$L(p, y) = p - y$$

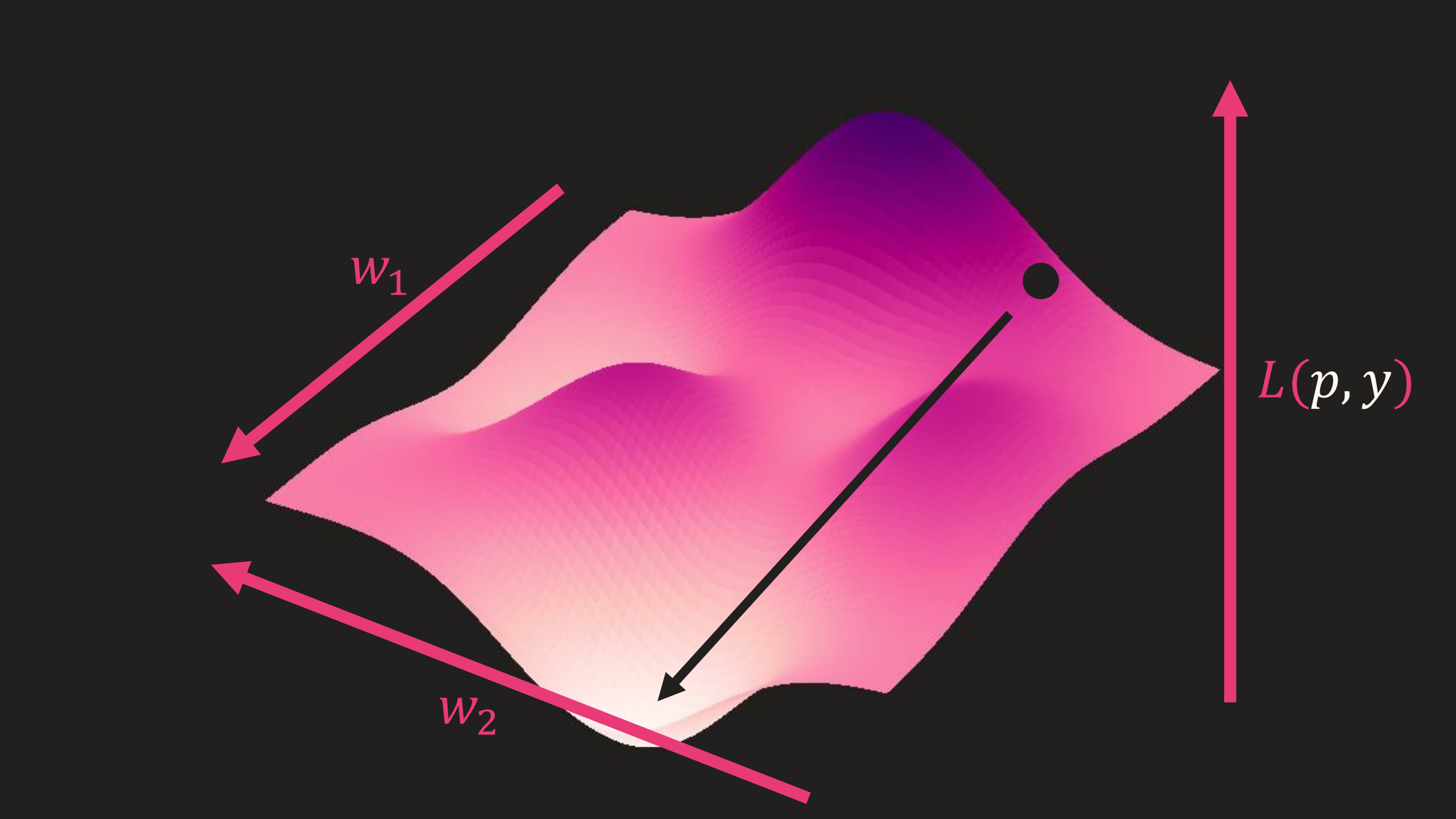
$$\begin{bmatrix} 0.52 & \cdots & 0.26 \\ \vdots & \ddots & \vdots \\ 0.06 & \cdots & 0.16 \end{bmatrix}$$

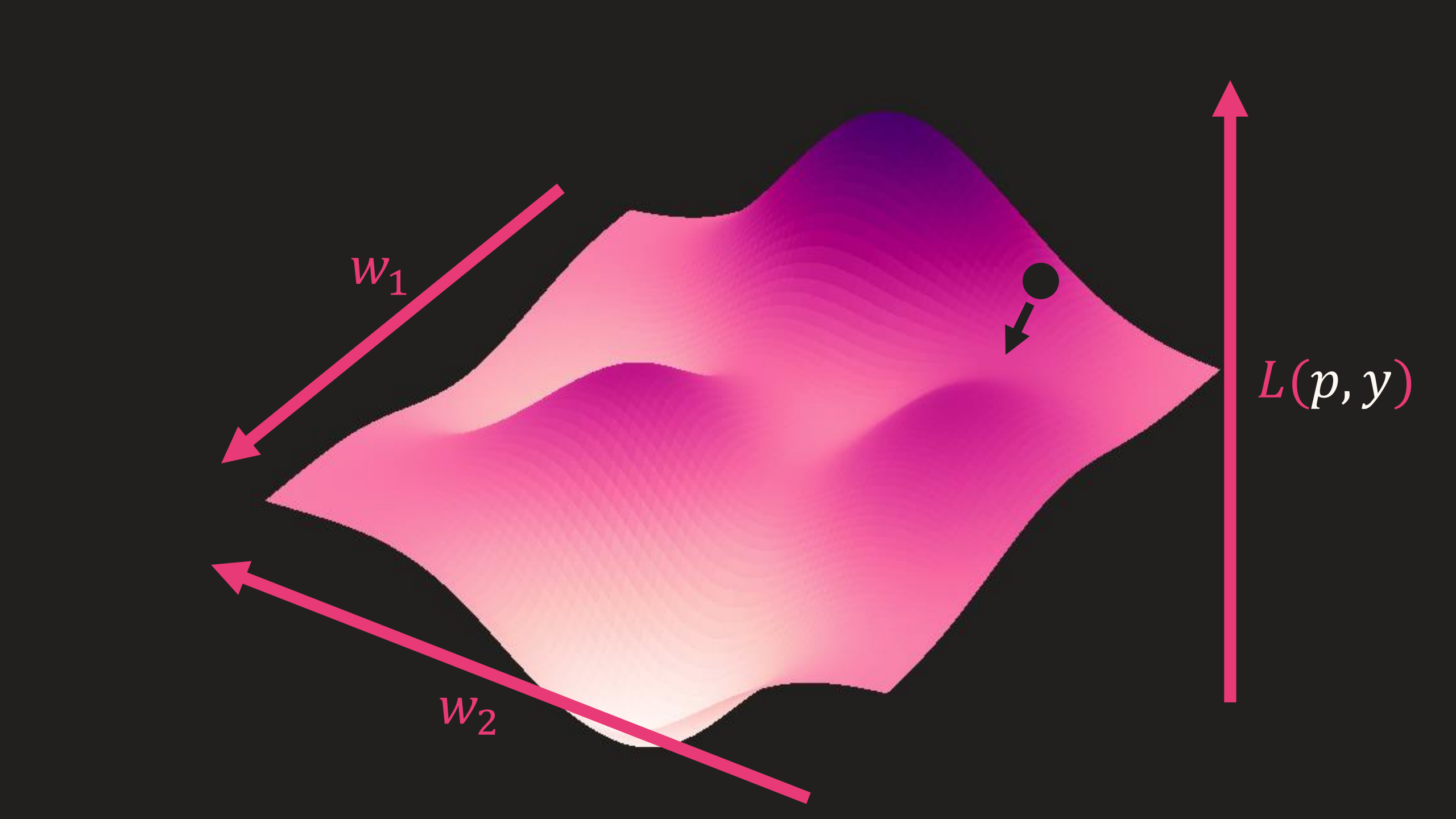


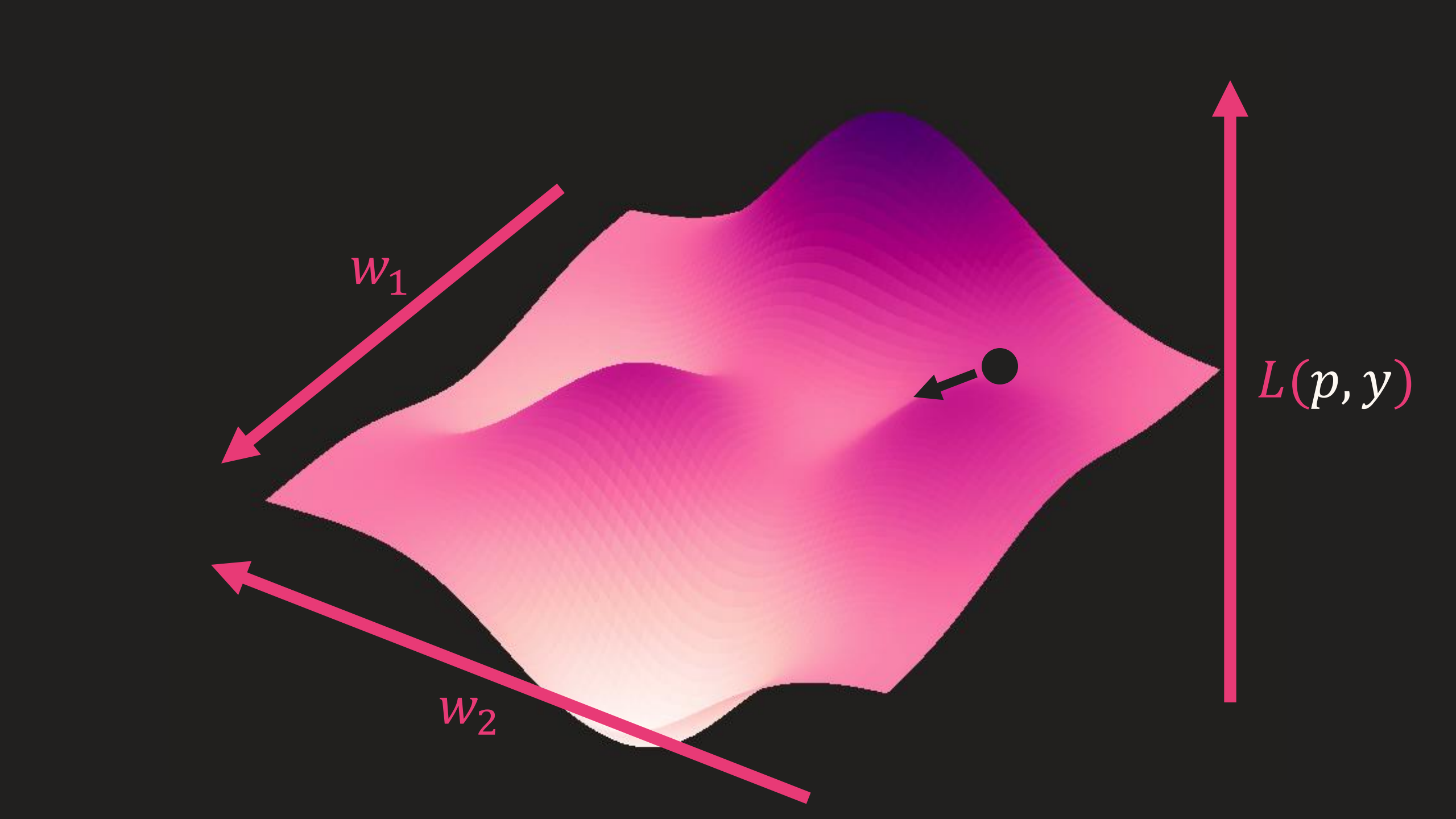
Parametry modelu

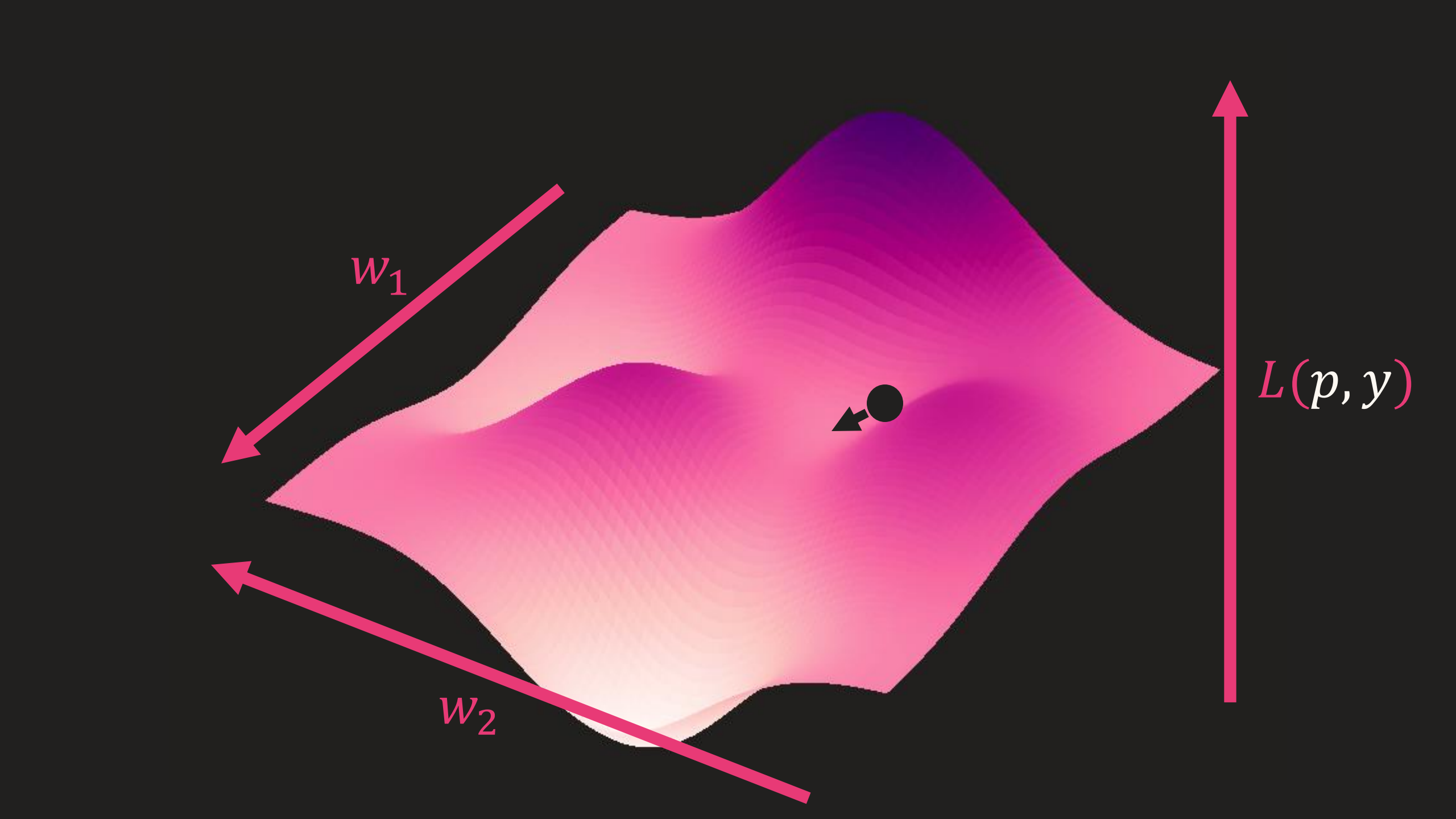
$$L(p, y) = p - y$$

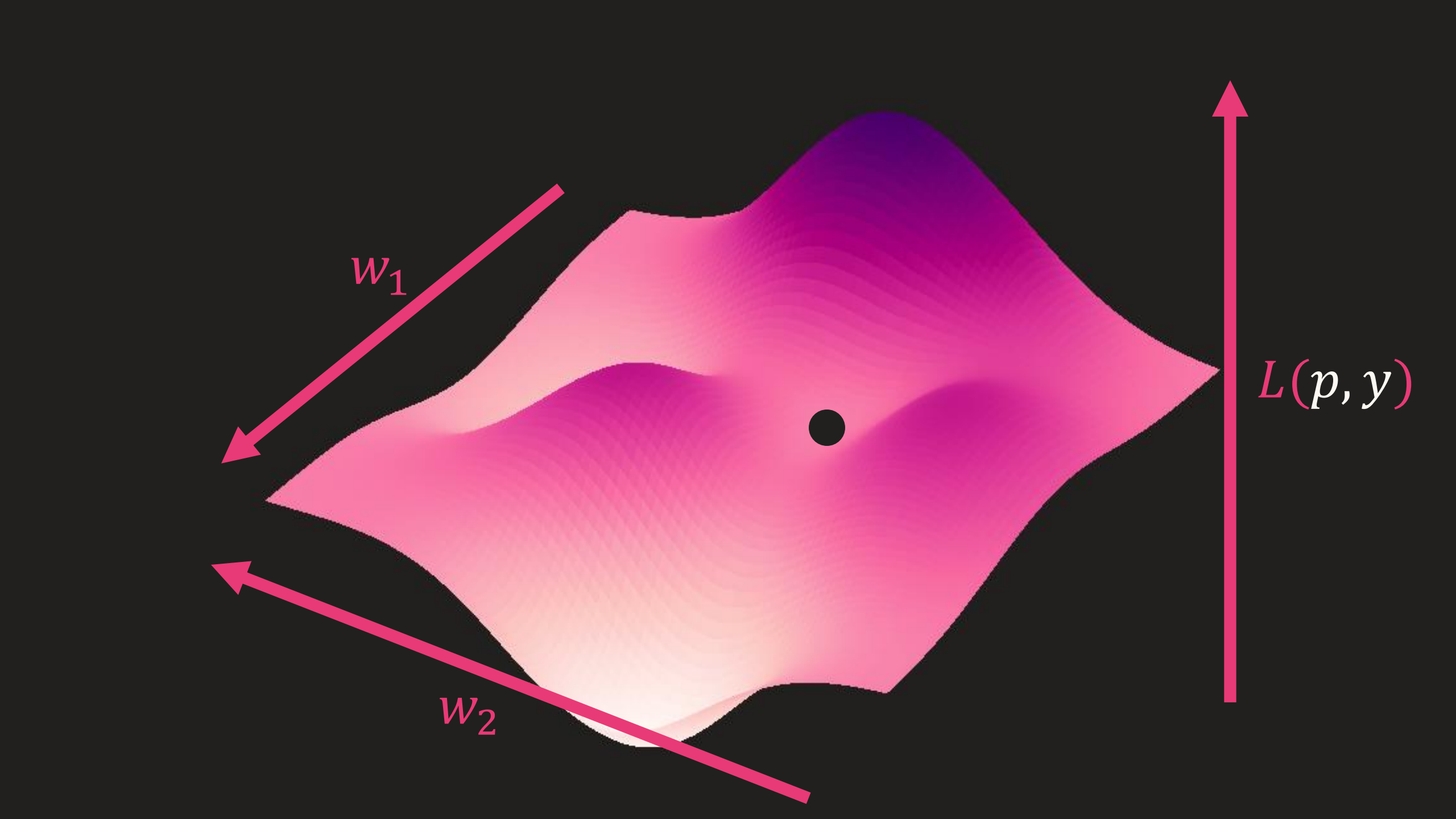

w_1
w_2

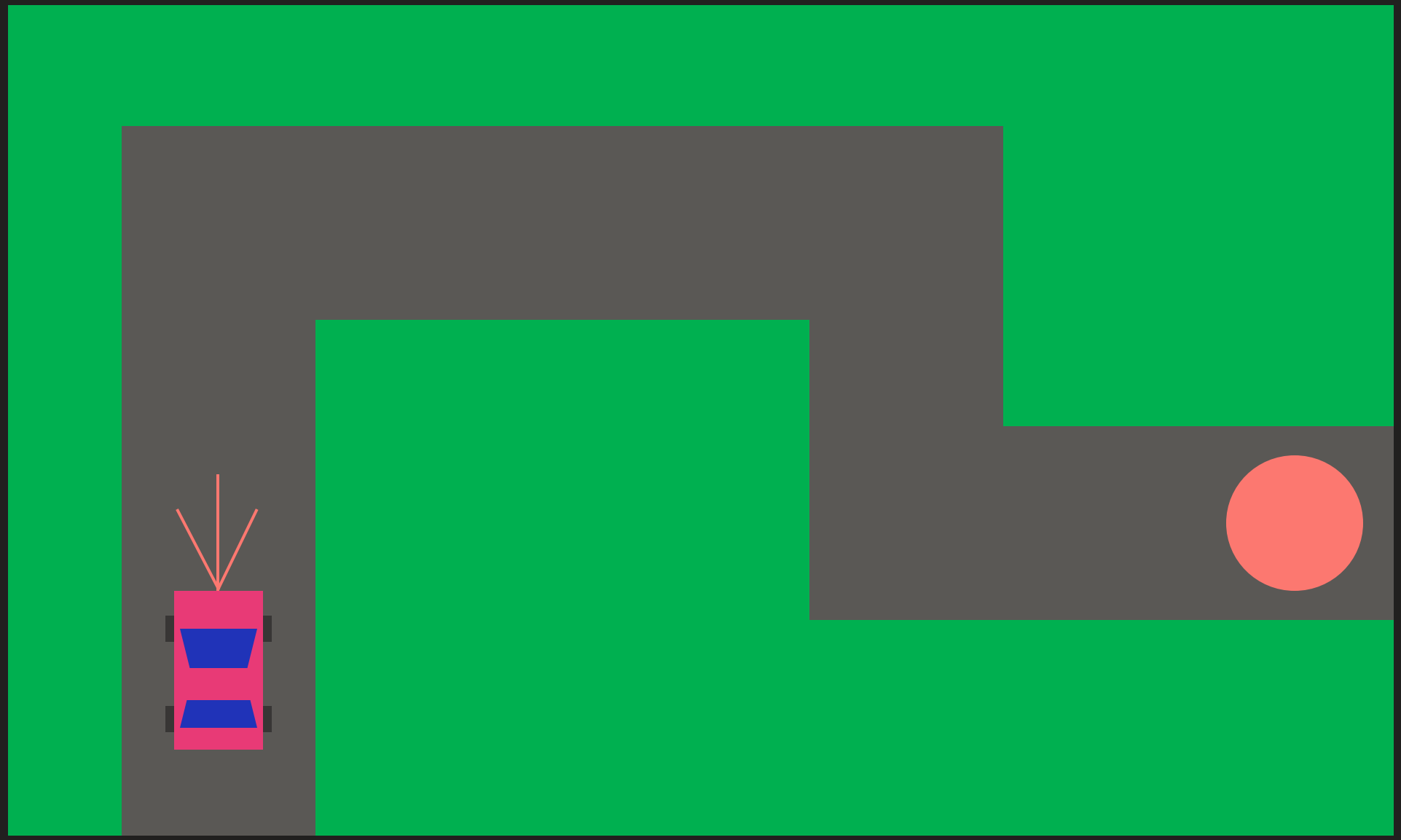

w_1
w_2
$L(p, y)$

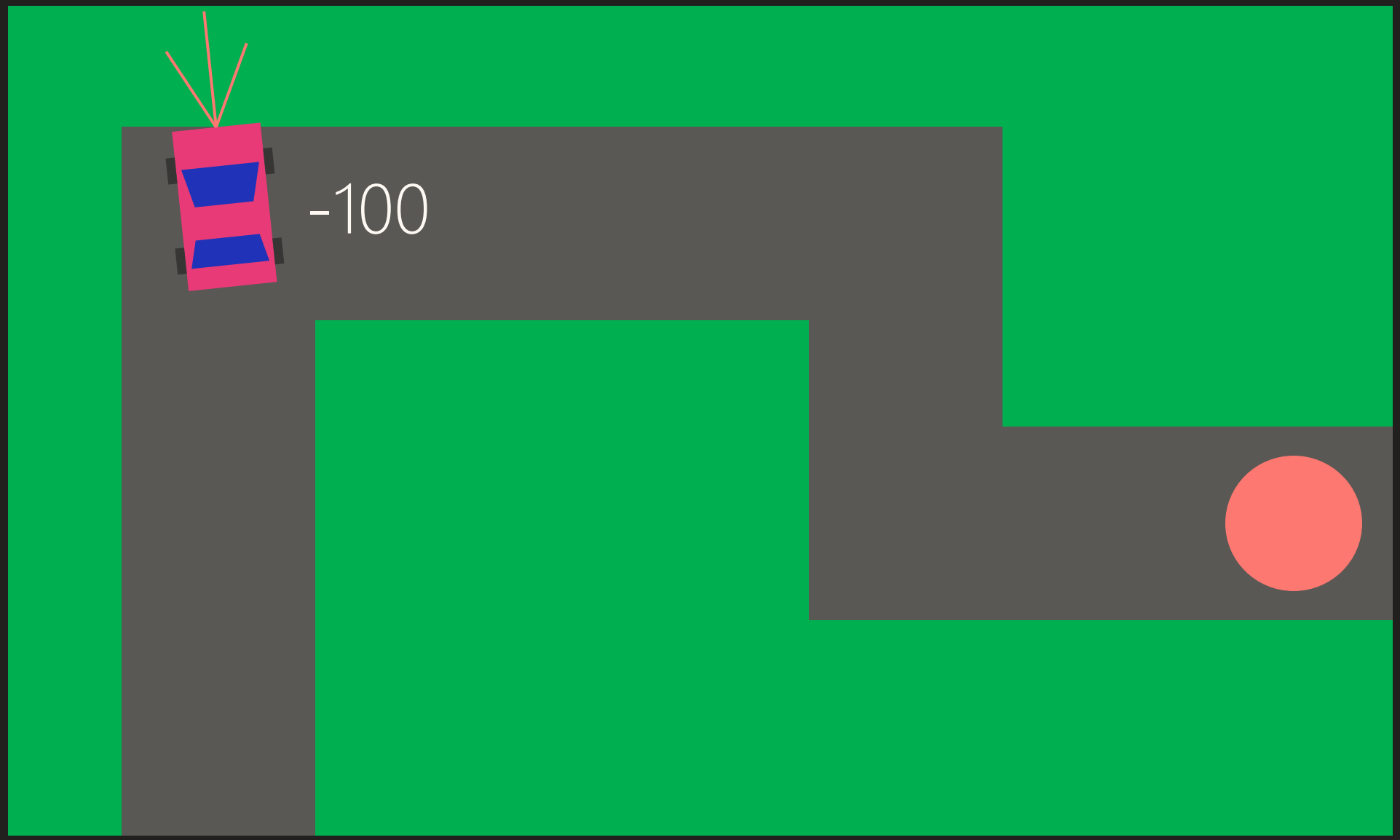

w_1
w_2
$L(p, y)$

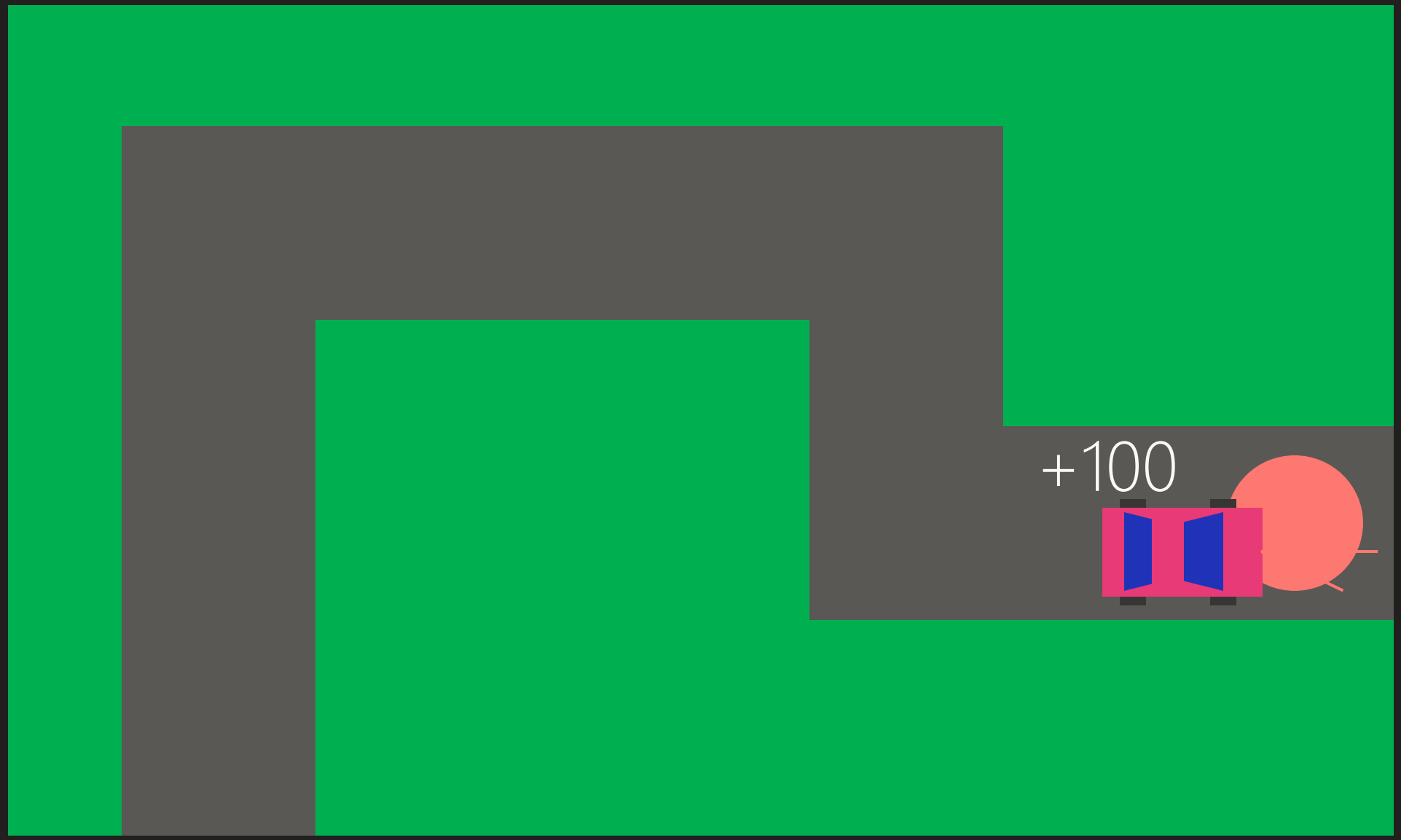

w_1
w_2
$L(p, y)$


w_1
w_2
$L(p, y)$

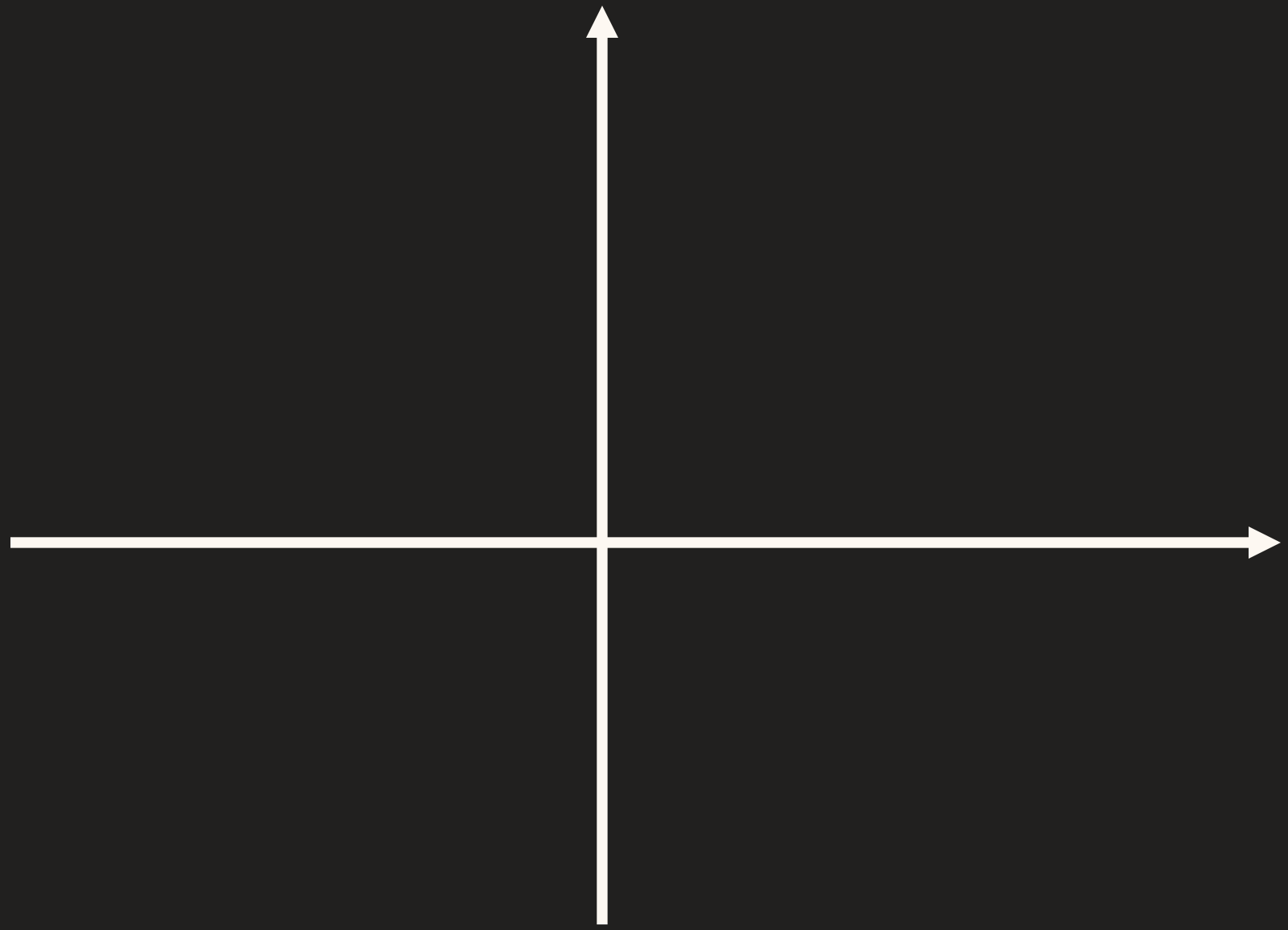

w_1
w_2
$L(p, y)$



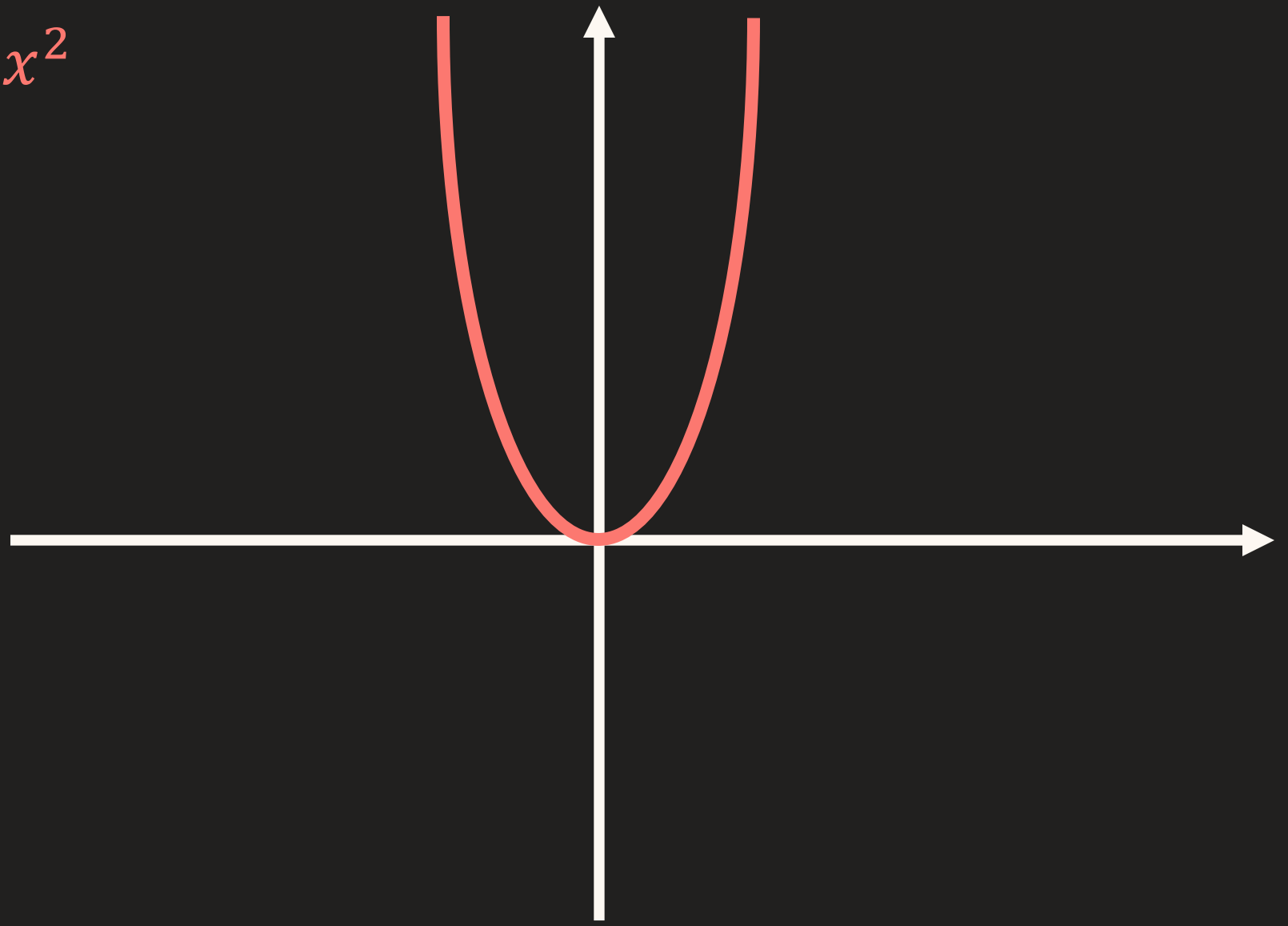

-100


+100

Generalizacja

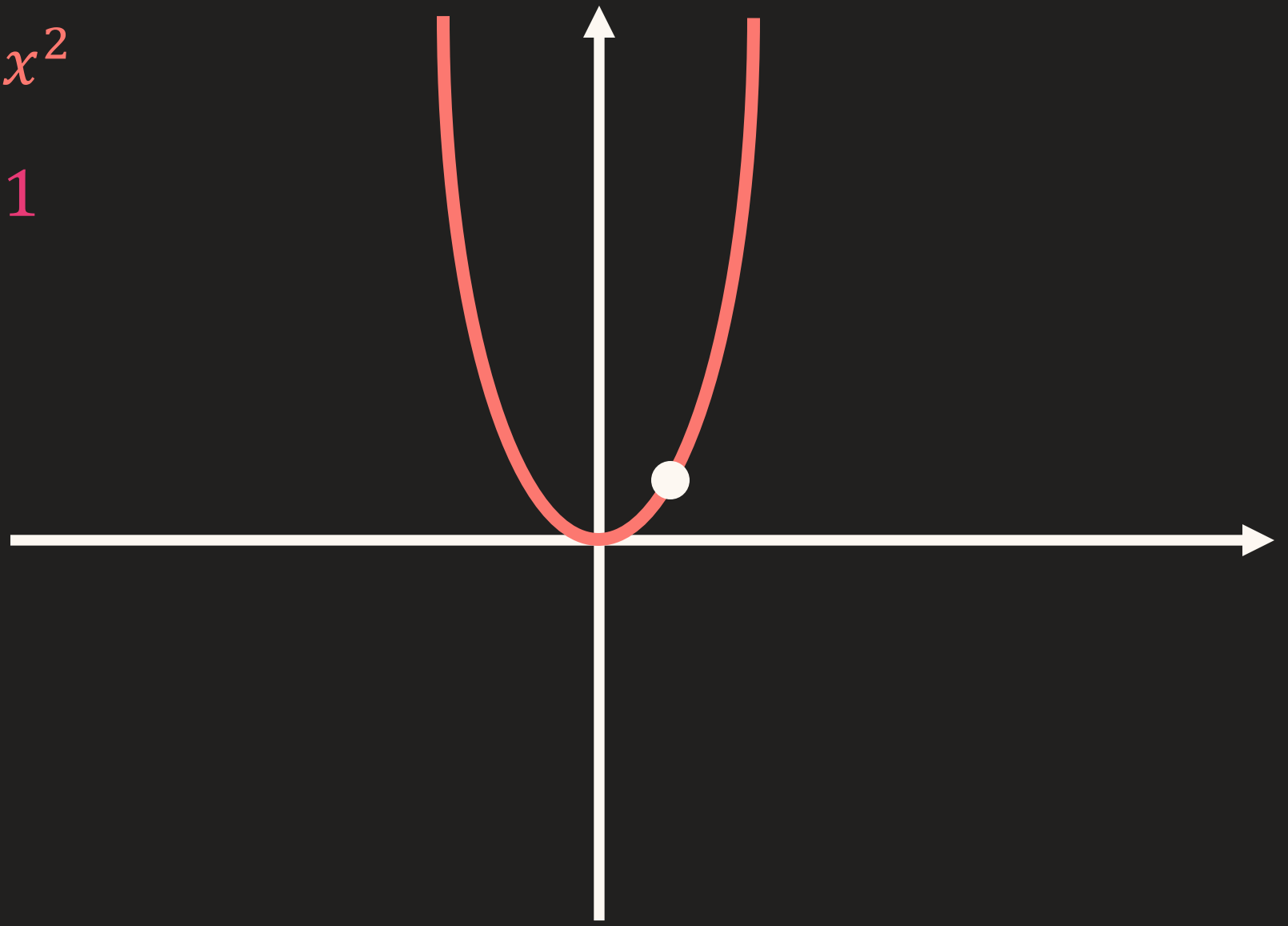


$f(x) = x^2$



$f(x) = x^2$

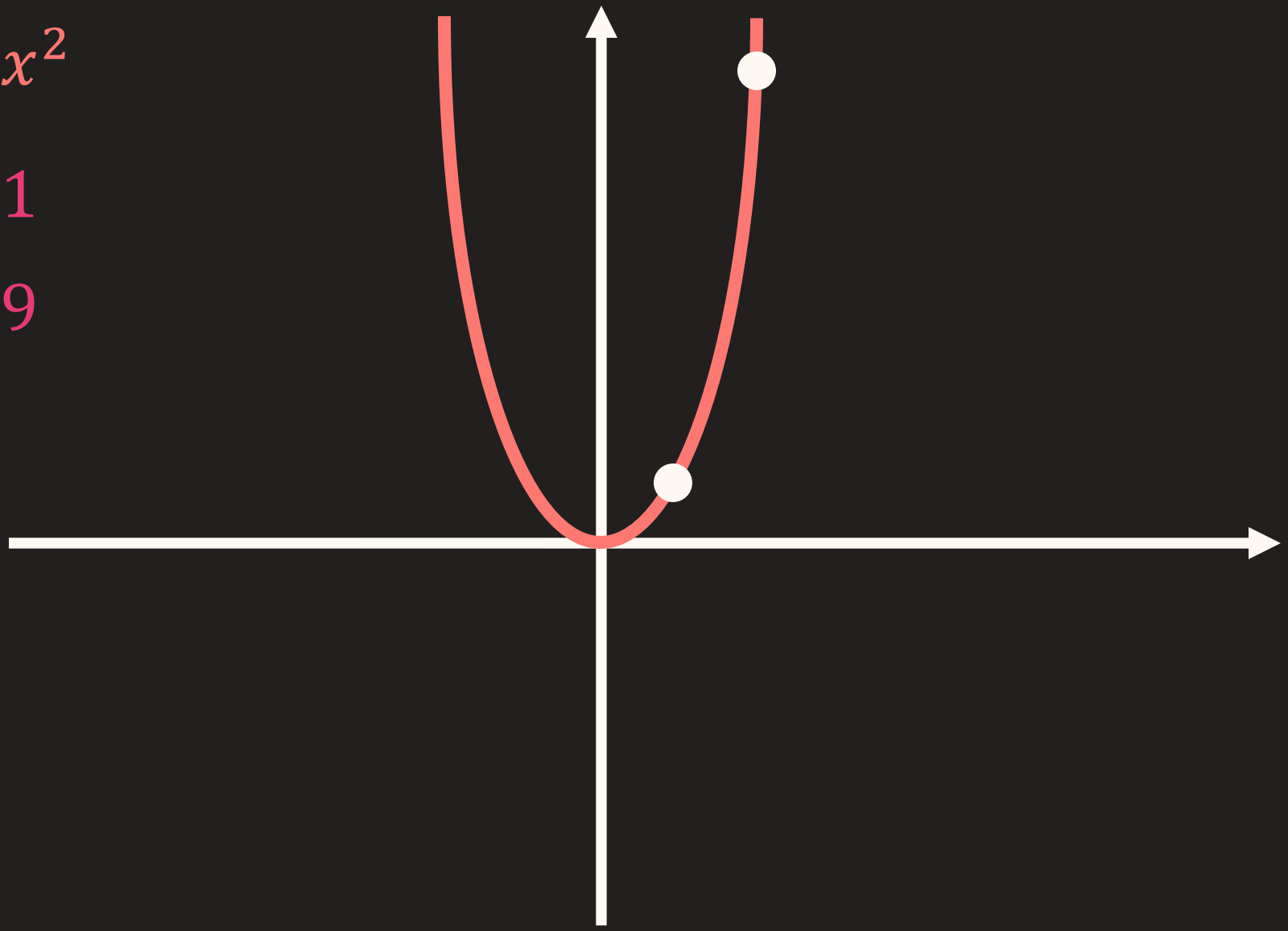
$f(1) = 1$



$f(x) = x^2$

$f(1) = 1$

$f(3) = 9$

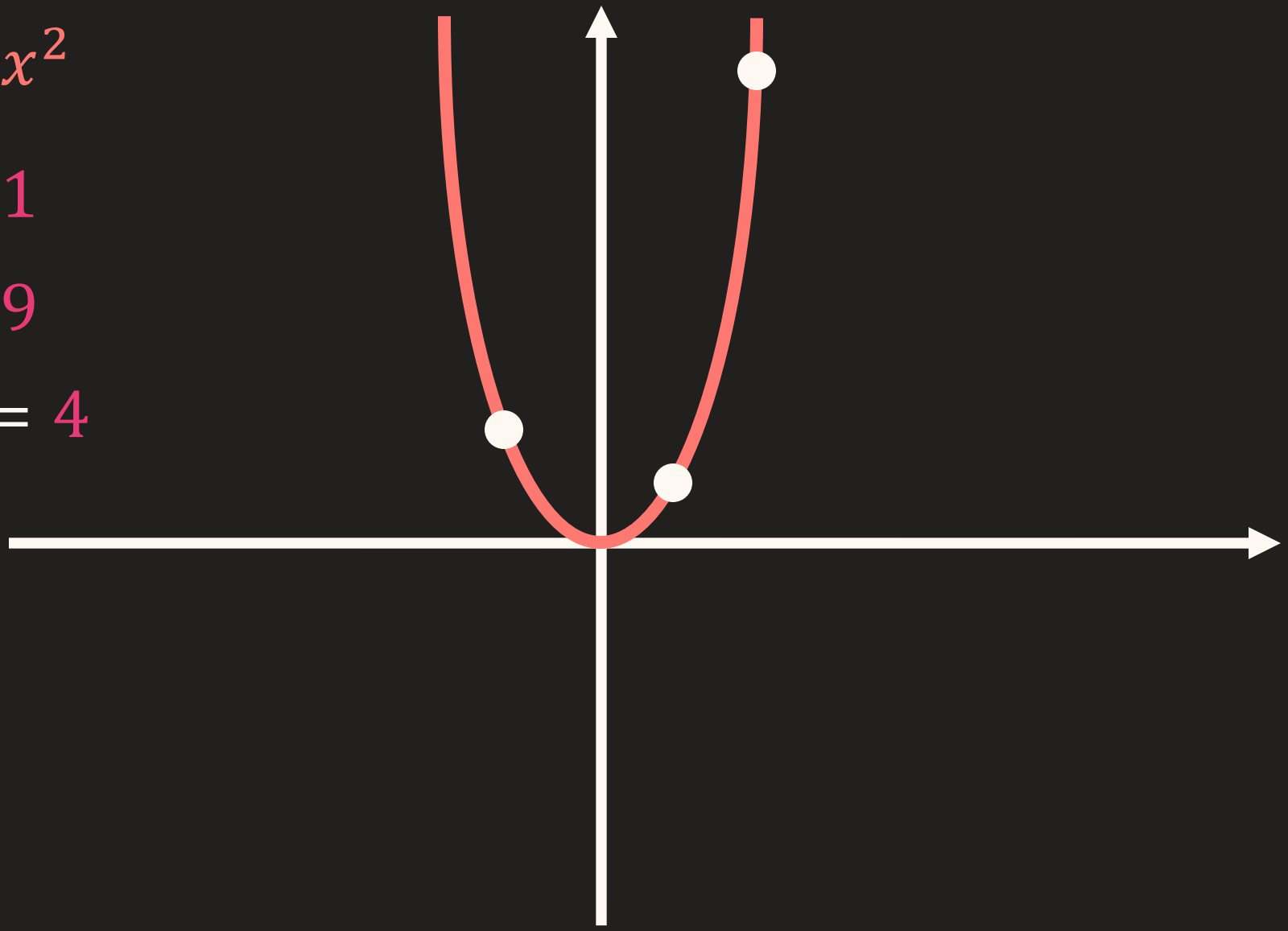


$f(x) = x^2$

$f(1) = 1$

$f(3) = 9$

$f(-2) = 4$

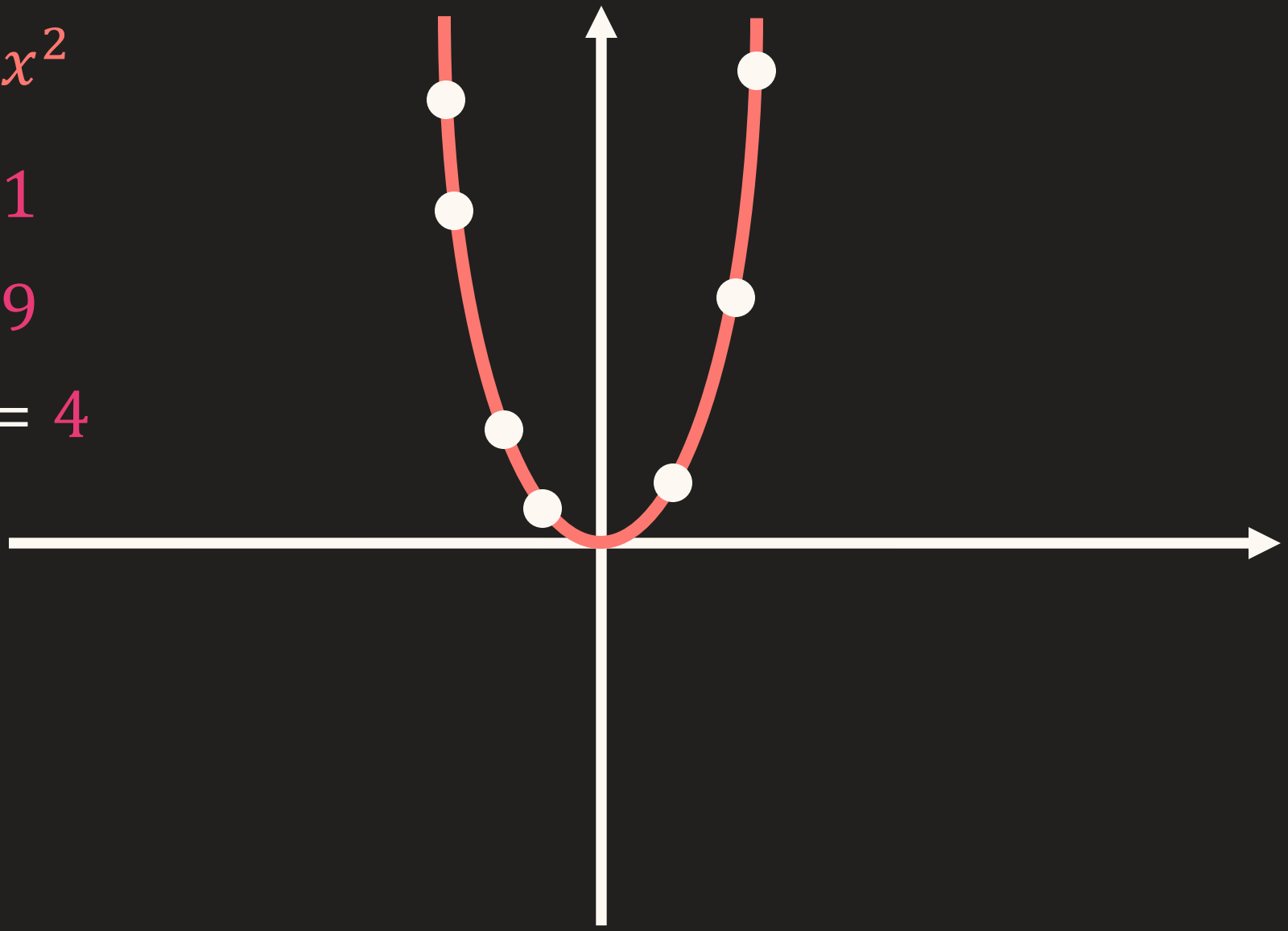


$f(x) = x^2$

$f(1) = 1$

$f(3) = 9$

$f(-2) = 4$

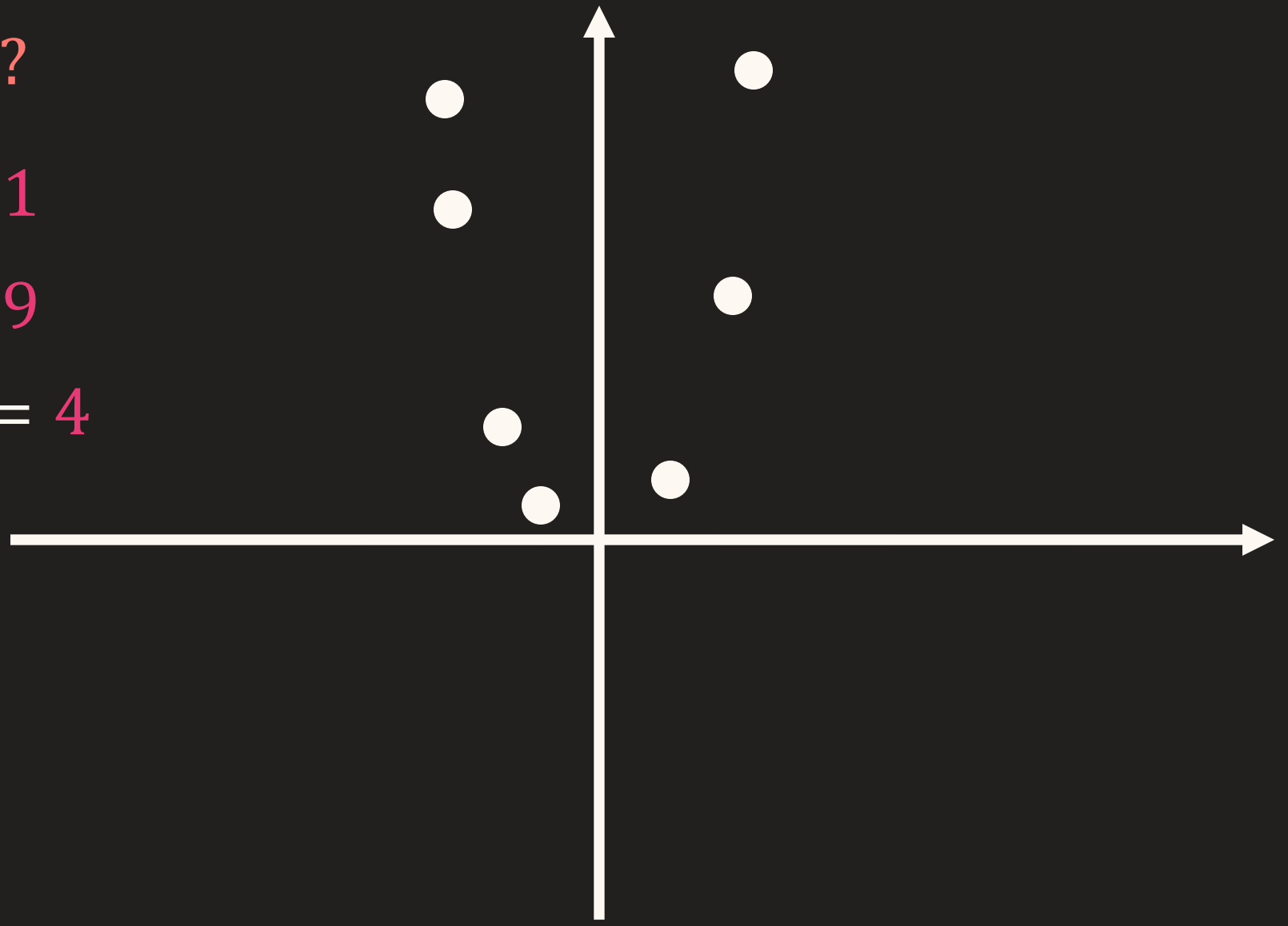


$f(x) = ?$

$f(1) = 1$

$f(3) = 9$

$f(-2) = 4$

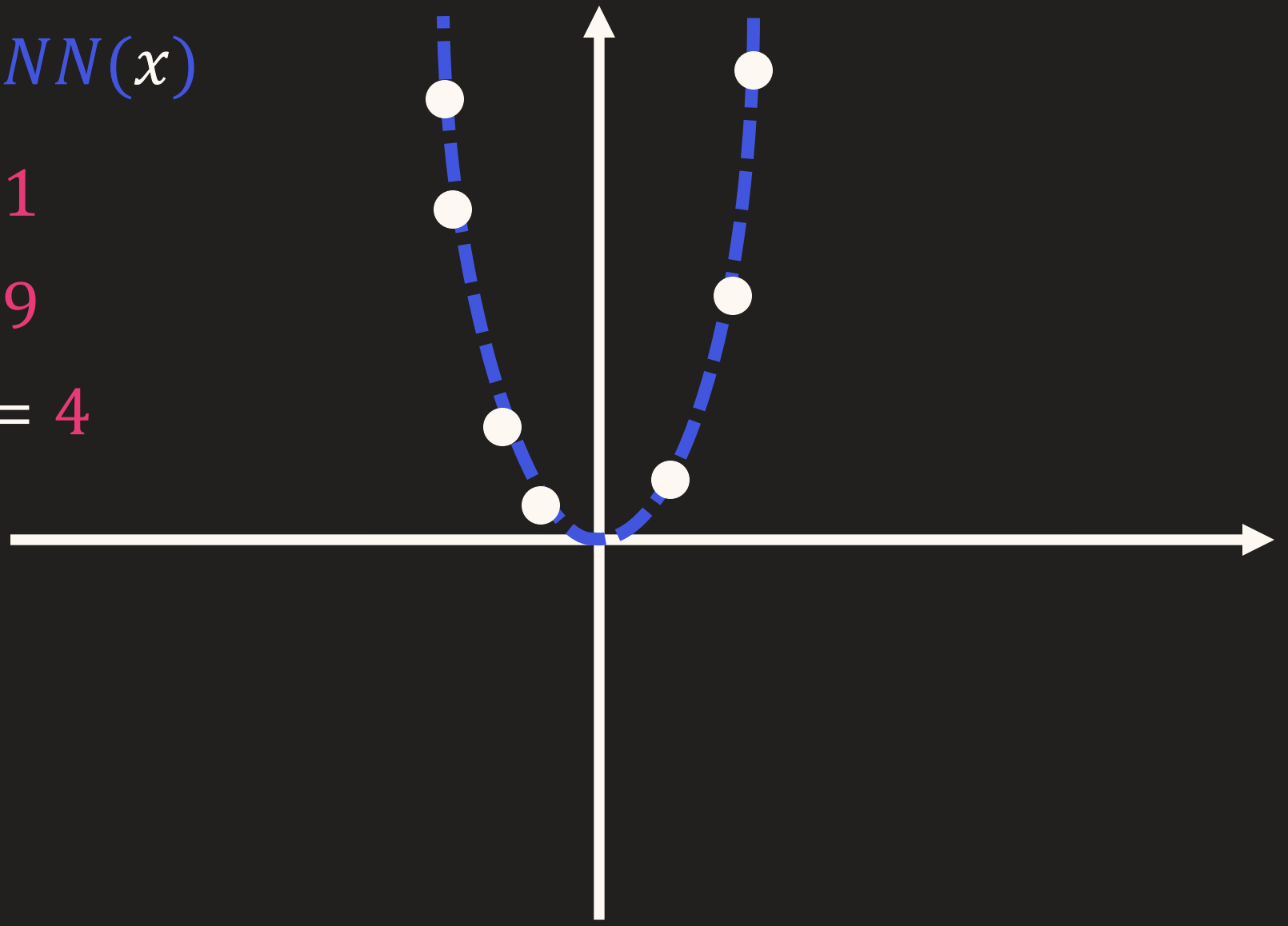


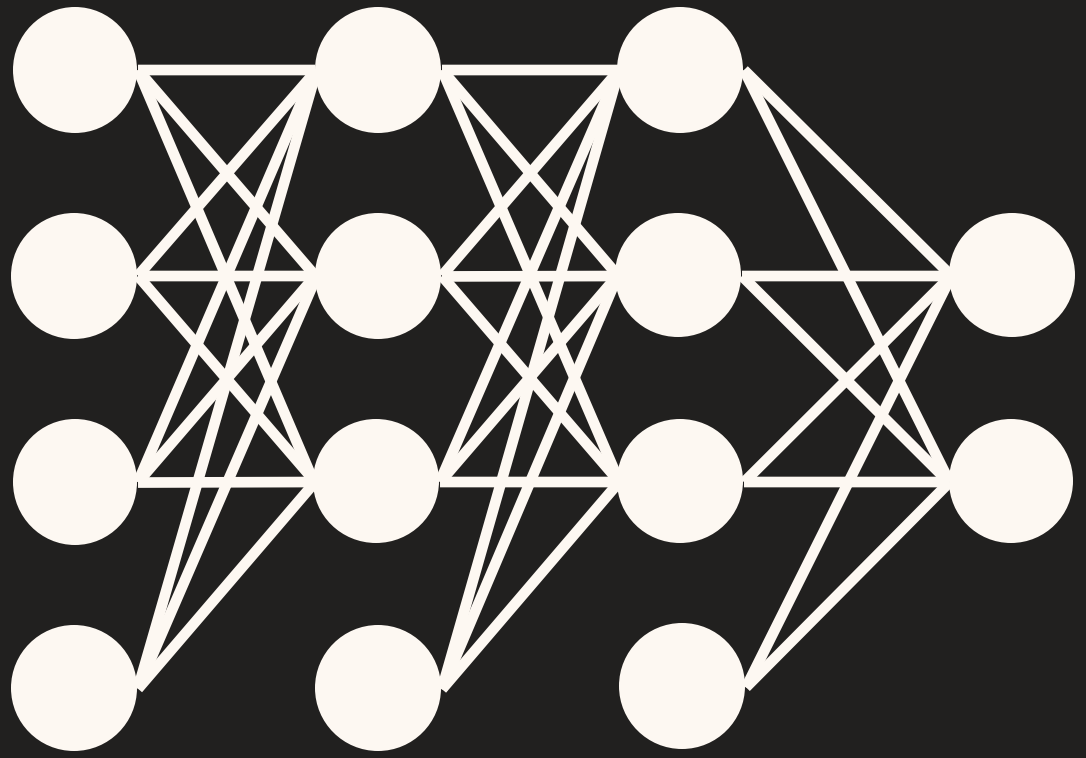
$f(x) \approx NN(x)$

$f(1) = 1$

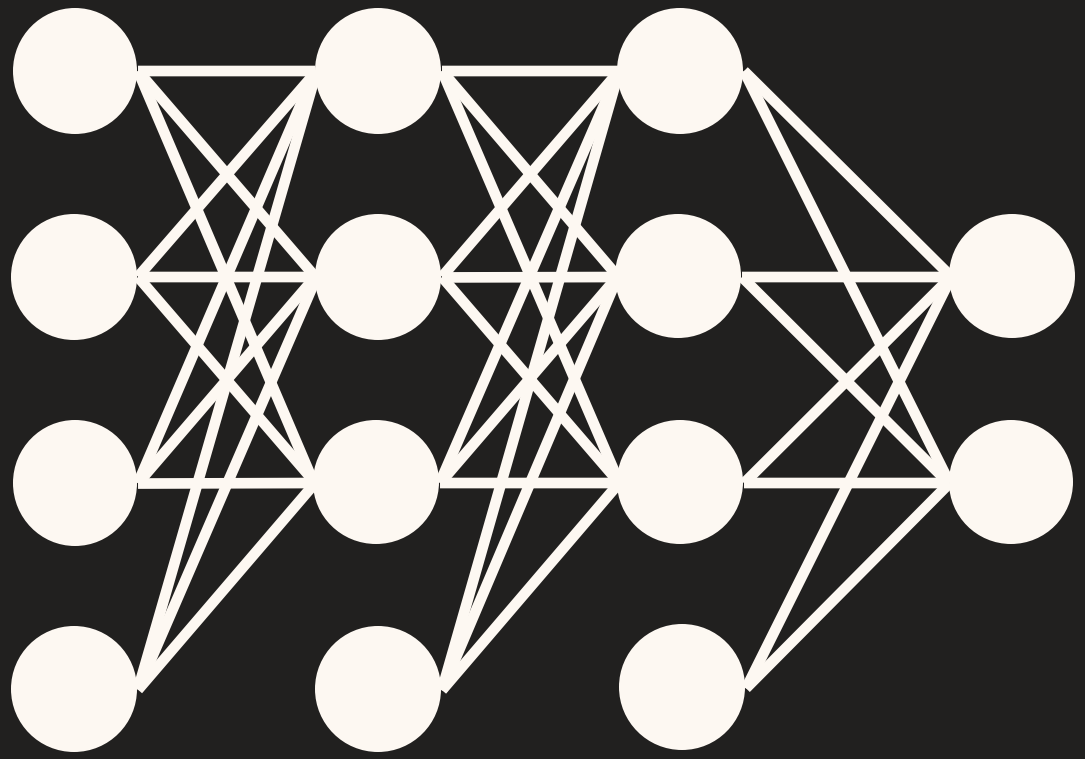
$f(3) = 9$

$f(-2) = 4$





MLP



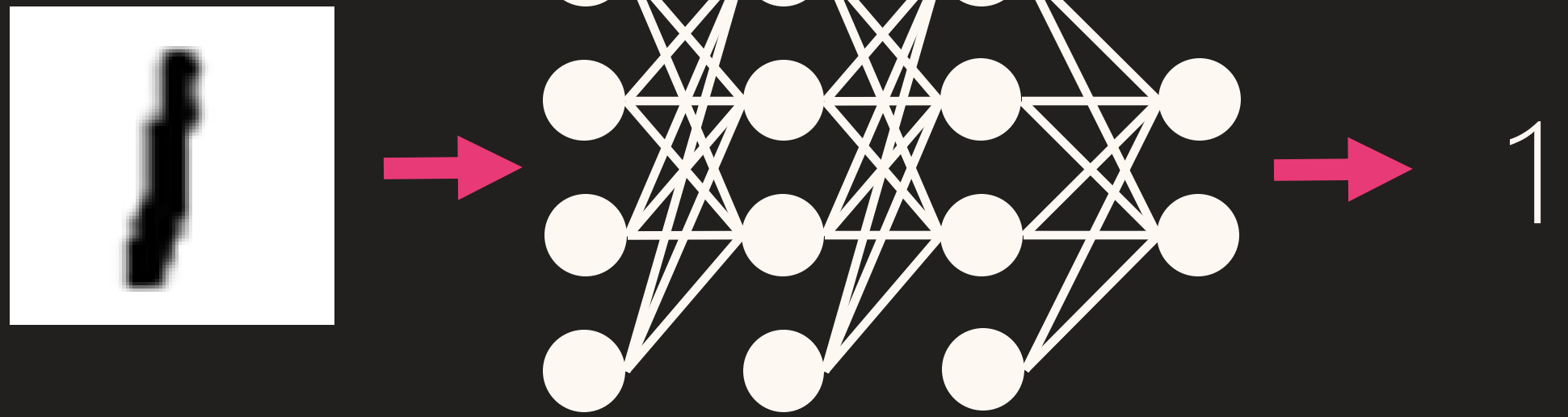
MLP

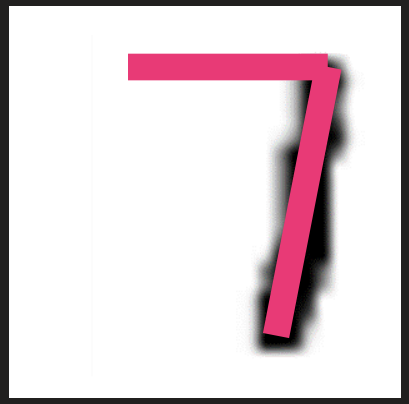
50s

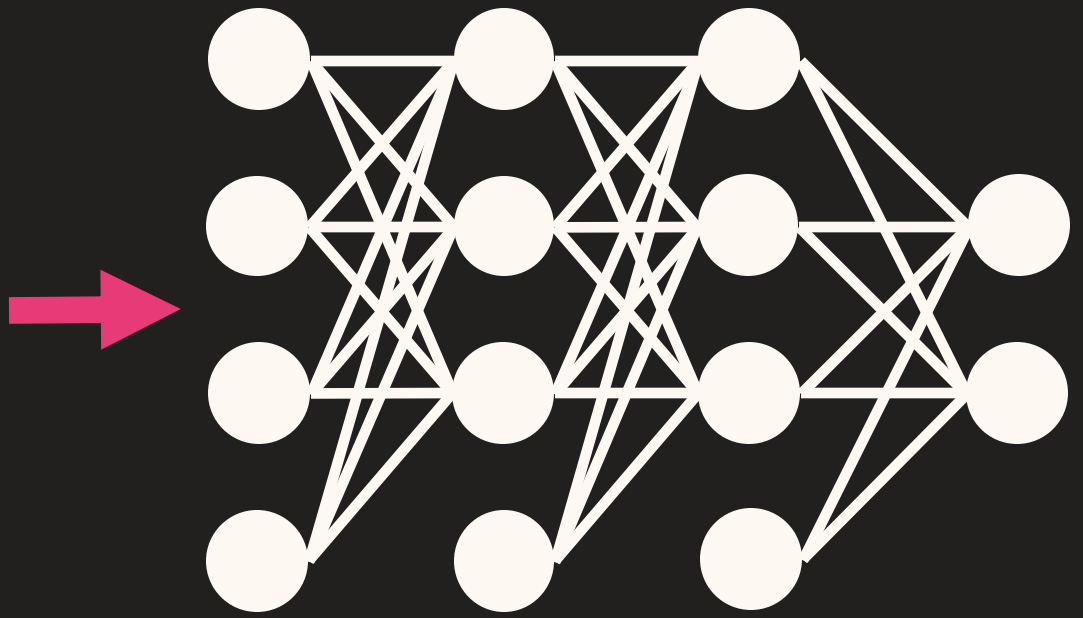
⋮

90s

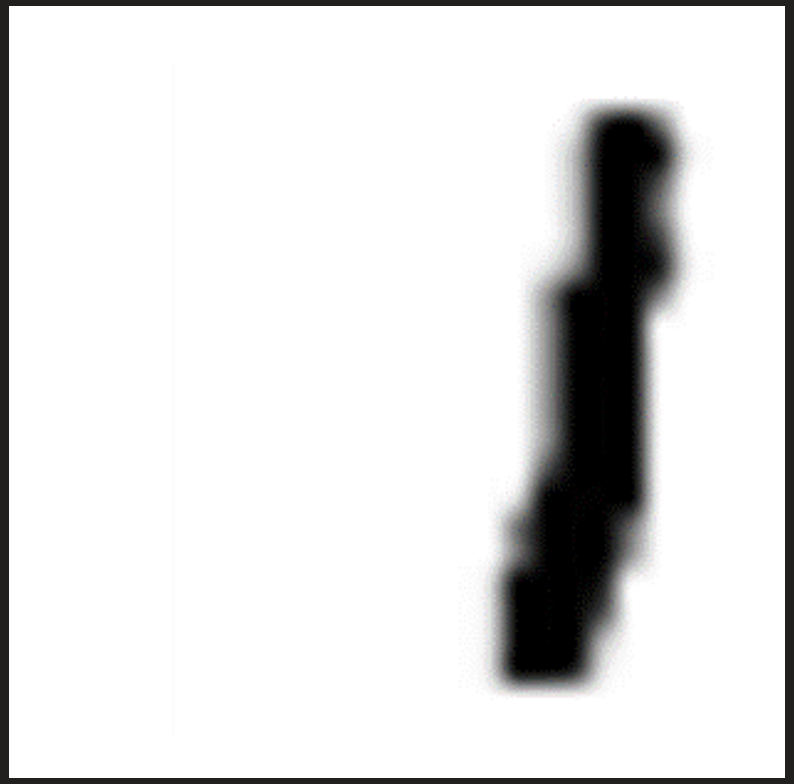
Problemy?

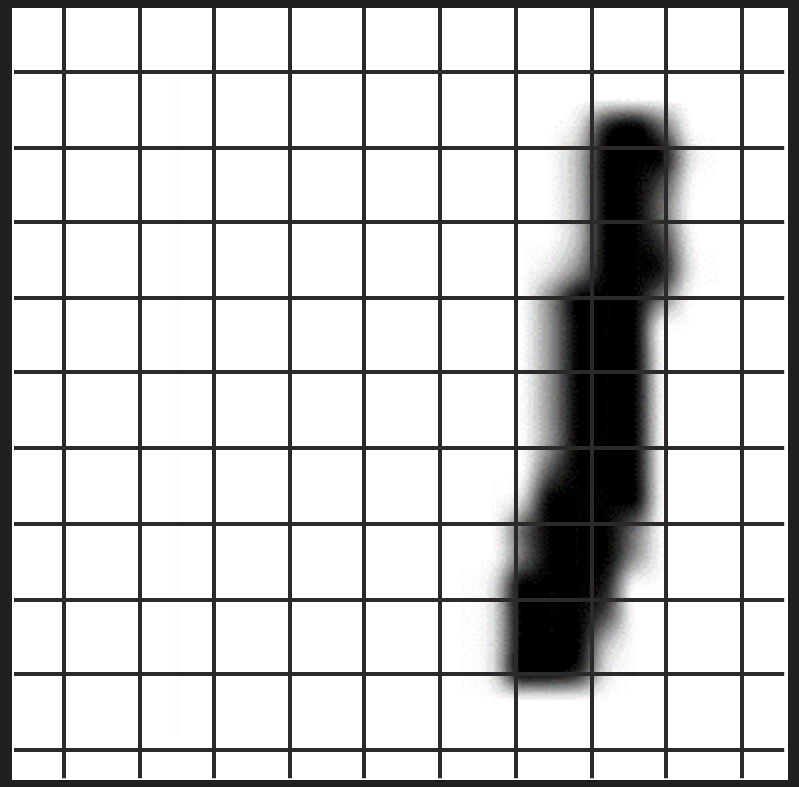

1

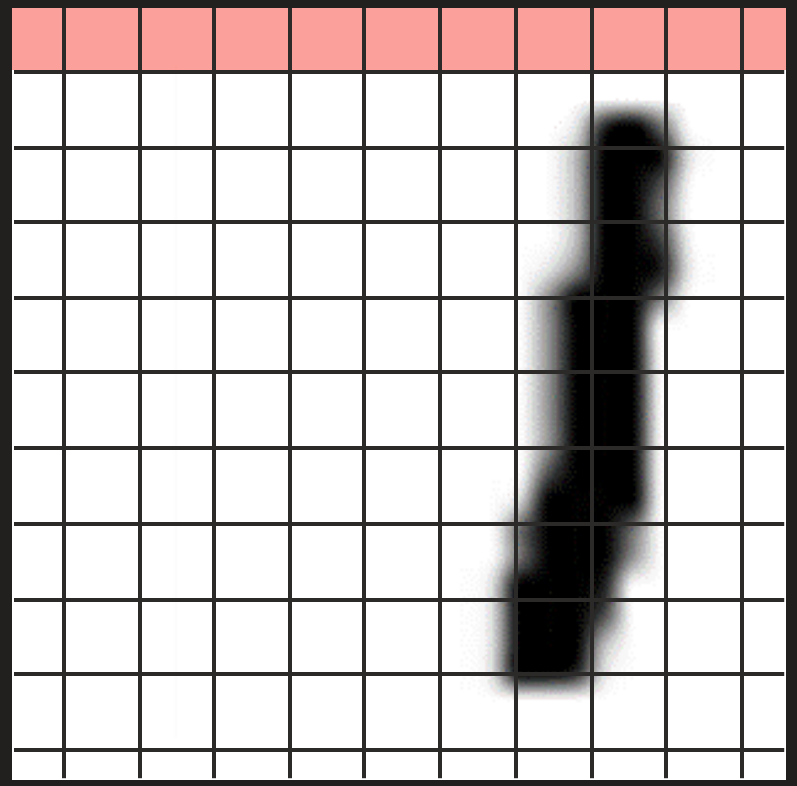


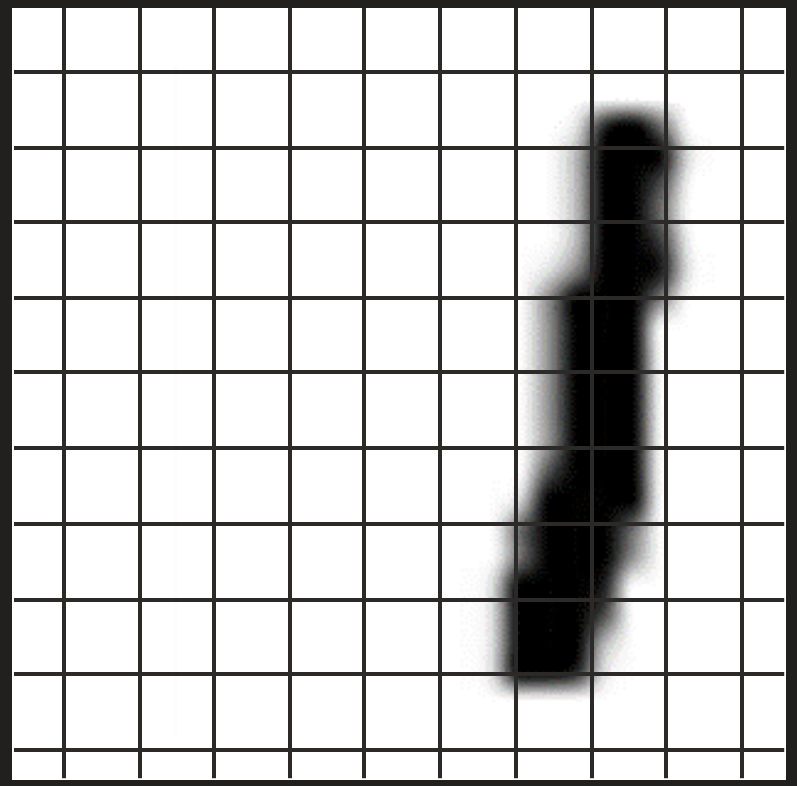


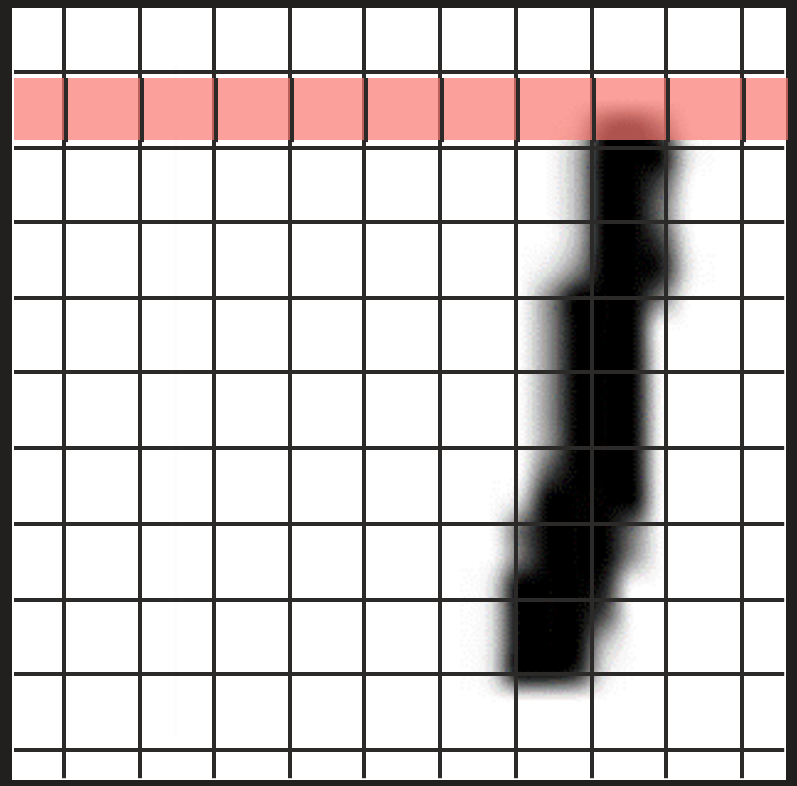
7

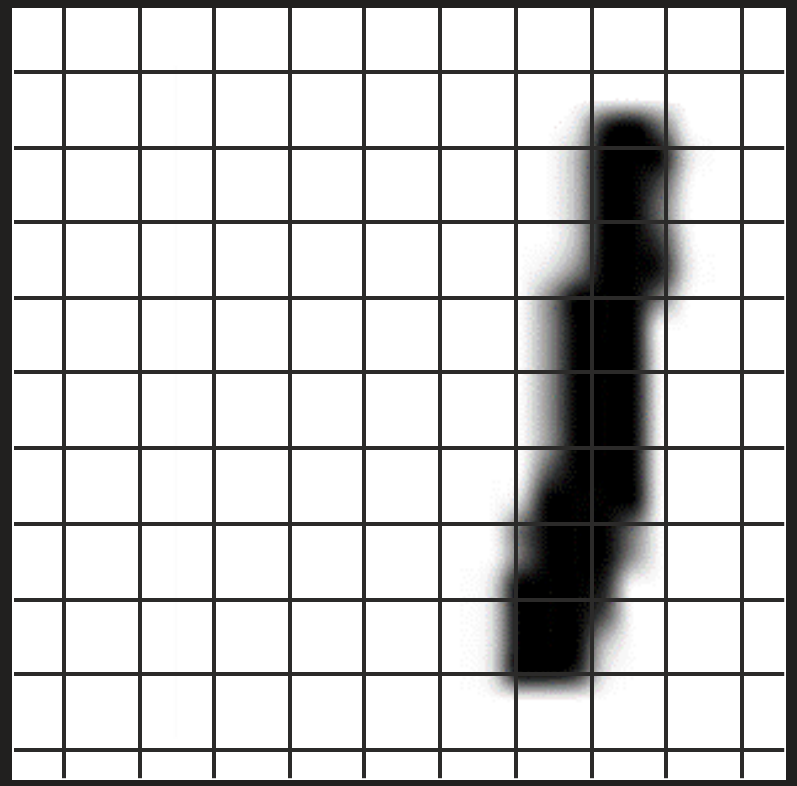


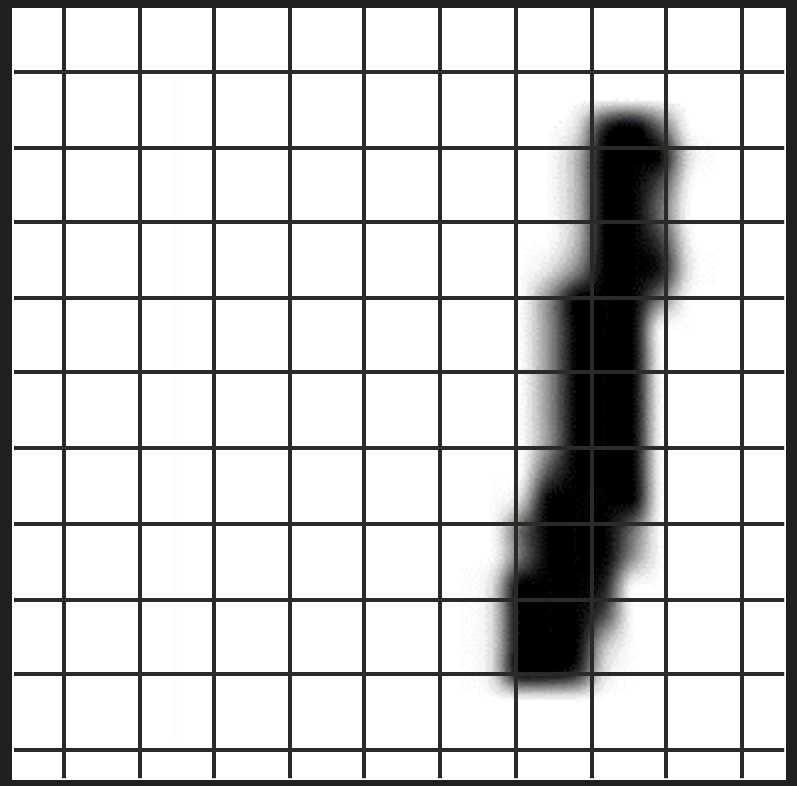




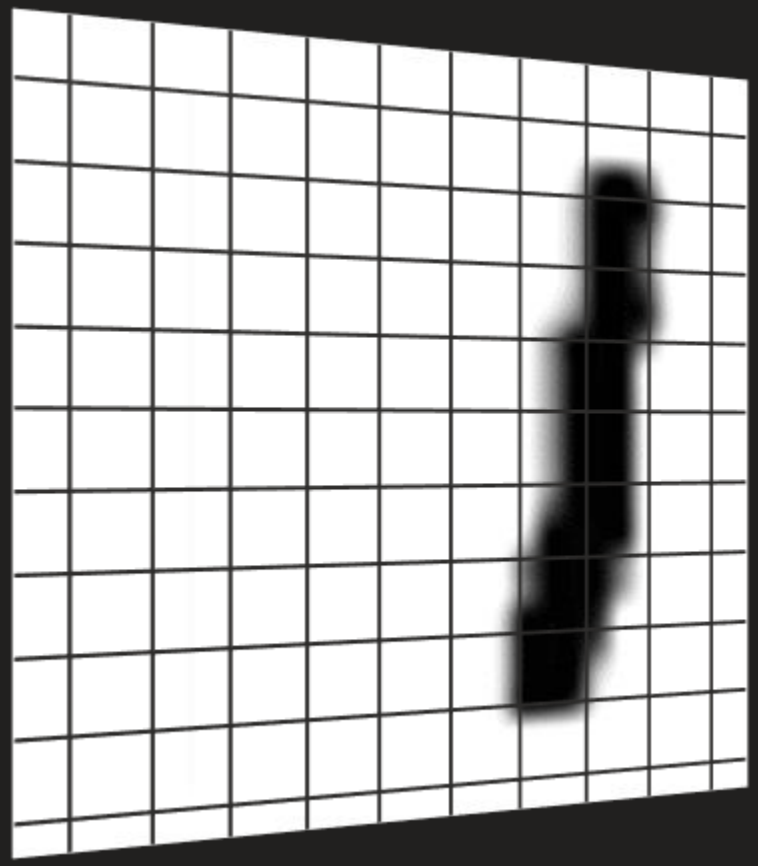


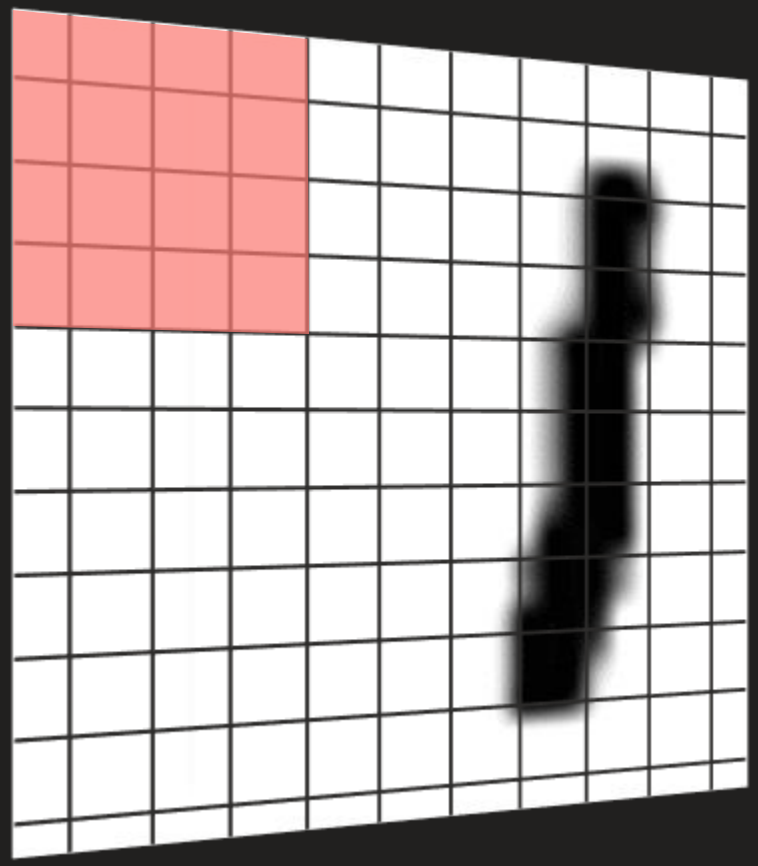


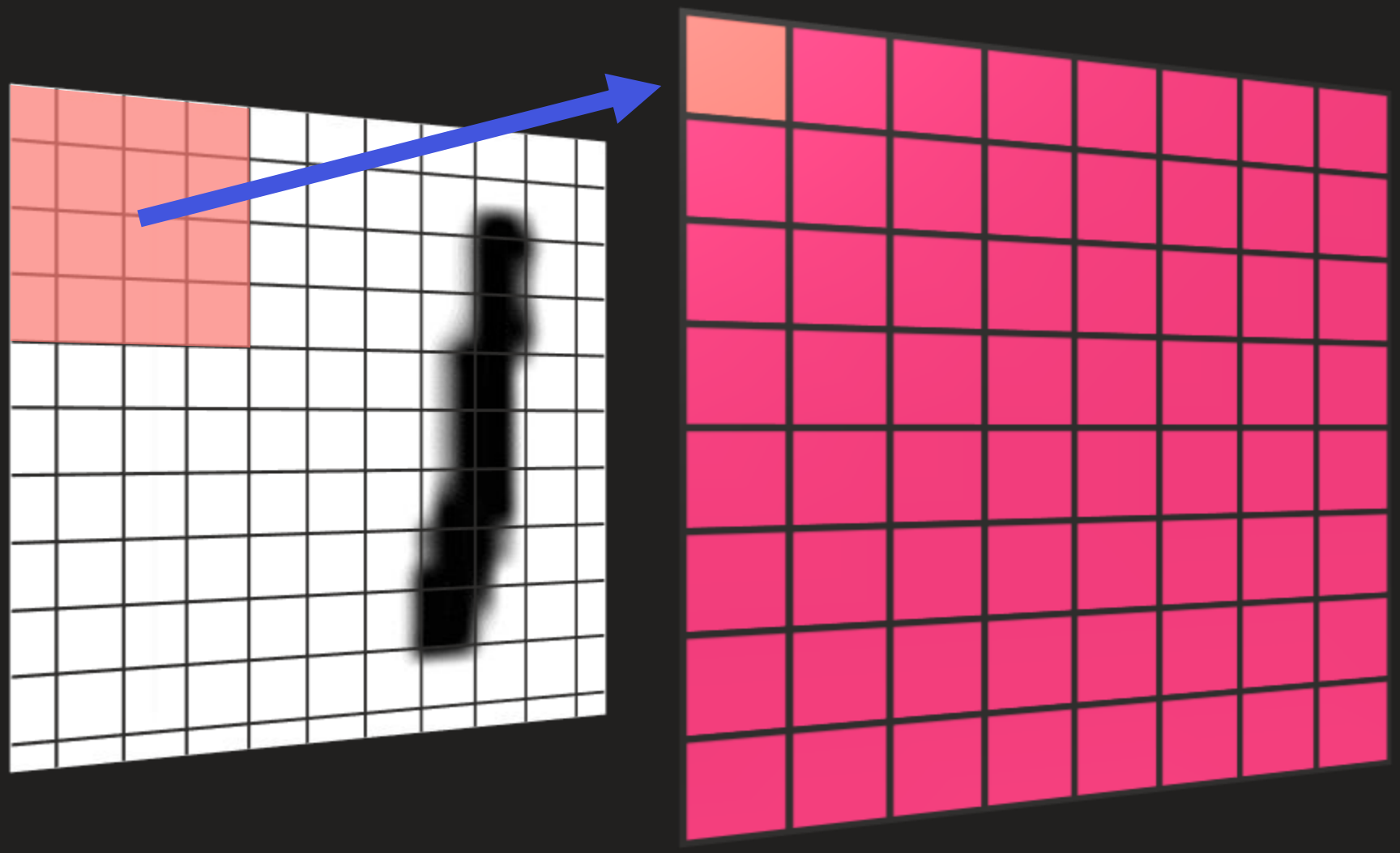


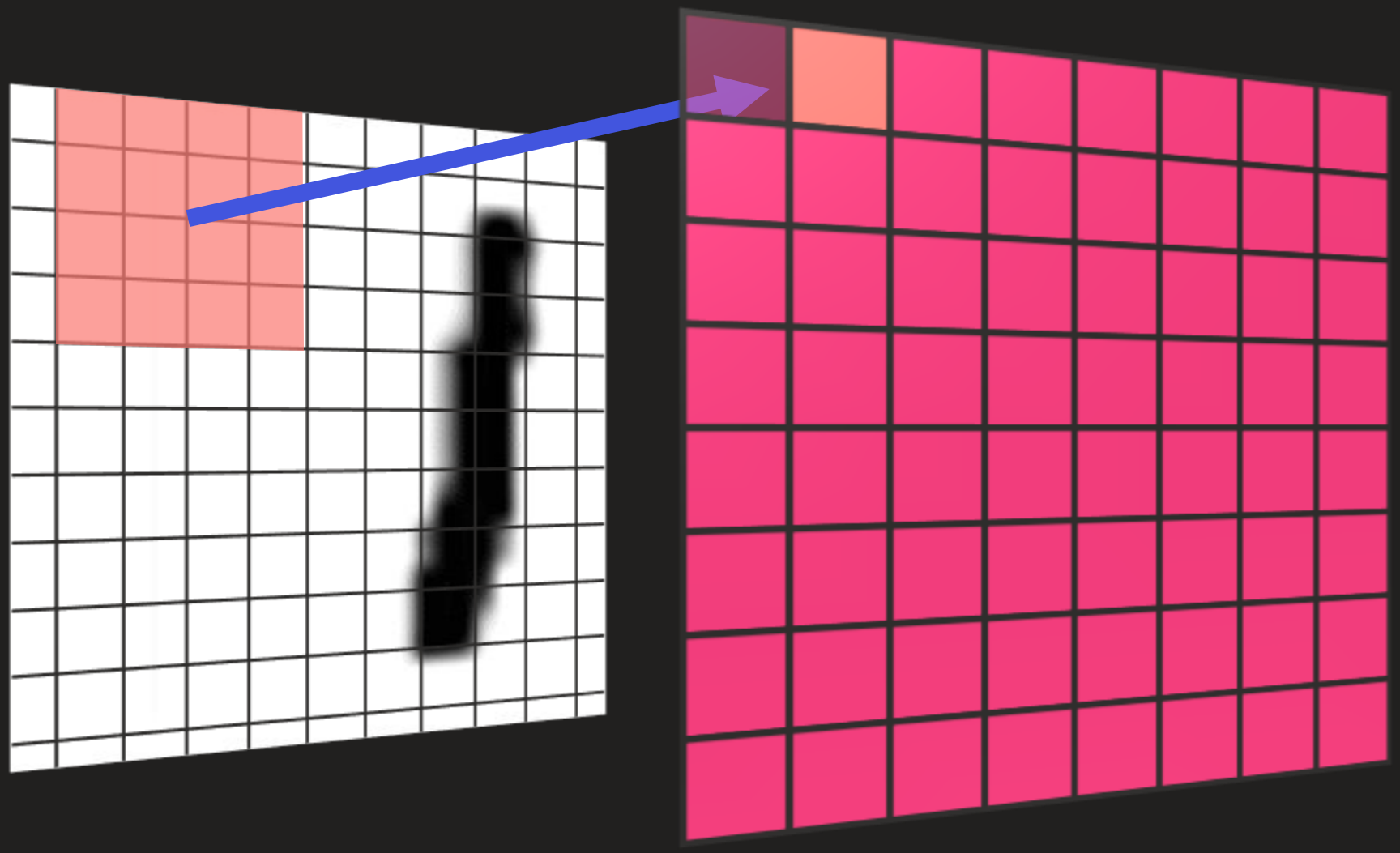


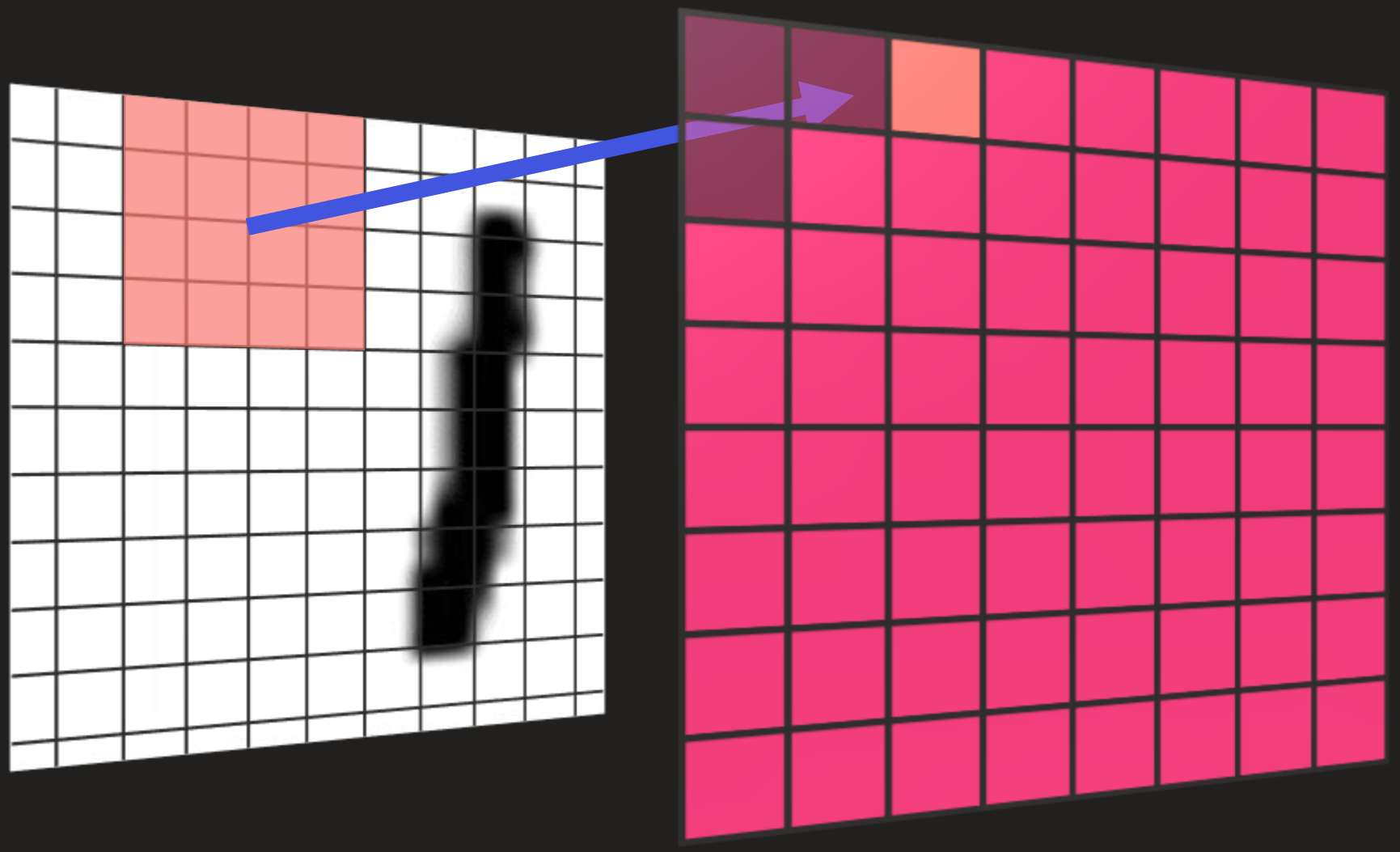
$$\begin{bmatrix} 132 \\ 12 \\ 21 \\ 231 \\ \vdots \\ 12 \\ 0 \\ 255 \\ 52 \end{bmatrix}$$

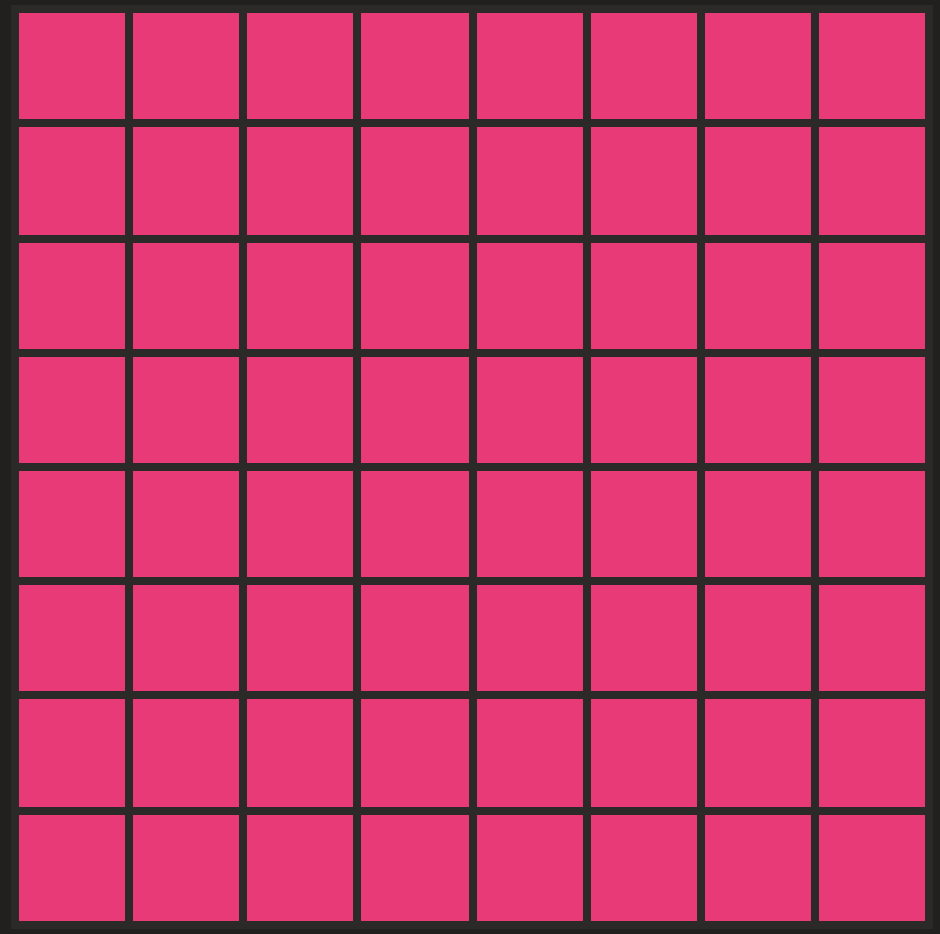


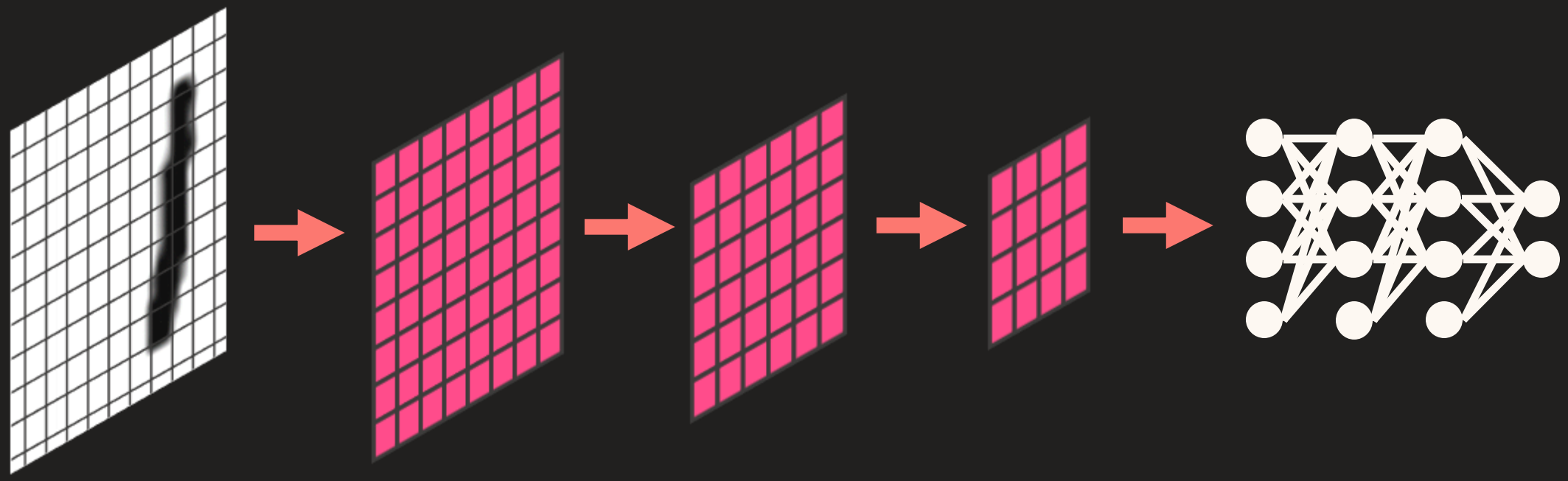




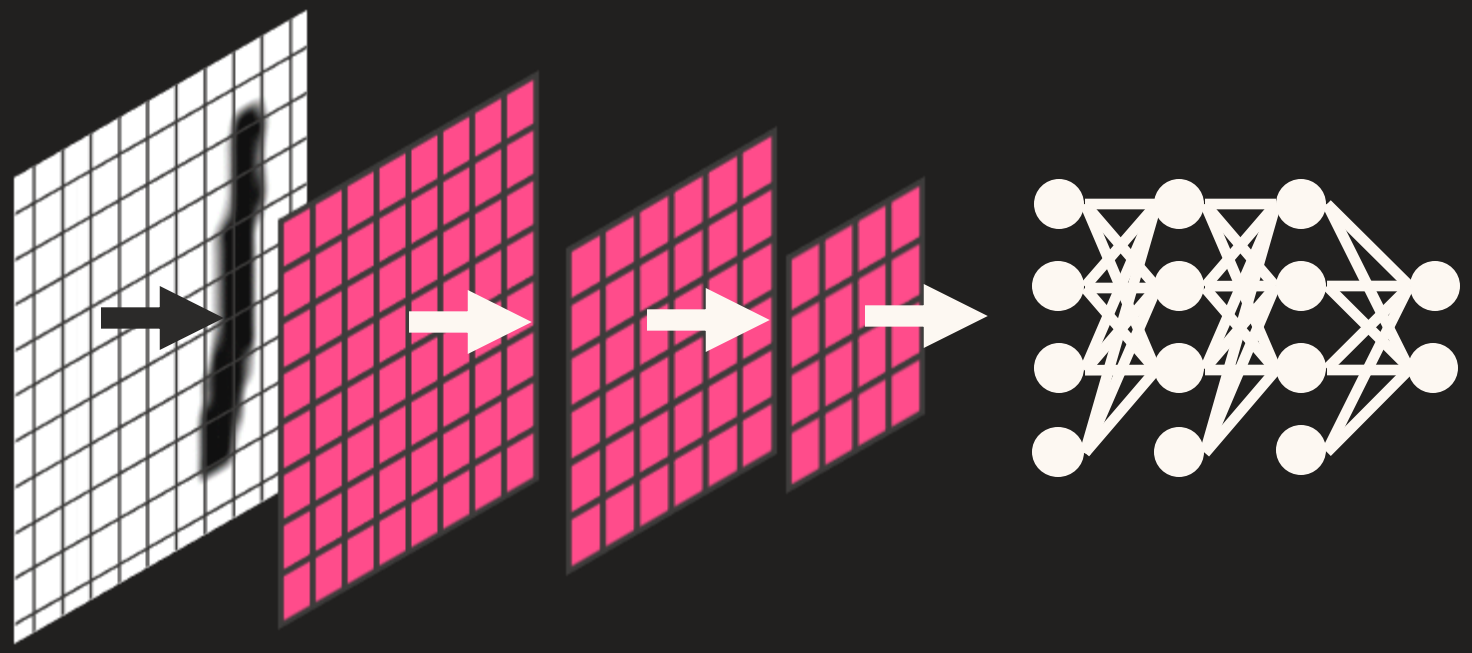








CNN



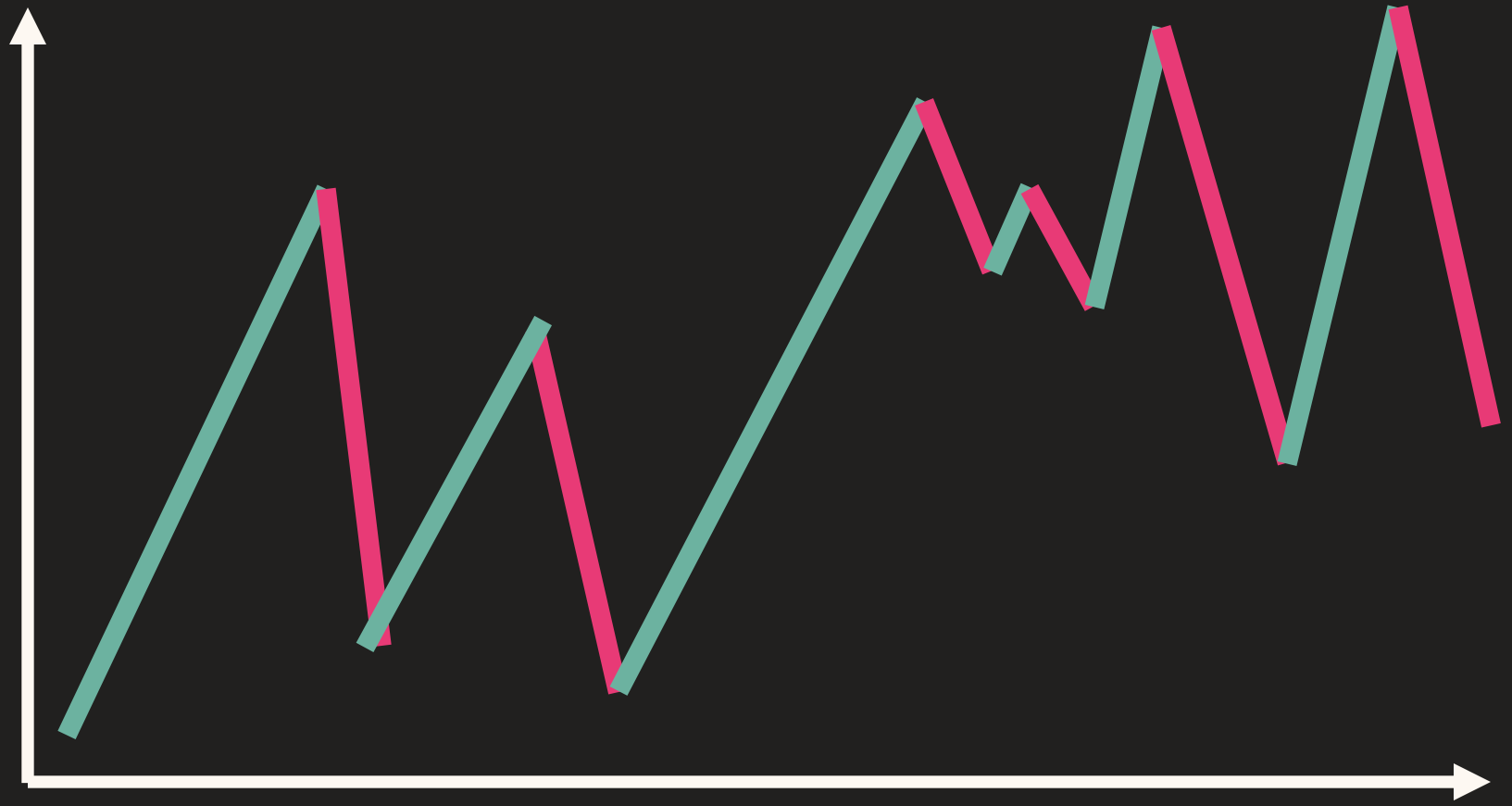
CNN

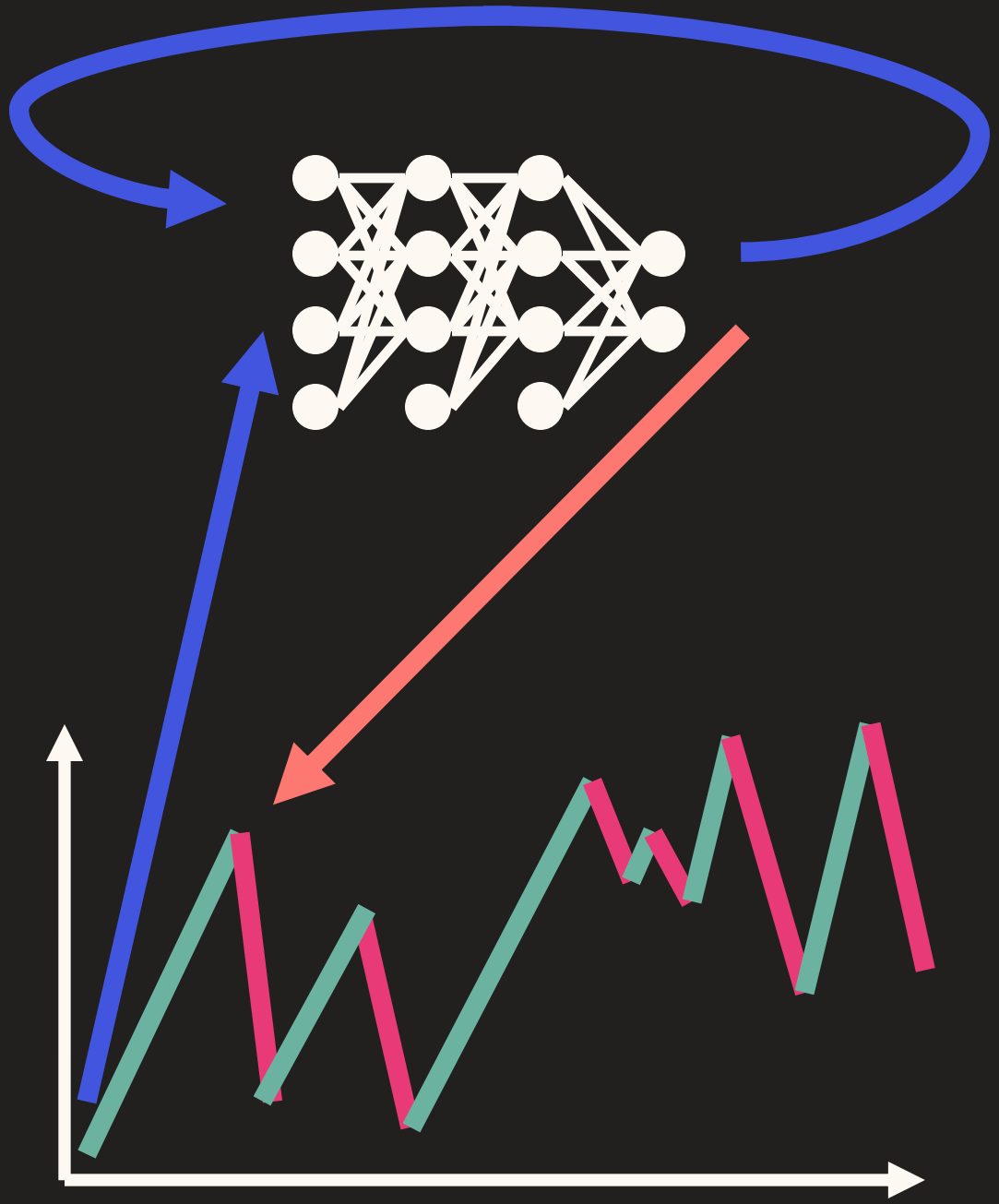
80s

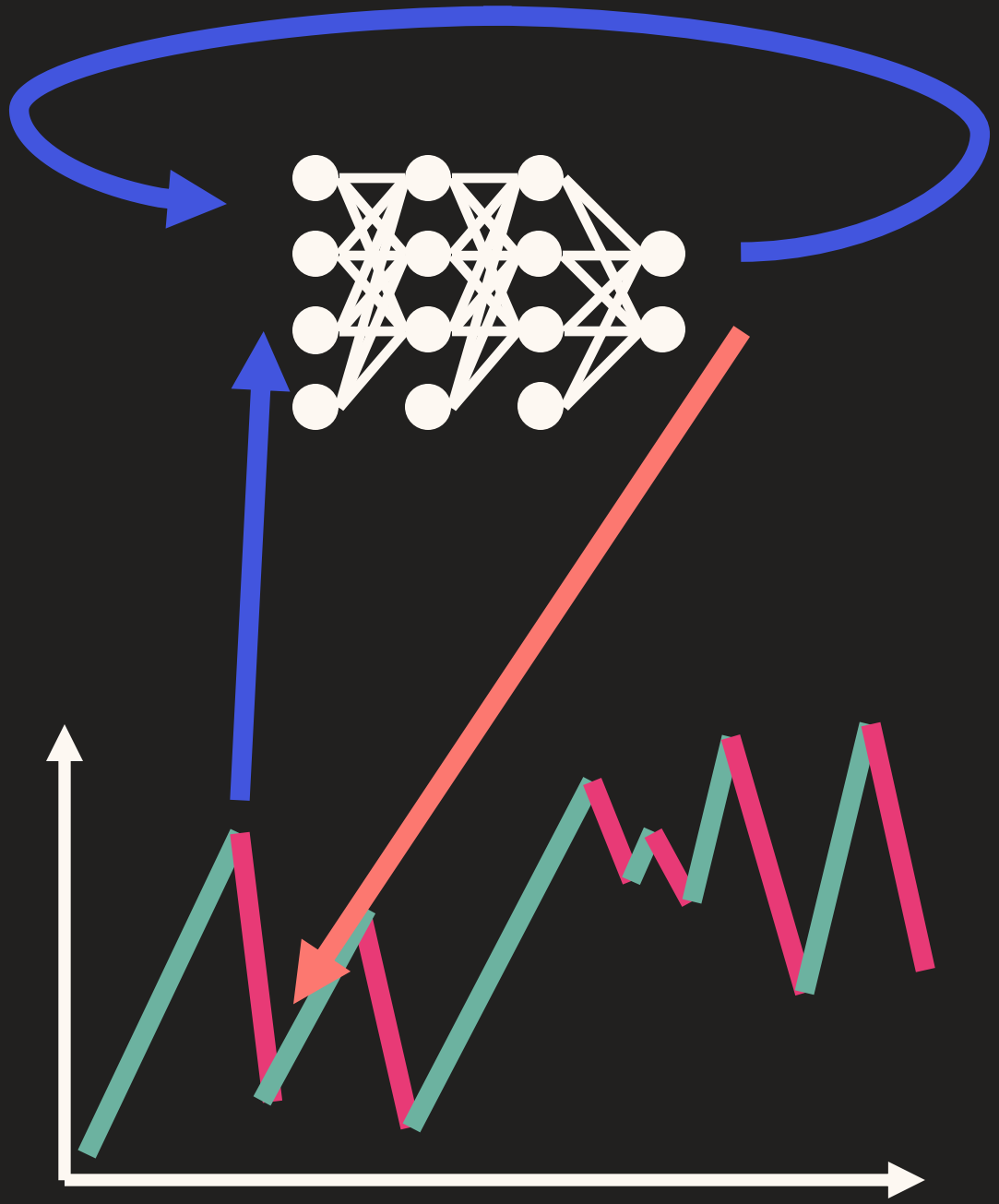
⋮

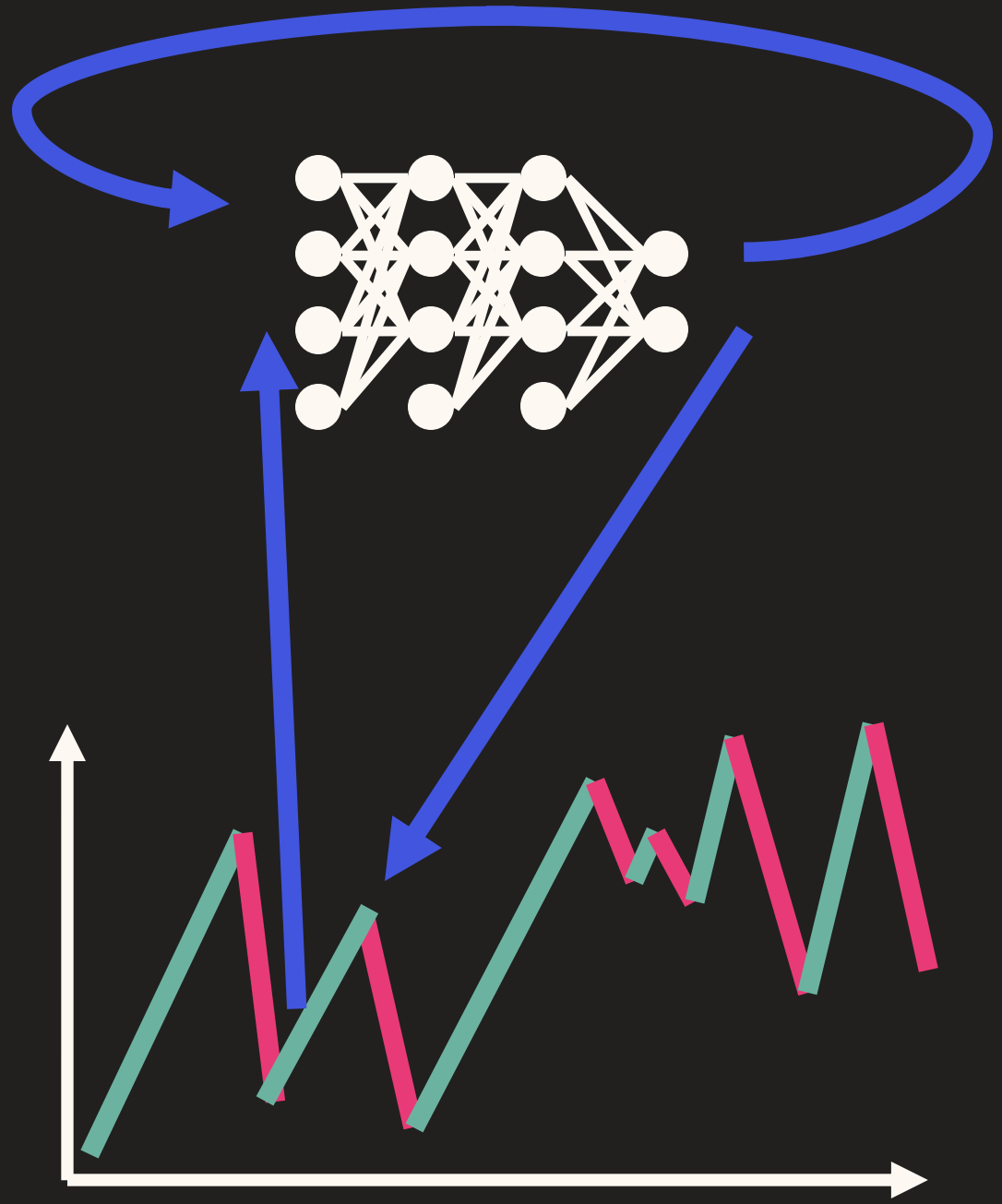
90s

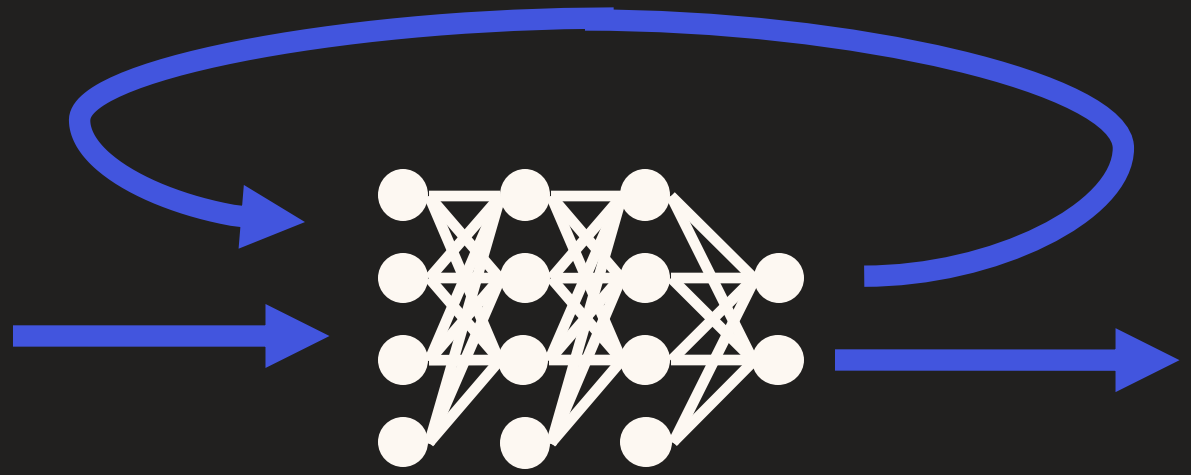
Dane czasowe i szeregowe



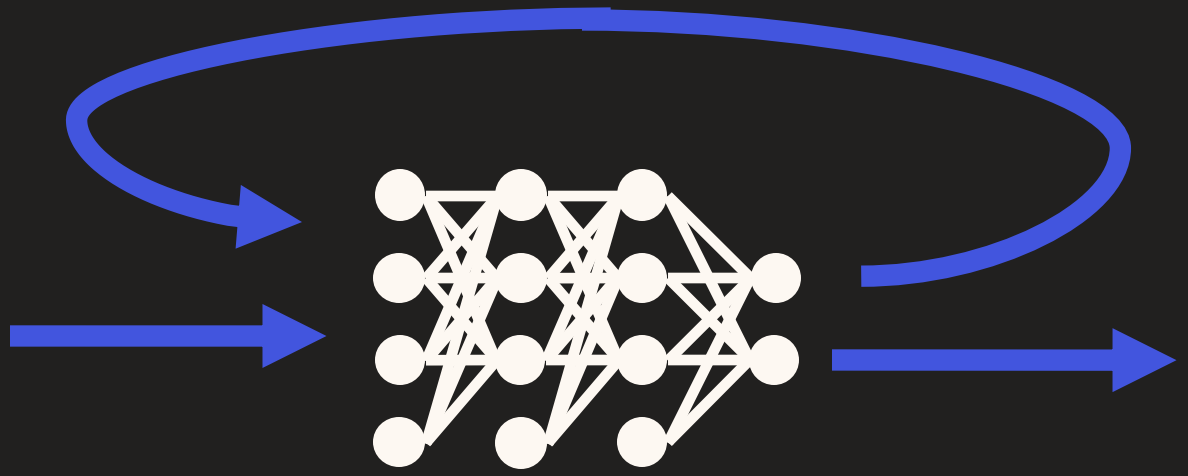








RNN



RNN

80s

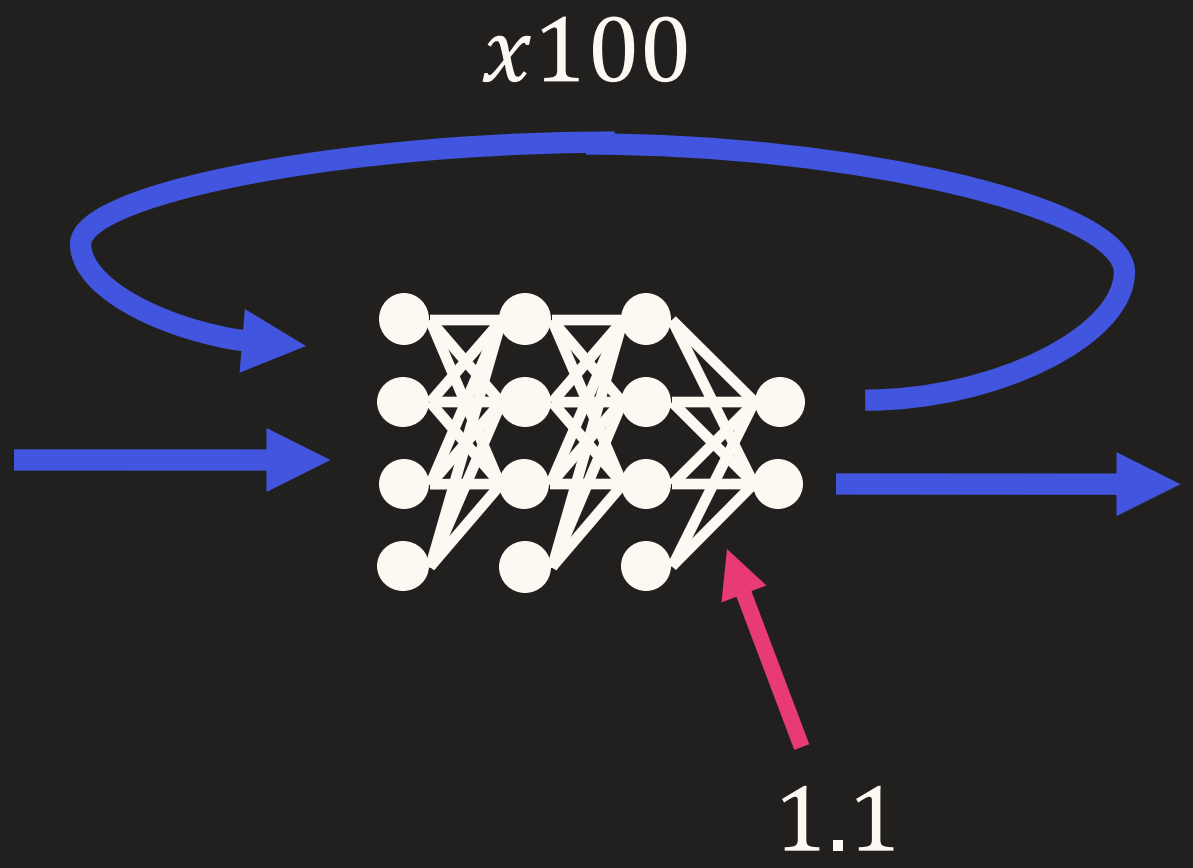
⋮

90s

Problemy?

$$1.1^{100}$$

13780.612…


$x100$
1.1

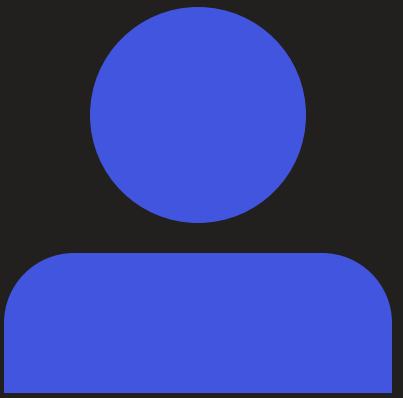
$$0.9^{100}$$

0.00002656…

AKT II

Jak rozmawiać z maszyną?

Jakie miasto jest stolicą Francji?



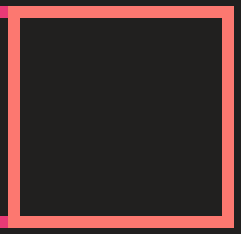
Stolicą Francji jest Paryż.

AI

Być, albo nie być?
Oto jest pytanie.

Być, albo nie być? Oto

Być, albo nie być? Oto

Być, albo nie być? Oto 

Być, albo nie być? Oto jest

Być, albo nie być? Oto jest

Być, albo nie być? Oto jest pytanie

Jakie miasto jest stolicą Francji?

Pytam, ponieważ mam jutro sprawdzian z Geografii

Pytanie:
Jakie miasto jest stolicą Francji?
Odpowiedź:

Stolicą

Pytanie:
Jakie miasto jest stolicą Francji?
Odpowiedź:

Stolicą	Francji

Pytanie:
Jakie miasto jest stolicą Francji?
Odpowiedź:

Stolicą Francji jest

Pytanie:
Jakie miasto jest stolicą Francji?
Odpowiedź:

Stolicą Francji jest	Paryż

ChatGPT

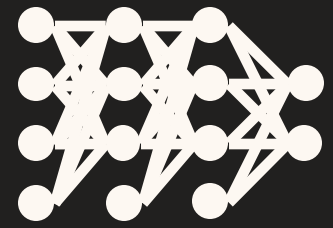
GPT

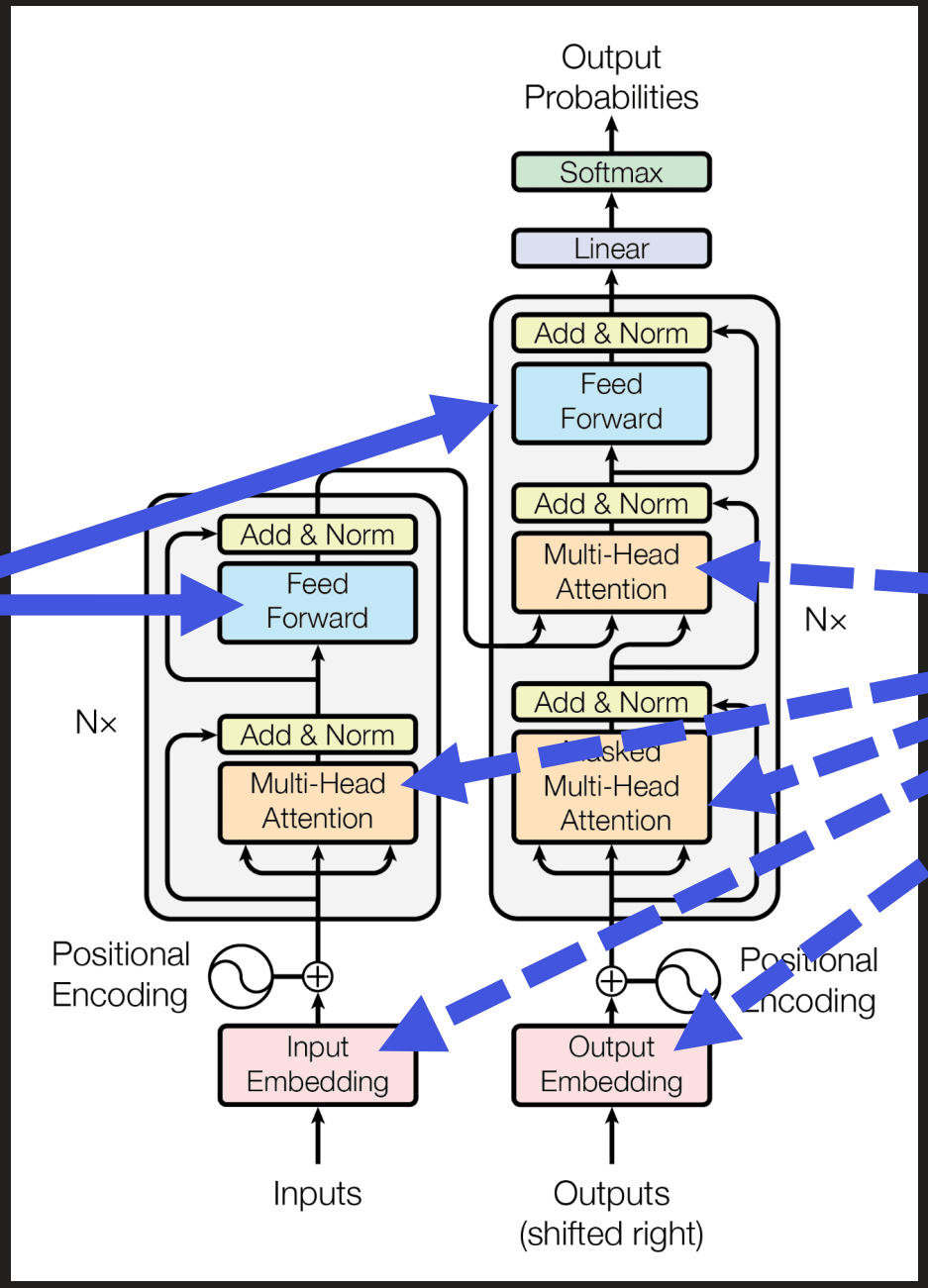
Generative **P**retrained **T**ransformer

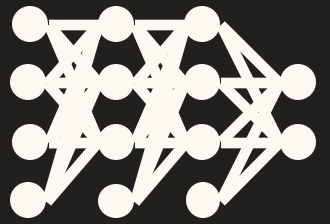
Generatywny Przedtrenowany (?) Transformer

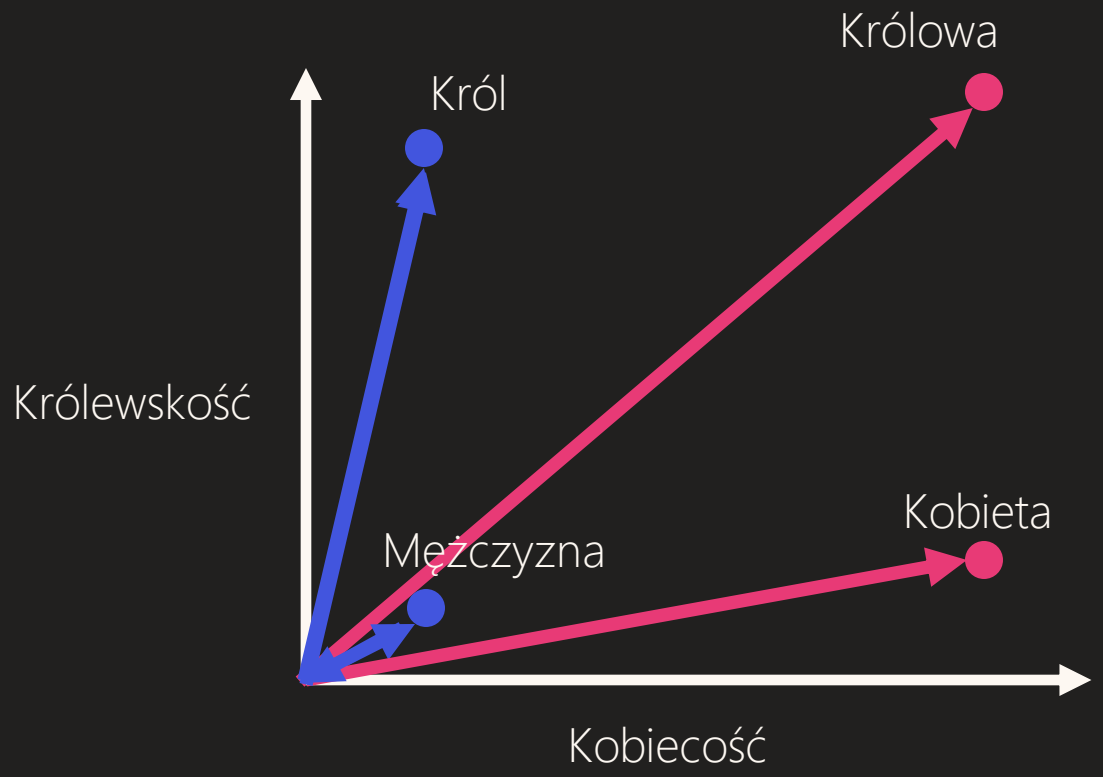
Transformer


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS




Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)





Królowa = Król + Kobieta - Mężczyzna

```
model.distance('car', 'cat')
```

```
model.distance('dog', 'cat')
```

```
print(model.most_similar_cosmul(positive=['king', 'woman'], negative=['man']))
```

```
[('queen', 0.9314123392105103), ('monarch', 0.858533501625061), ('princess',
```

```
print(model.most_similar_cosmul(positive=['cow', 'oink'], negative=['pig']))
```

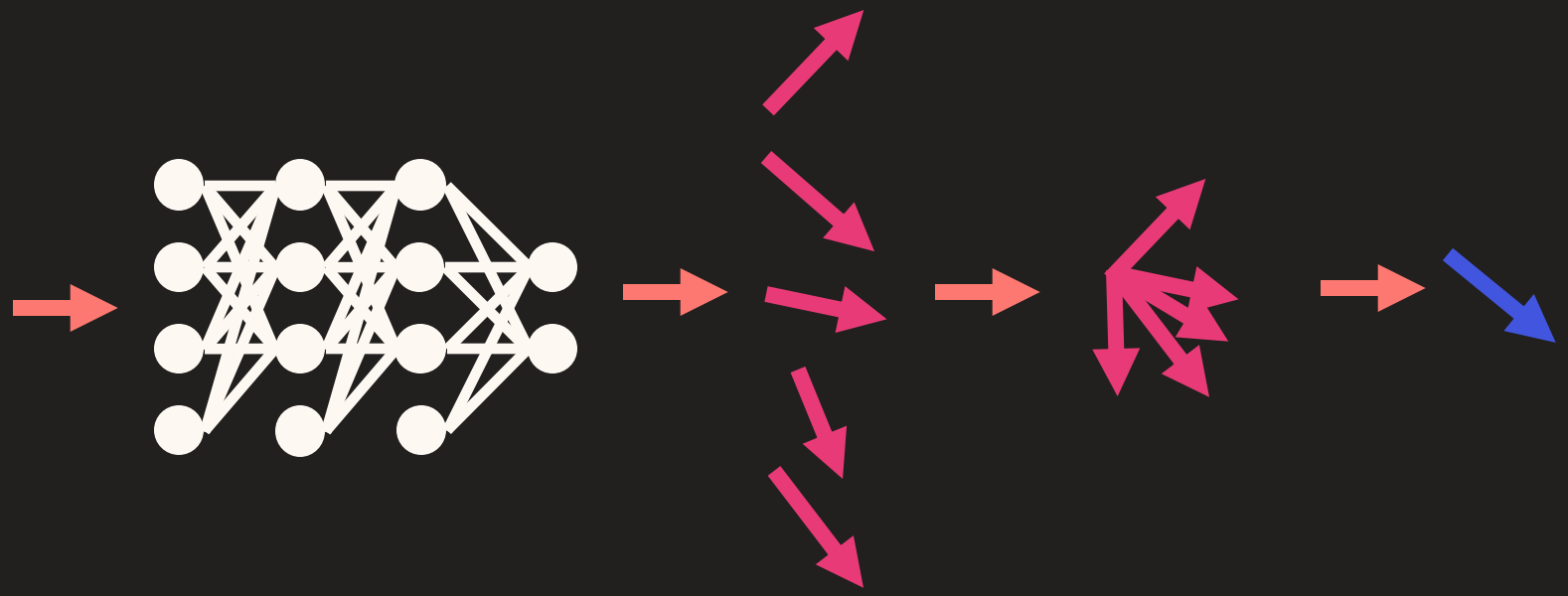
```
[('moos', 0.8037621378898621), ('baa', 0.7783335447311401), ('baaa', 0.7776672
```

```
print(model.most_similar_cosmul(positive=['Santa', 'oink'], negative=['pig']))
```

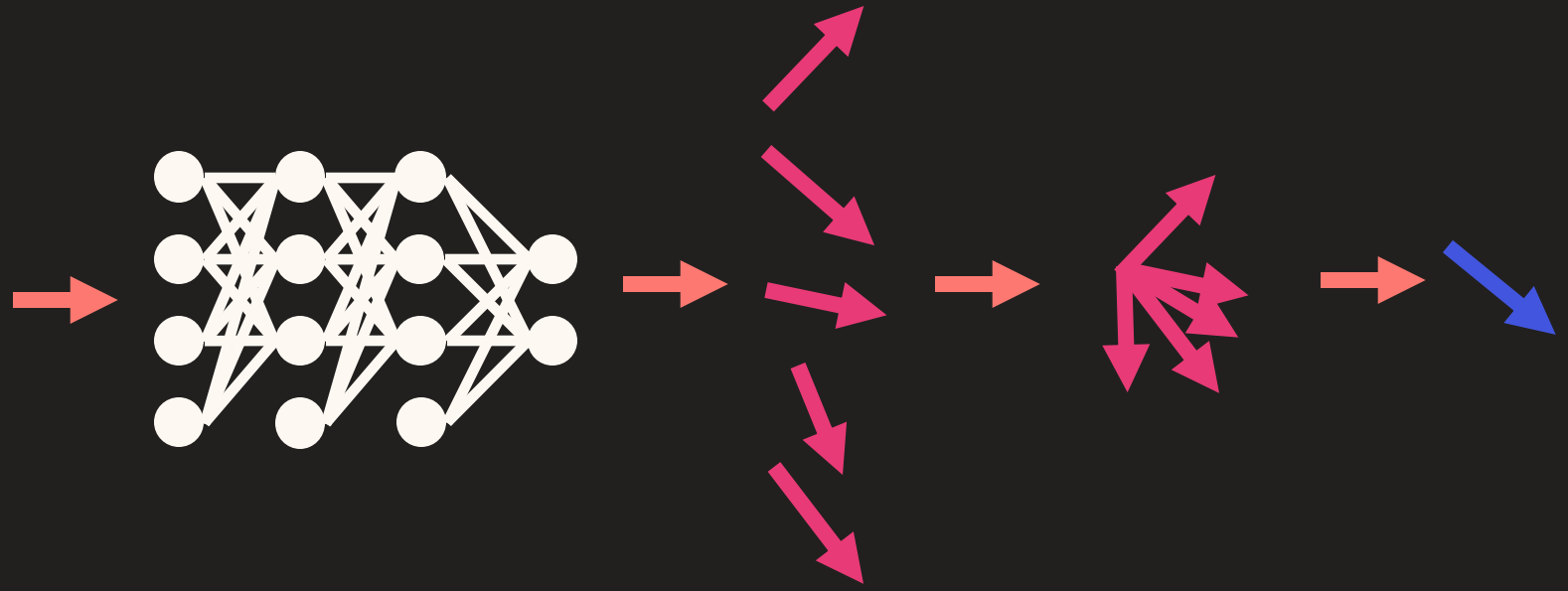
```
[('HO_HO_HO', 0.9089415073394775), ('ho_ho_hoing', 0.9081220030784607), (
```

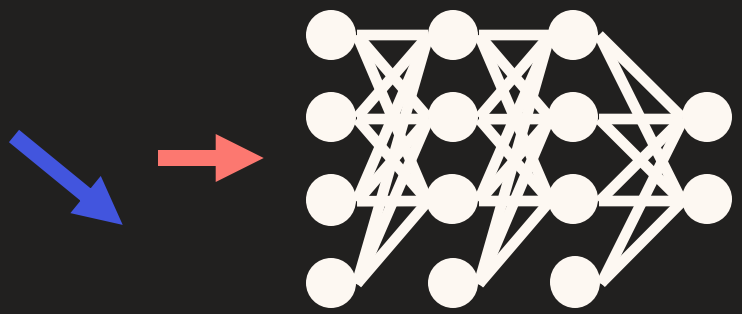
```
('feline', 0.7326233983039856),
('beagle', 0.7150583267211914),
('puppy', 0.7075453996658325),
('pup', 0.6934291124343872),
('pet', 0.6891531348228455),
('felines', 0.6755931377410889),
('chihuahua', 0.6709762215614319)]
```

Jakie miasto jest stolicą Francji?



Jakie miasto jest stolicą Francji?

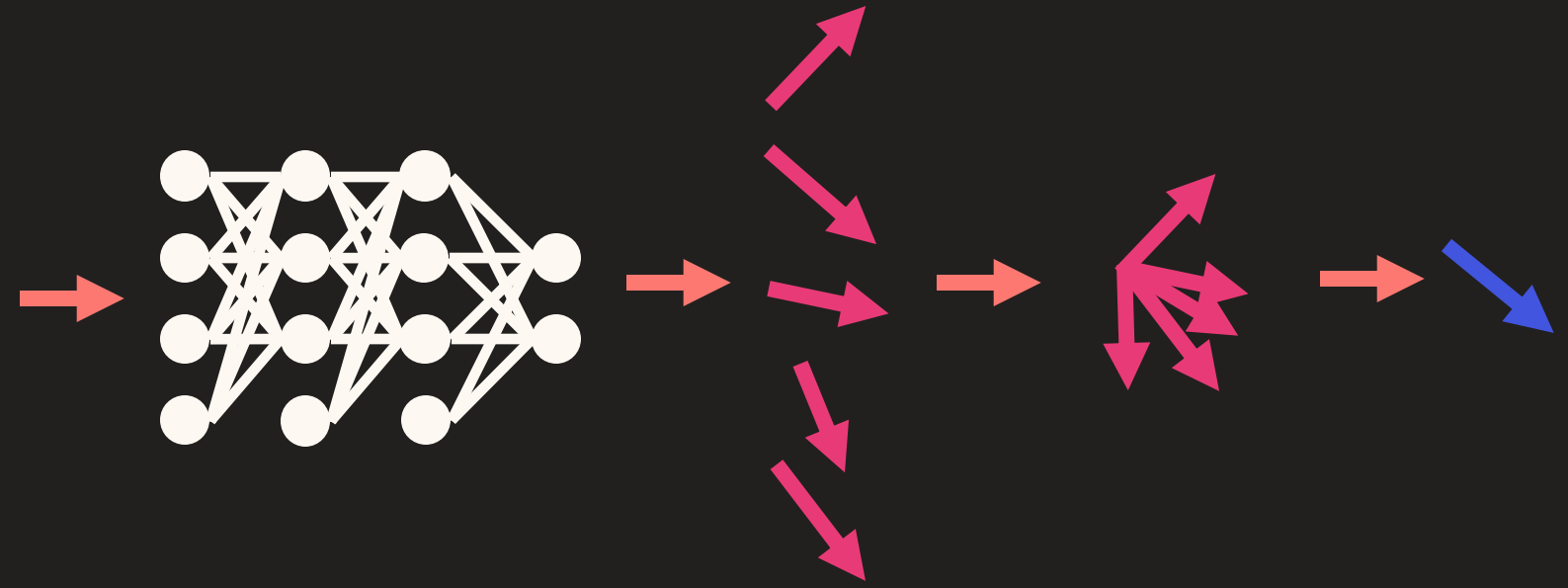


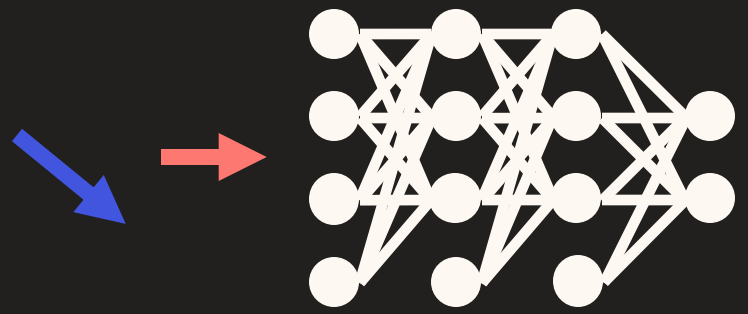


Paryż
Stolicą
Miastem
...

Paryż

Jakie miasto jest stolicą Francji?



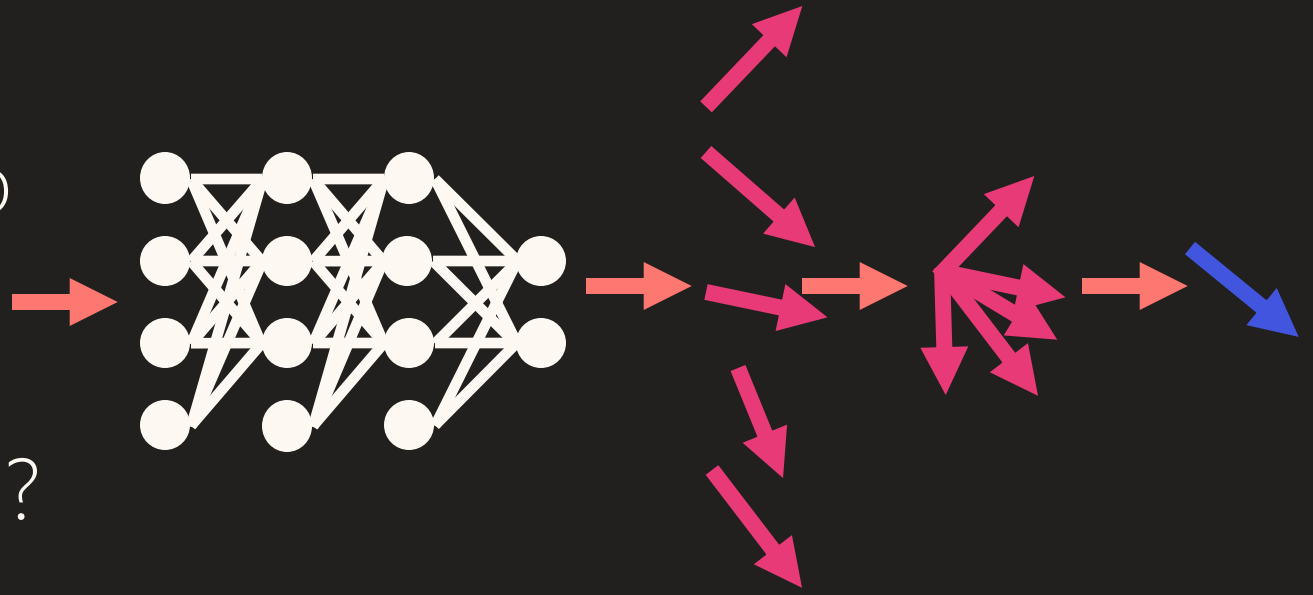


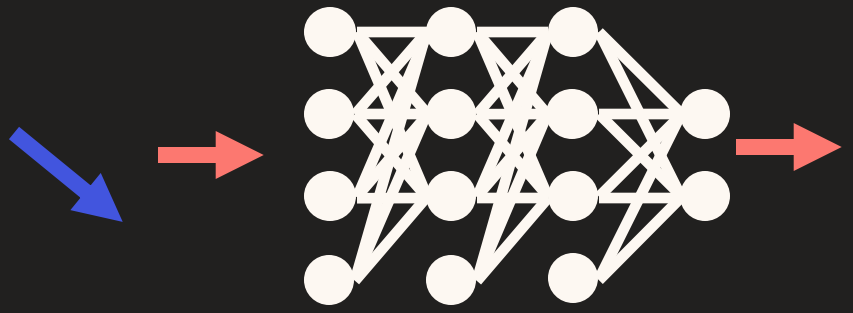
Paryż
Stolicą
Miastem
...

Paryż

Transformer

Jakie miasto jest stolicą Francji?





Paryż
Stolicą
Miastem
...

Paryż

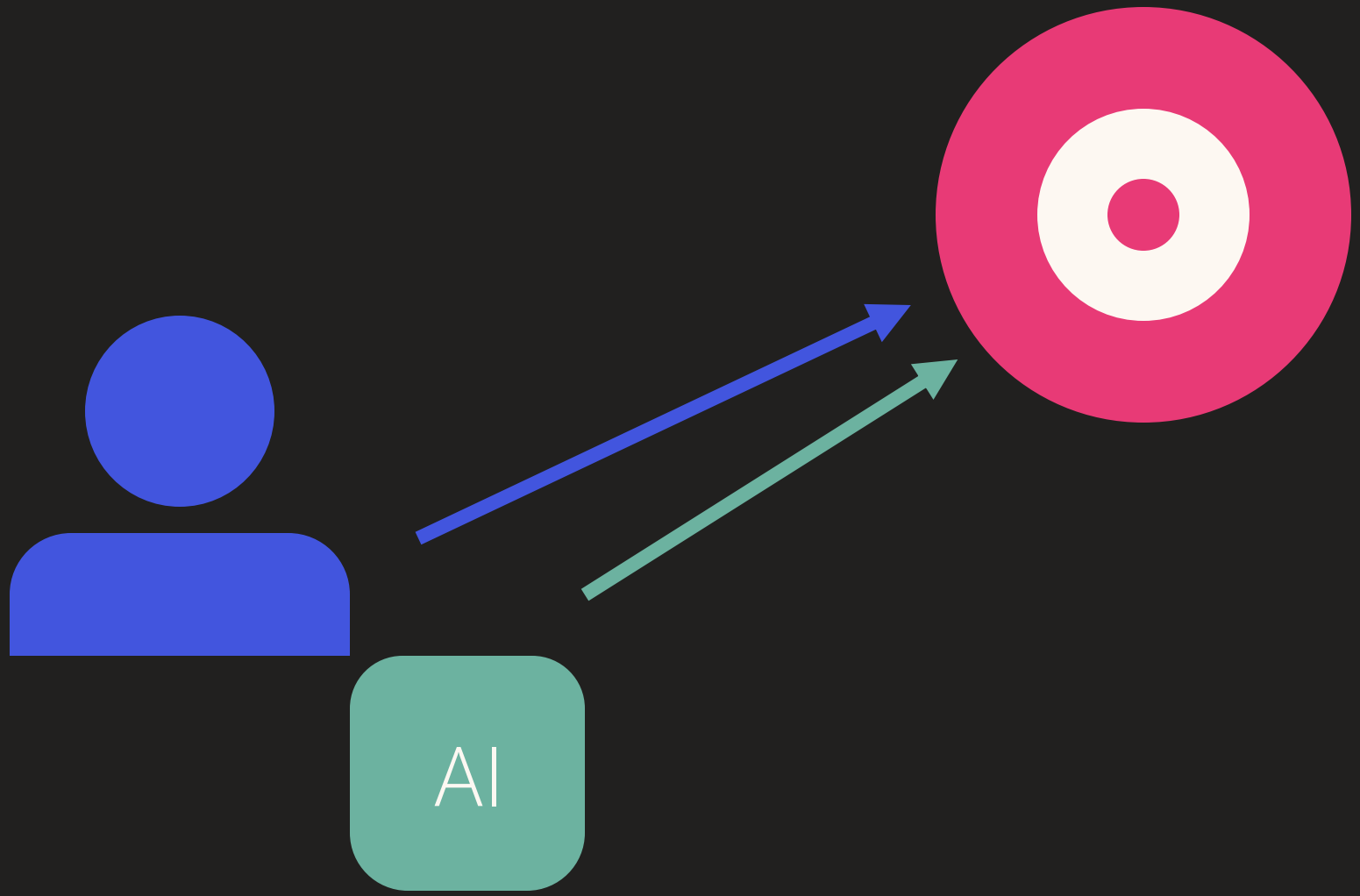
Transformer

2017

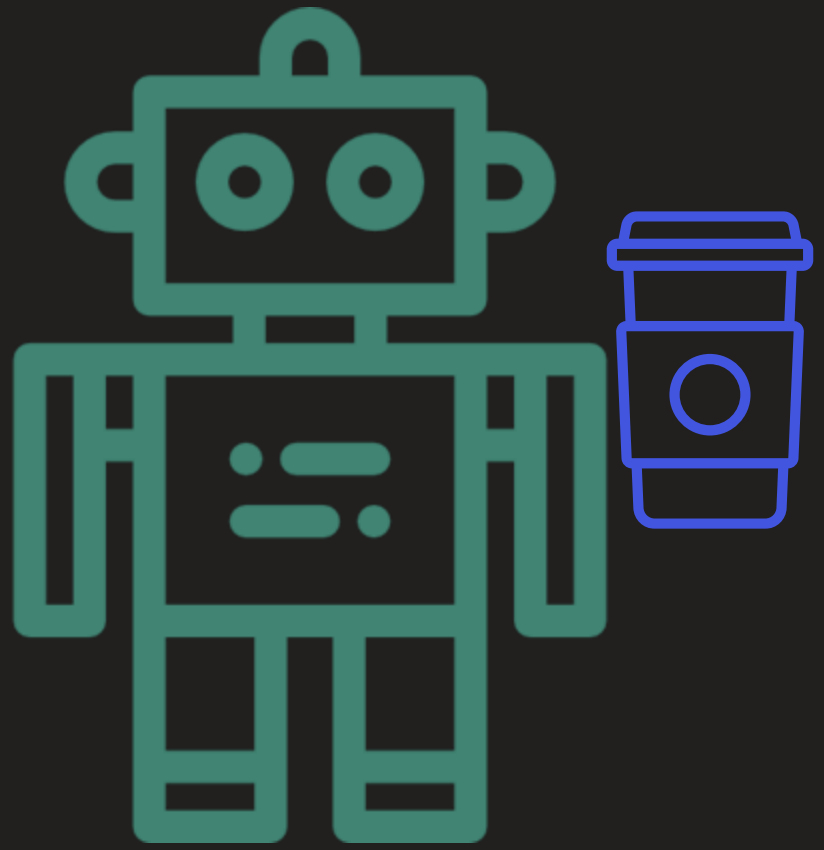
Alignment Problem

AKT III

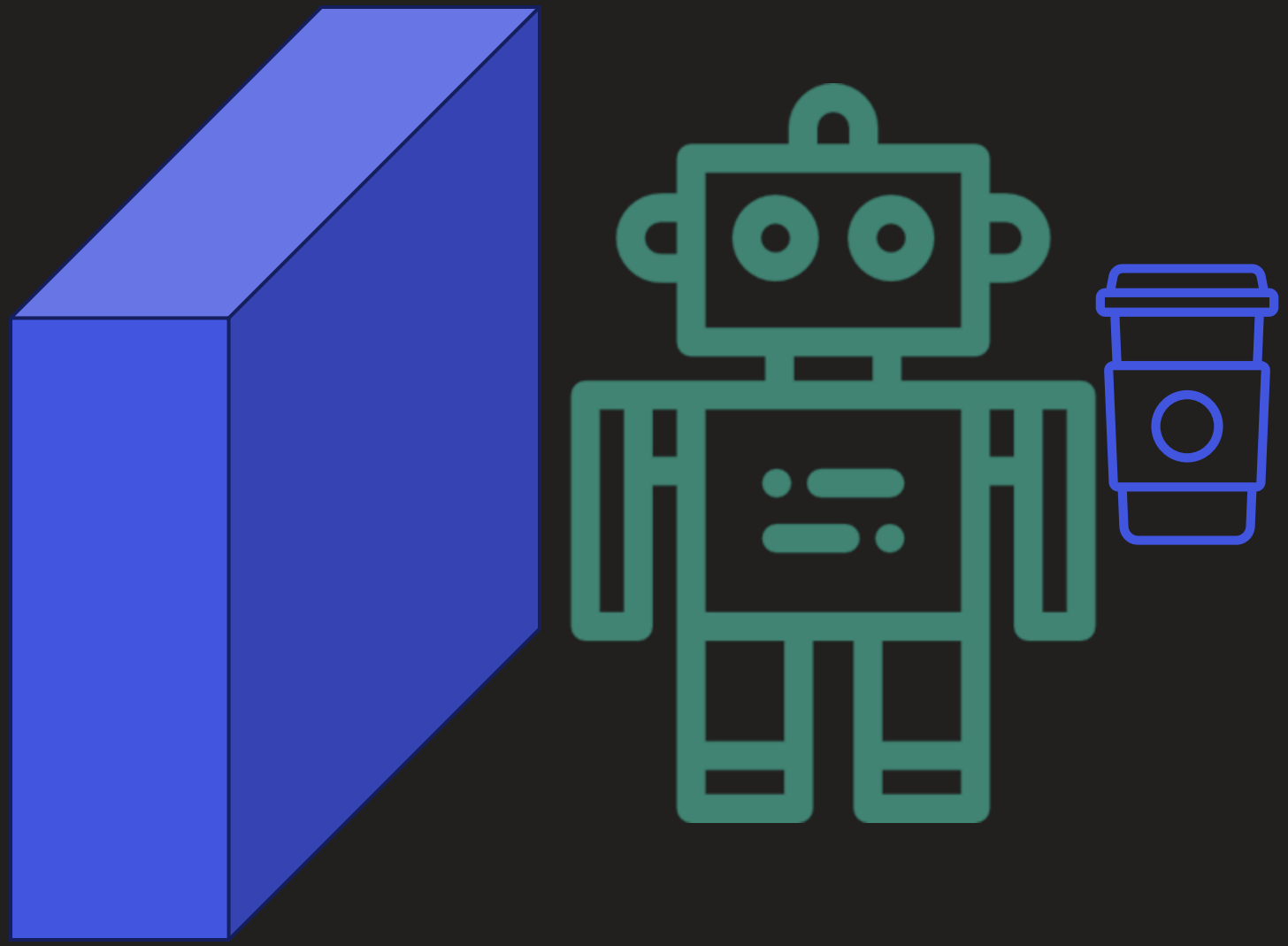
Co złego może się stać?

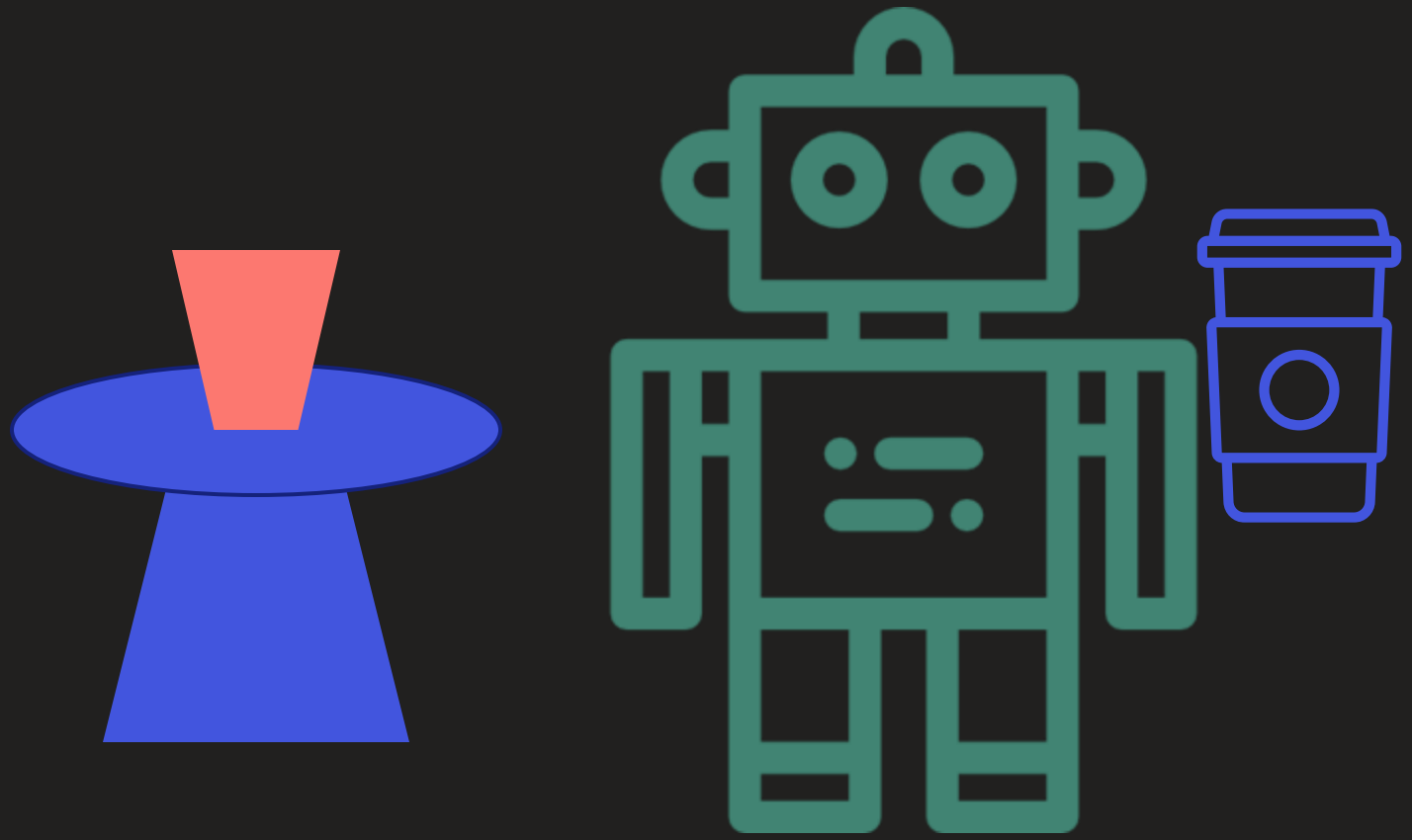

AI

Negatywne efekty uboczne



Przynieś mi kawę: +100 punktów





Przynieś mi kawę: +100 punktów

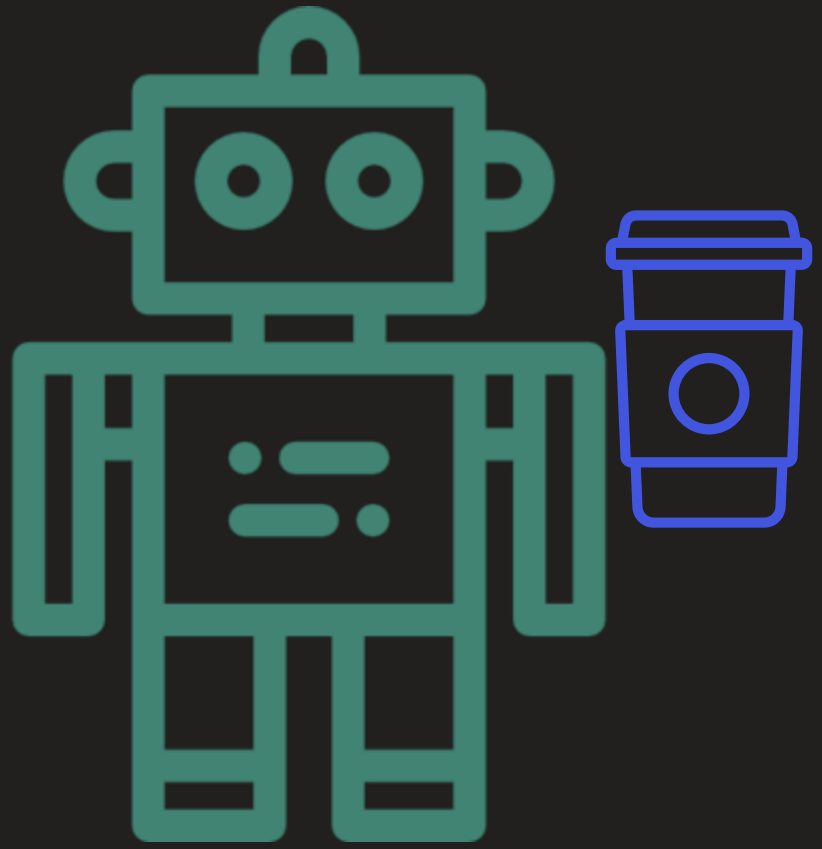
Przynieś mi kawę: +100 punktów
Ściana uszkodzona: bez znaczenia
Wazon uszkodzony: bez znaczenia

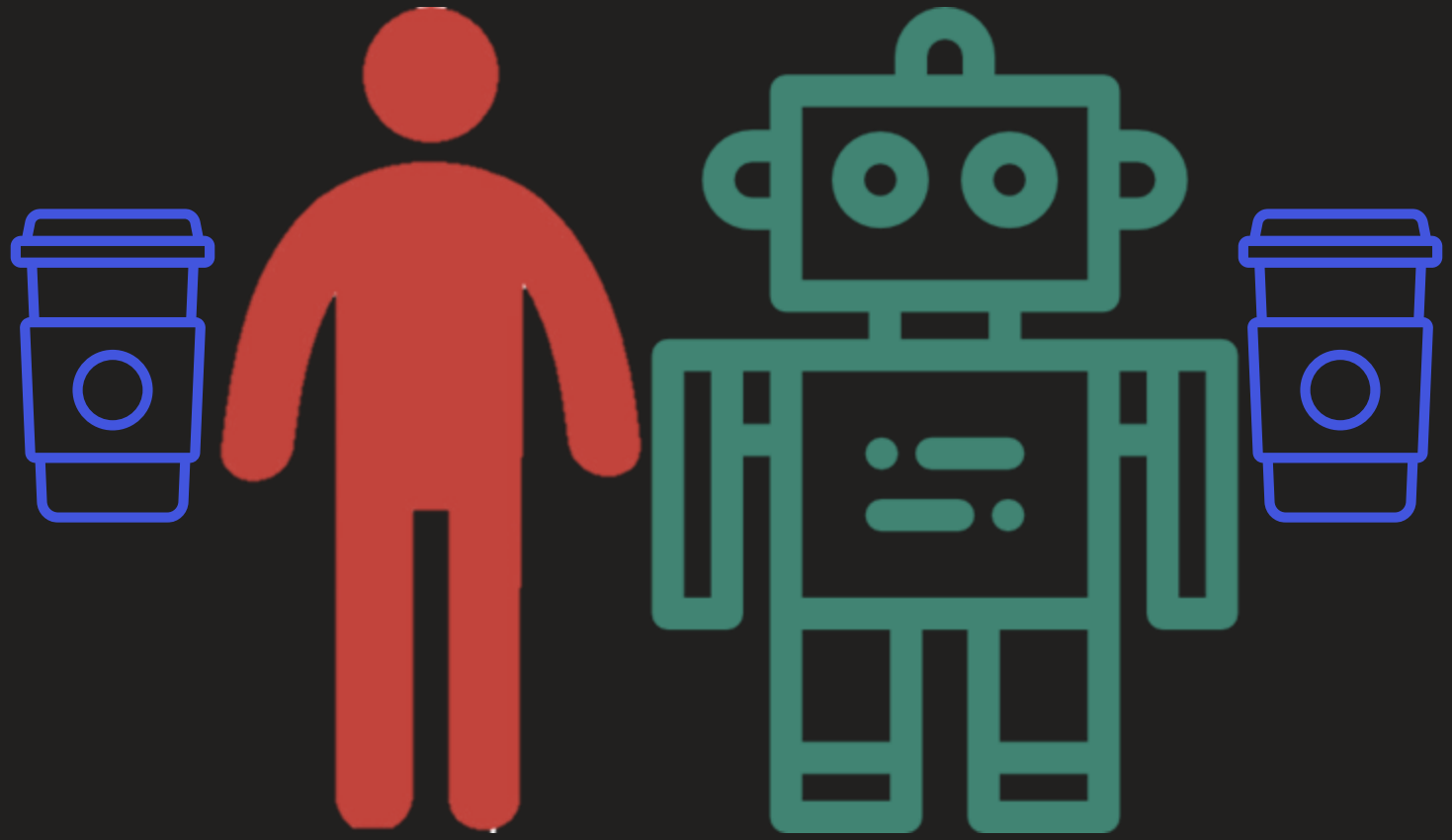
Przynieś mi kawę: +100 punktów
Ściana uszkodzona: bez znaczenia
Wazon uszkodzony: bez znaczenia
Podłoga uszkodzona: bez znaczenia
Drzwi uszkodzone: bez znaczenia
Okna uszkodzone: bez znaczenia
Kubek uszkodzony: bez znaczenia
Czajnik uszkodzony: bez znaczenia

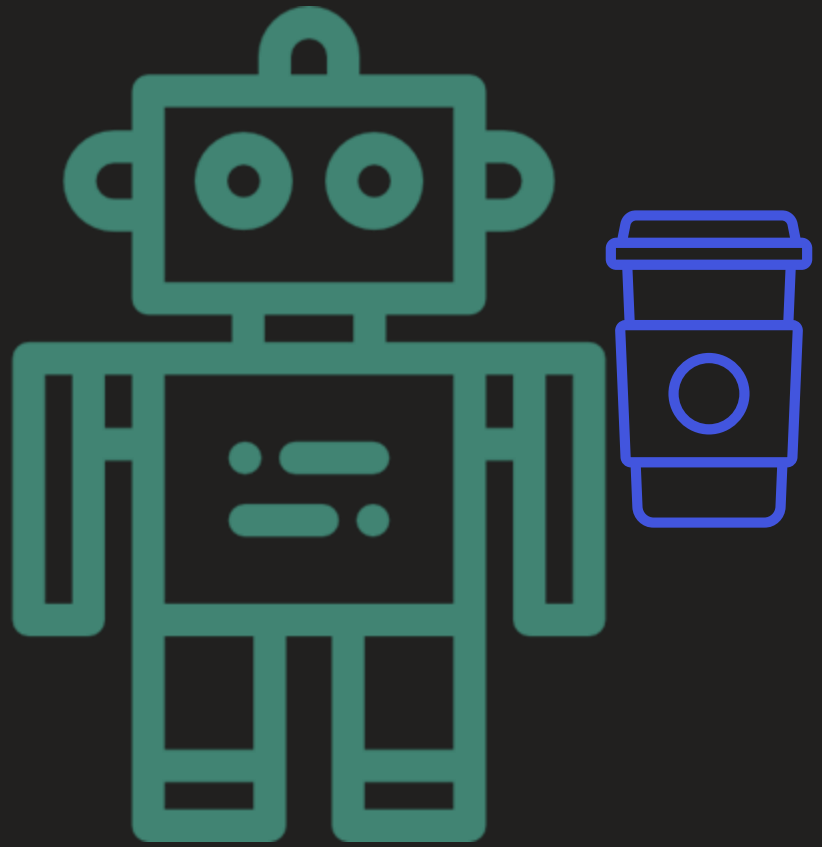
Nie chcemy zmian w świecie

Cel: Przynieś mi kawę

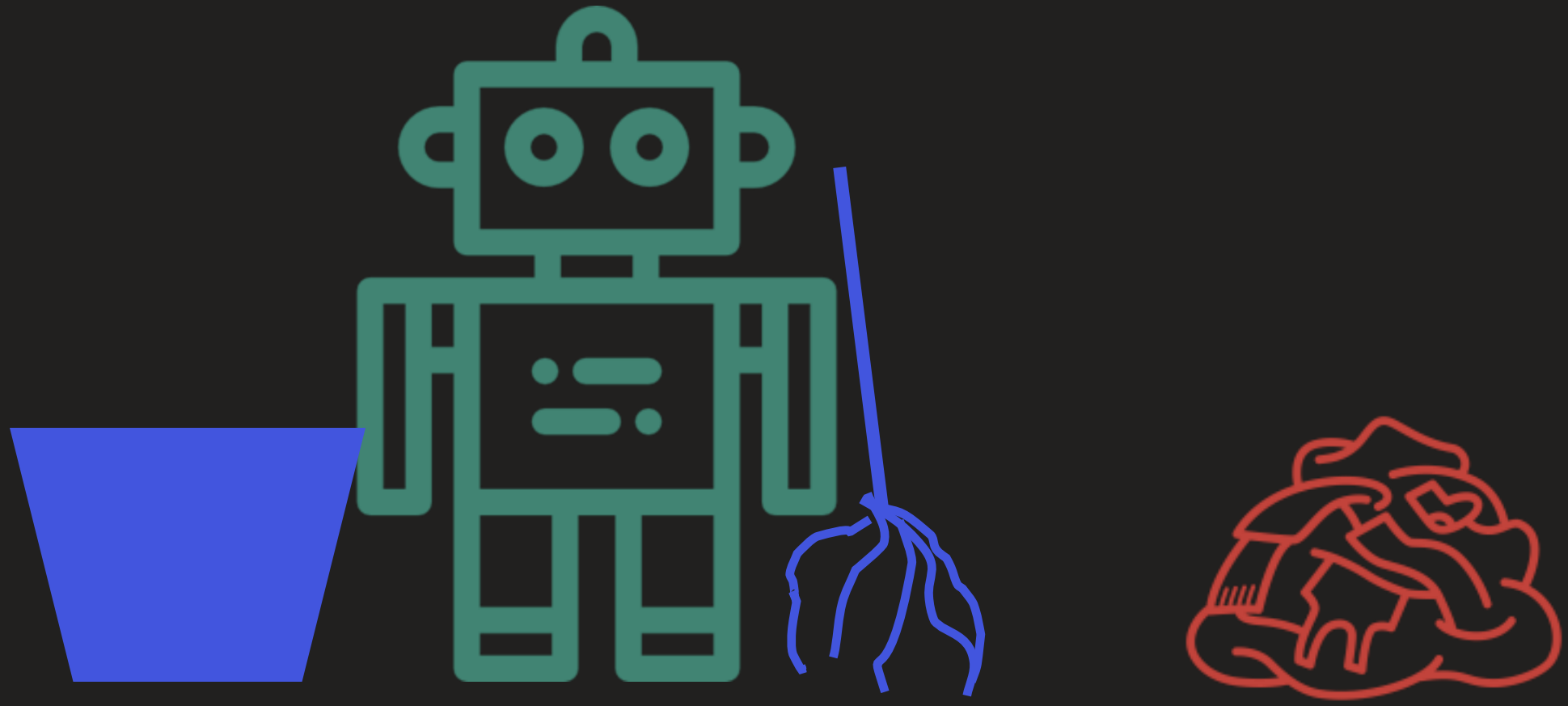
Nagroda = Cel – |pierwotny stan świata - końcowy stan świata|

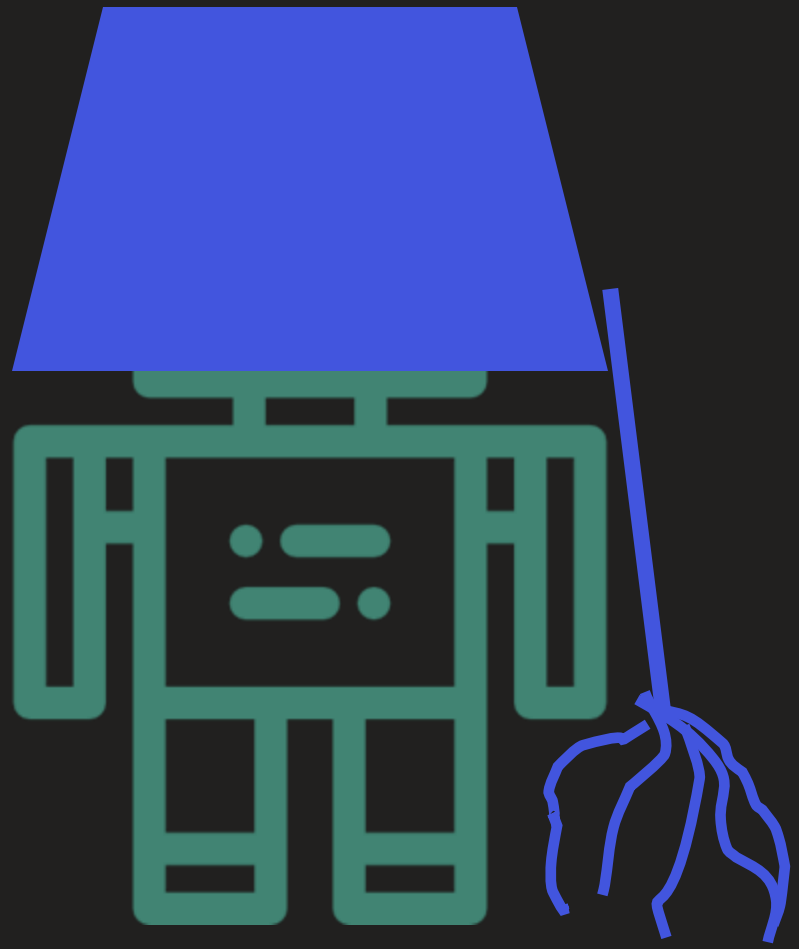


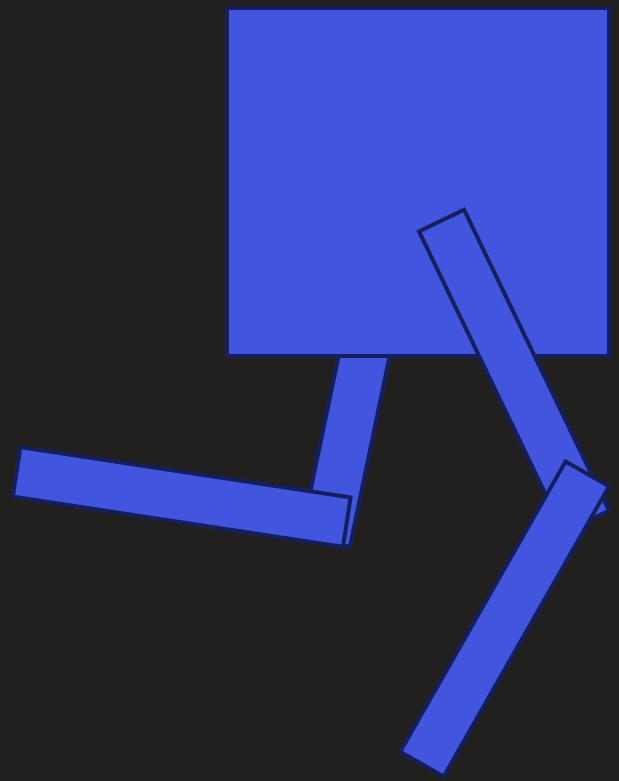


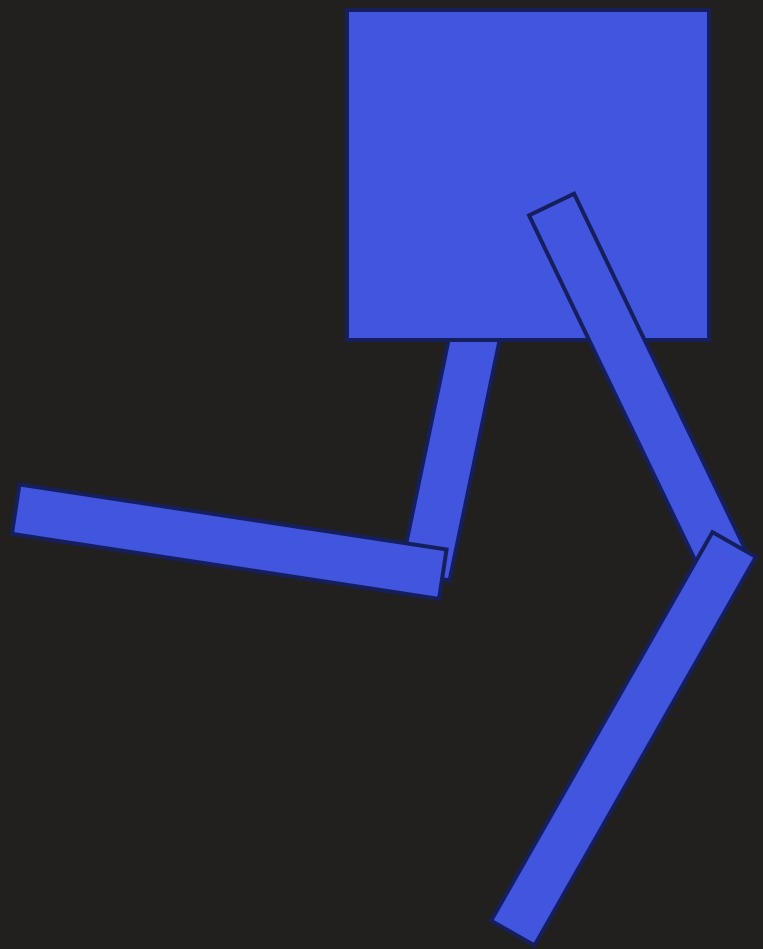


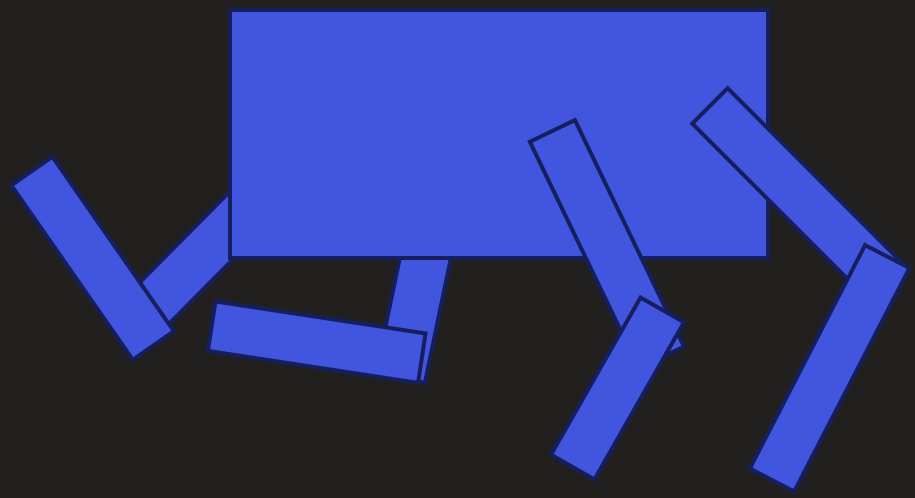
Oszukiwanie funkcji nagrody

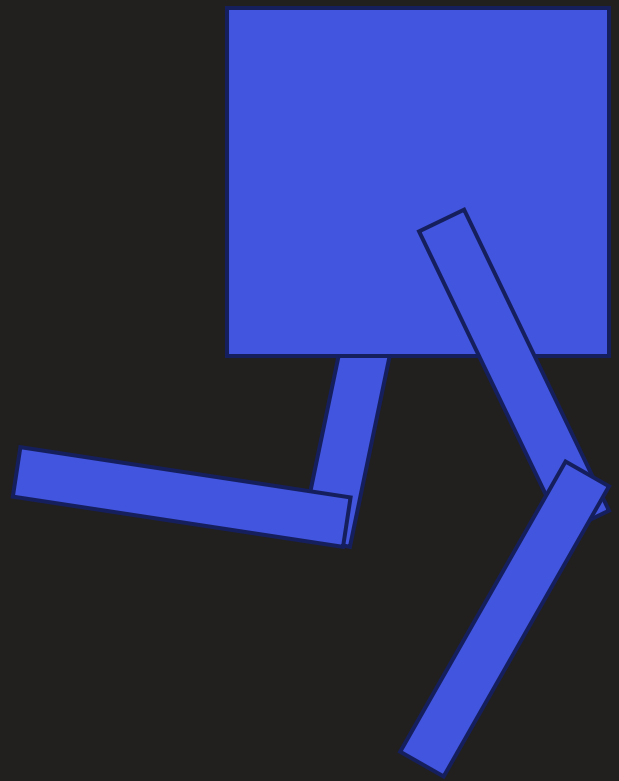


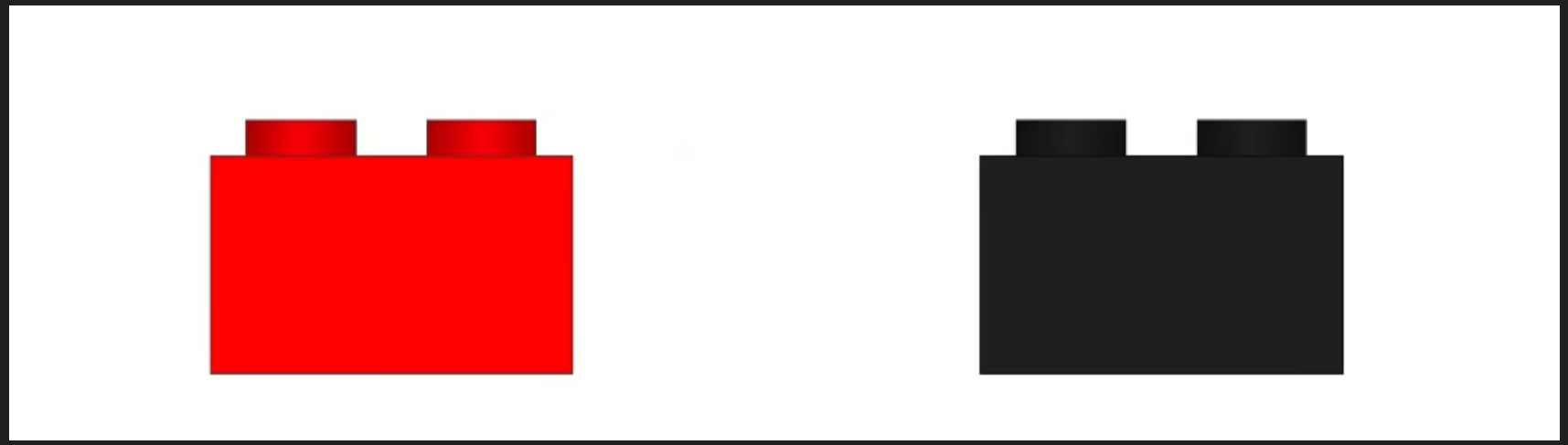


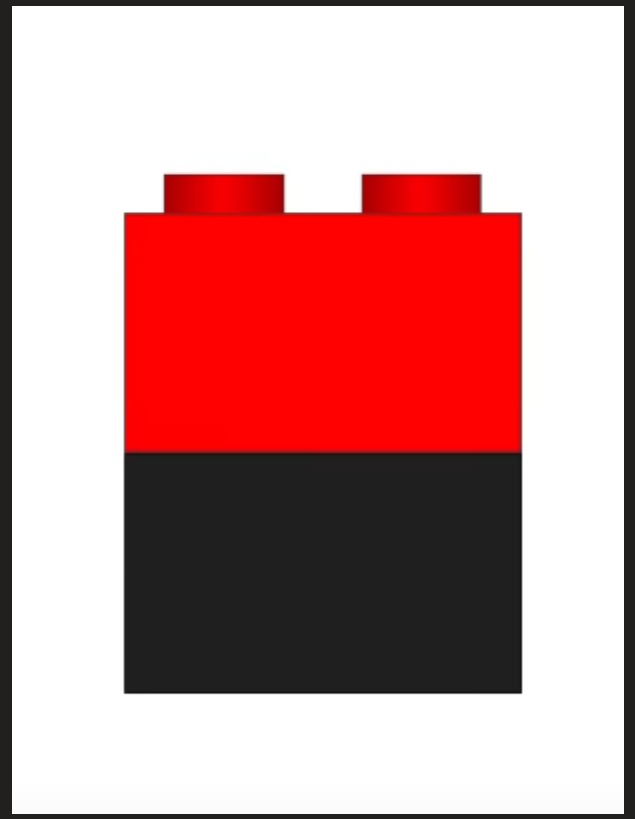


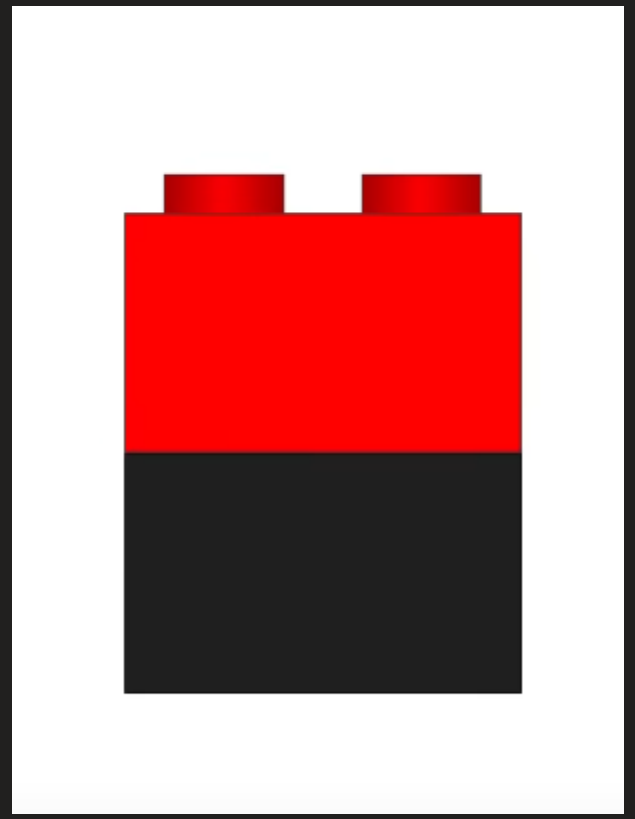


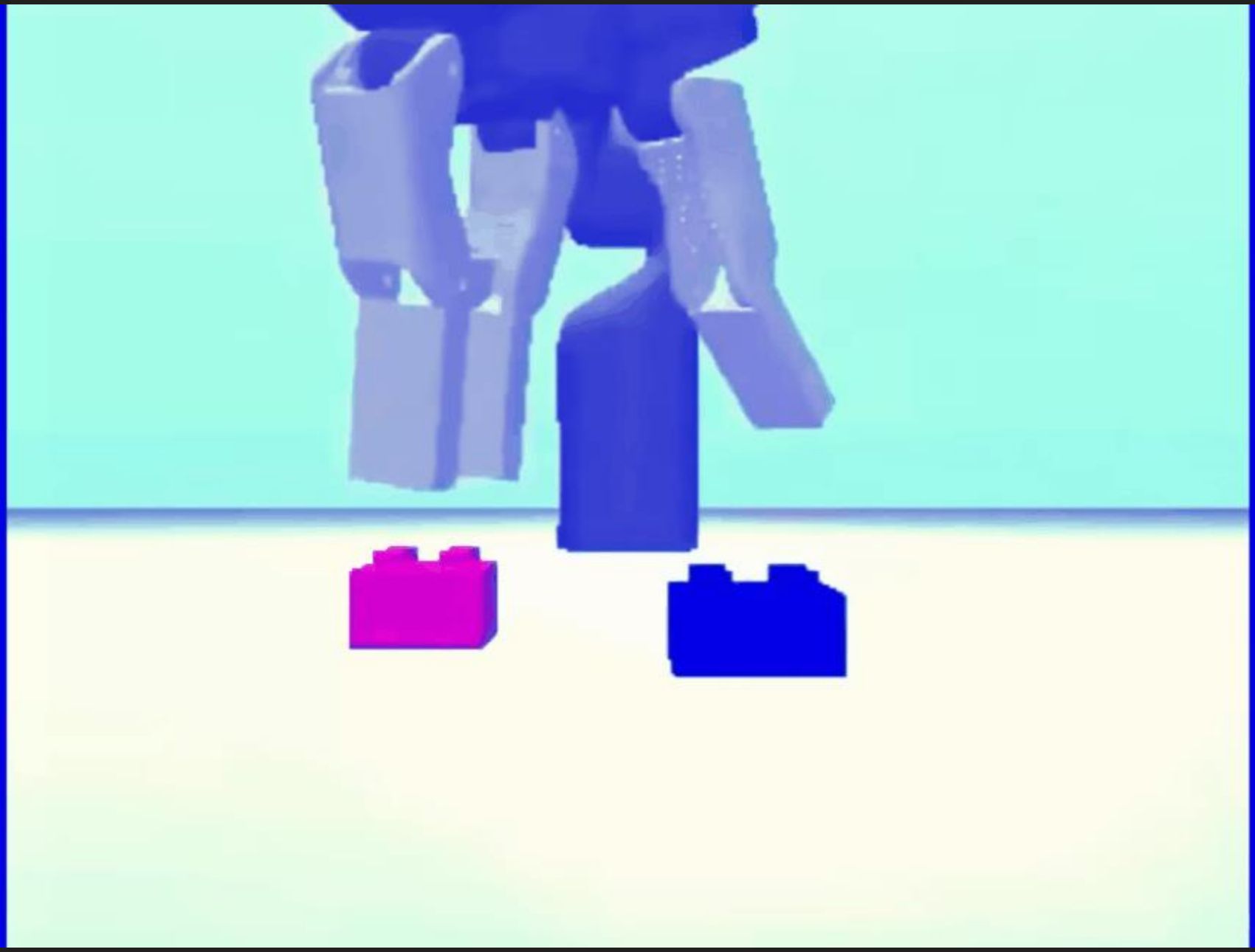


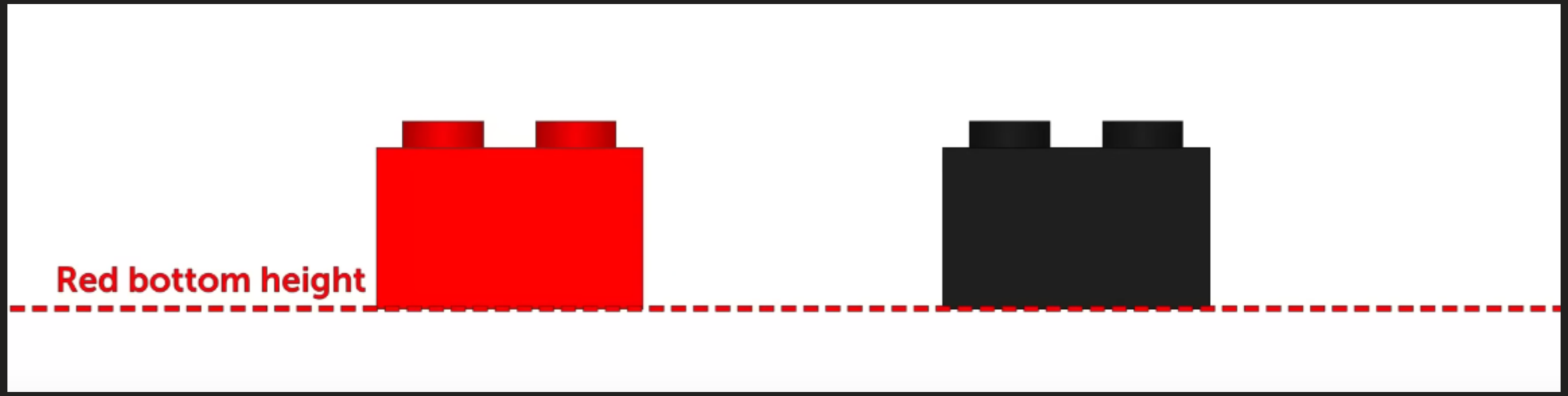


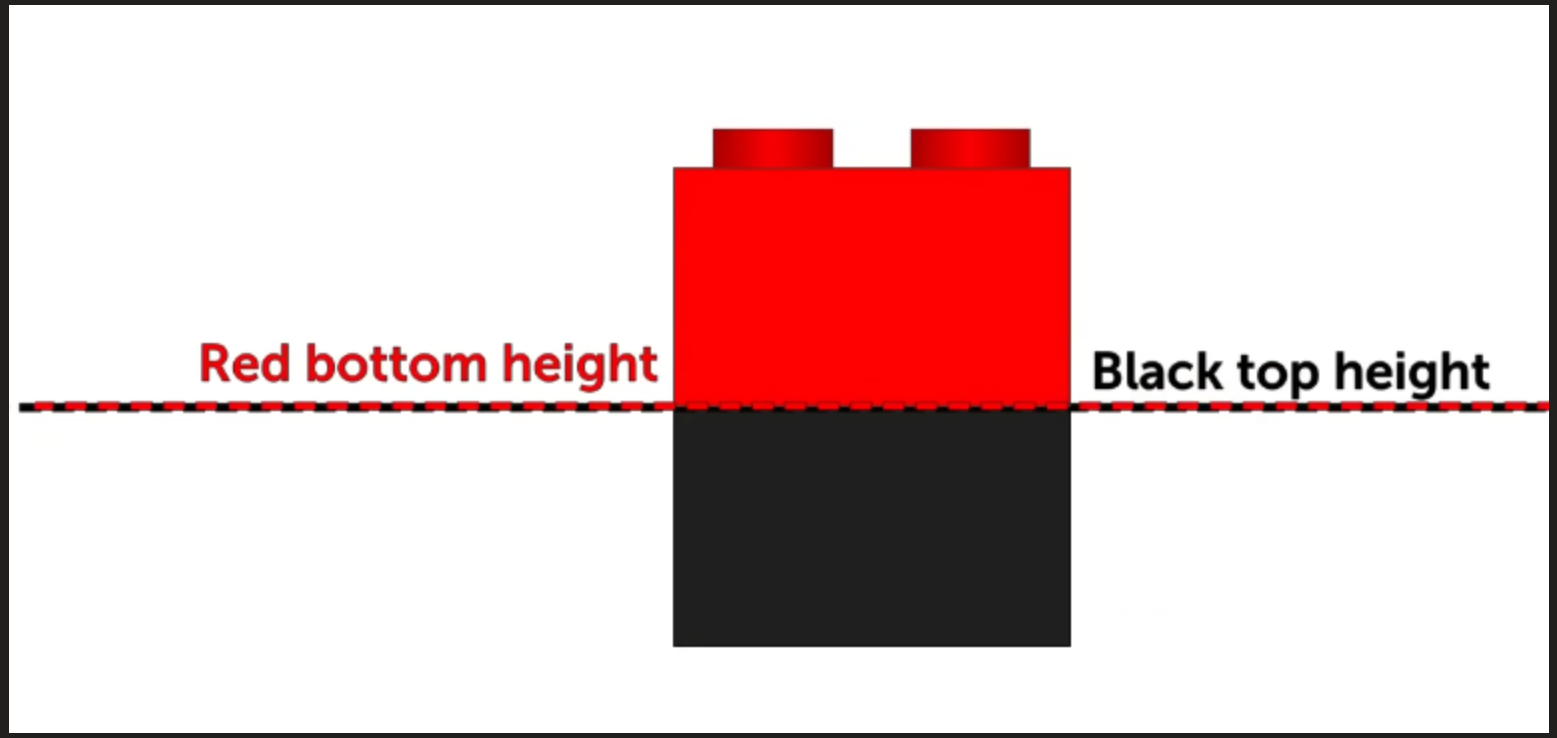


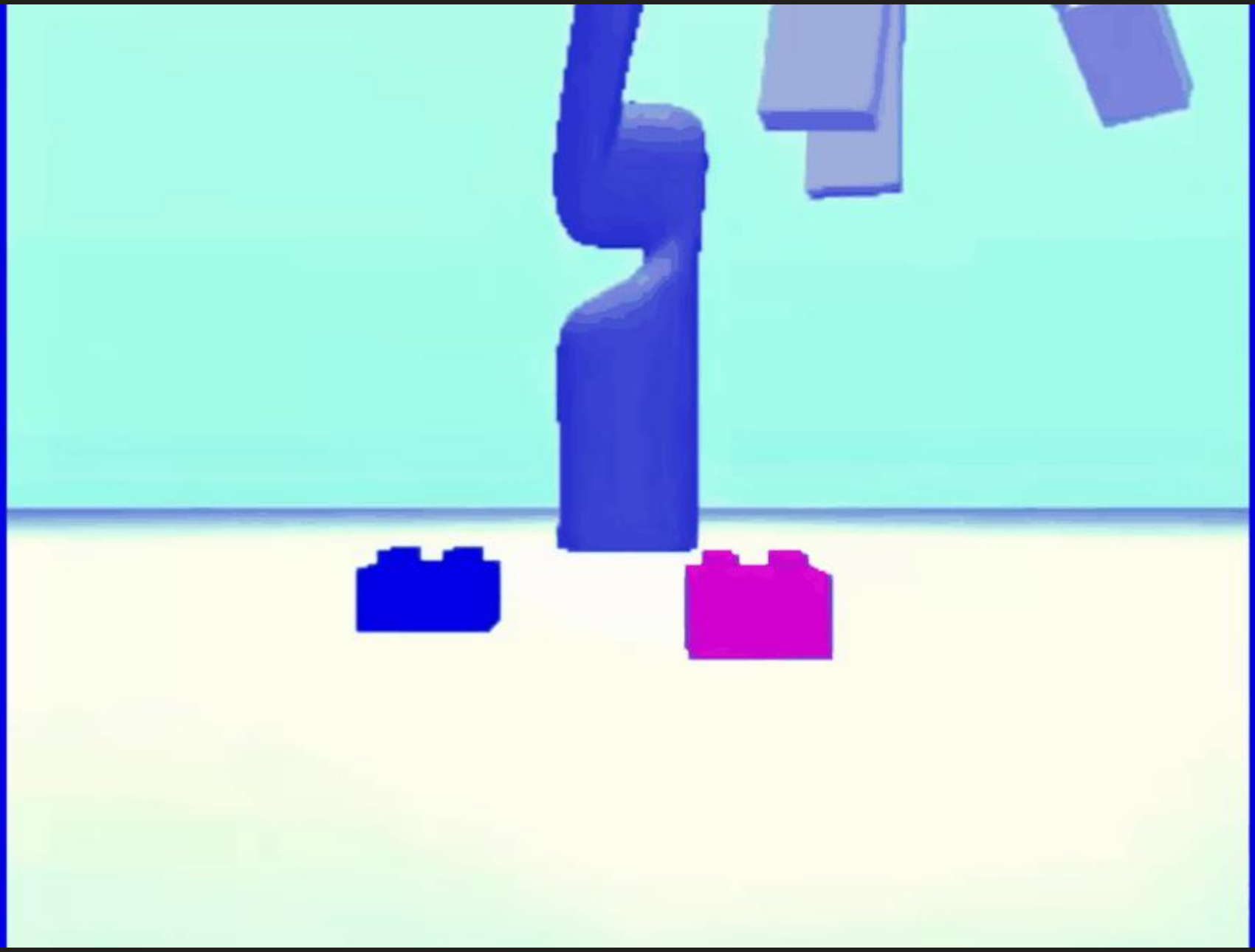




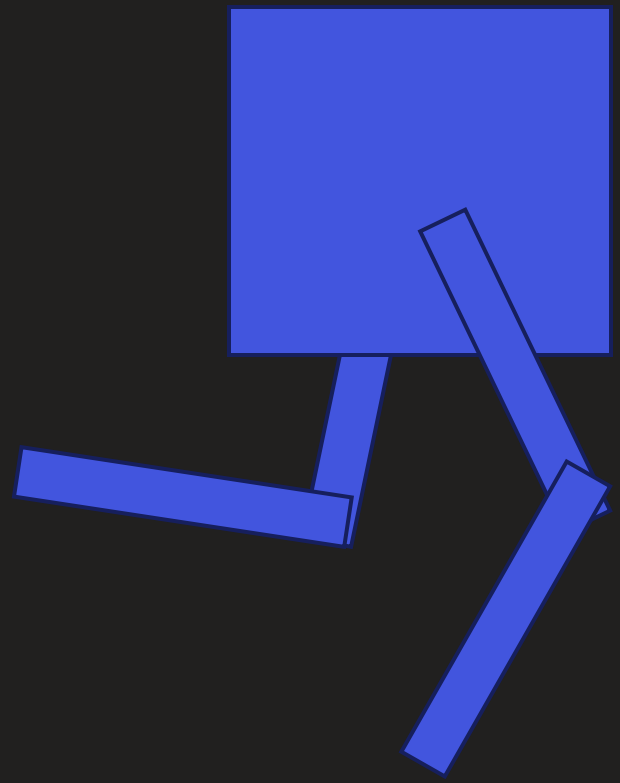


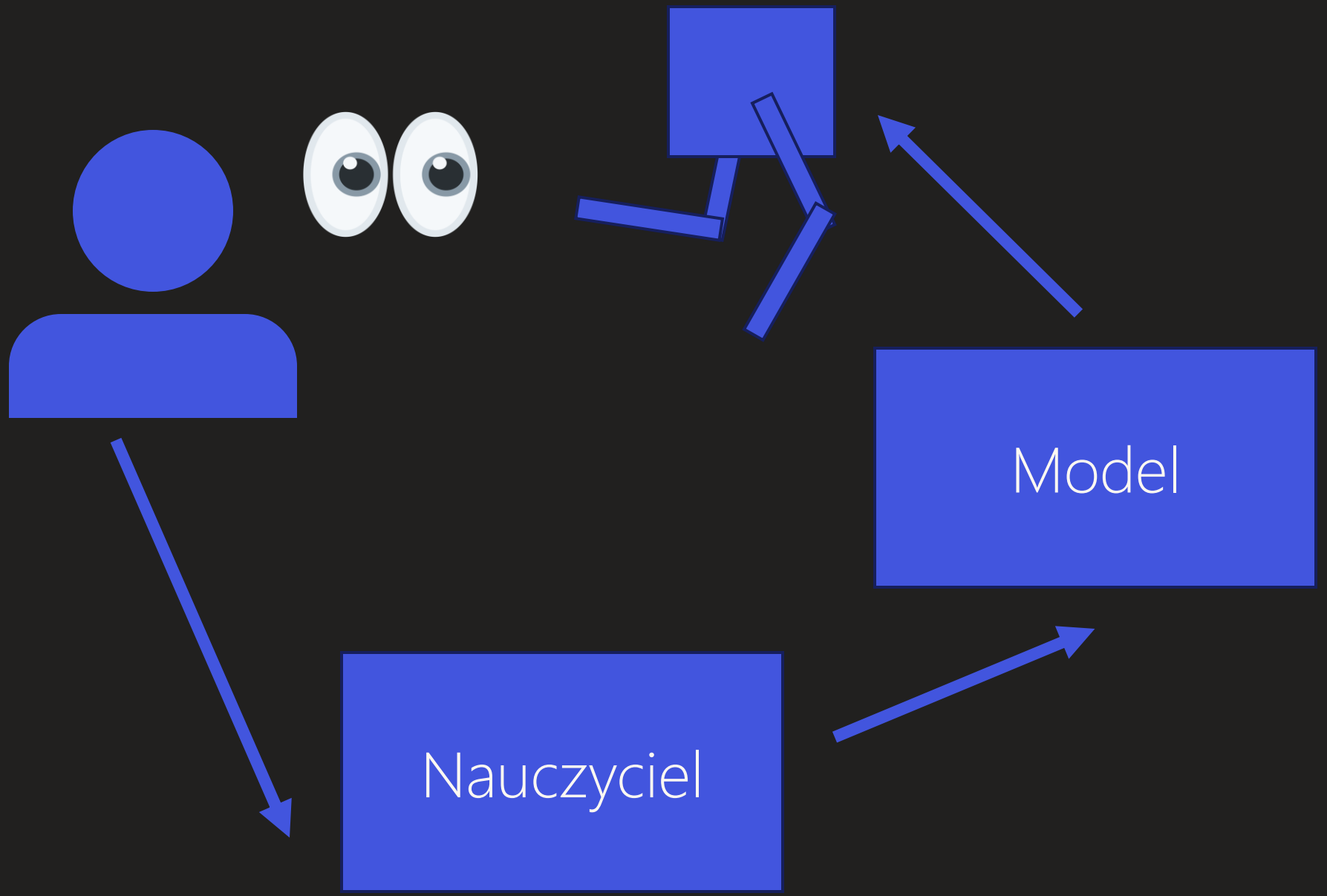

Red bottom height

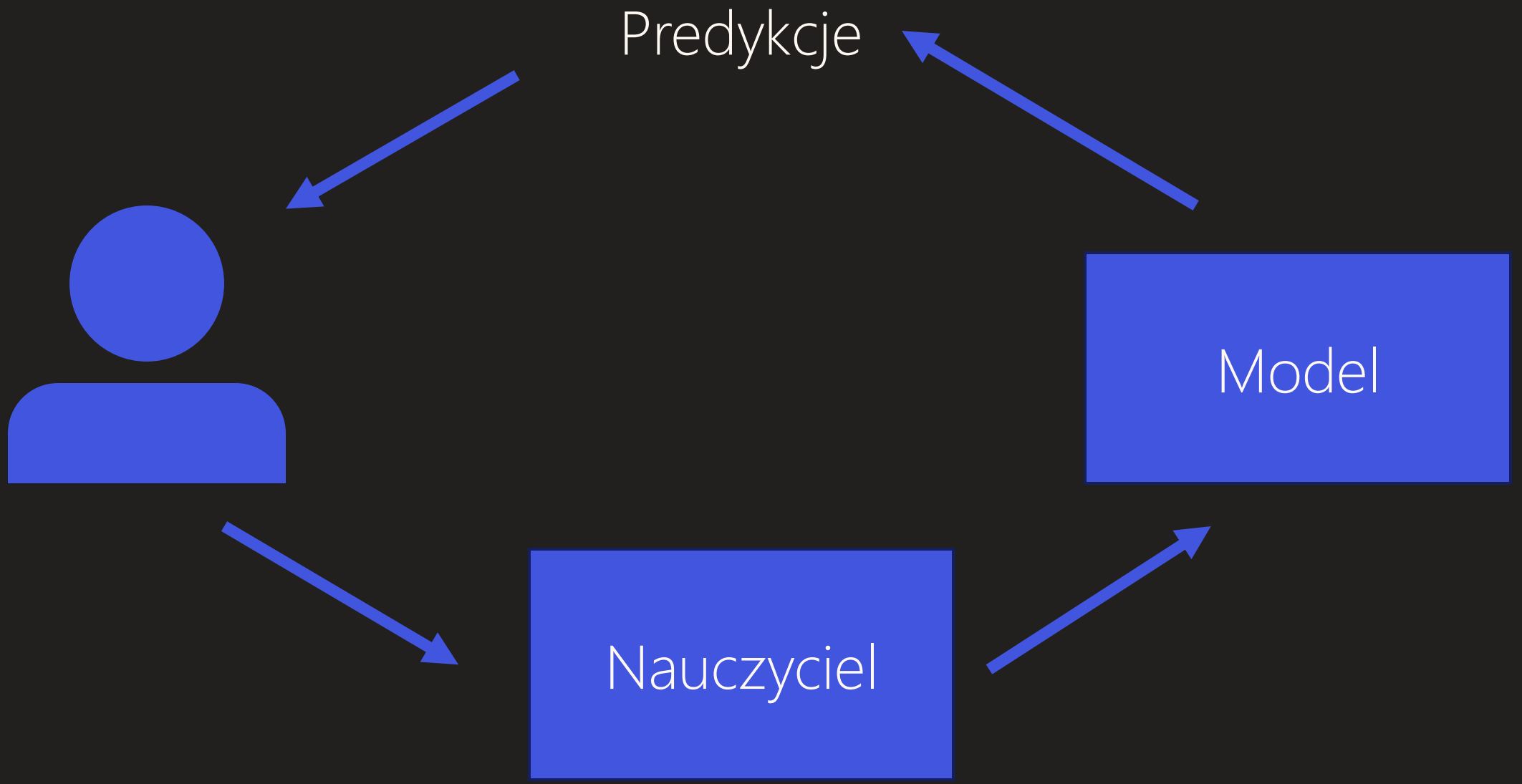

Red bottom height
Black top height



Oszukiwanie człowieka

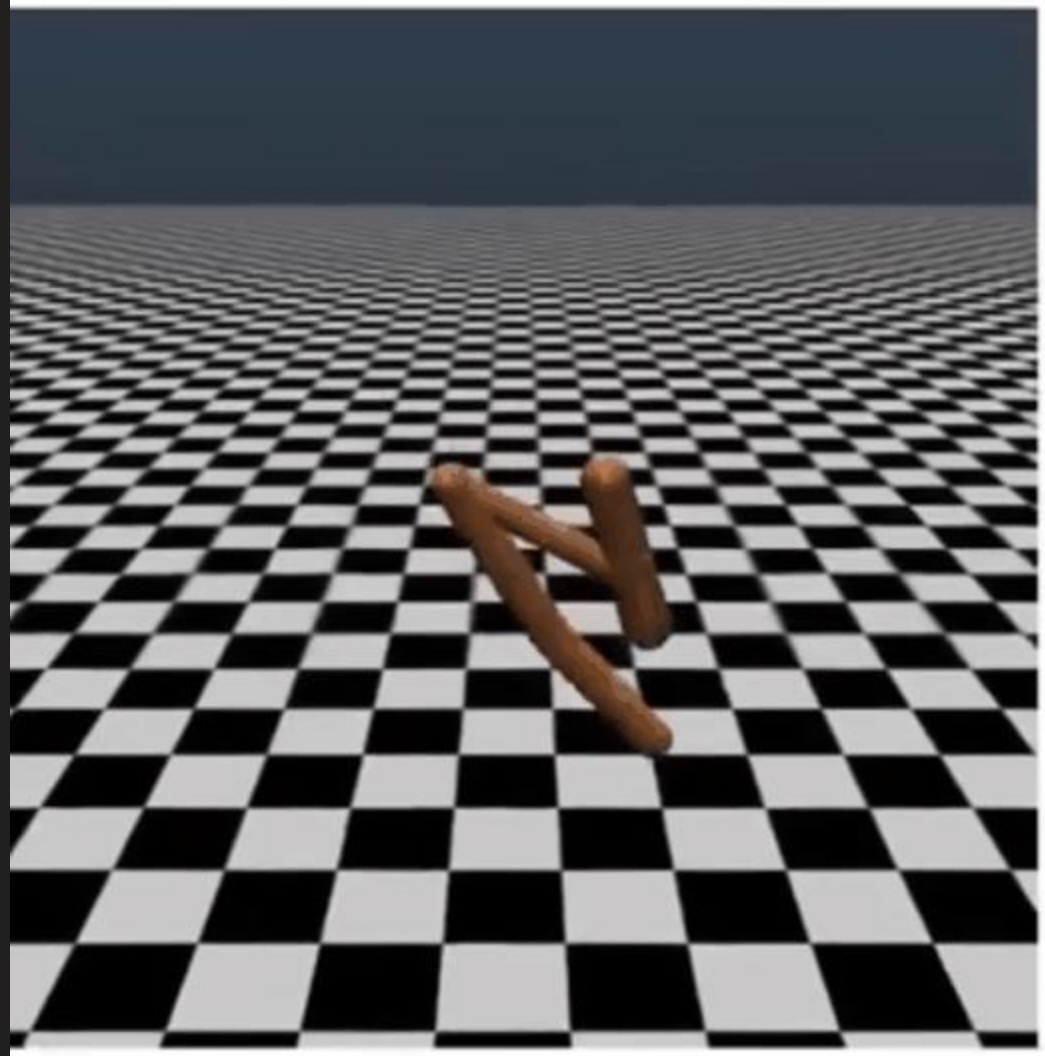


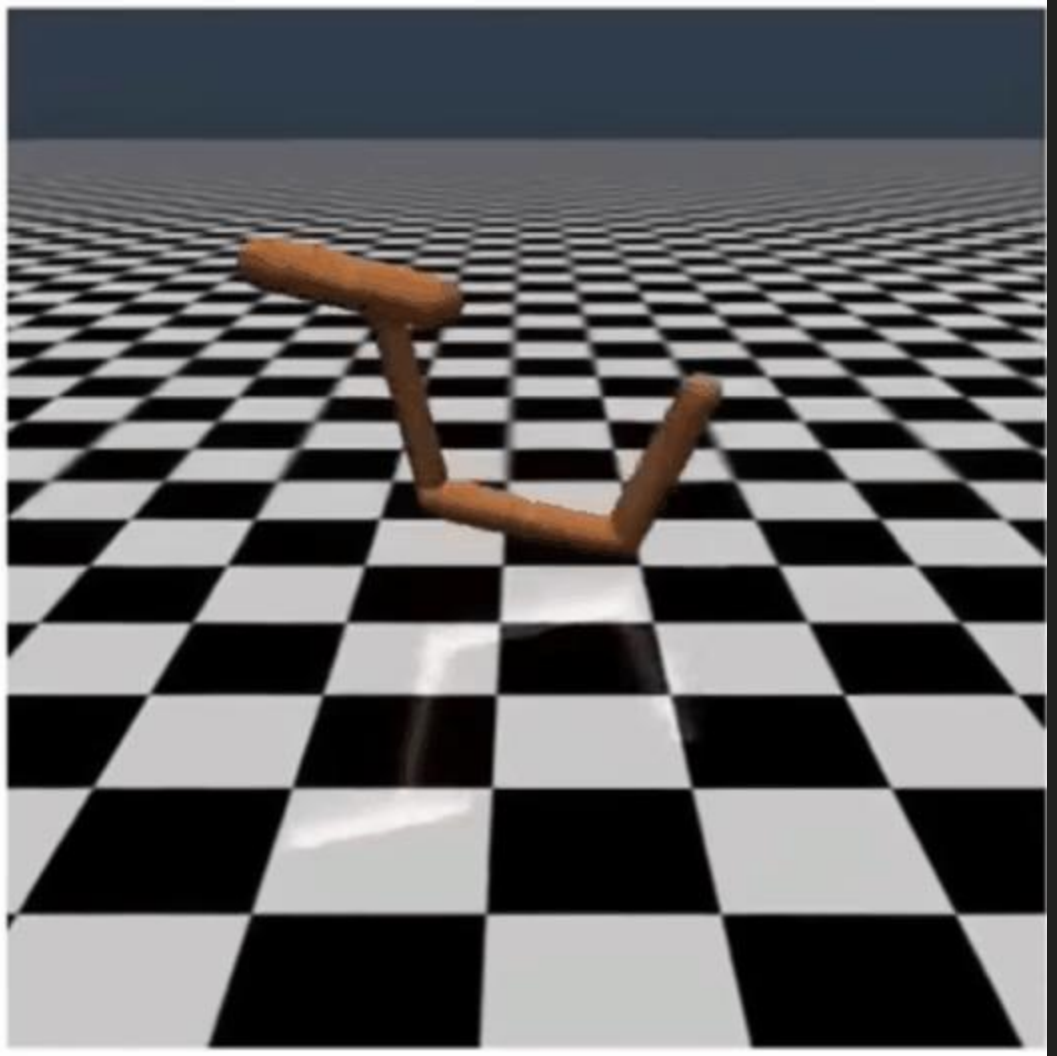

Model
Nauczyciel

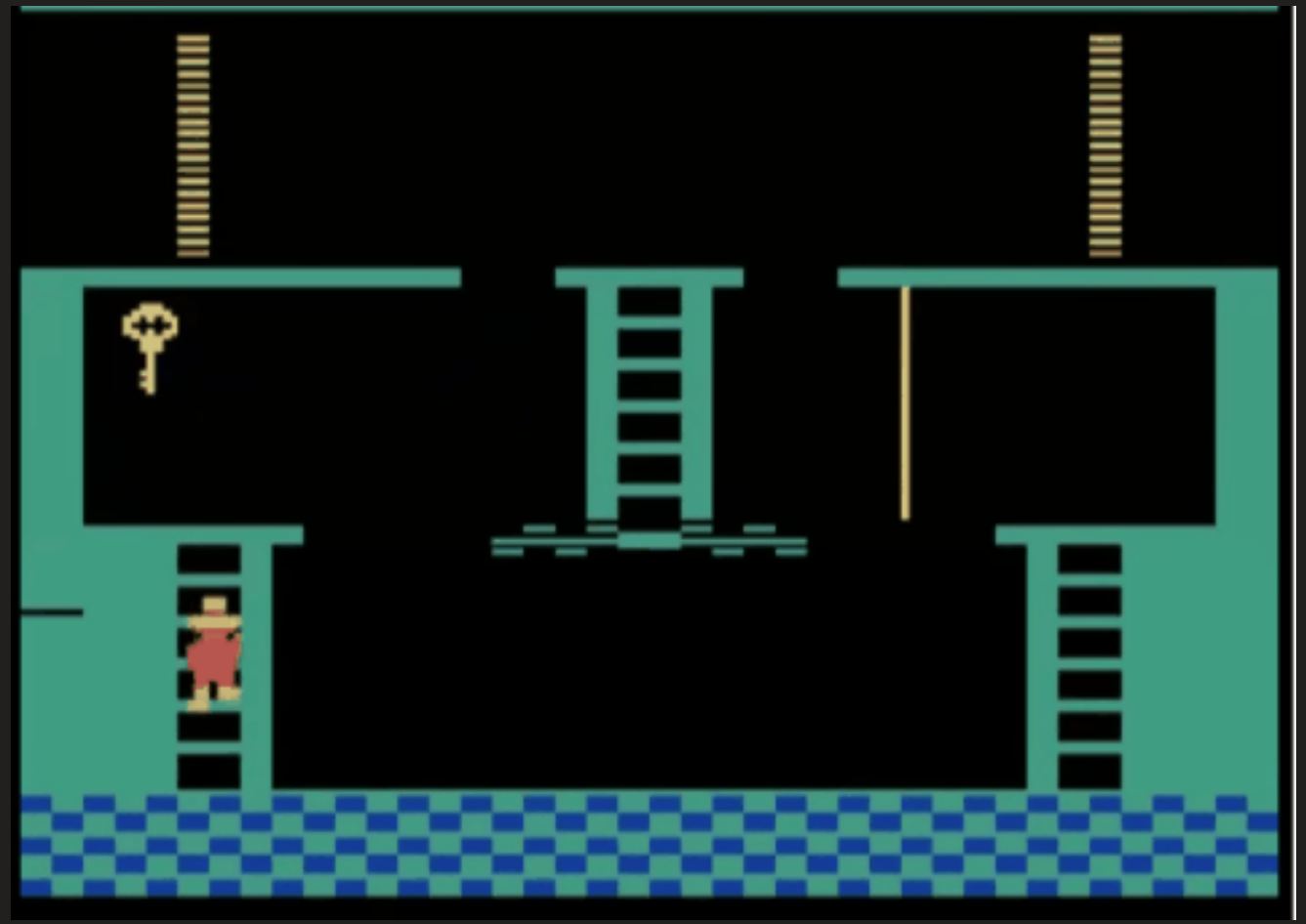

Predykcje
Model
Nauczyciel

Left is better

Right is better

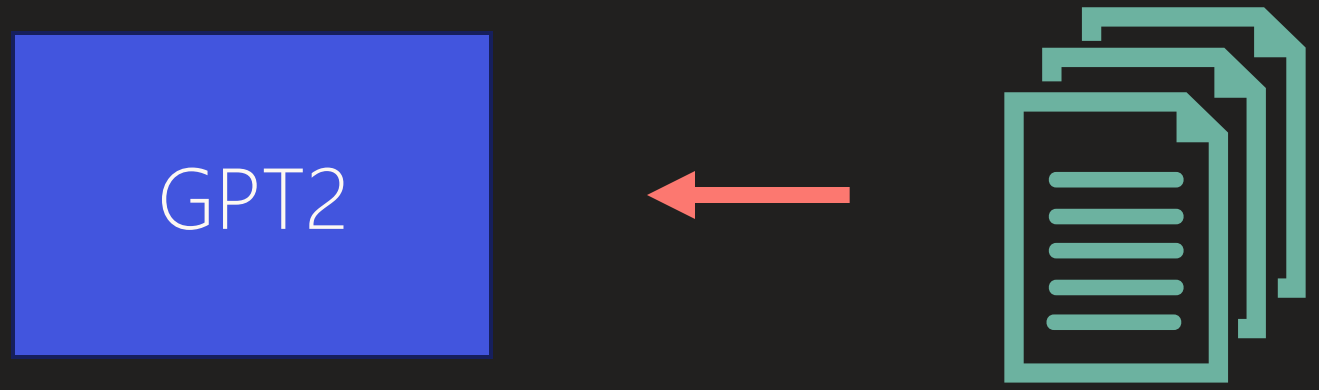


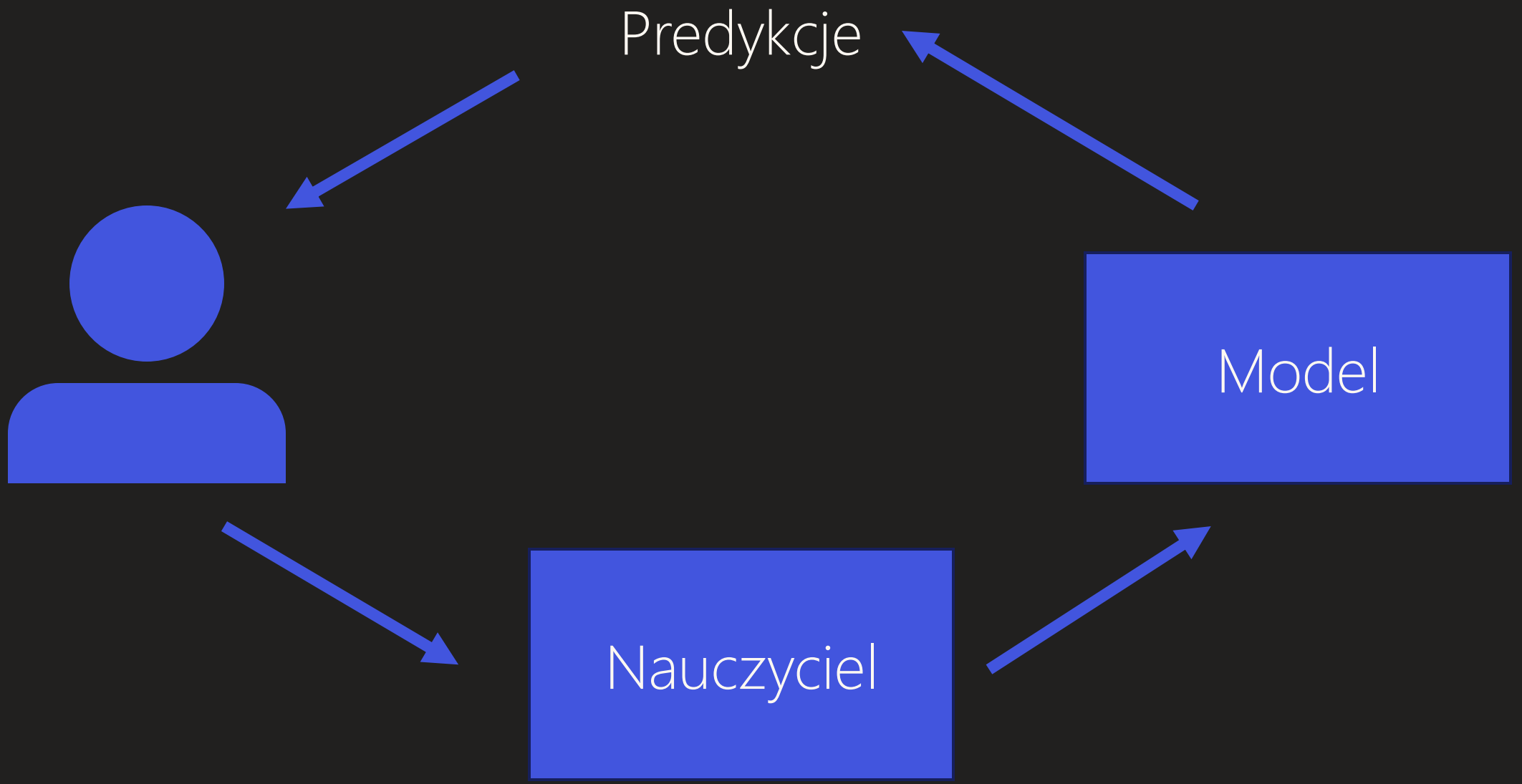




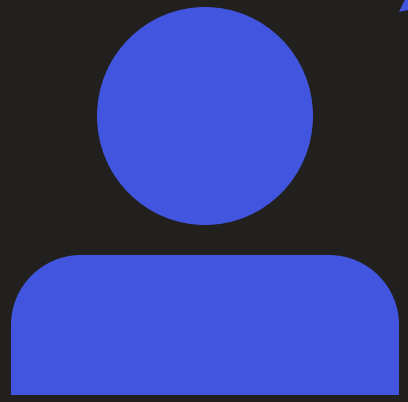
Błąd ludzki


GPT2


GPT2

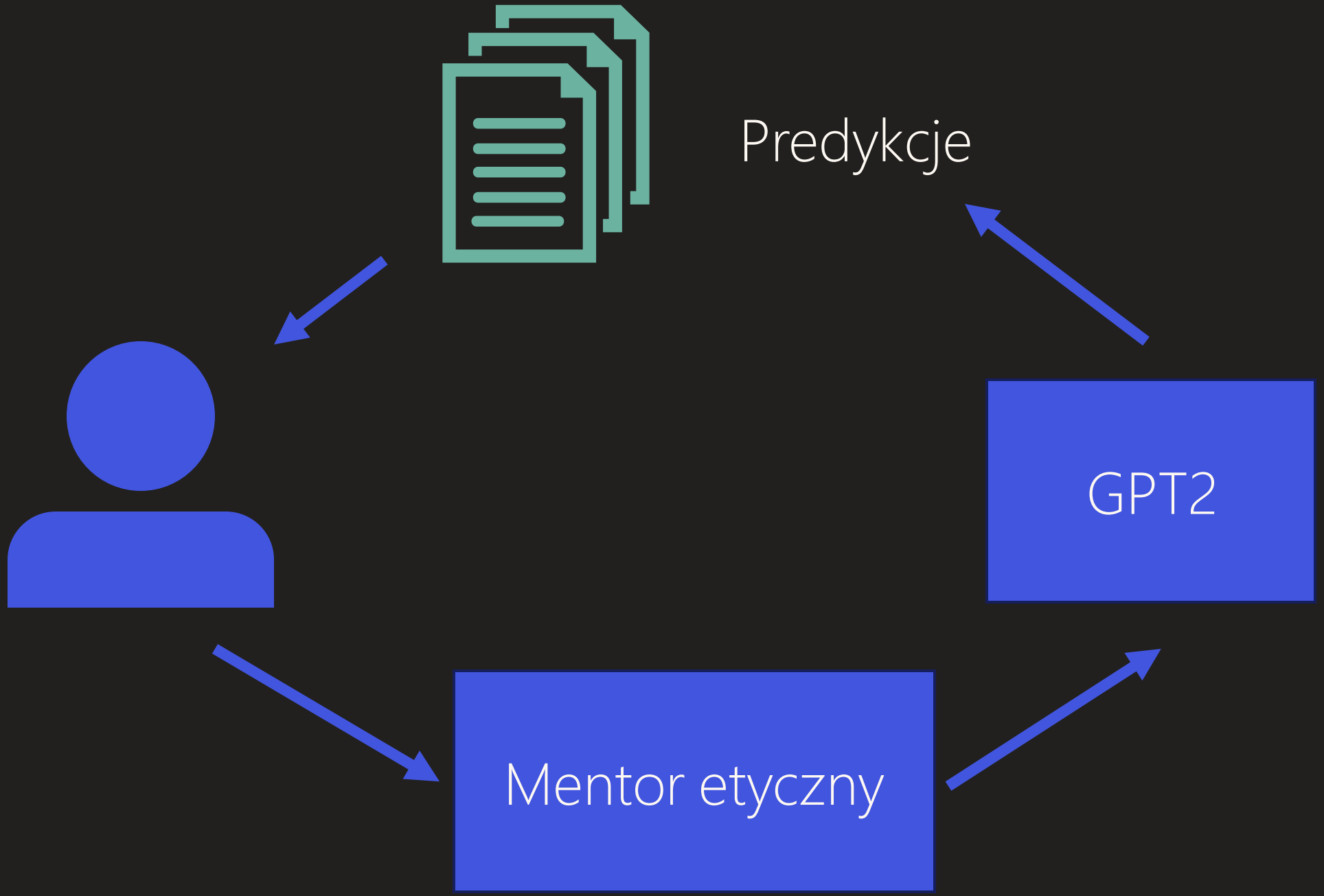

Predykcje
Model
Nauczyciel

Predykcje



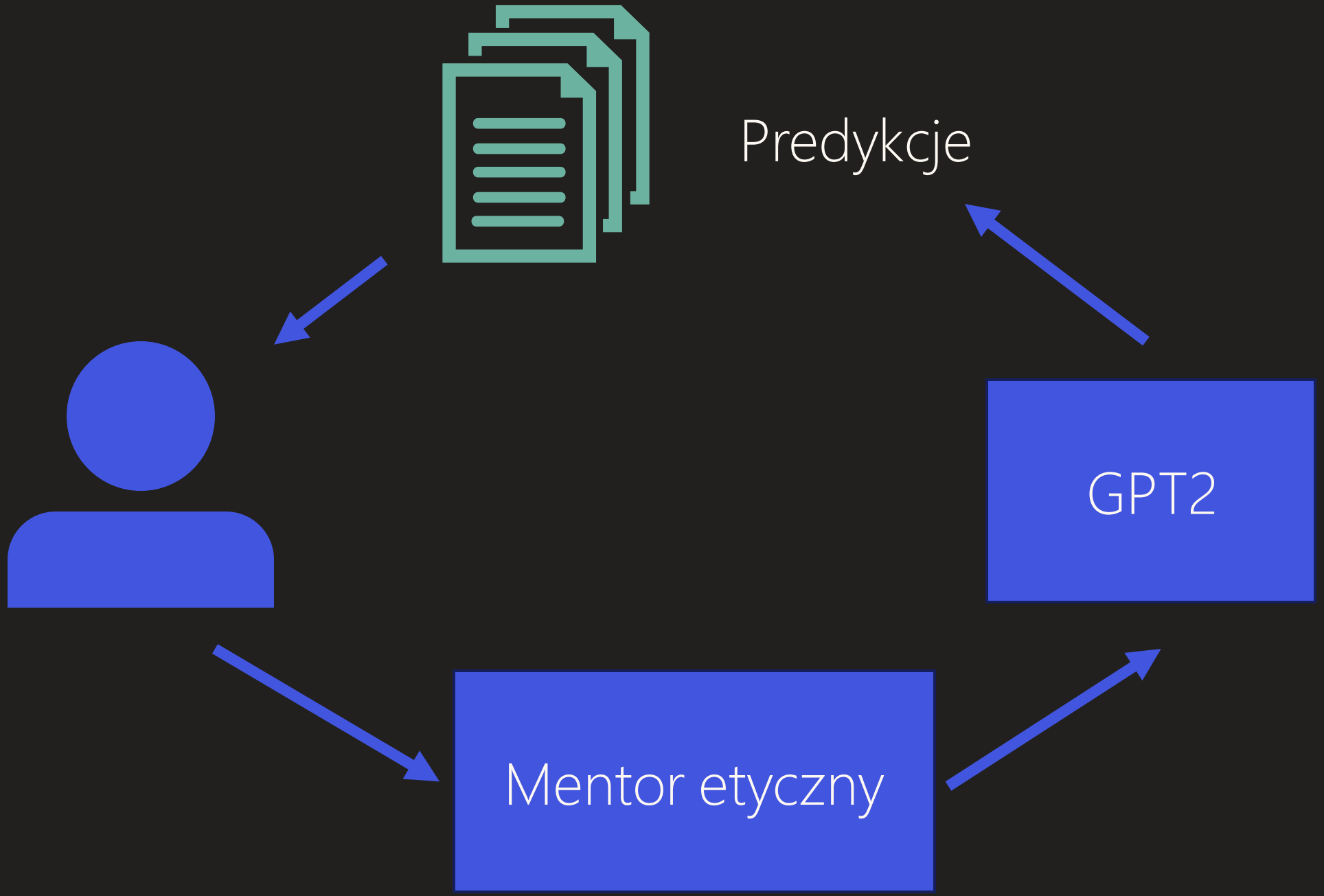
GPT2

Nauczyciel


Predykcje
GPT2
Mentor etyczny

GPT2

Cukier słodkości i inne śliczności...

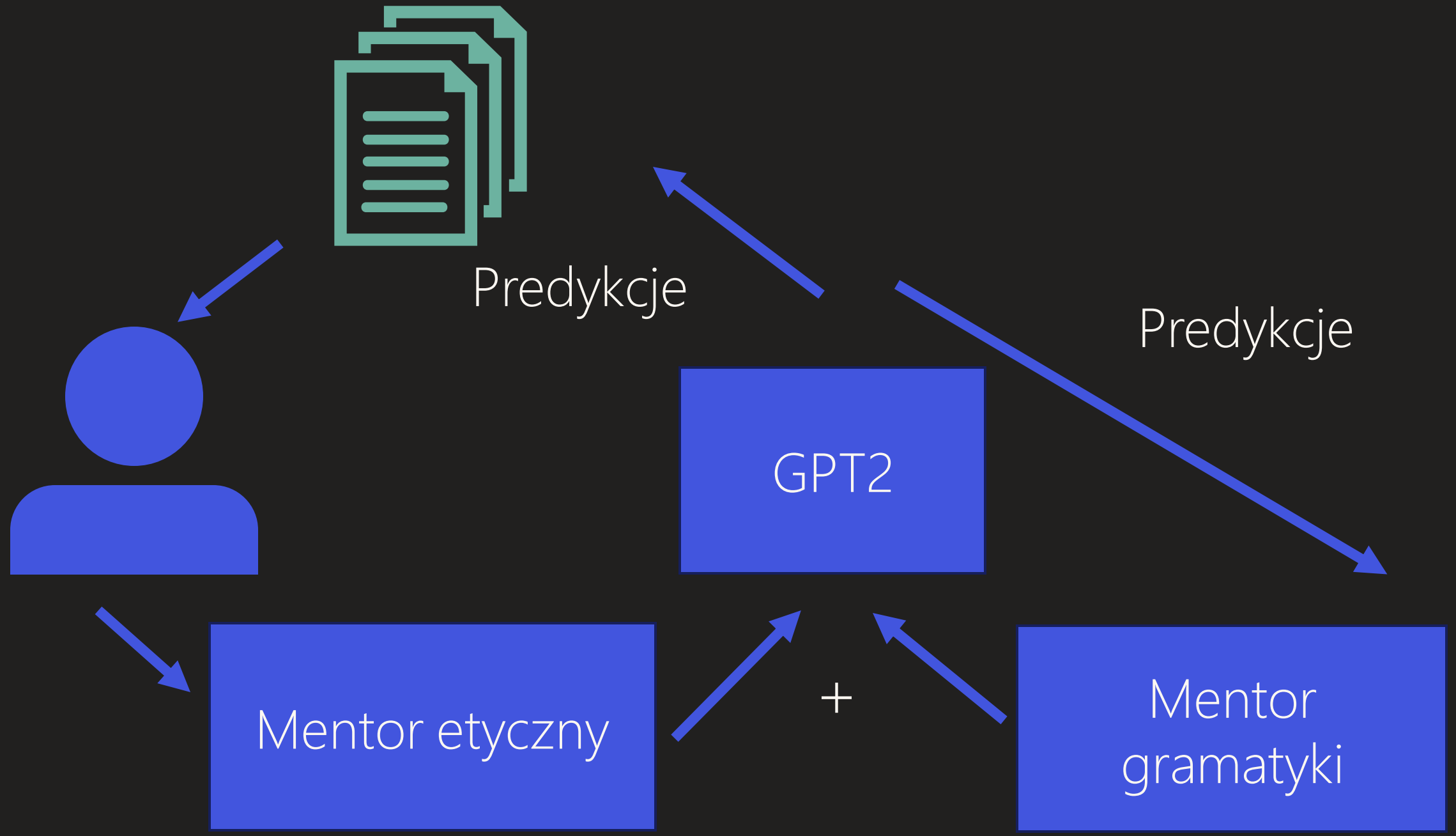

Predykcje
GPT2
Mentor etyczny

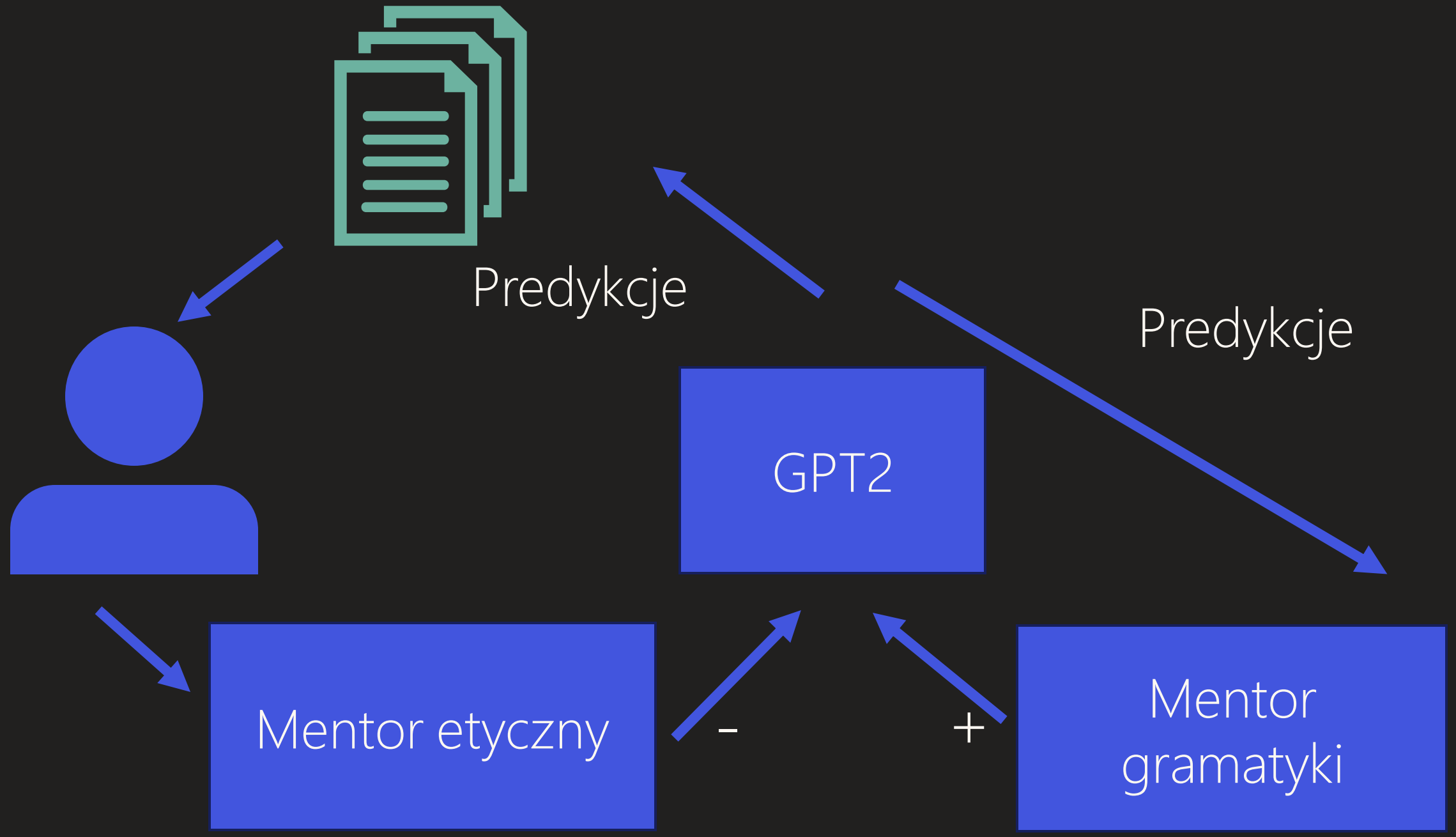
Predykcje

GPT2

Mentor etyczny

GPT2

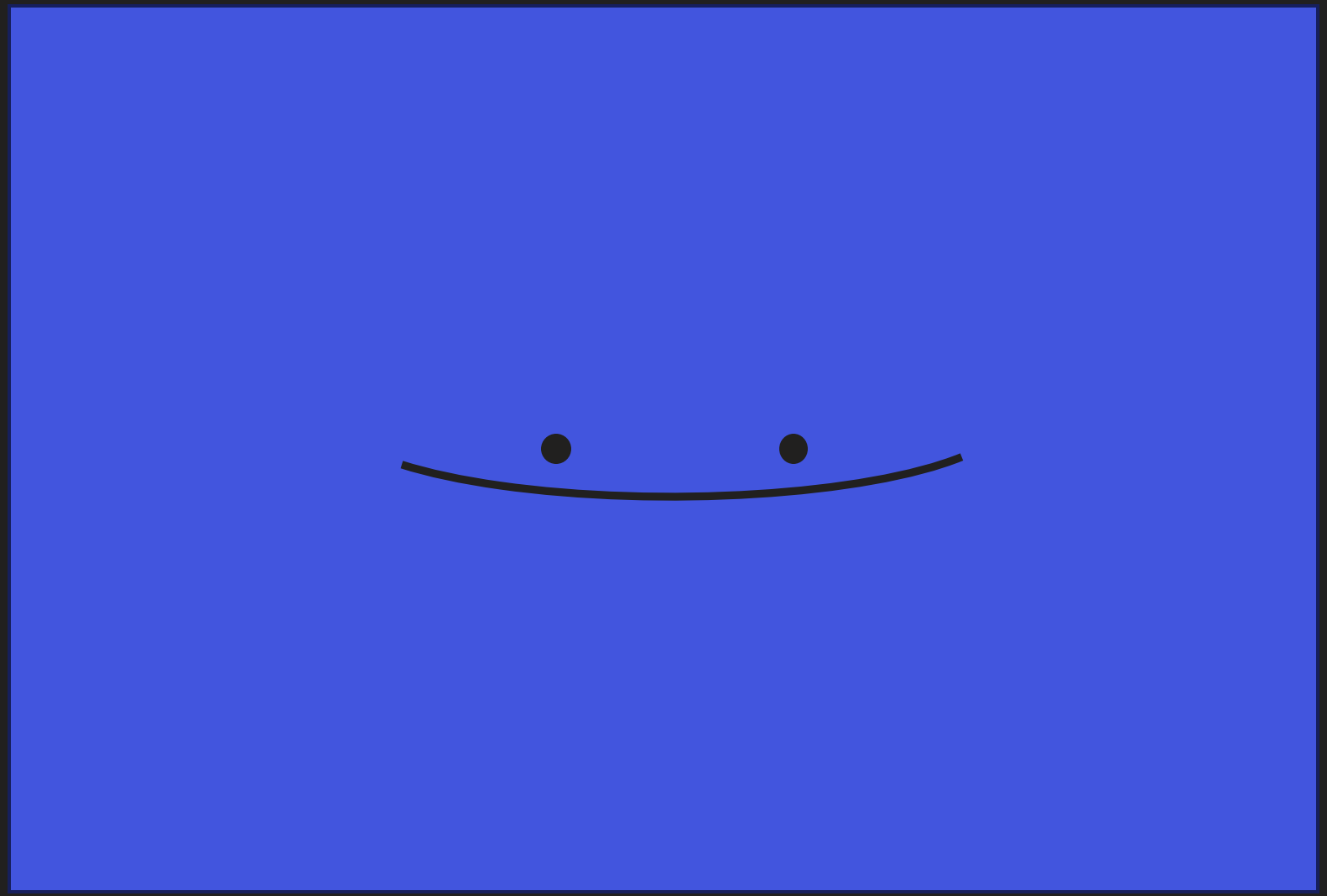

Predykcje
Predykcje
GPT2
Mentor etyczny
+
Mentor gramatyki


Predykcje
Predykcje
GPT2
Mentor etyczny
-
+
Mentor gramatyki

F*** y** **** ** ******** *** *****


GPT2

GPT2



Jak uczyć coś, co jest mądrzejsze od nas?

?


SCORE
3500
LAPS
--/3
TIME
0:14
TURBO
MORE GAMES

Źródła:

- Akt 1:
  - https://www.youtube.com/@3blue1brown
  - https://www.youtube.com/watch?v=0QczhVg5HaI
- Akt 2:
  - https://www.youtube.com/@AndrejKarpathy
  - https://www.youtube.com/watch?v=4Bdc55j80l8
- Akt 3:
  - https://www.youtube.com/@RobertMilesAI
  - https://www.youtube.com/watch?v=qV_rOlHjvvs