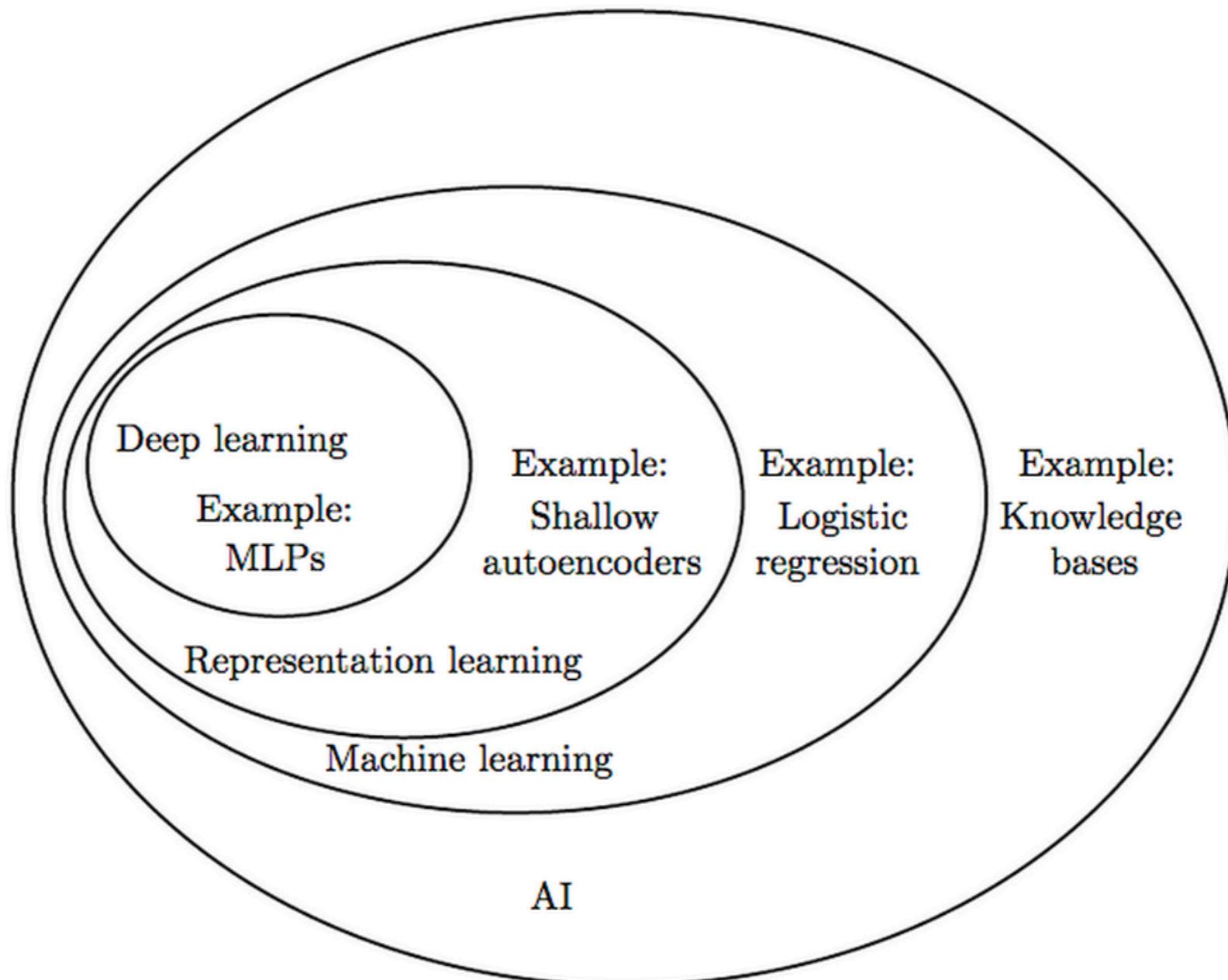


Deep Learning

Wstęp Teoretyczny

Co to jest Deep Learning?

Deep Learning vs Machine Learning

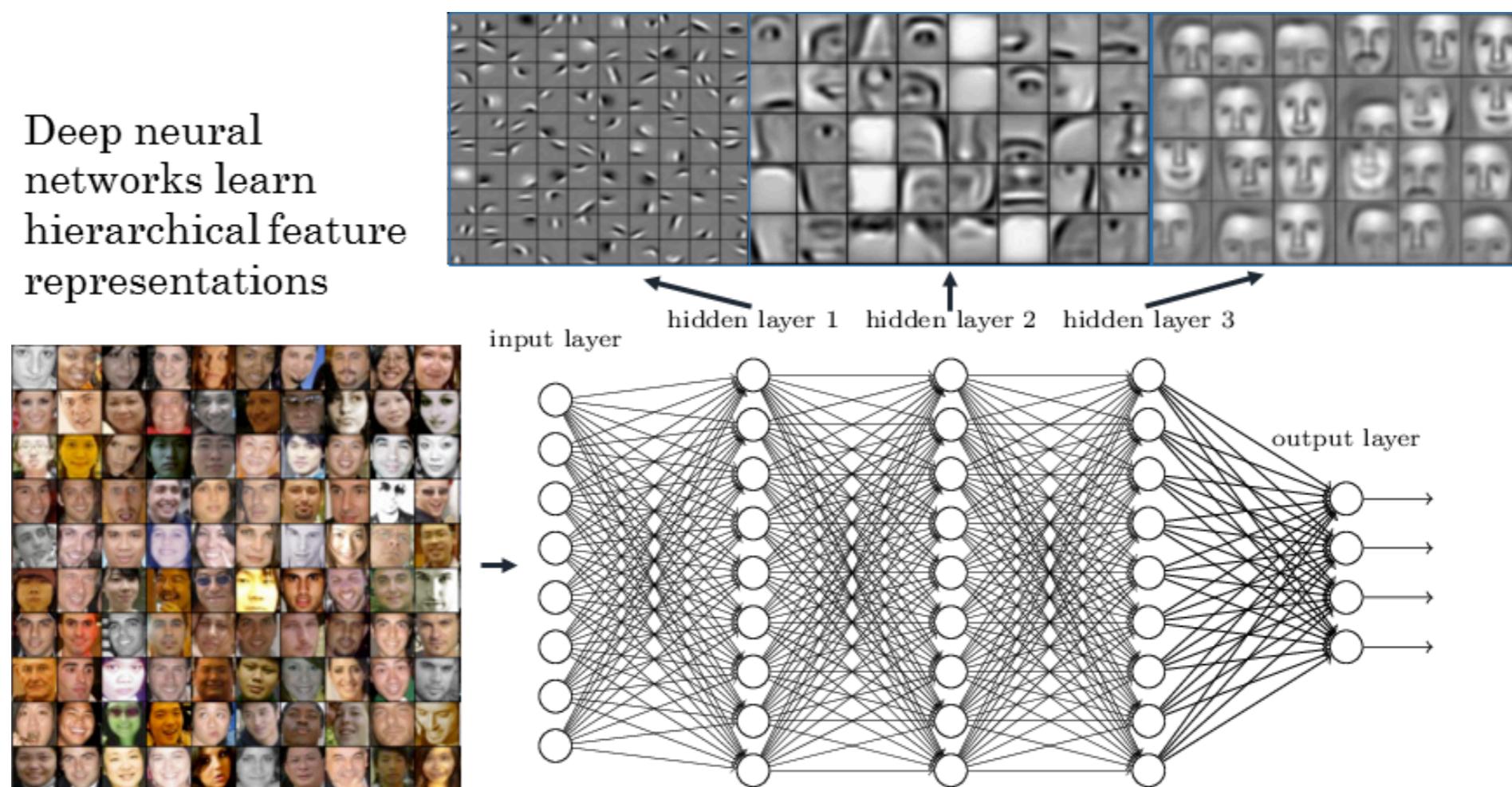


Co to jest Deep Learning?

Pierwsza warstwa wyłapuje najprostsze fakty dotyczące otrzymanych danych. Kolejne warstwy korzystają z wniosków poprzednich i tym samym są one w stanie nauczyć się bardziej skomplikowanych rzeczy. Ilość warstw Sieci Neuronowej nazywamy “głębokością”. Im więcej warstw tym głębsza sieć i większe możliwości.

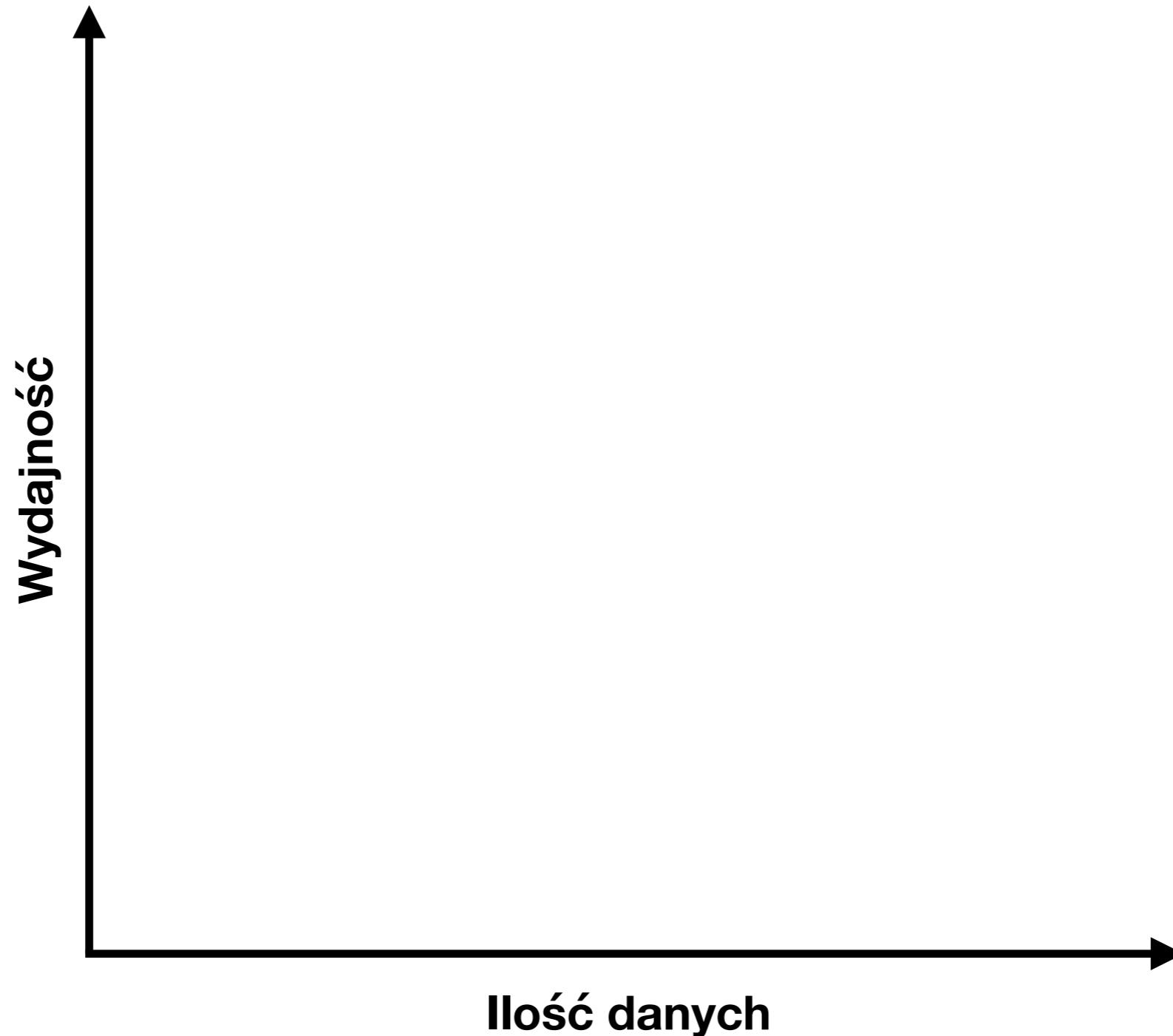
Co to jest Deep Learning?

Pierwsza warstwa wyłapuje najprostsze fakty dotyczące otrzymanych danych. Kolejne warstwy korzystają z wniosków poprzednich i tym samym są one w stanie nauczyć się bardziej skomplikowanych rzeczy. Ilość warstw Sieci Neuronowej nazywamy “głębokością”. Im więcej warstw tym głębsza sieć i większe możliwości.



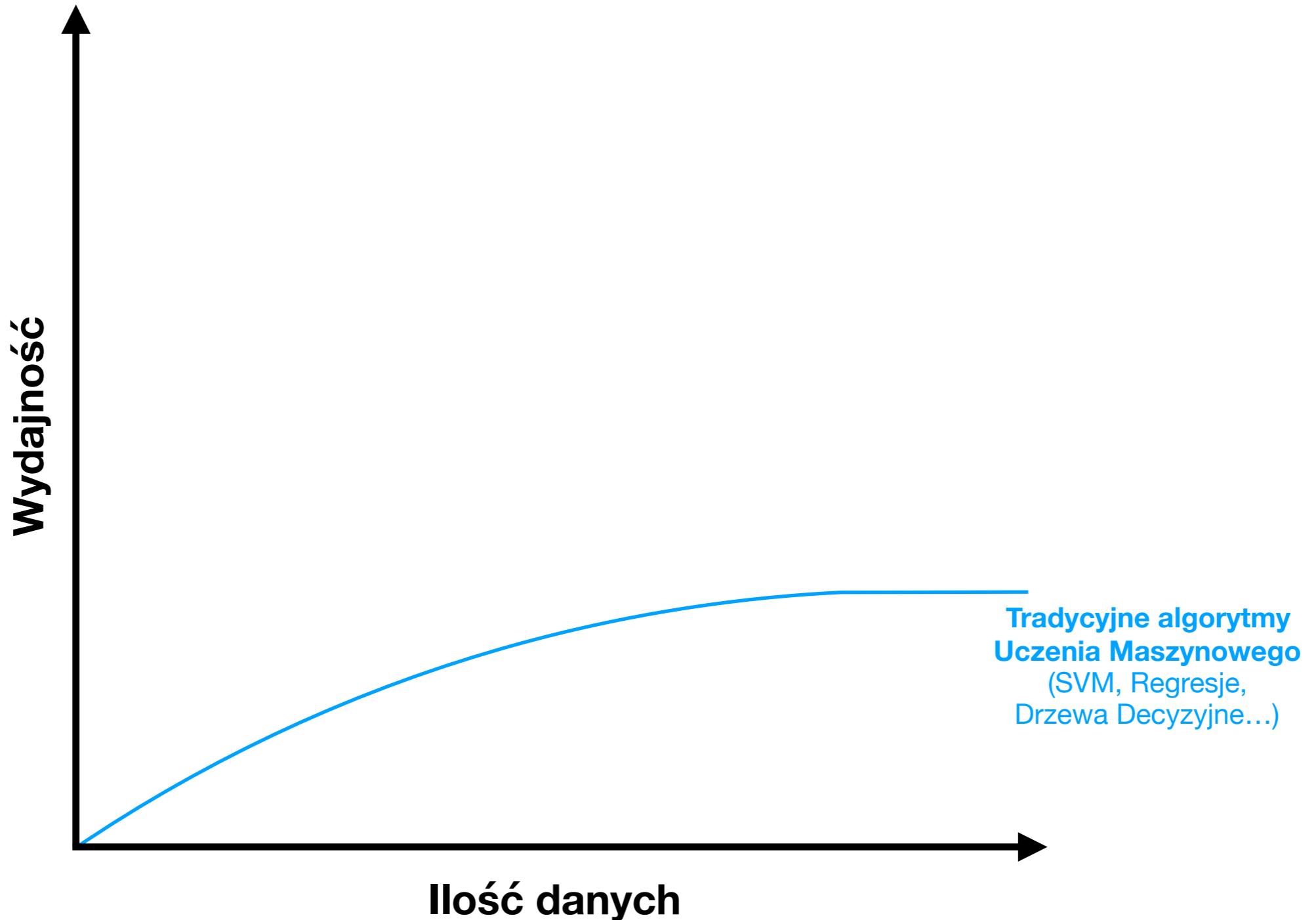
Deep Learning vs Machine Learning

Sposób przedstawienia Andrew Ng



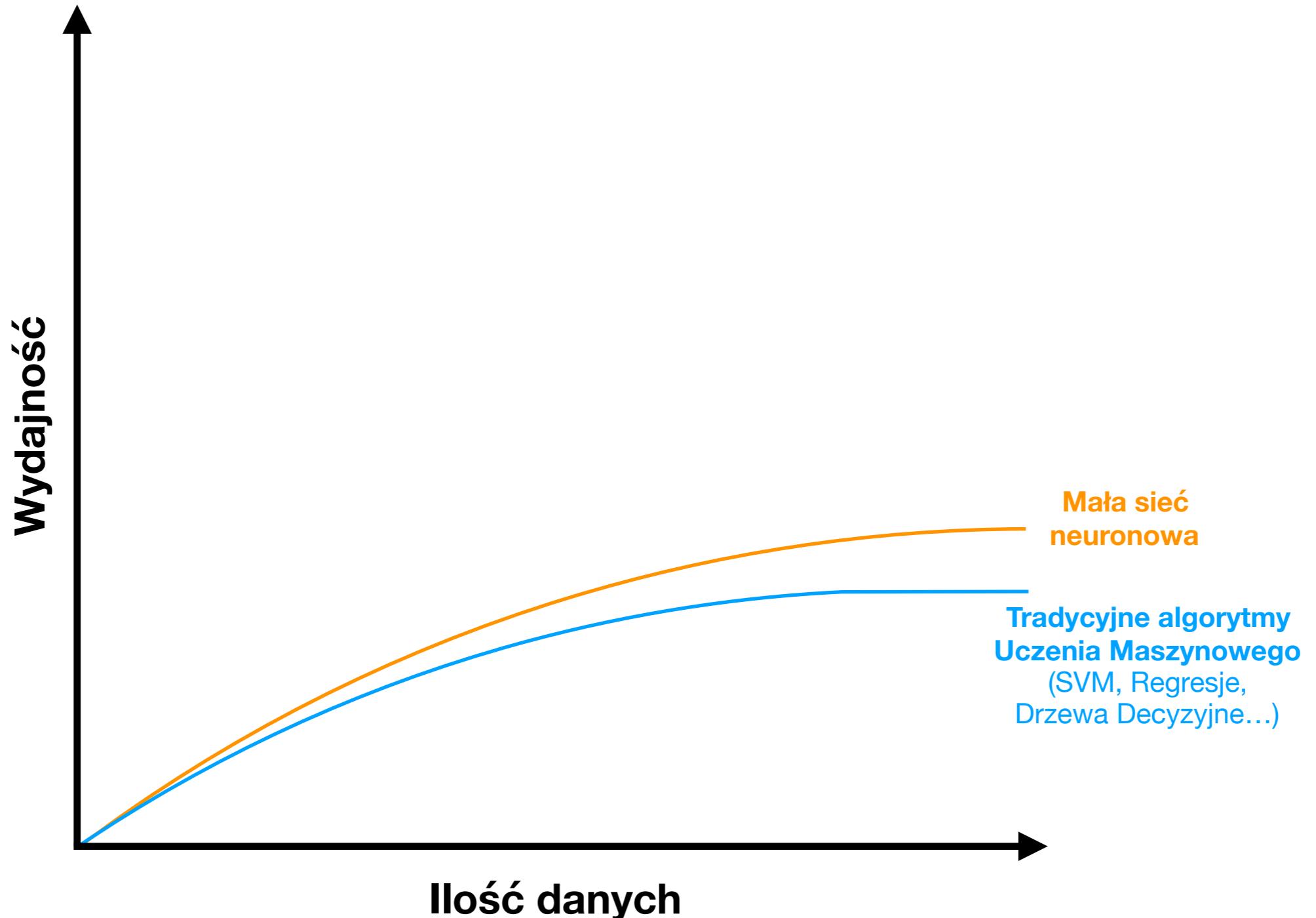
Deep Learning vs Machine Learning

Sposób przedstawienia Andrew Ng



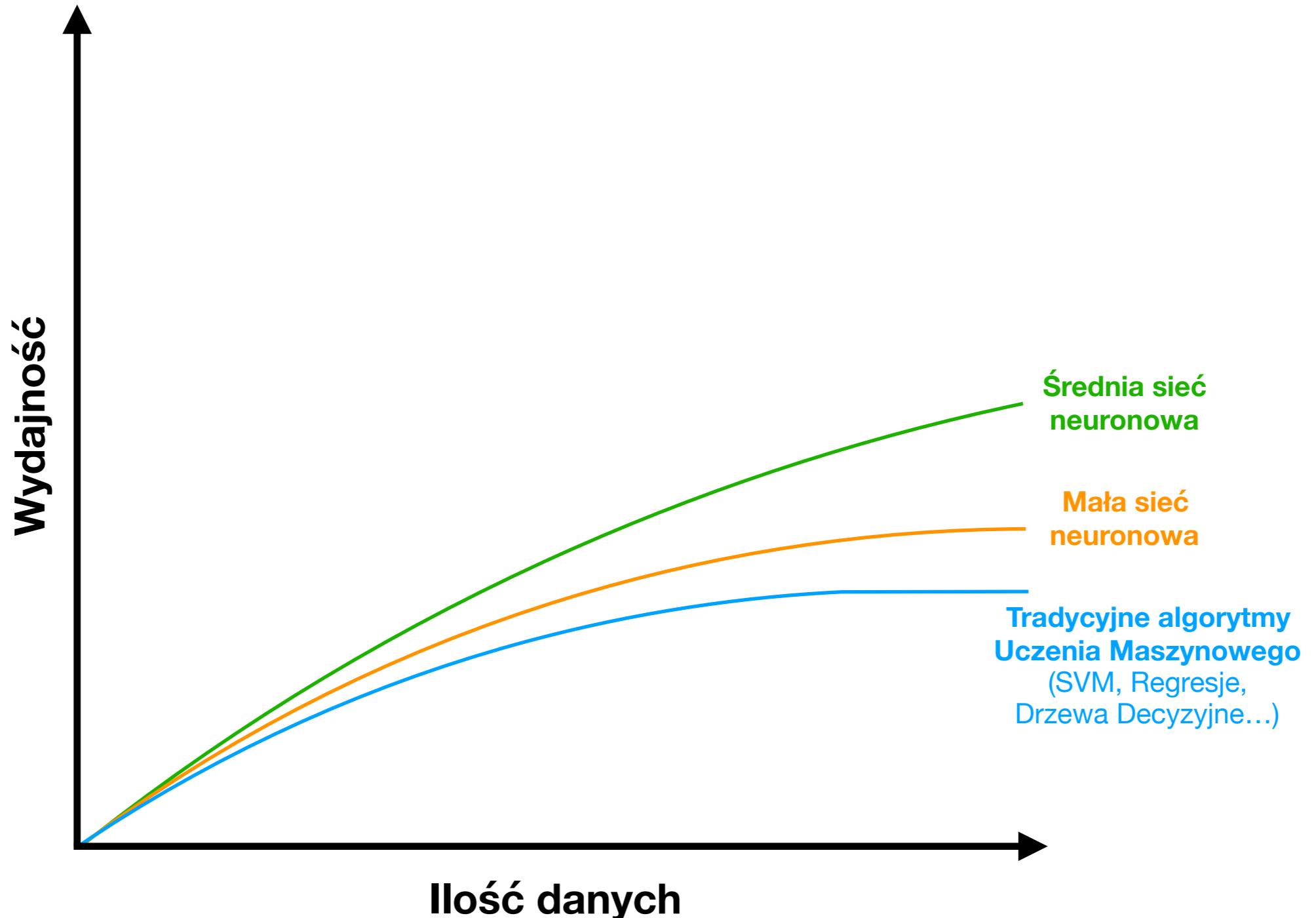
Deep Learning vs Machine Learning

Sposób przedstawienia Andrew Ng



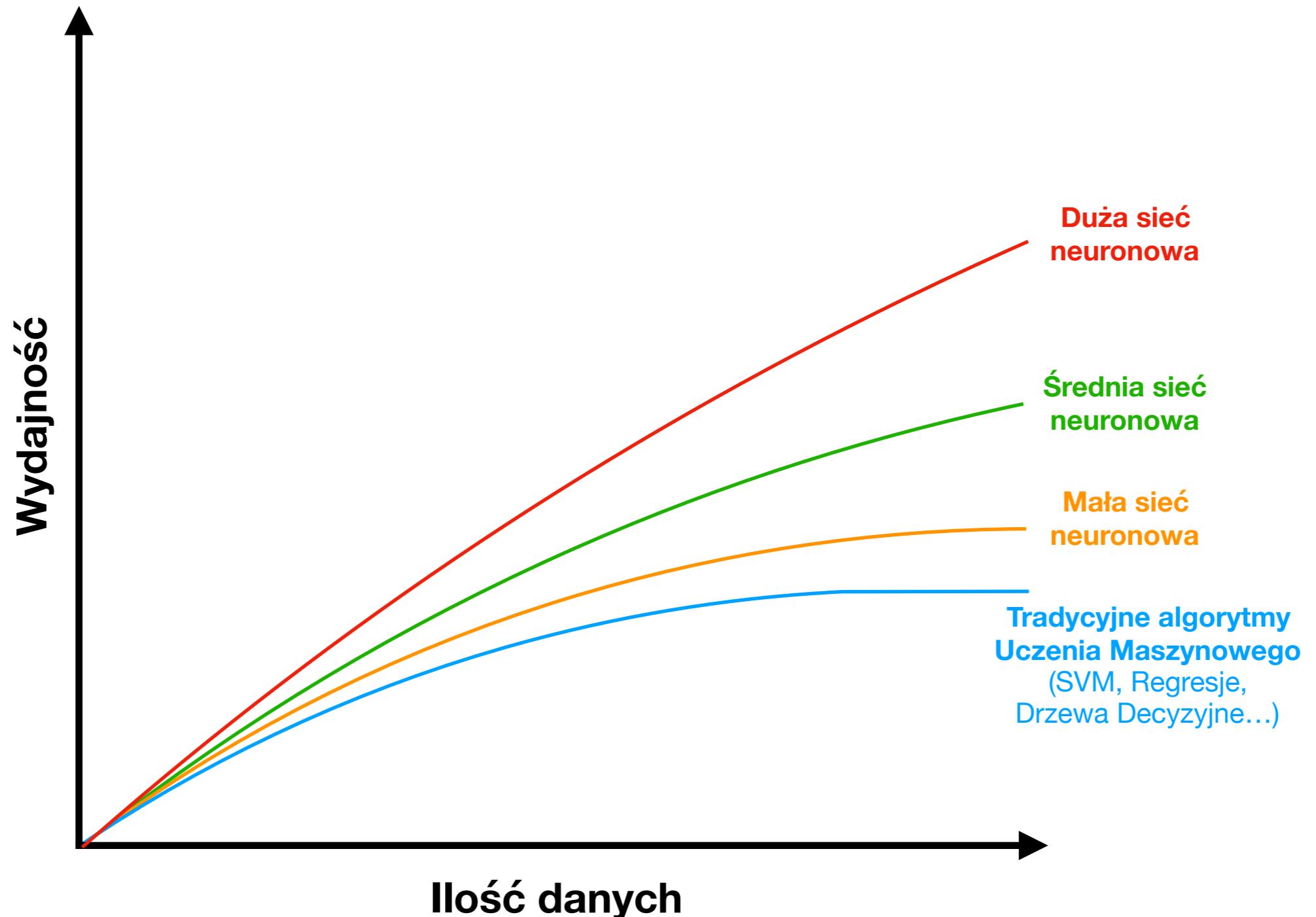
Deep Learning vs Machine Learning

Sposób przedstawienia Andrew Ng

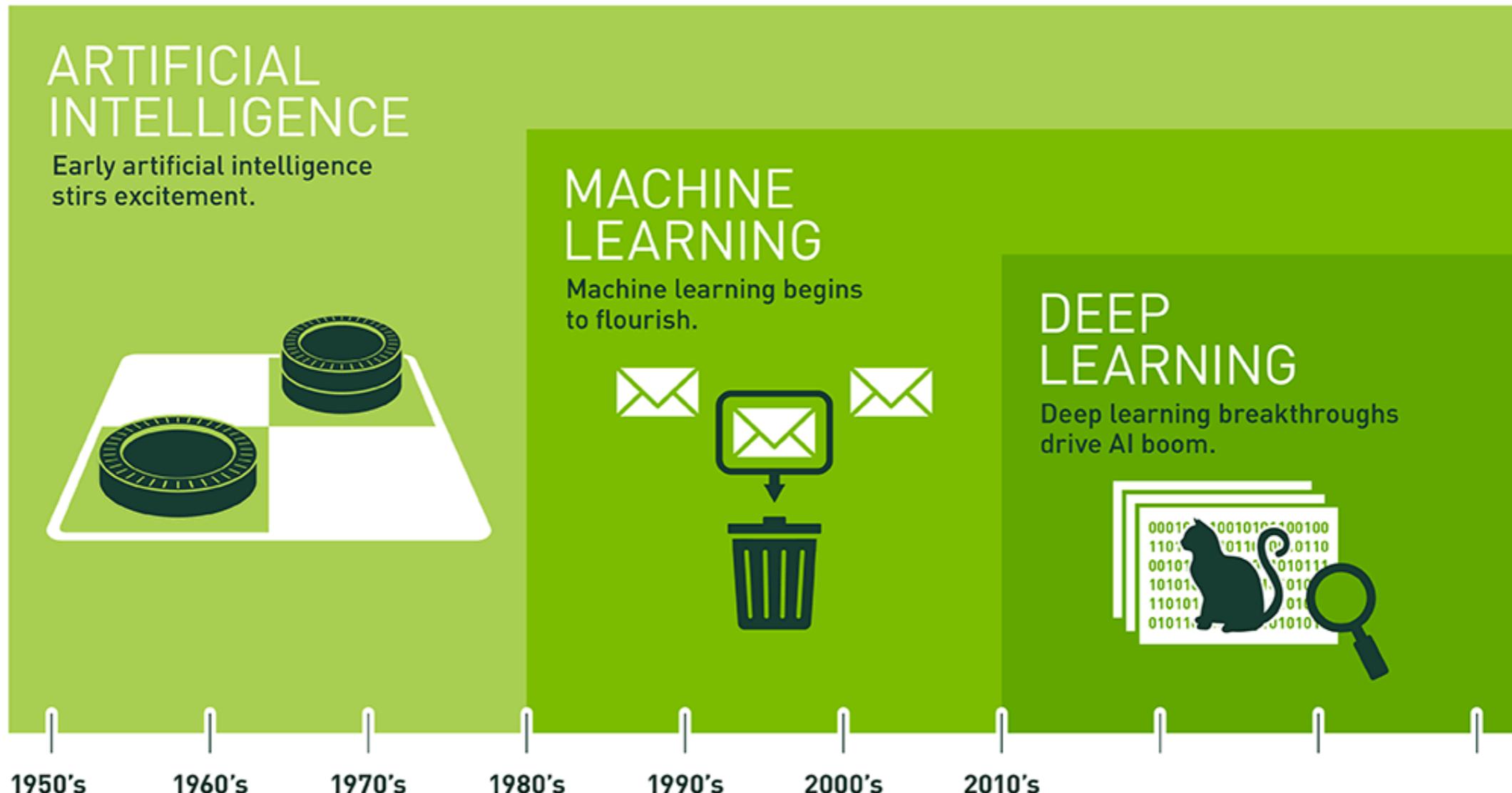


Deep Learning vs Machine Learning

Sposób przedstawienia Andrew Ng

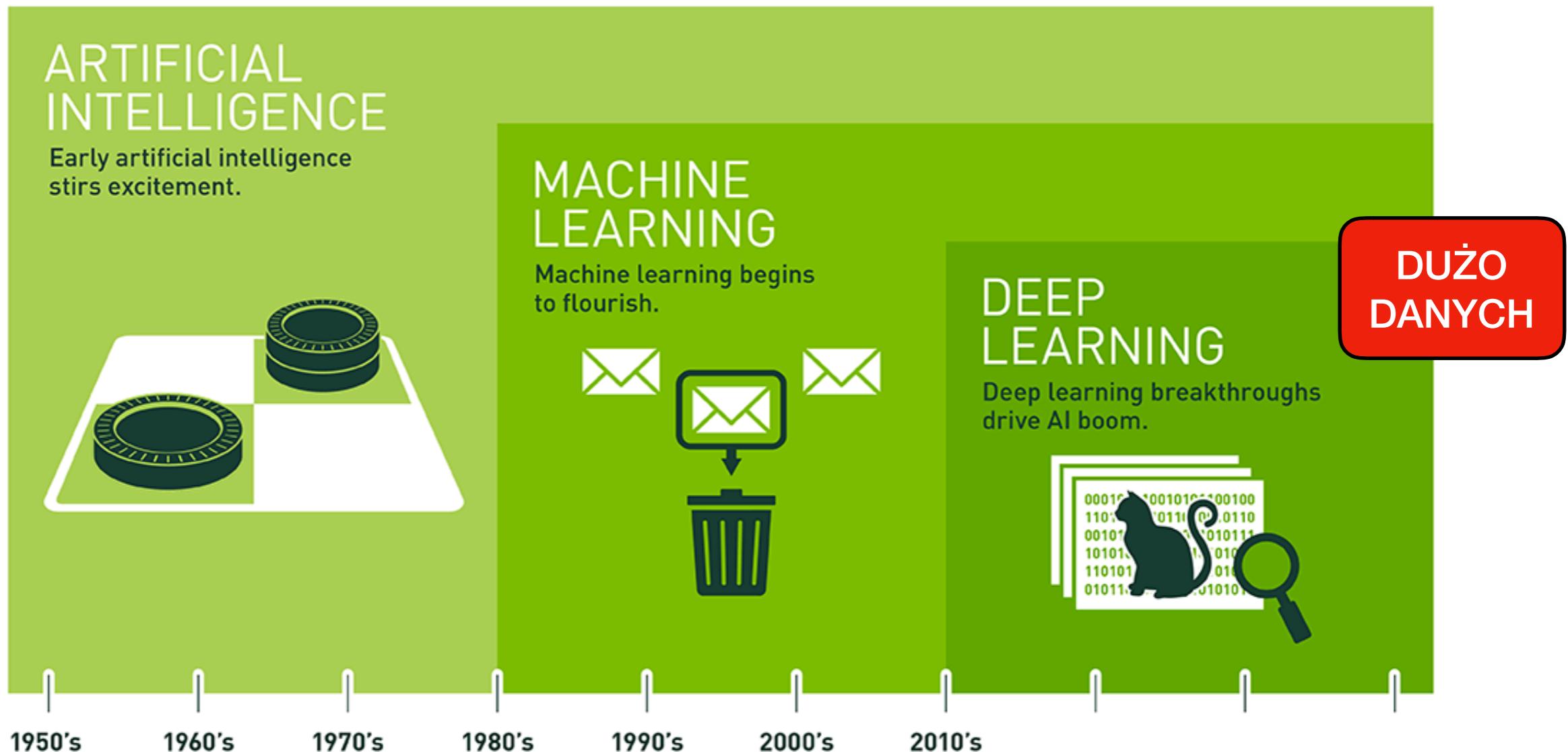


Historia



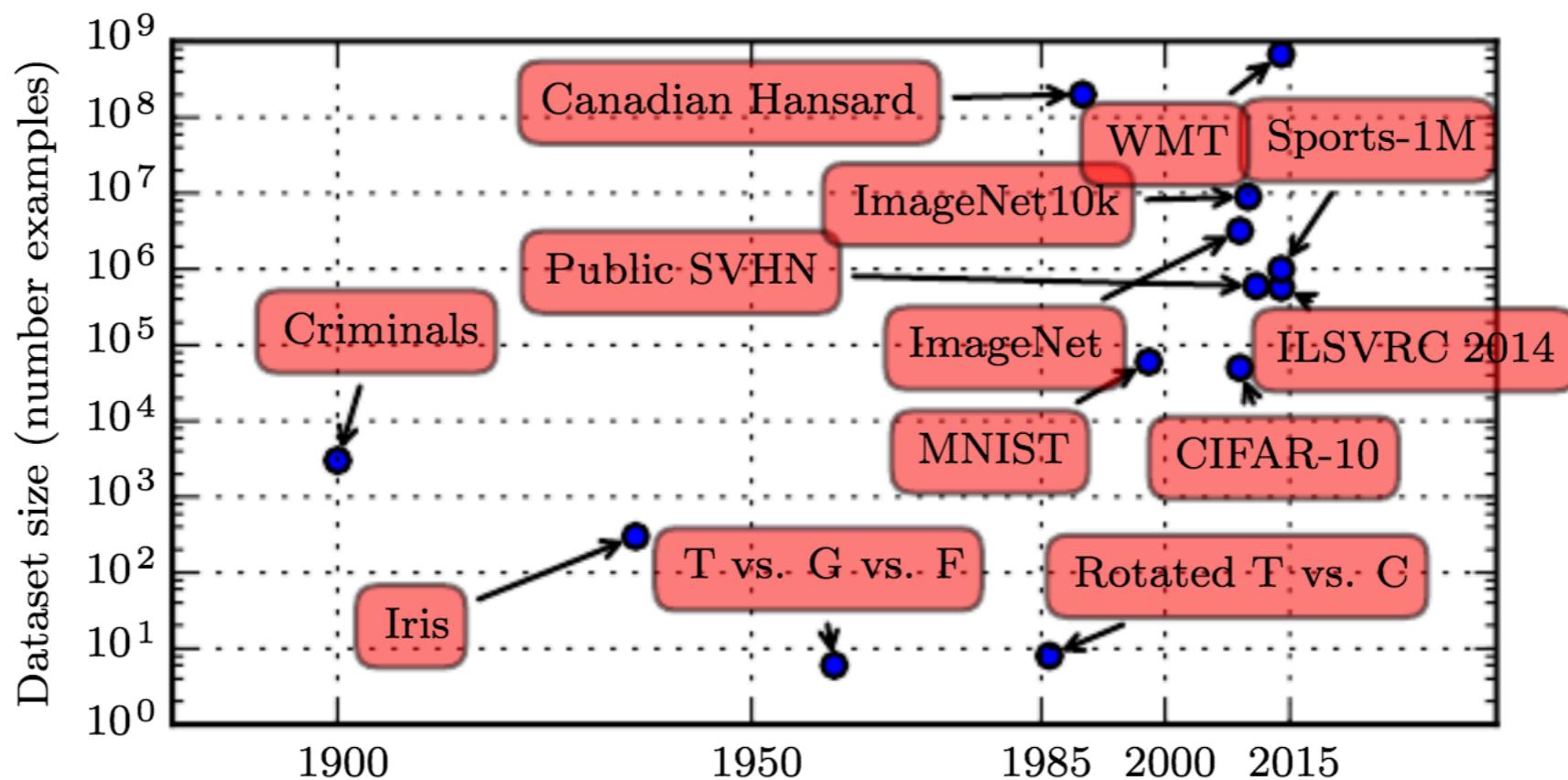
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Historia



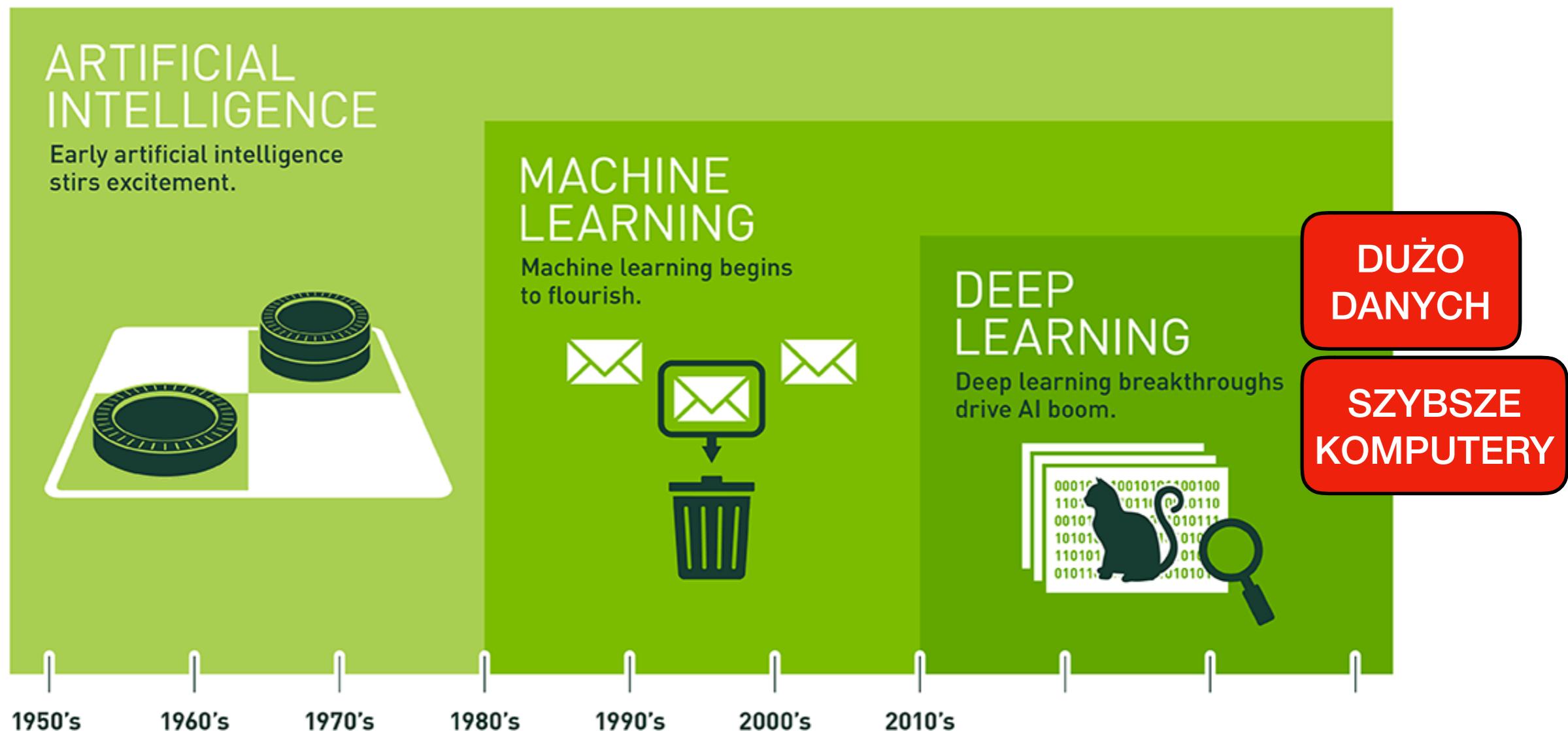
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Historia



Źródło: “Deep Learning”, Ian Goodfellow, Yoshua Bengio, Aaron Courville

Uczenie Maszynowe vs Głębokie Uczenie



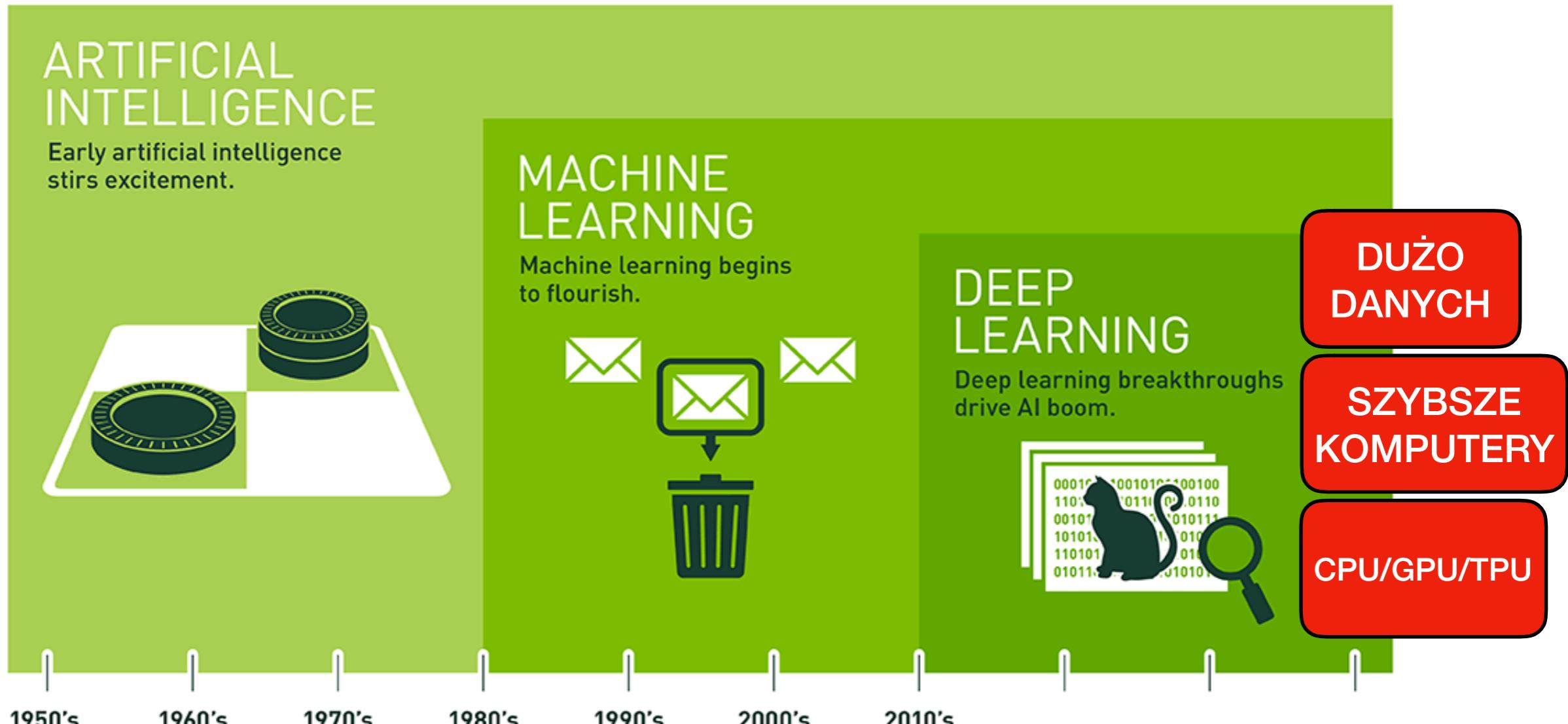
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Historia



Sunway TaihuLight Supercomputer

Historia



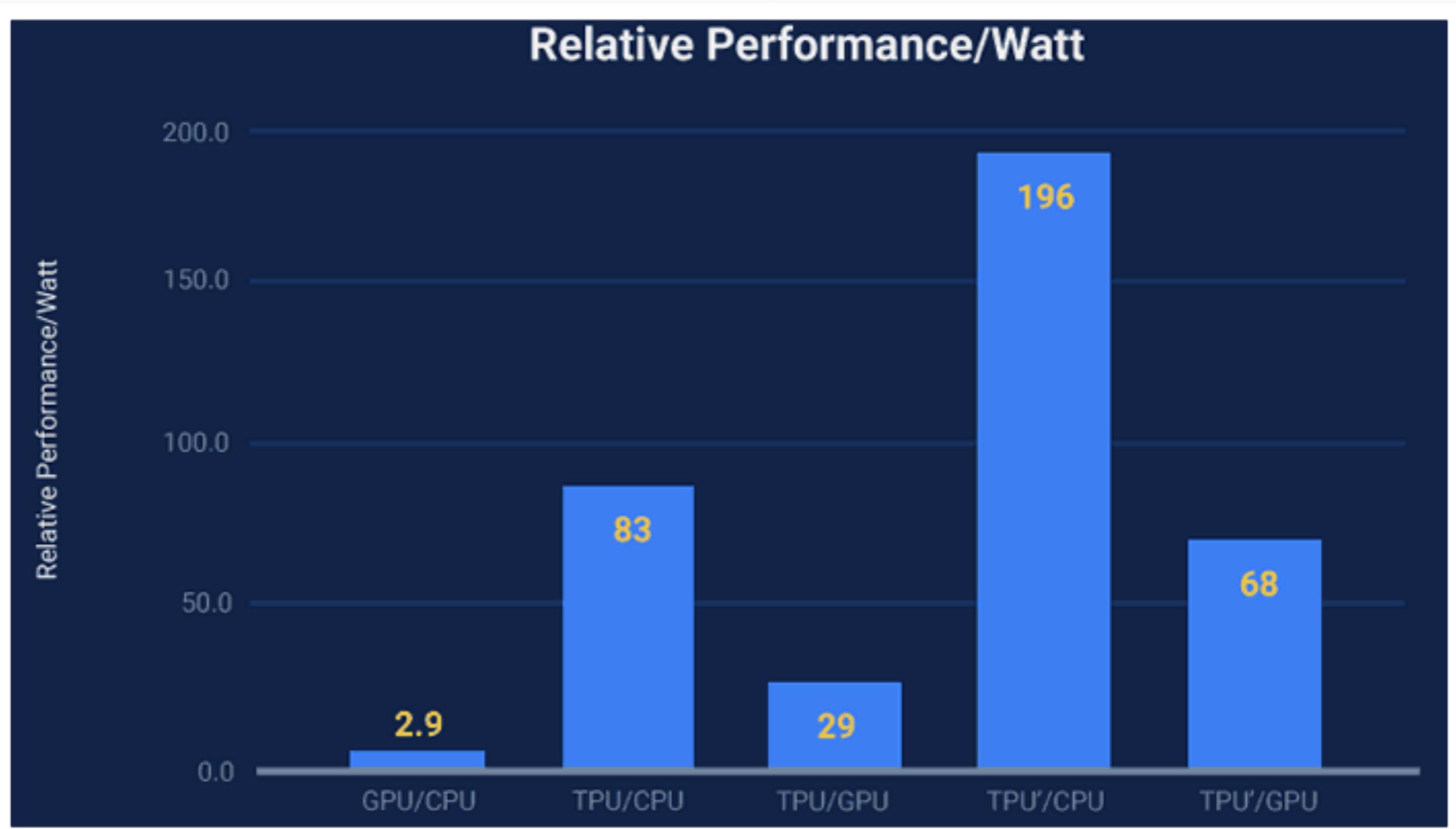
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Historia

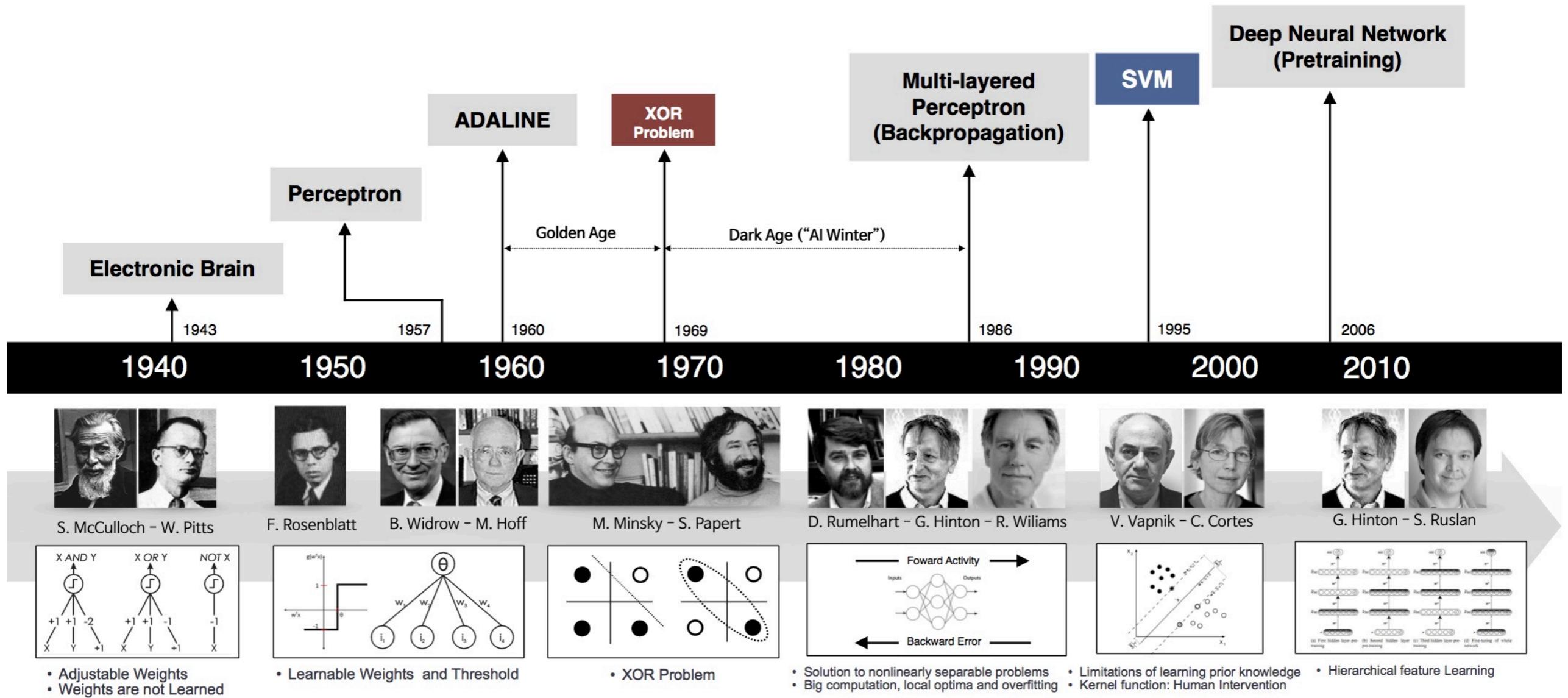
Cloud TPU



Historia

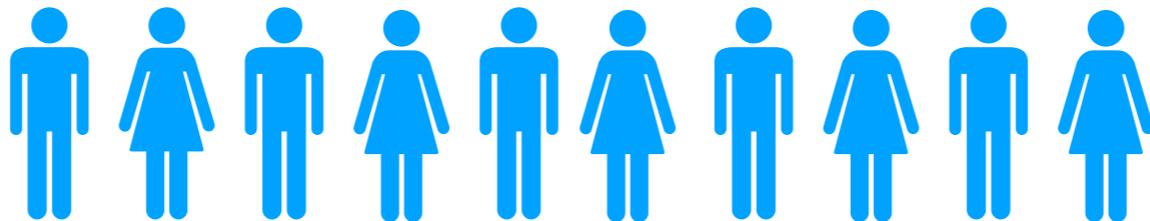


Historia



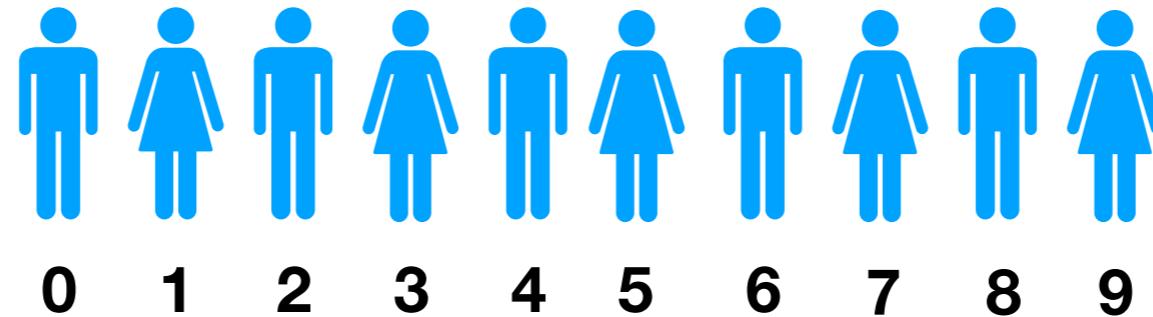
Teoria

Studenci:



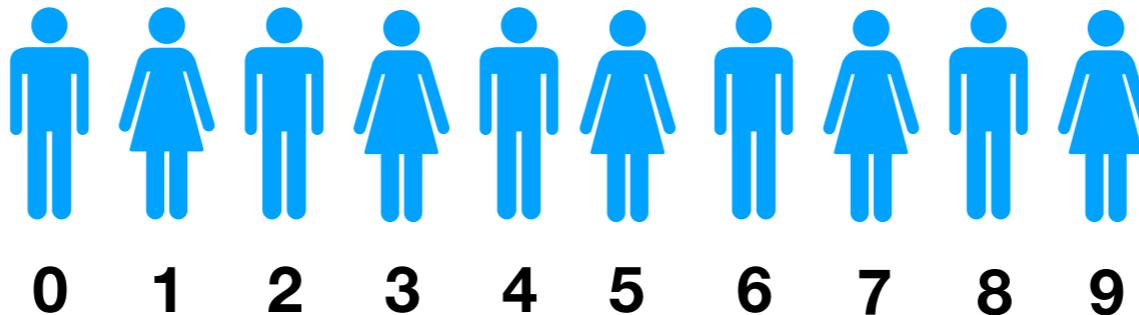
Teoria

Studenci:



Teoria

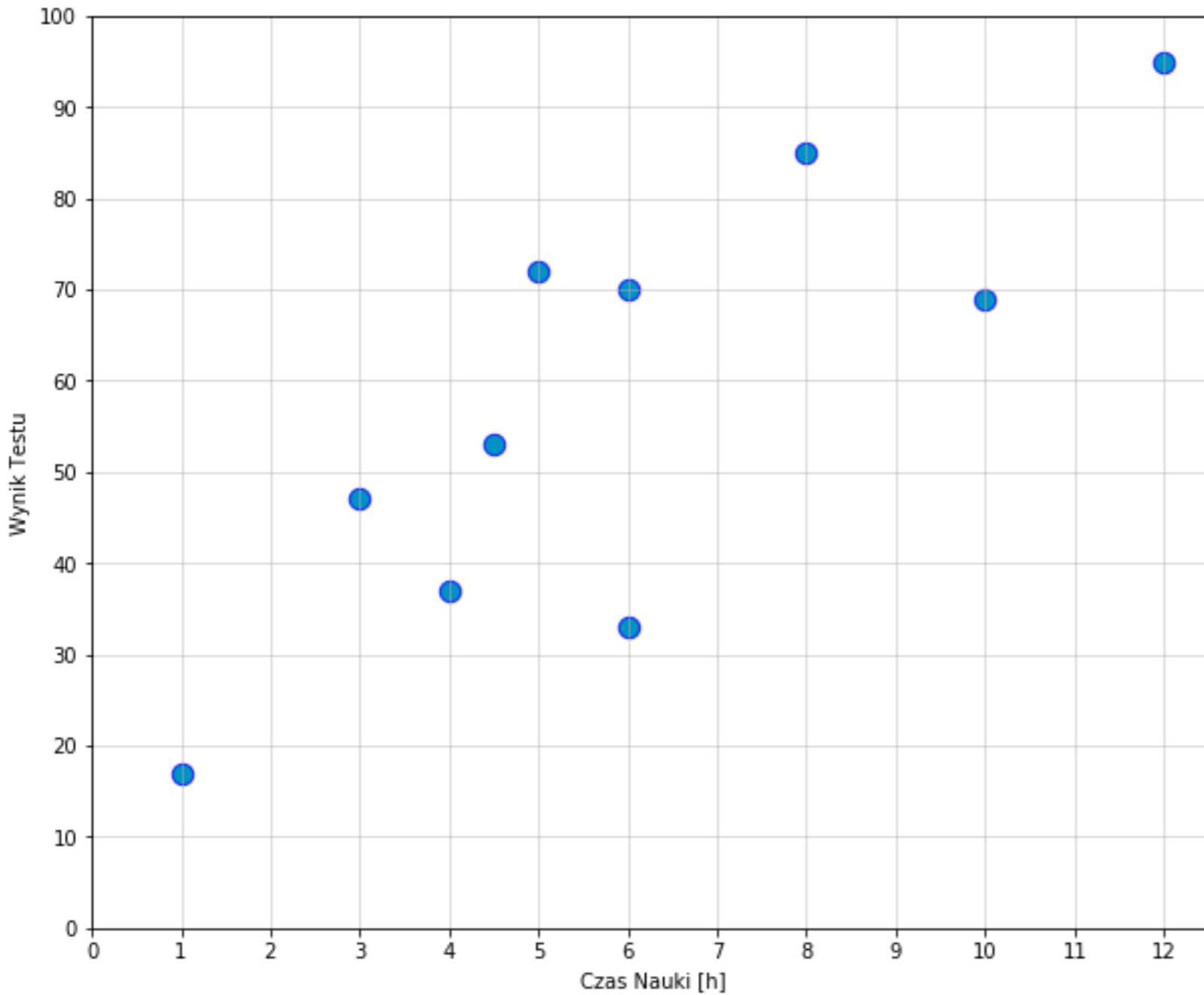
Studenci:



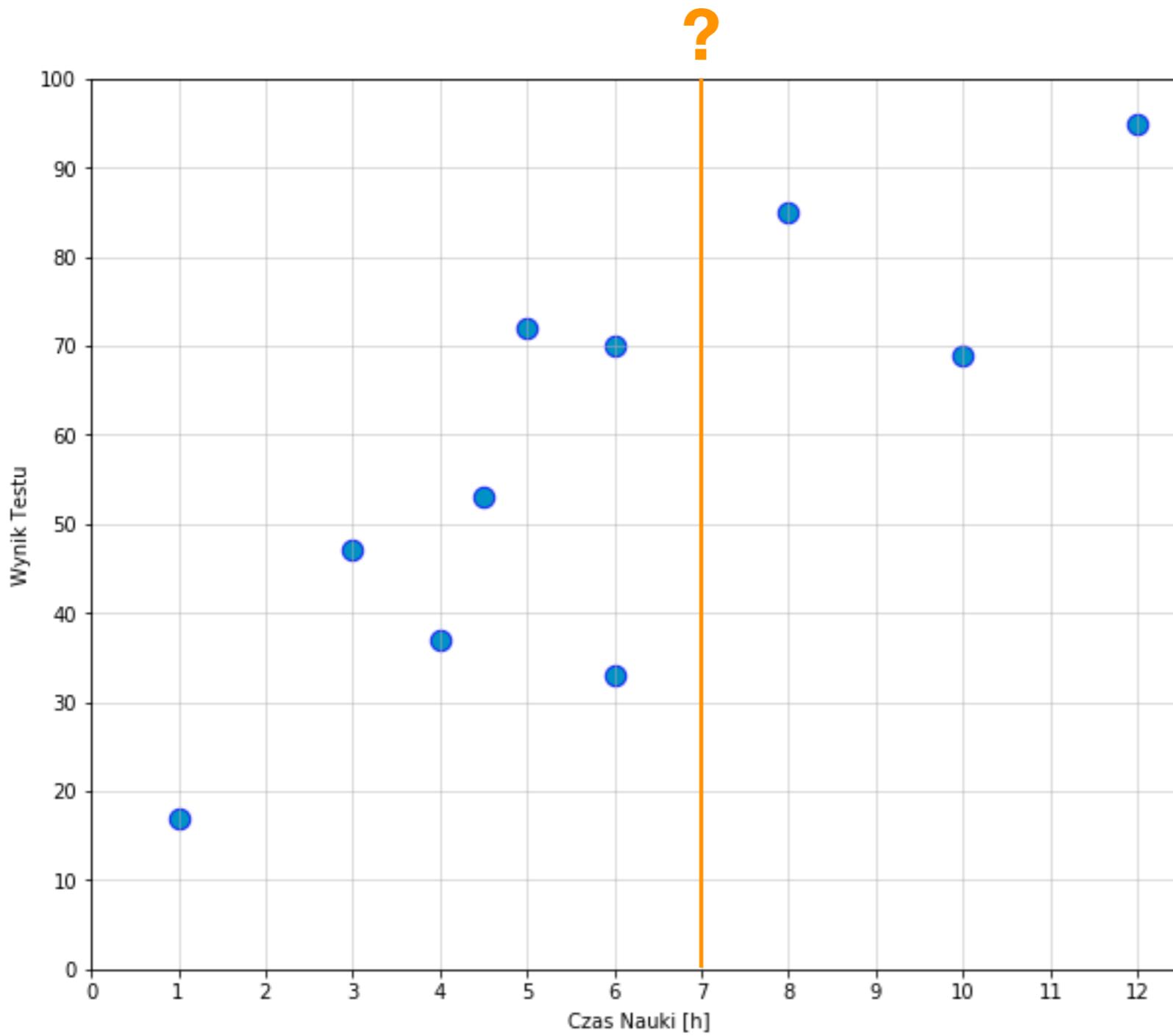
Dane:

Id	Czas Nauki [h]	Czas Snu [h]	Wynik Testu
0	6	8	70
1	1	3	17
2	5	7.5	72
3	3	6	47
4	4.5	6.5	53
5	10	4	69
6	12	8	95
7	8	8	85
8	6	2	33
9	4	5	37

Teoria

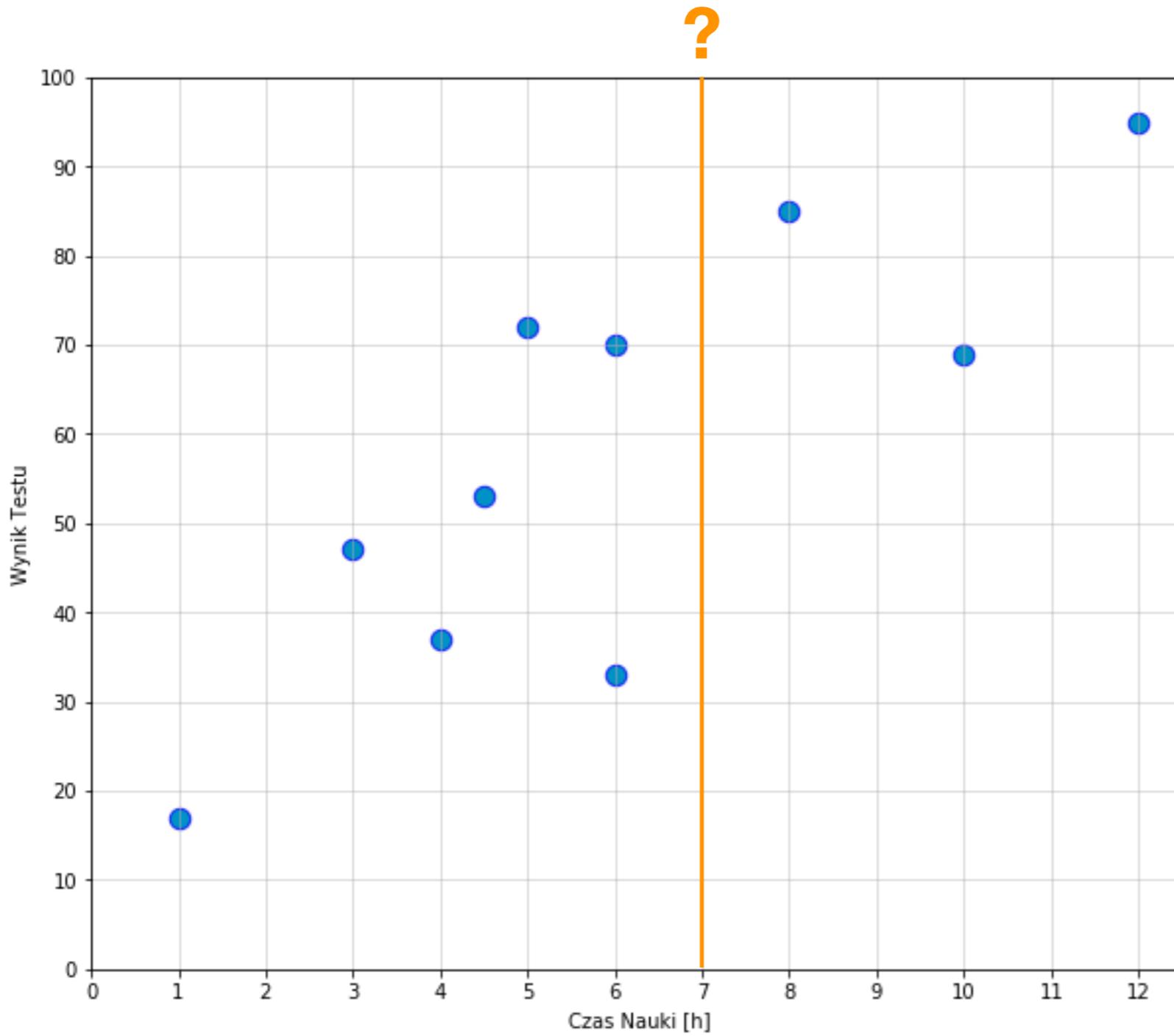


Teoria



Ille punktów otrzymam, jeżeli poświęczę na naukę 7 godzin?

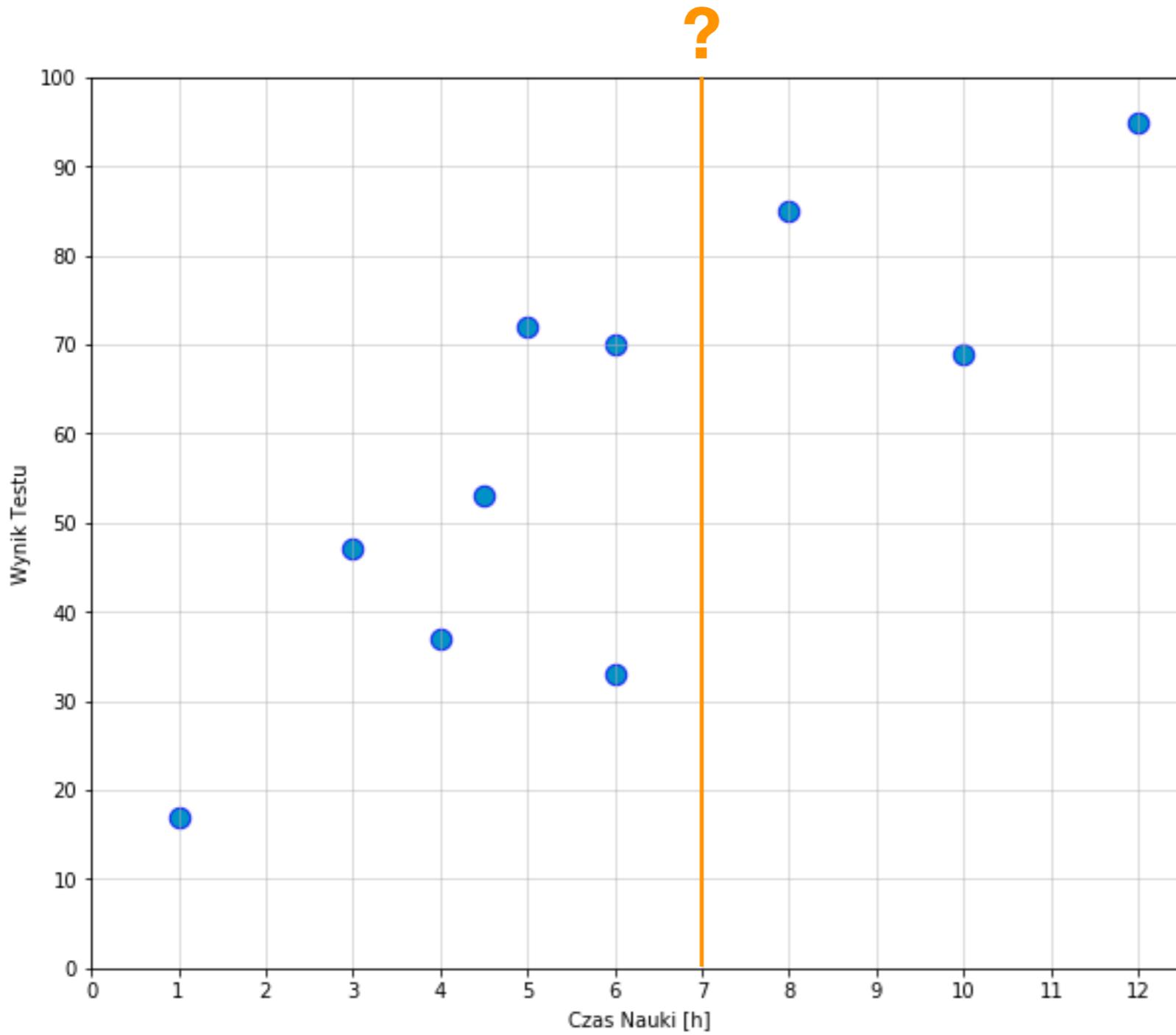
Teoria



Ille punktów otrzymam, jeżeli poświęczę na naukę 7 godzin?

x = 7 godzin

Teoria

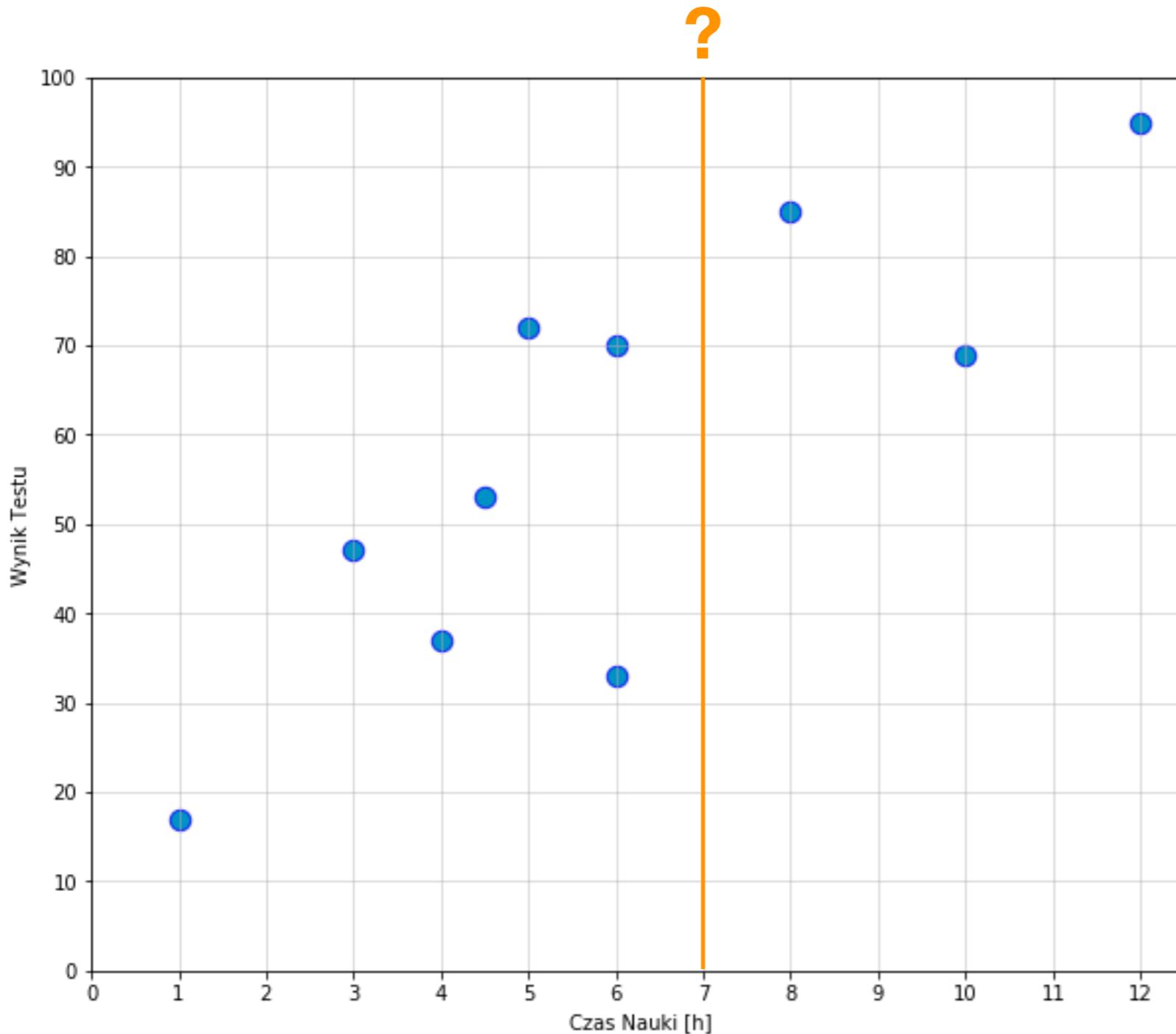


Ille punktów otrzymam, jeżeli poświęczę na naukę 7 godzin?

$x = 7 \text{ godzin}$

$f(x)$

Teoria



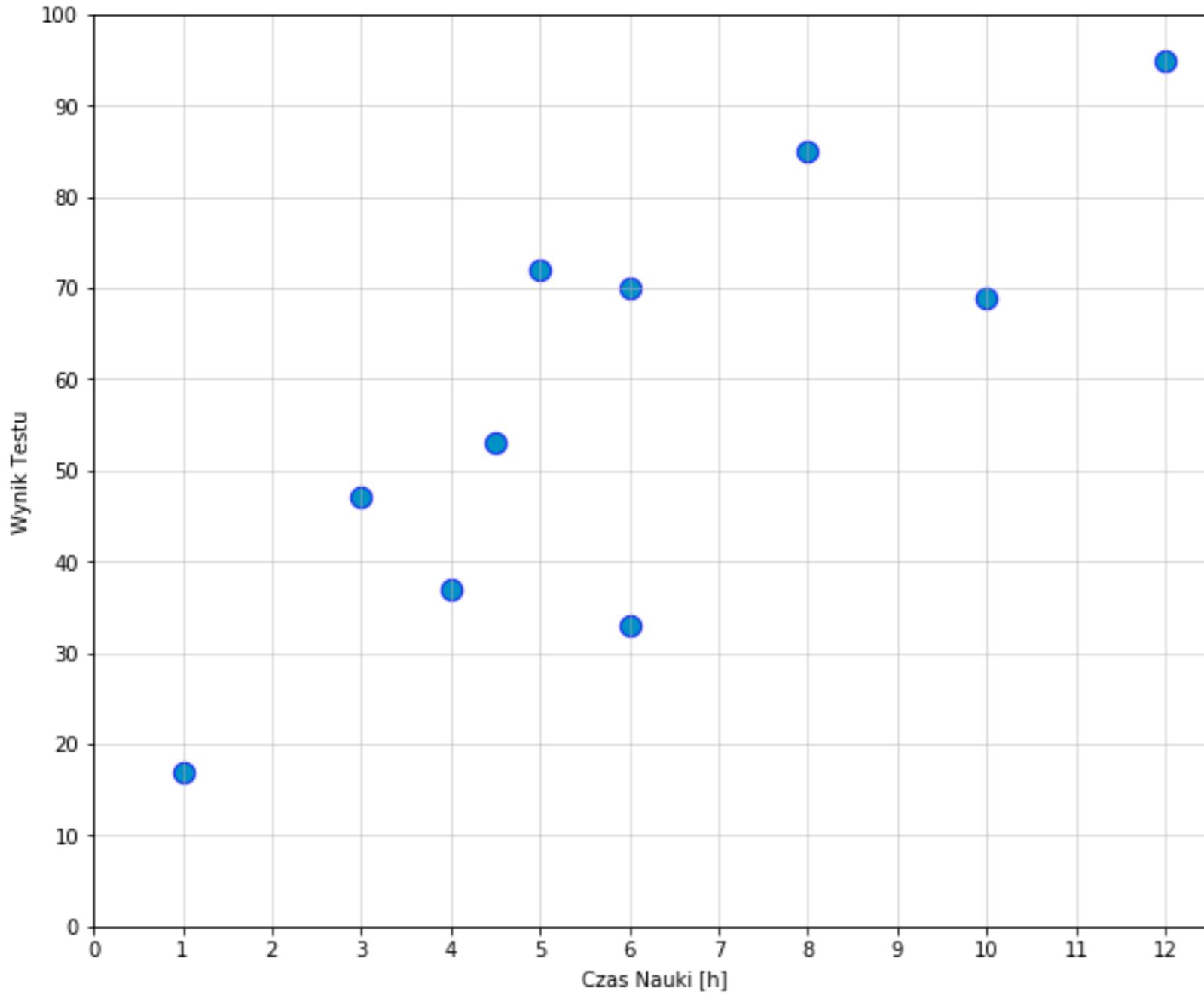
Ille punktów otrzymam, jeżeli poświęczę na naukę 7 godzin?

$x = 7 \text{ godzin}$

$f(x)$

$\hat{y} = \text{Wynik Testu}$

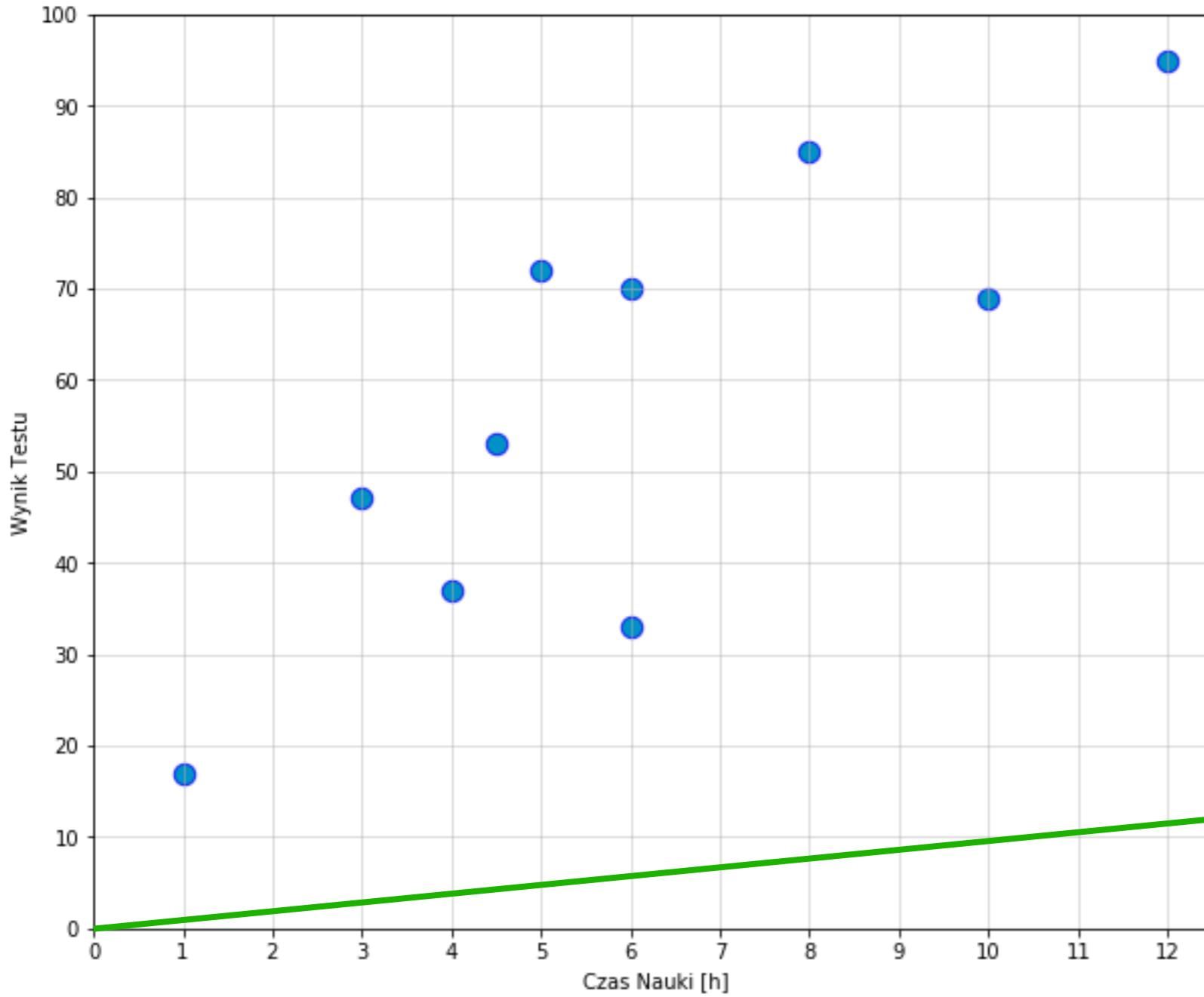
Teoria



Założmy, że $f(x)$ to funkcja liniowa.

$$f(x) : \hat{y} = x$$

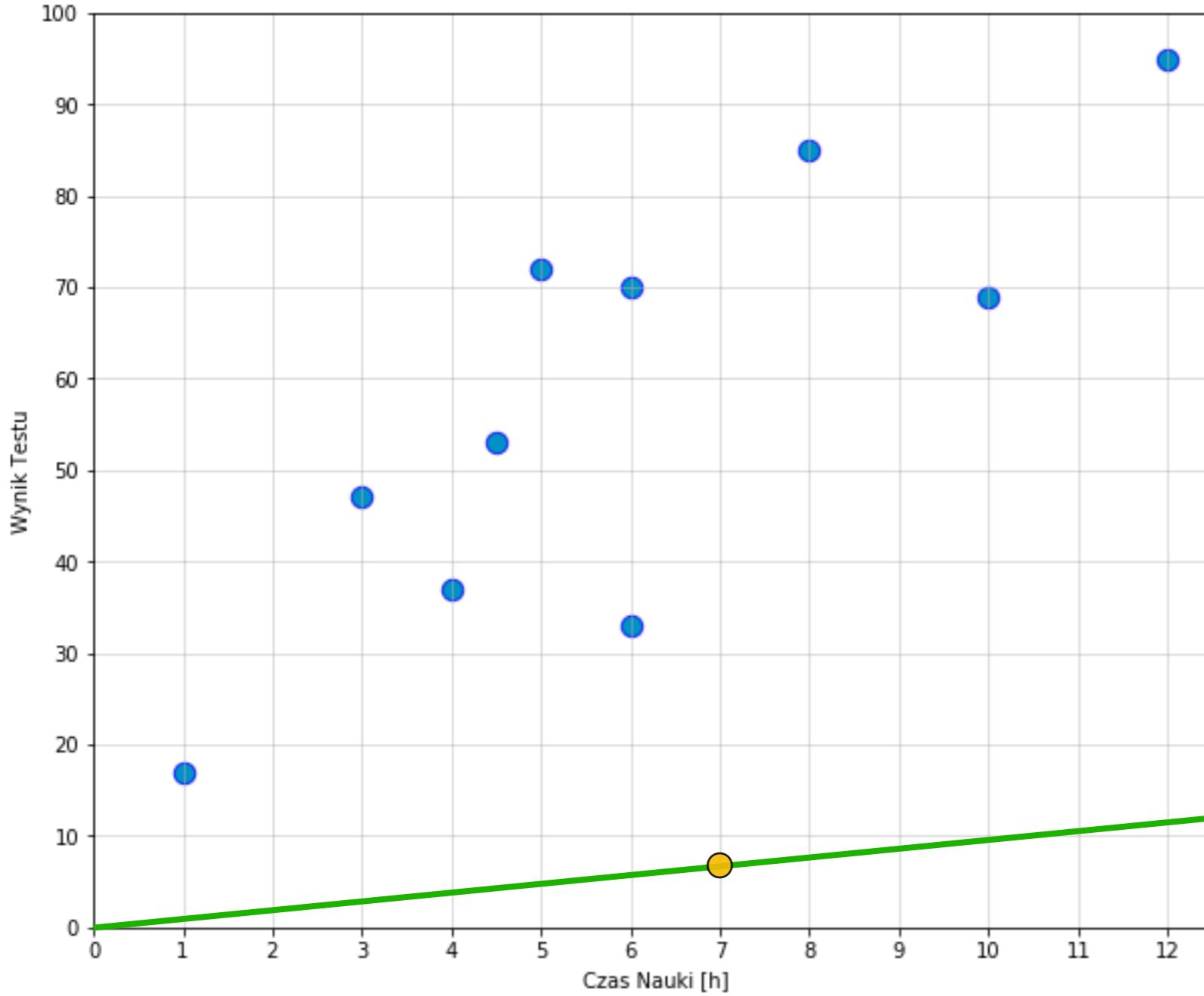
Teoria



Założmy, że $f(x)$ to funkcja liniowa.

$$f(x) : \hat{y} = x$$

Teoria



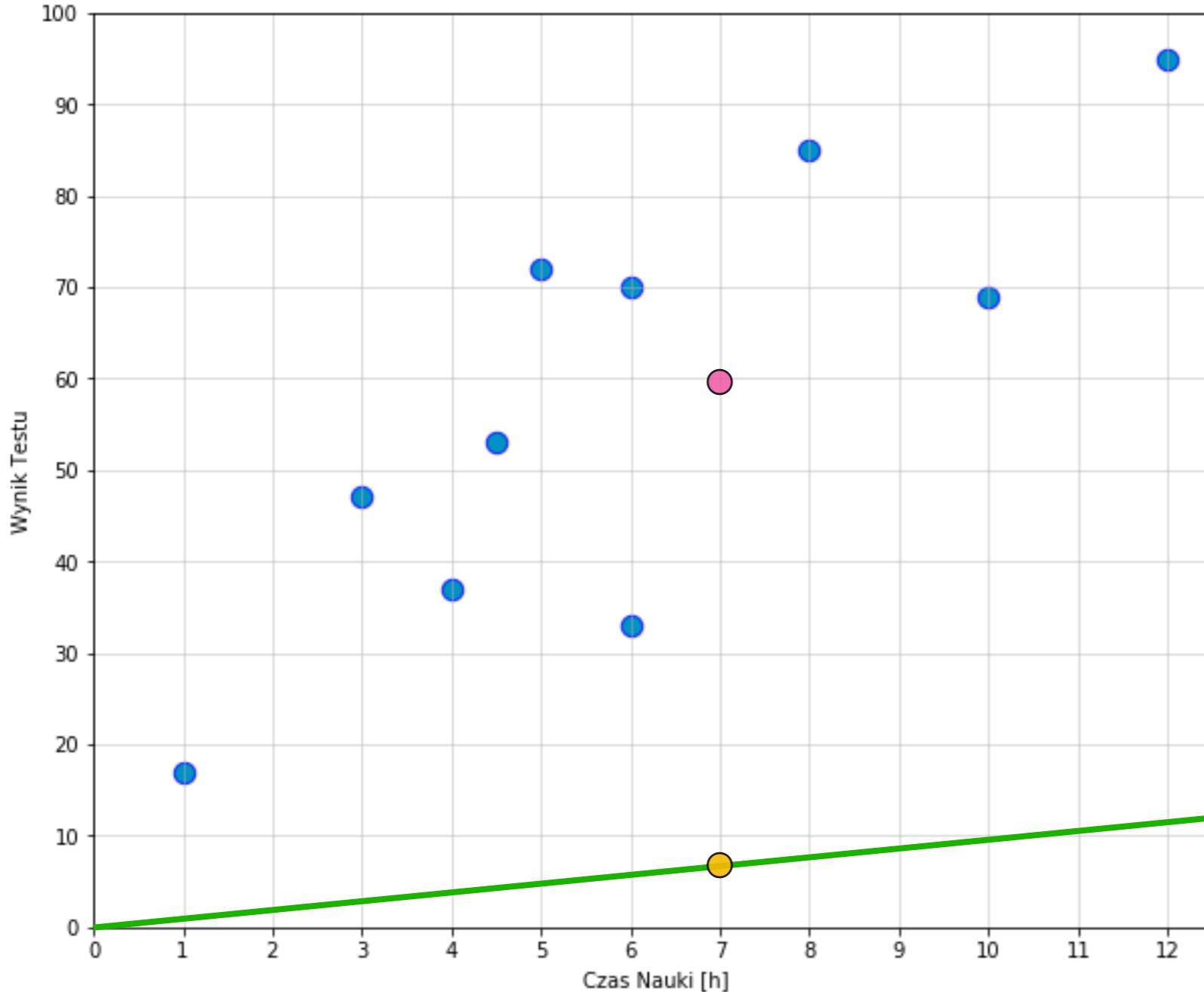
Założmy, że $f(x)$ to funkcja liniowa.

$$f(x) : \hat{y} = x$$



$$f(7) = 7$$

Teoria



Założymy, że $f(x)$ to funkcja liniowa.

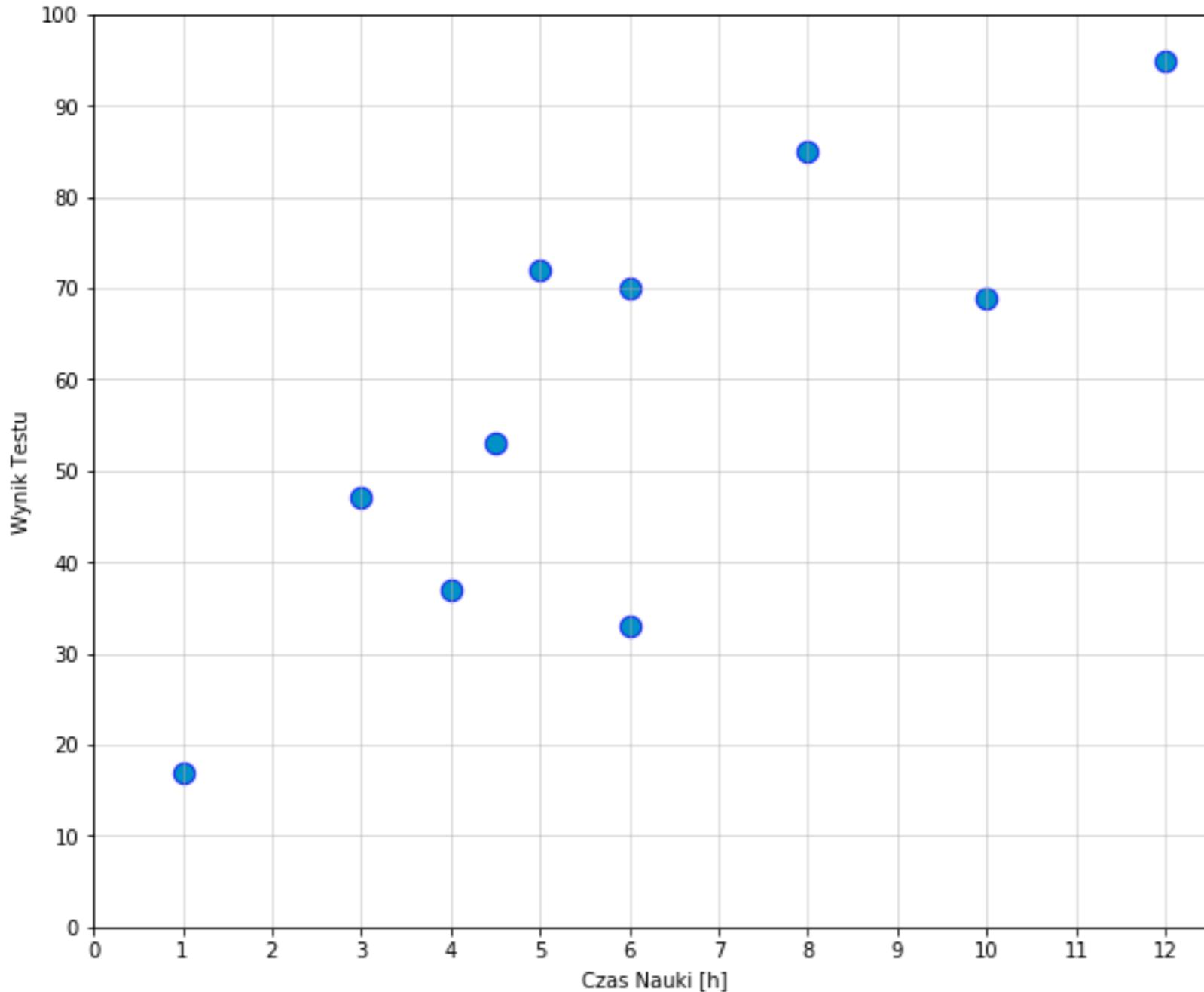
$$f(x) : \hat{y} = x$$



$$f(7) = 7$$

Oczekiwaliśmy pewnie wyniku bliżej pozostałych punktów.

Teoria

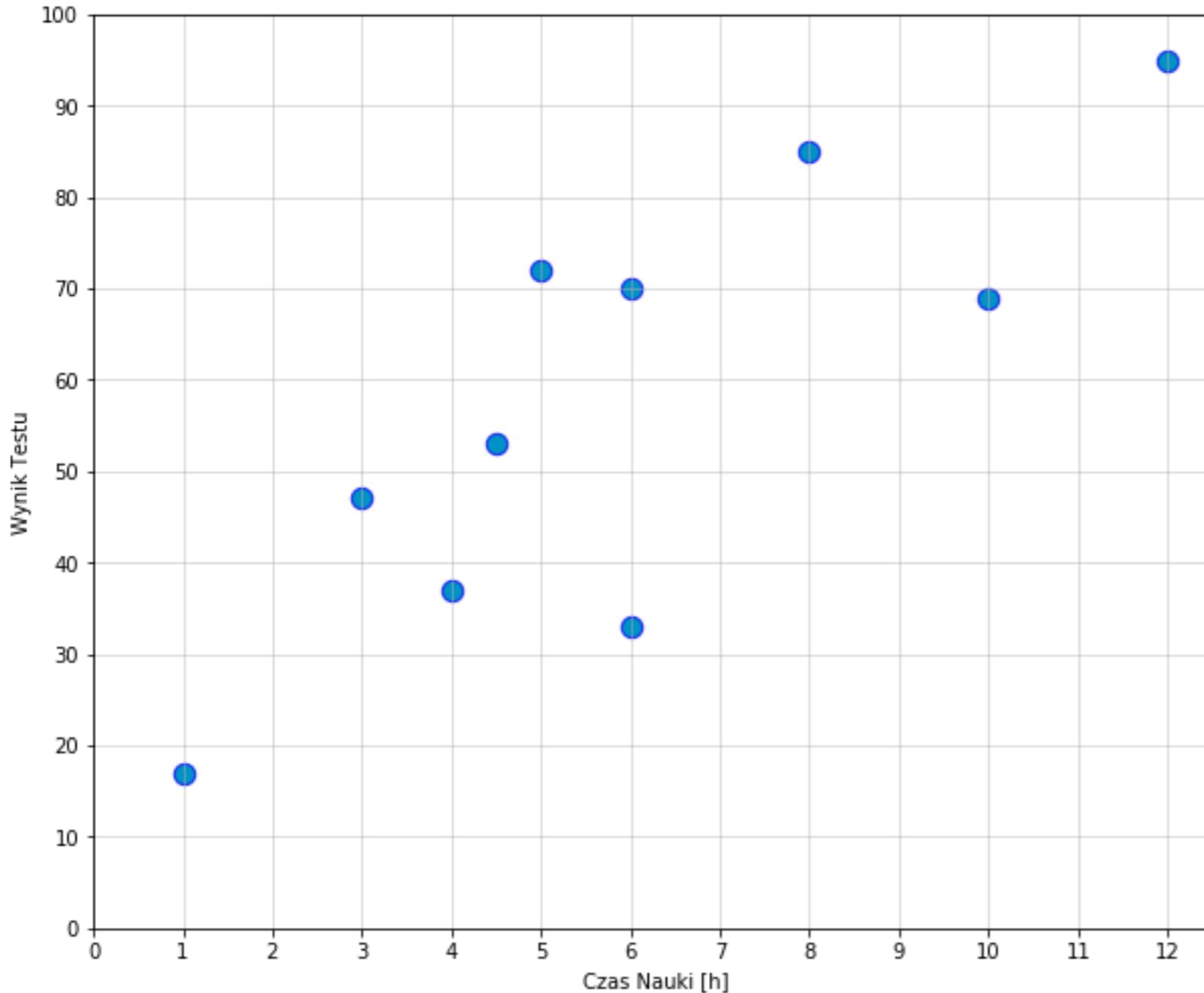


Założmy, że $f(x)$ to funkcja liniowa.

$$f(x) : \hat{y} = x$$

Ta funkcja jest nie odpowiednia - zmodyfikujmy ją i dodajmy parametry.

Teoria

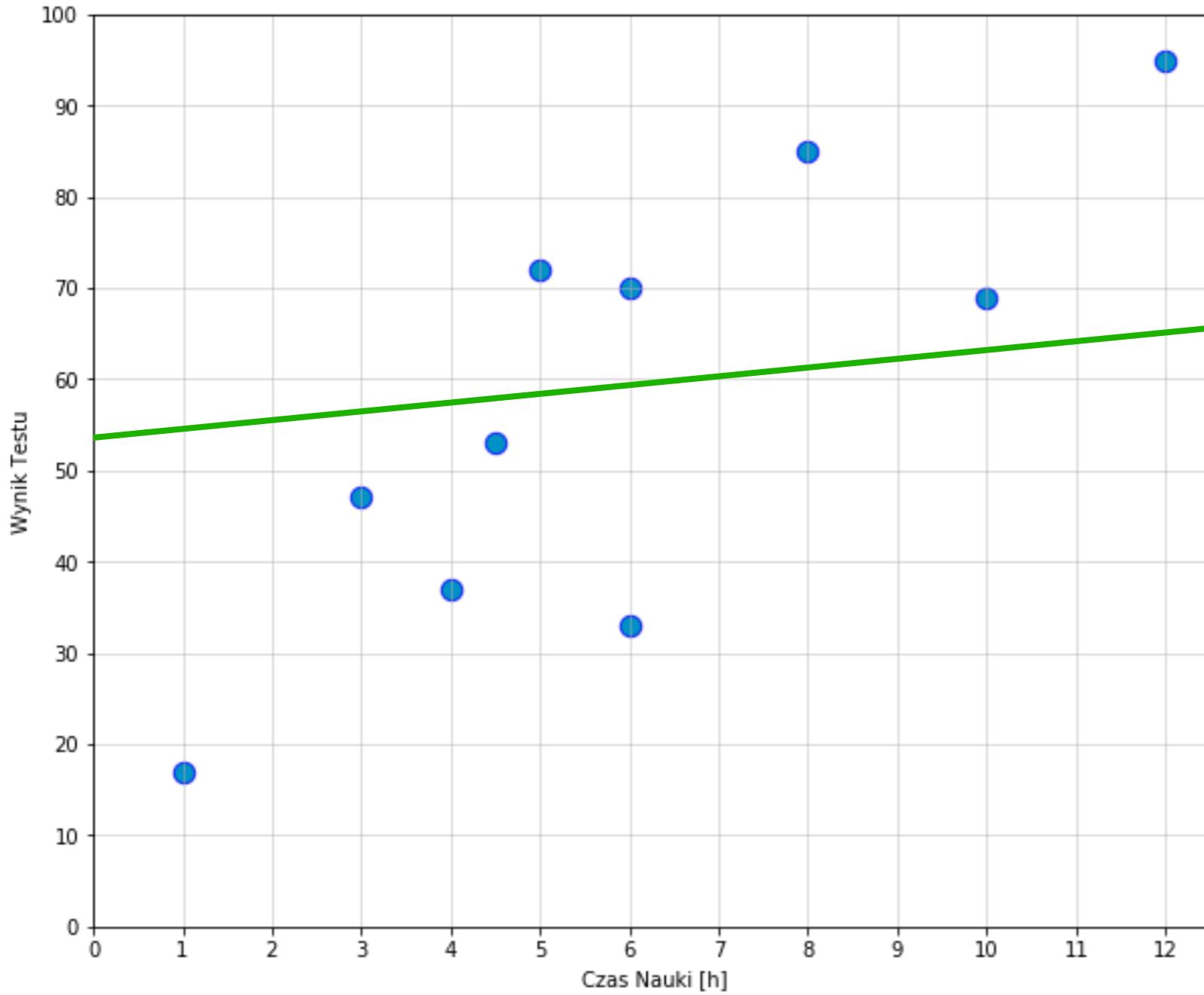


Założmy, że $f(x)$ to funkcja liniowa.

$$f(x): \hat{y} = wx + b$$

Ta funkcja jest nie odpowiednia - zmodyfikujmy ją i dodajmy parametry.

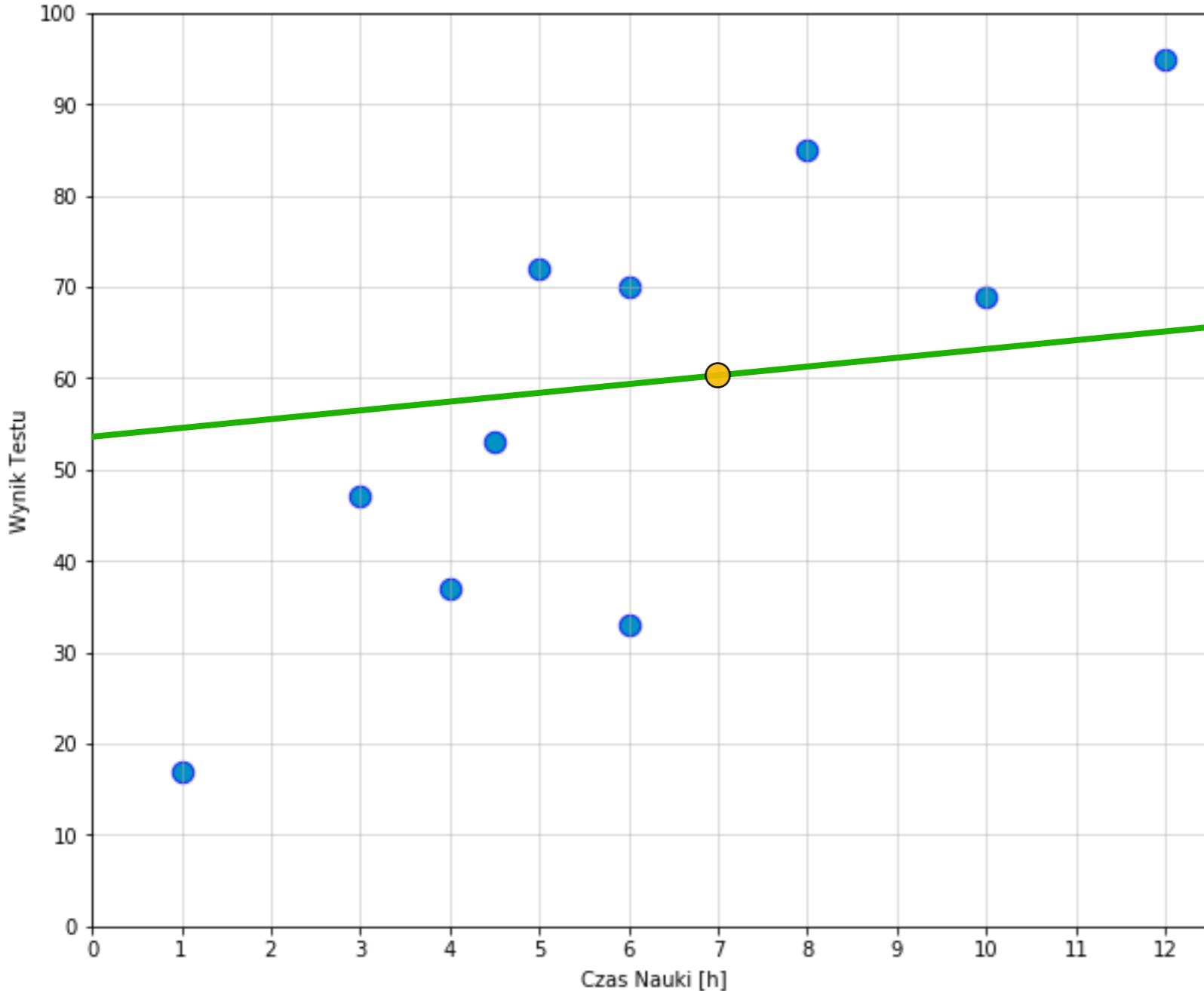
Teoria



Założmy, że $f(x)$ to funkcja liniowa.

$$f(x): \hat{y} = 1x + 57$$

Teoria



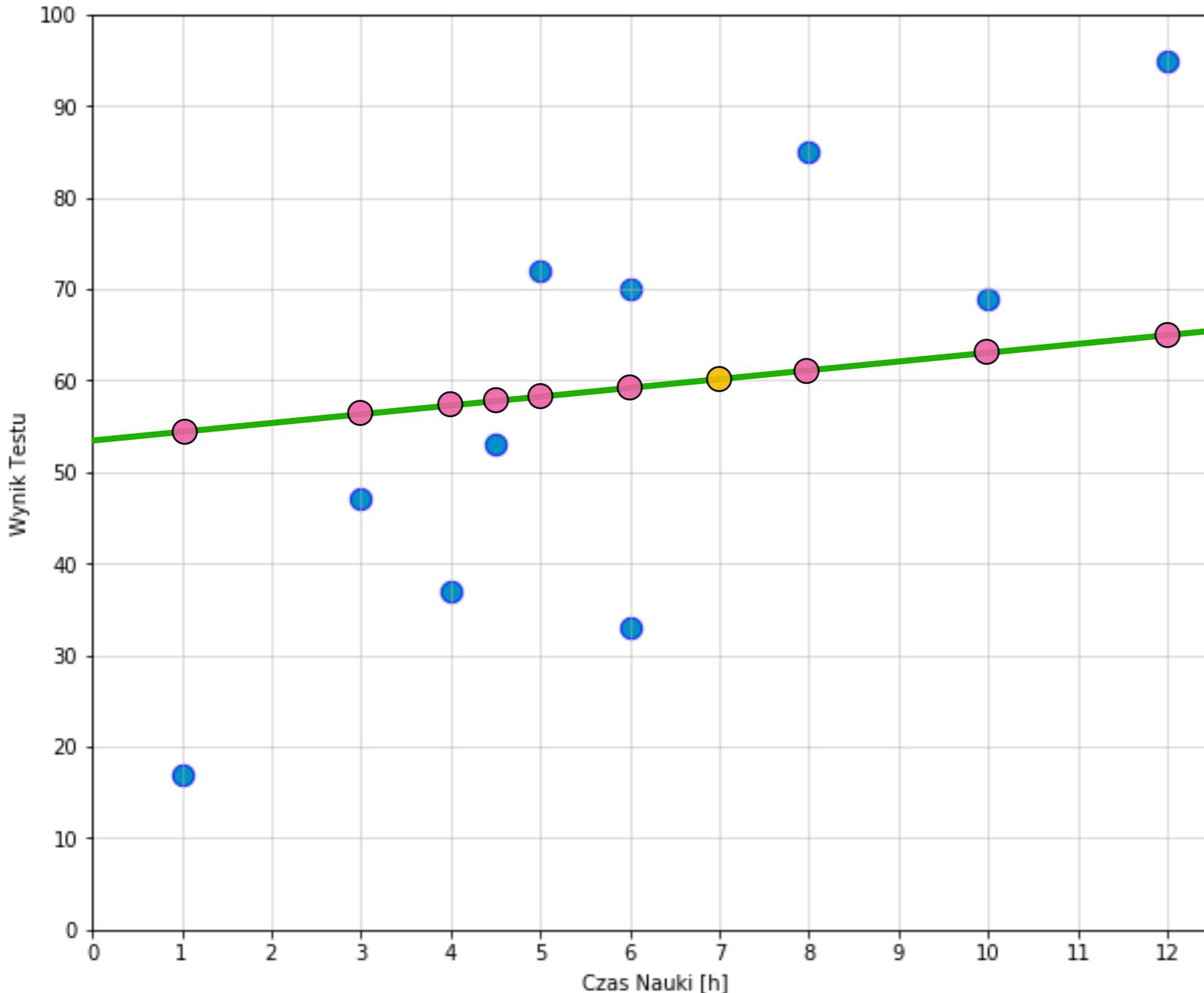
Założmy, że $f(x)$ to funkcja liniowa.

$$f(x): \hat{y} = 1x + 53$$



$$f(7) = 60$$

Teoria



Założymy, że $f(x)$ to funkcja liniowa.

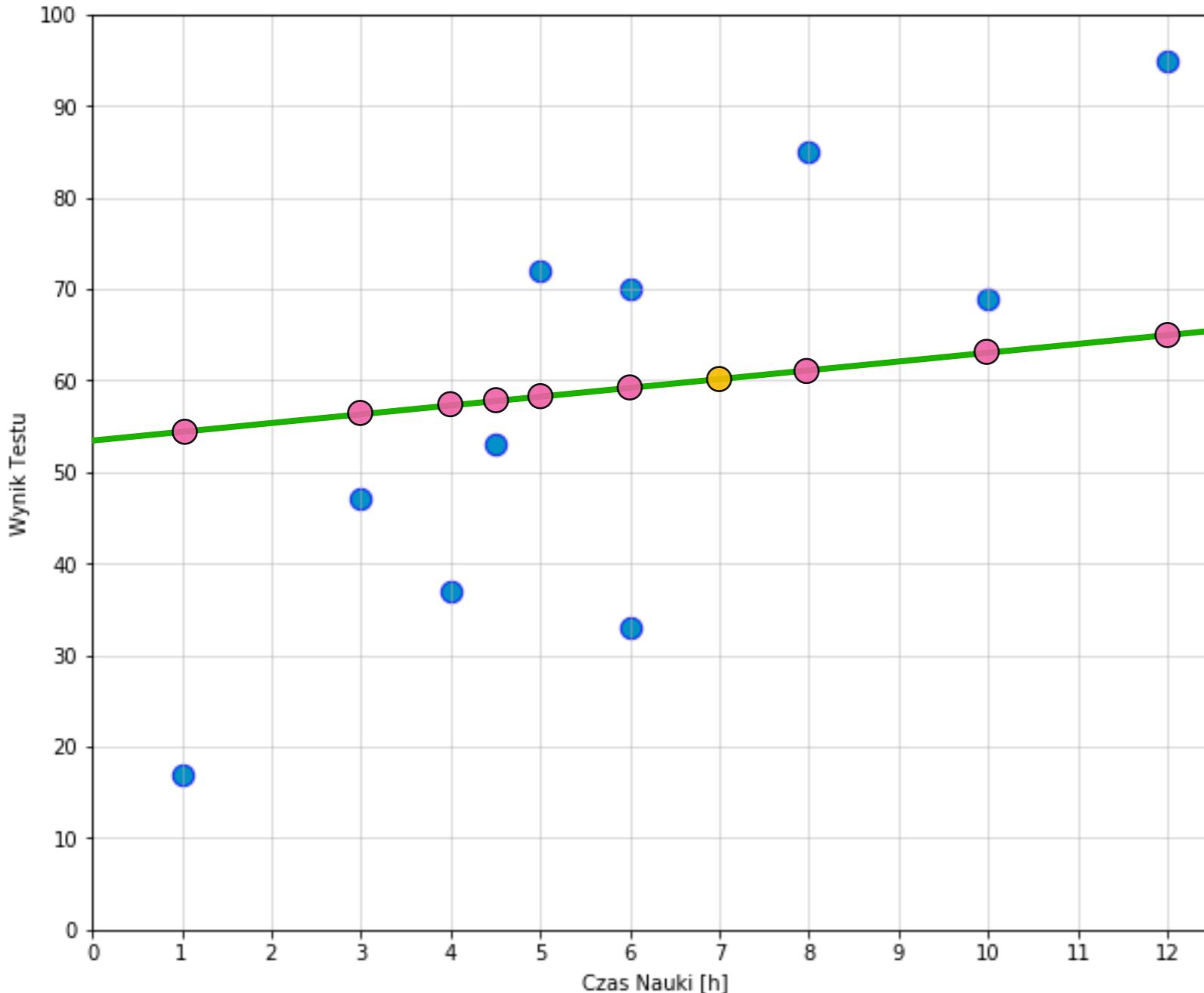
$$f(x): \hat{y} = 1x + 53$$



$$f(7) = 60$$

Ale jeżeli użyjemy tej funkcji do dokonania predykcji dla pozostałych wartości to punkty się nie pokryją.

Teoria



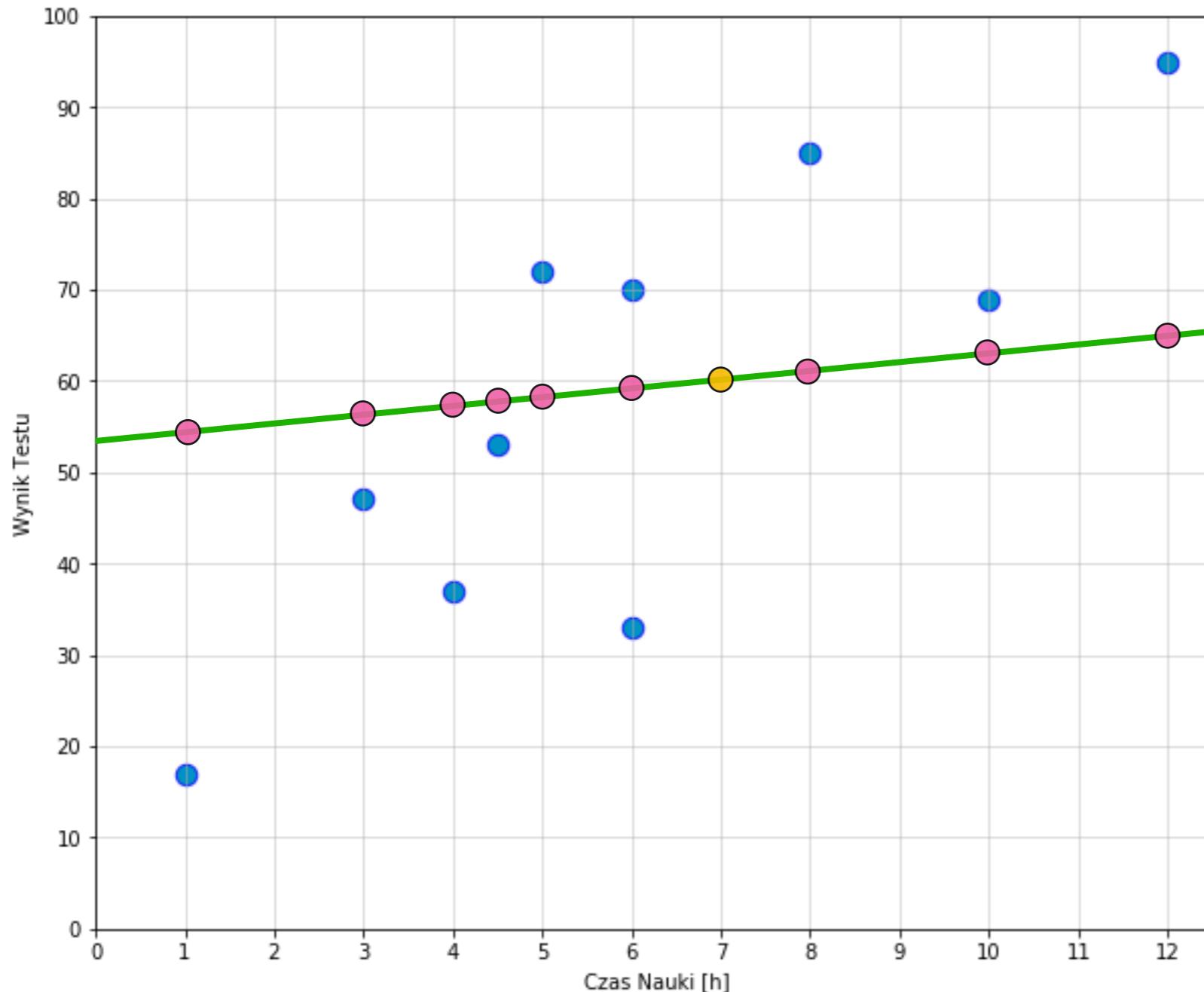
Założymy, że $f(x)$ to funkcja liniowa.

$$f(x): \hat{y} = 1x + 53$$

Musimy znaleźć inne wartości parametrów:
 w, b

Teoria

$$f(x) : \hat{y} = 1x + 53$$

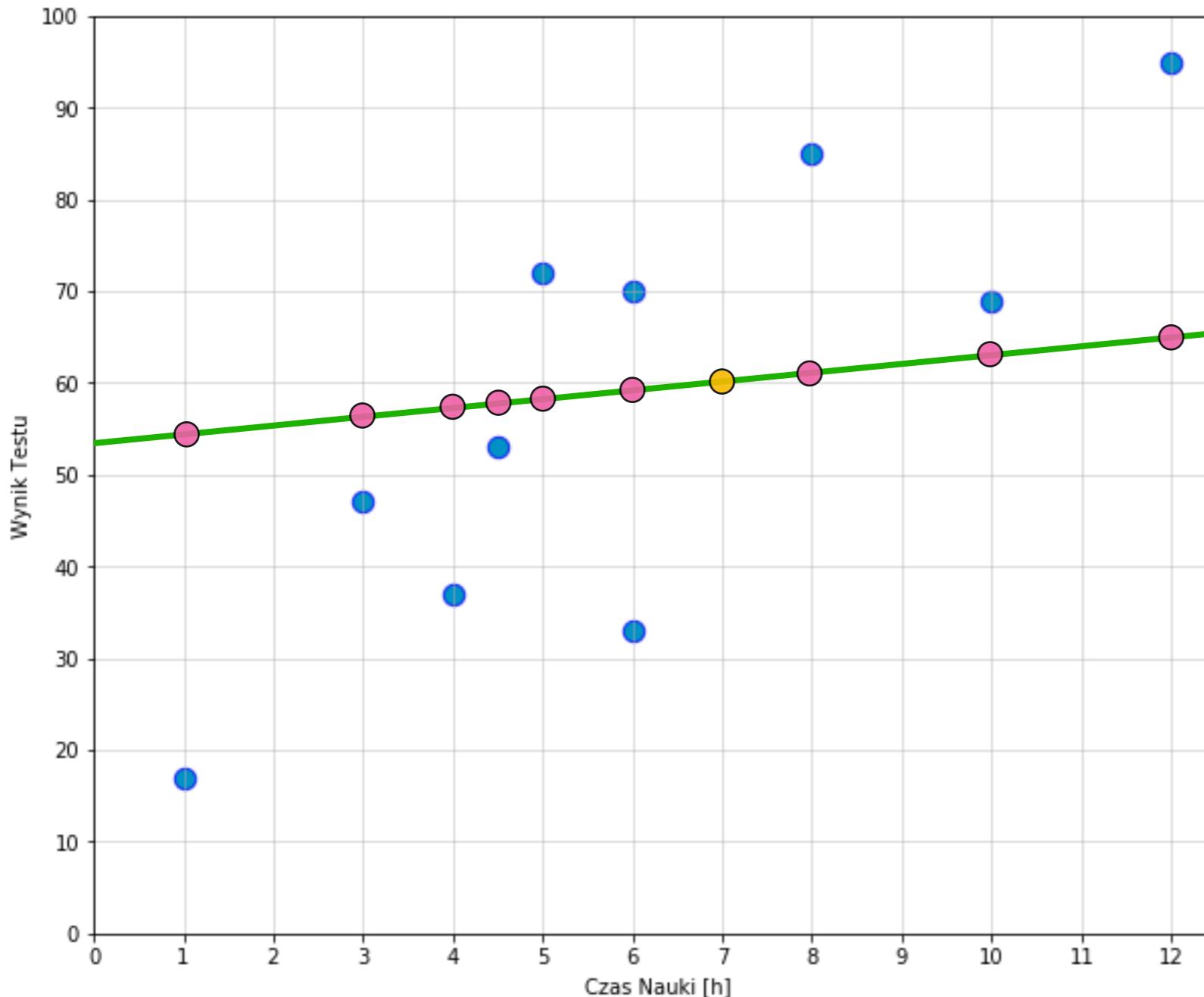


Założmy, że $f(x)$ to funkcja liniowa.

Musimy znaleźć inne wartości parametrów:
 w, b

Teoria

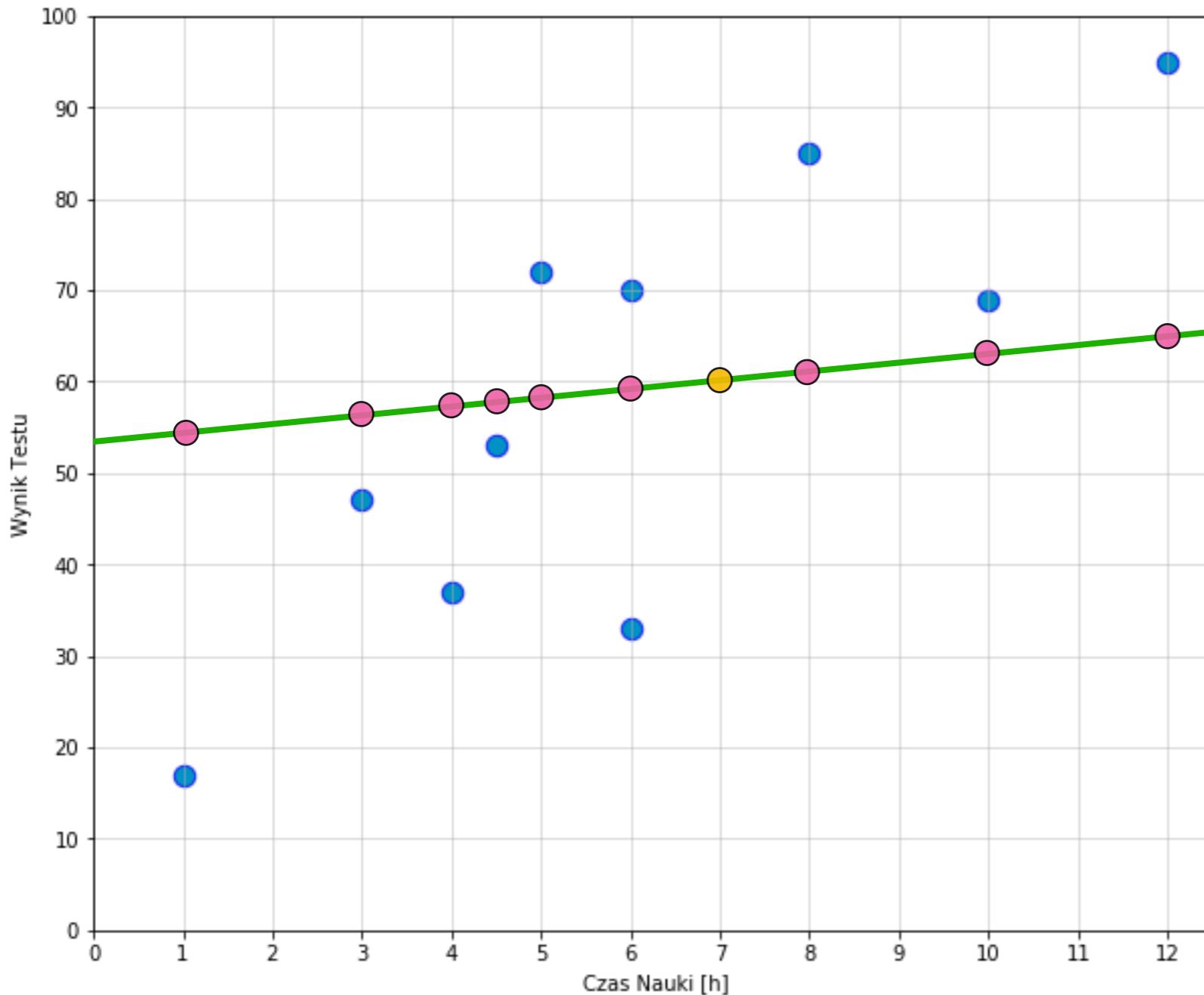
$$f(x) : \hat{y} = 1x + 53$$



W jaki sposób szukać parametrów w i b aby wynik był satysfakcjonujący?

Teoria

$$f(x) : \hat{y} = 1x + 53$$

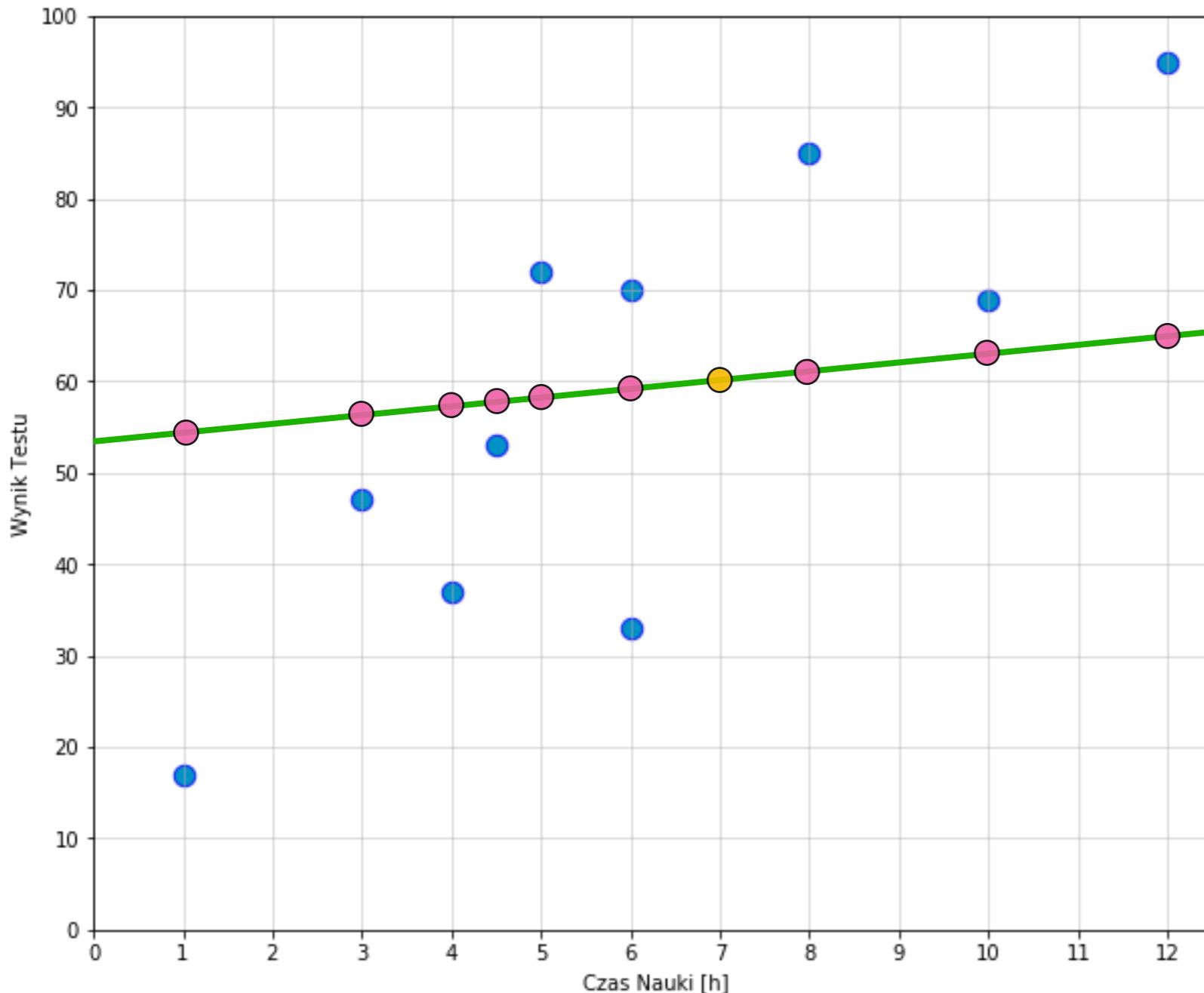


W jaki sposób szukać parametrów **w** i **b** aby wynik był satysfakcjonujący?

$$f(1) = 54$$

Teoria

$$f(x) : \hat{y} = 1x + 53$$



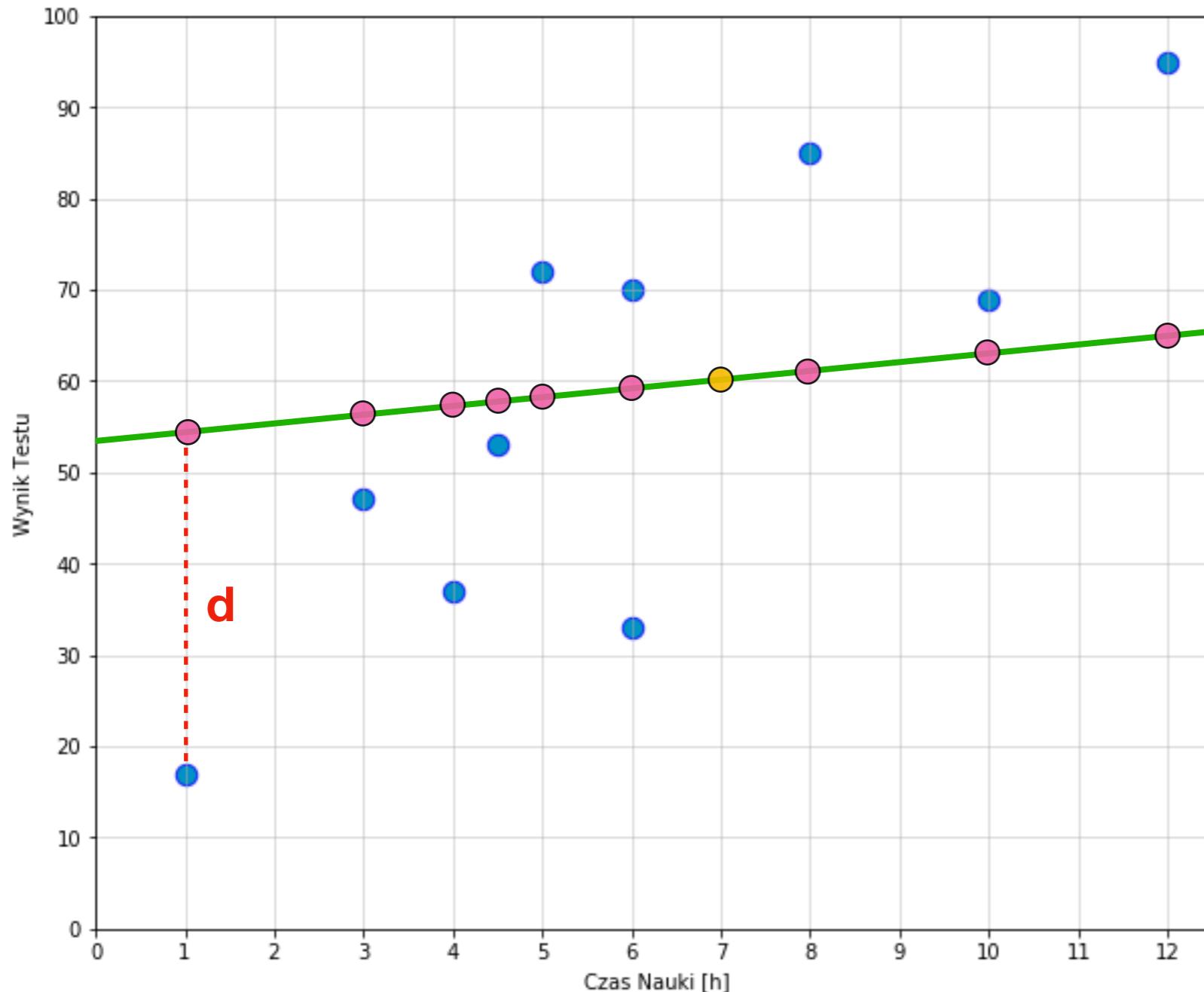
W jaki sposób szukać parametrów **w** i **b** aby wynik był satysfakcjonujący?

$$f(1) = 54$$

$$y = 17$$

Teoria

$$f(x) : \hat{y} = 1x + 53$$



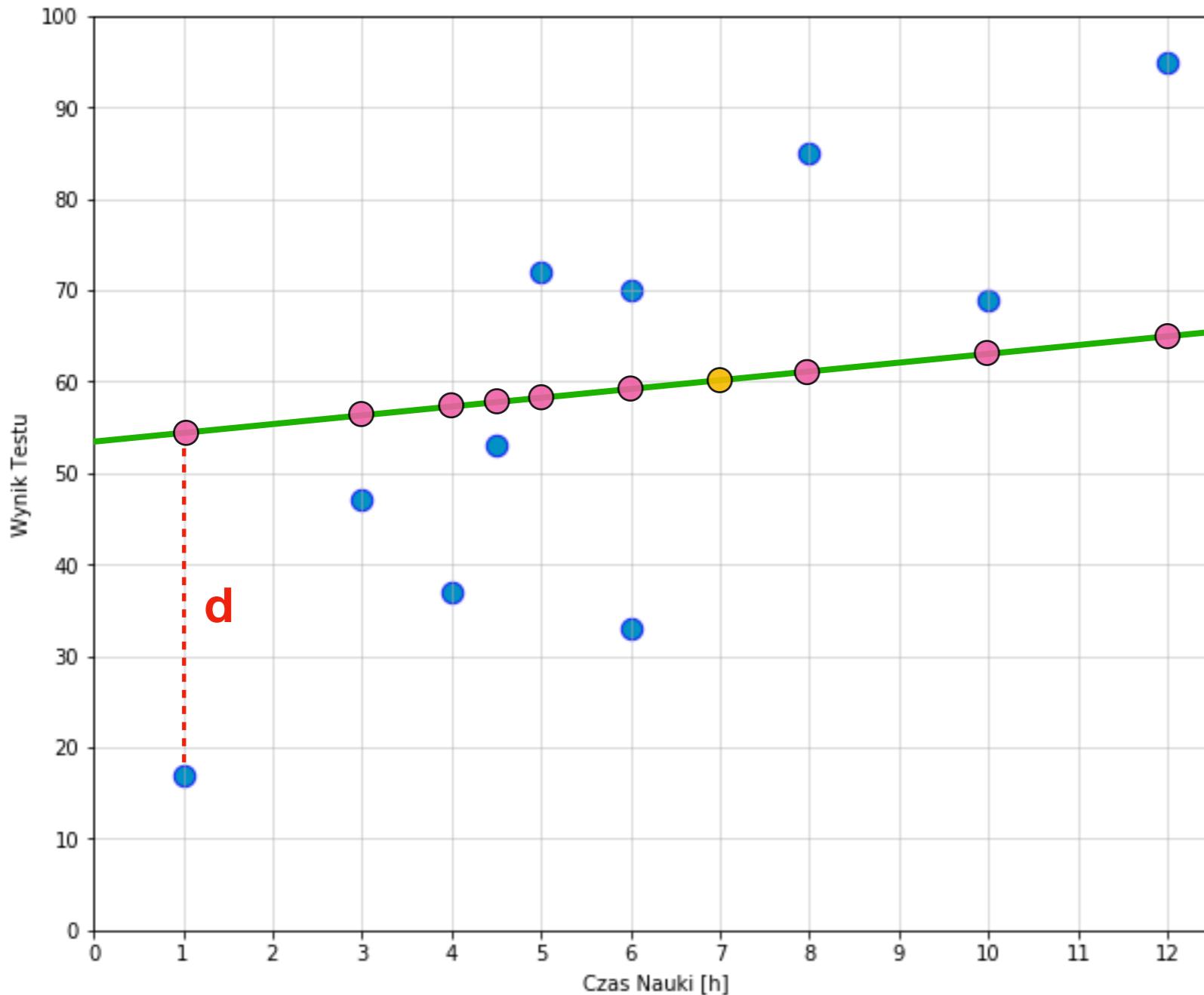
W jaki sposób szukać parametrów w i b aby wynik był satysfakcjonujący?

$$f(1) = 54$$

$$y = 17$$

Teoria

$$f(x) : \hat{y} = 1x + 53$$



W jaki sposób szukać parametrów w i b aby wynik był satysfakcjonujący?

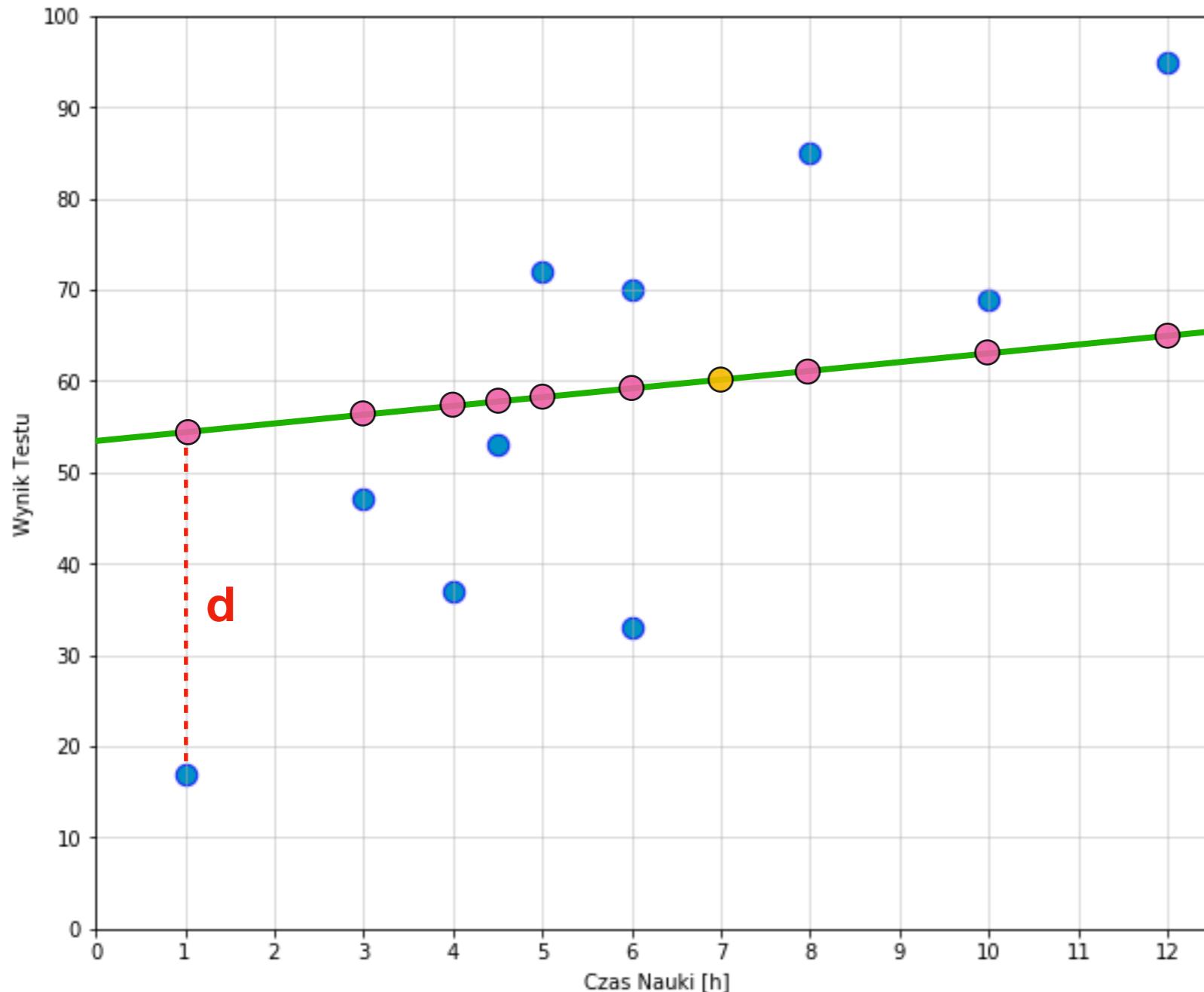
$$f(1) = 54$$

$$y = 17$$

$$d = |y - f(1)|$$

Teoria

$$f(x) : \hat{y} = 1x + 53$$



W jaki sposób szukać parametrów w i b aby wynik był satysfakcjonujący?

$$f(1) = 54$$

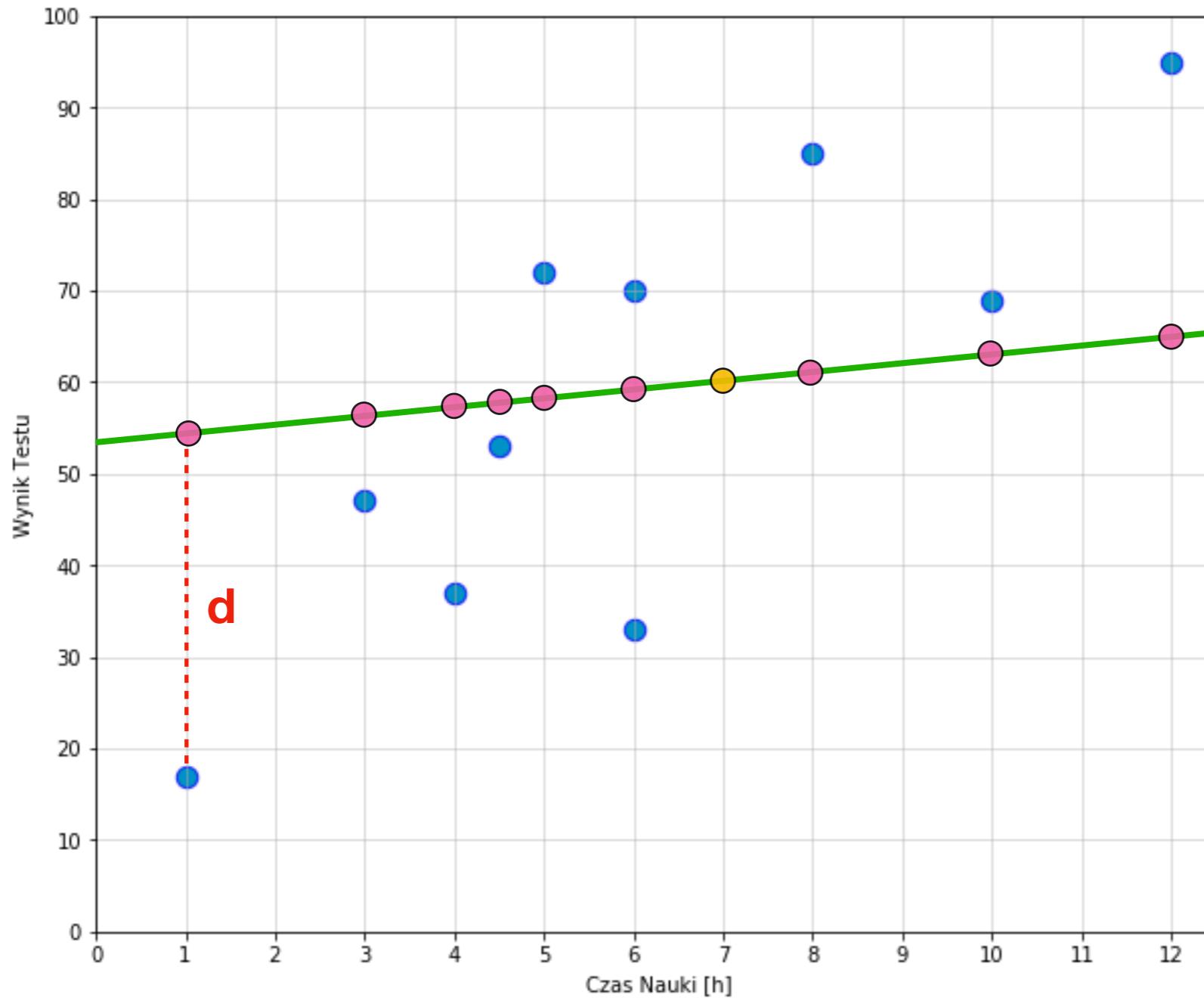
$$y = 17$$

$$d = |y - f(1)|$$

$$d = 37$$

Teoria

$$f(x) : \hat{y} = 1x + 53$$



W jaki sposób szukać parametrów w i b aby wynik był satysfakcjonujący?

$$f(1) = 54$$

$$y = 17$$

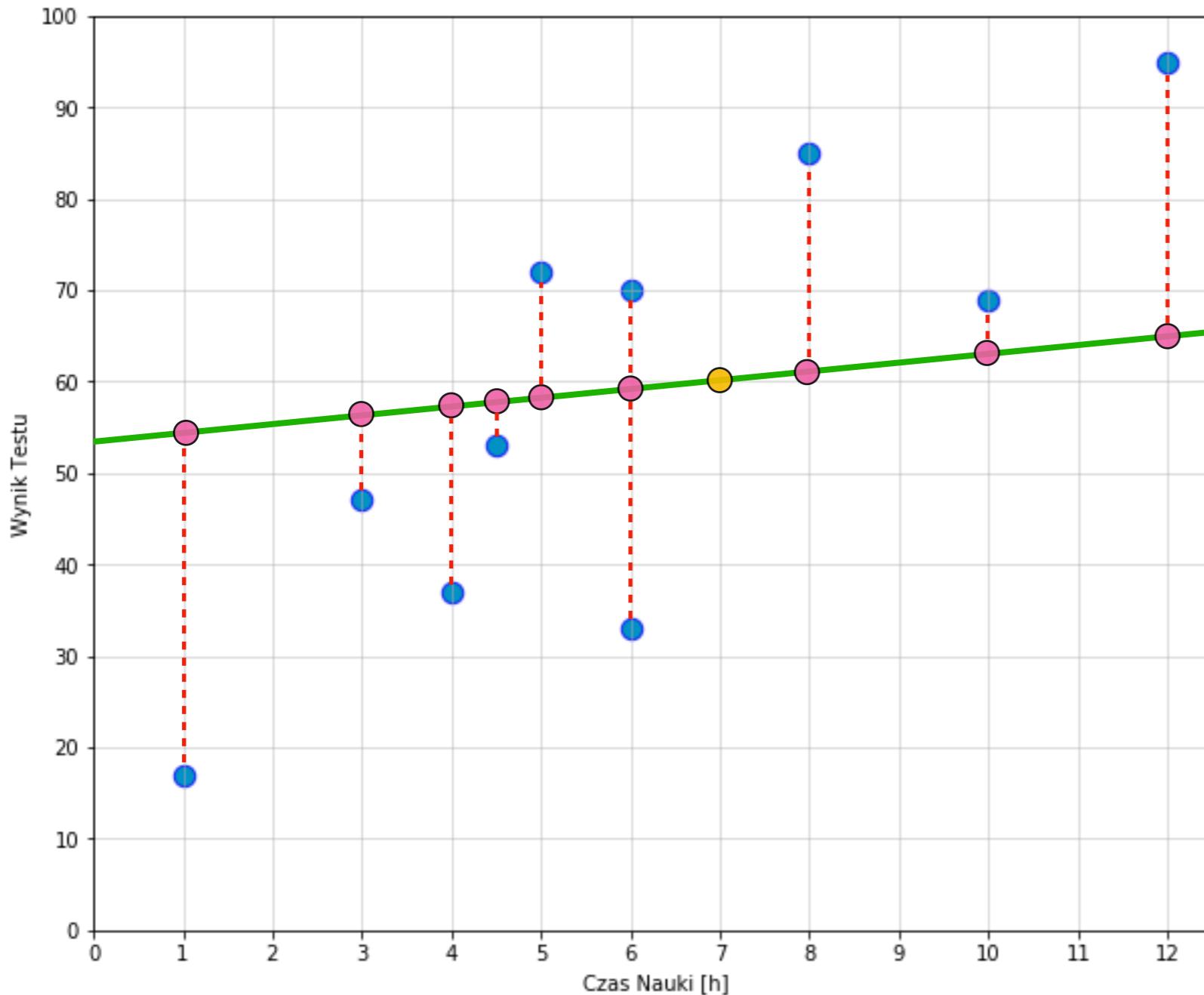
$$d = |y - f(1)|$$

$$d = 37$$

Pomyliliśmy się o 37, chcielibyśmy się mylić o 0!

Teoria

$$f(x) : \hat{y} = 1x + 57$$



I to nie jedyna
pomyłka!

Teoria

$$f(x) : \hat{y} = 1x + 57$$

Id	Czas Nauki	Wynik Testu
0	6	70
1	1	17
2	5	72
3	3	47
4	4.5	53
5	10	69
6	12	95
7	8	85
8	6	33
9	4	37

Teoria

$$f(x) : \hat{y} = 1x + 57$$

Id	Czas Nauki	Wynik Testu
0	$x_0 = 6$	70
1	$x_1 = 1$	17
2	$x_2 = 5$	72
3	$x_3 = 3$	47
4	$x_4 = 4.5$	53
5	$x_5 = 10$	69
6	$x_6 = 12$	95
7	$x_7 = 8$	85
8	$x_8 = 6$	33
9	$x_9 = 4$	37

Teoria

$$f(x) : \hat{y} = 1x + 57$$

Id	Czas Nauki	Wynik Testu
0	$x_0 = 6$	$y_0 = 70$
1	$x_1 = 1$	$y_1 = 17$
2	$x_2 = 5$	$y_2 = 72$
3	$x_3 = 3$	$y_3 = 47$
4	$x_4 = 4.5$	$y_4 = 53$
5	$x_5 = 10$	$y_5 = 69$
6	$x_6 = 12$	$y_6 = 95$
7	$x_7 = 8$	$y_7 = 85$
8	$x_8 = 6$	$y_8 = 33$
9	$x_9 = 4$	$y_9 = 37$

Teoria

$$f(x) : \hat{y} = 1x + 57$$

Id	Czas Nauki	Wynik Testu
0	$x_0 = 6$	$y_0 = 70$
1	$x_1 = 1$	$y_1 = 17$
2	$x_2 = 5$	$y_2 = 72$
3	$x_3 = 3$	$y_3 = 47$
4	$x_4 = 4.5$	$y_4 = 53$
5	$x_5 = 10$	$y_5 = 69$
6	$x_6 = 12$	$y_6 = 95$
7	$x_7 = 8$	$y_7 = 85$
8	$x_8 = 6$	$y_8 = 33$
9	$x_9 = 4$	$y_9 = 37$

Wejście Oczekiwane
do funkcji rezultaty

Teoria

$$f(x) : \hat{y} = 1x + 57$$

Id	Czas Nauki	Wynik Testu	f(x)
0	$x_0 = 6$	$y_0 = 70$	$\hat{y}_0 = 63$
1	$x_1 = 1$	$y_1 = 17$	$\hat{y}_1 = 58$
2	$x_2 = 5$	$y_2 = 72$	$\hat{y}_2 = 62$
3	$x_3 = 3$	$y_3 = 47$	$\hat{y}_3 = 60$
4	$x_4 = 4.5$	$y_4 = 53$	$\hat{y}_4 = 61.5$
5	$x_5 = 10$	$y_5 = 69$	$\hat{y}_5 = 67$
6	$x_6 = 12$	$y_6 = 95$	$\hat{y}_6 = 69$
7	$x_7 = 8$	$y_7 = 85$	$\hat{y}_7 = 65$
8	$x_8 = 6$	$y_8 = 33$	$\hat{y}_8 = 63$
9	$x_9 = 4$	$y_9 = 37$	$\hat{y}_9 = 61$

Wejście do funkcji Oczekiwane rezultaty Predykcje

Teoria

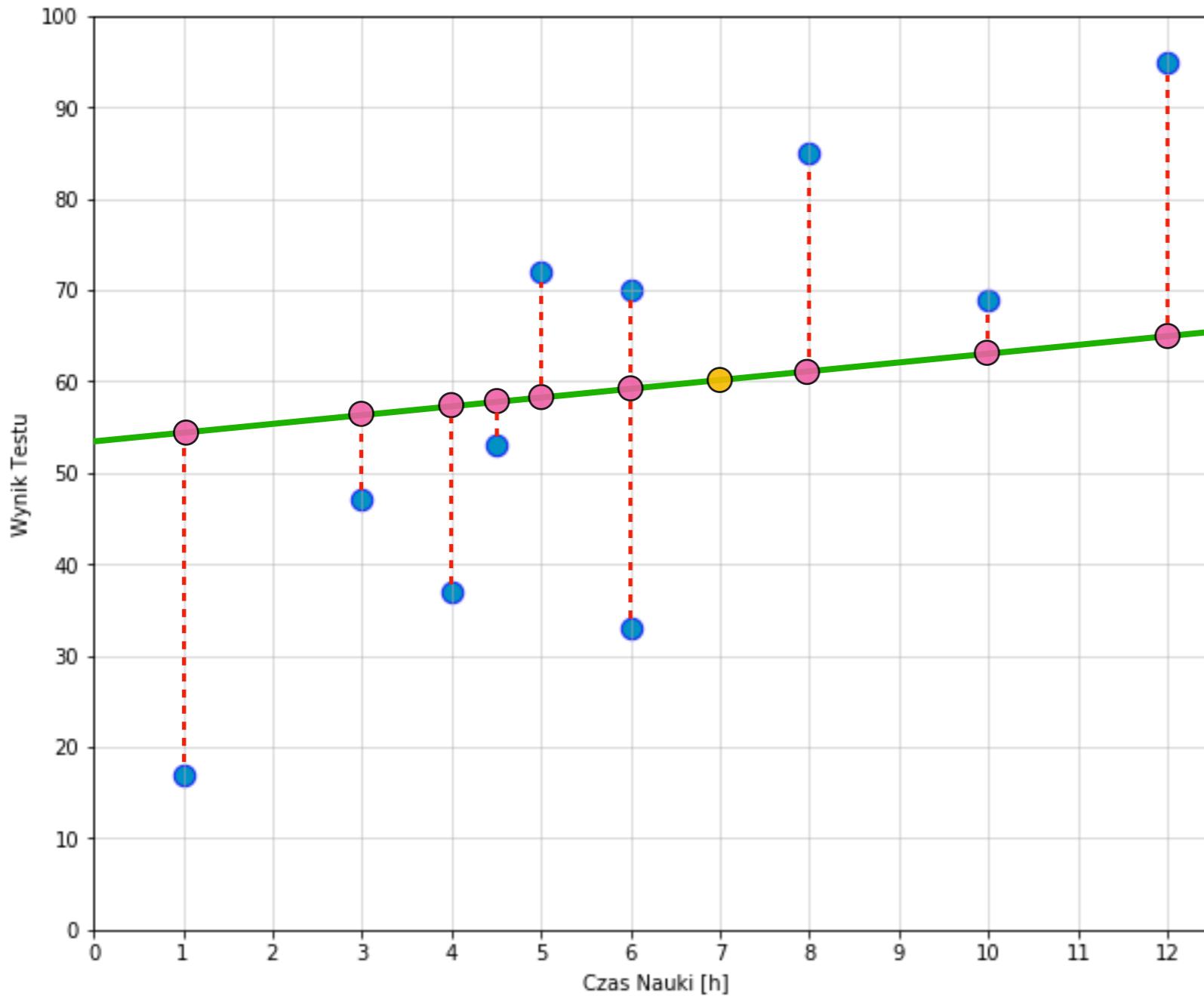
$$f(x) : \hat{y} = 1x + 57$$

Id	Czas Nauki	Wynik Testu	f(x)	d = y - \hat{y}
0	x ₀ = 6	y ₀ = 70	$\hat{y}_0 = 63$	d ₀ = 7
1	x ₁ = 1	y ₁ = 17	$\hat{y}_1 = 58$	d ₁ = 41
2	x ₂ = 5	y ₂ = 72	$\hat{y}_2 = 62$	d ₂ = 10
3	x ₃ = 3	y ₃ = 47	$\hat{y}_3 = 60$	d ₃ = 13
4	x ₄ = 4.5	y ₄ = 53	$\hat{y}_4 = 61.5$	d ₄ = 8.5
5	x ₅ = 10	y ₅ = 69	$\hat{y}_5 = 67$	d ₅ = 2
6	x ₆ = 12	y ₆ = 95	$\hat{y}_6 = 69$	d ₆ = 26
7	x ₇ = 8	y ₇ = 85	$\hat{y}_7 = 65$	d ₇ = 20
8	x ₈ = 6	y ₈ = 33	$\hat{y}_8 = 63$	d ₈ = 30
9	x ₉ = 4	y ₉ = 37	$\hat{y}_9 = 61$	d ₉ = 34

Wejście do funkcji Oczekiwane rezultaty Predykcje Błędy

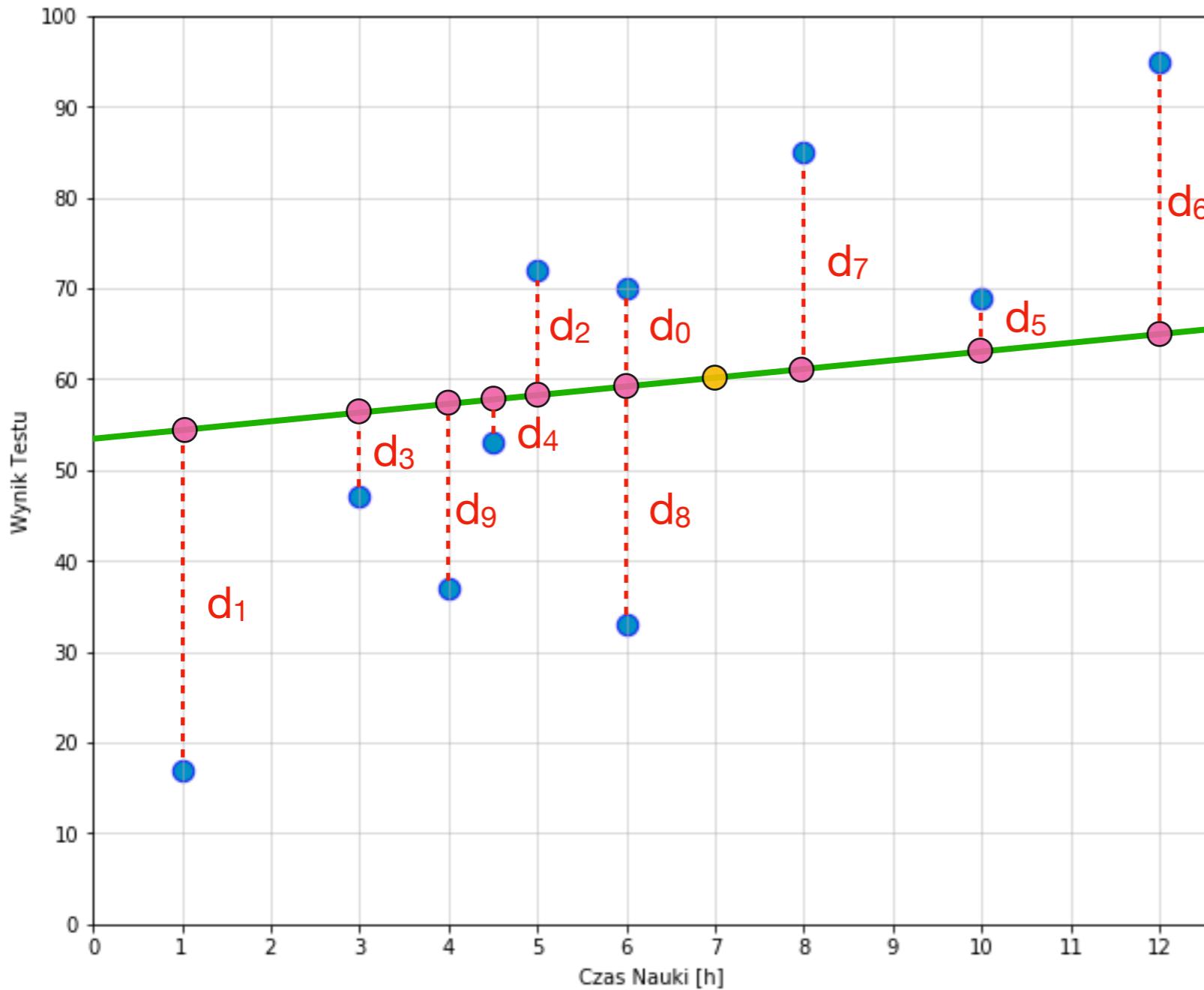
Teoria

$$f(x) : \hat{y} = 1x + 57$$



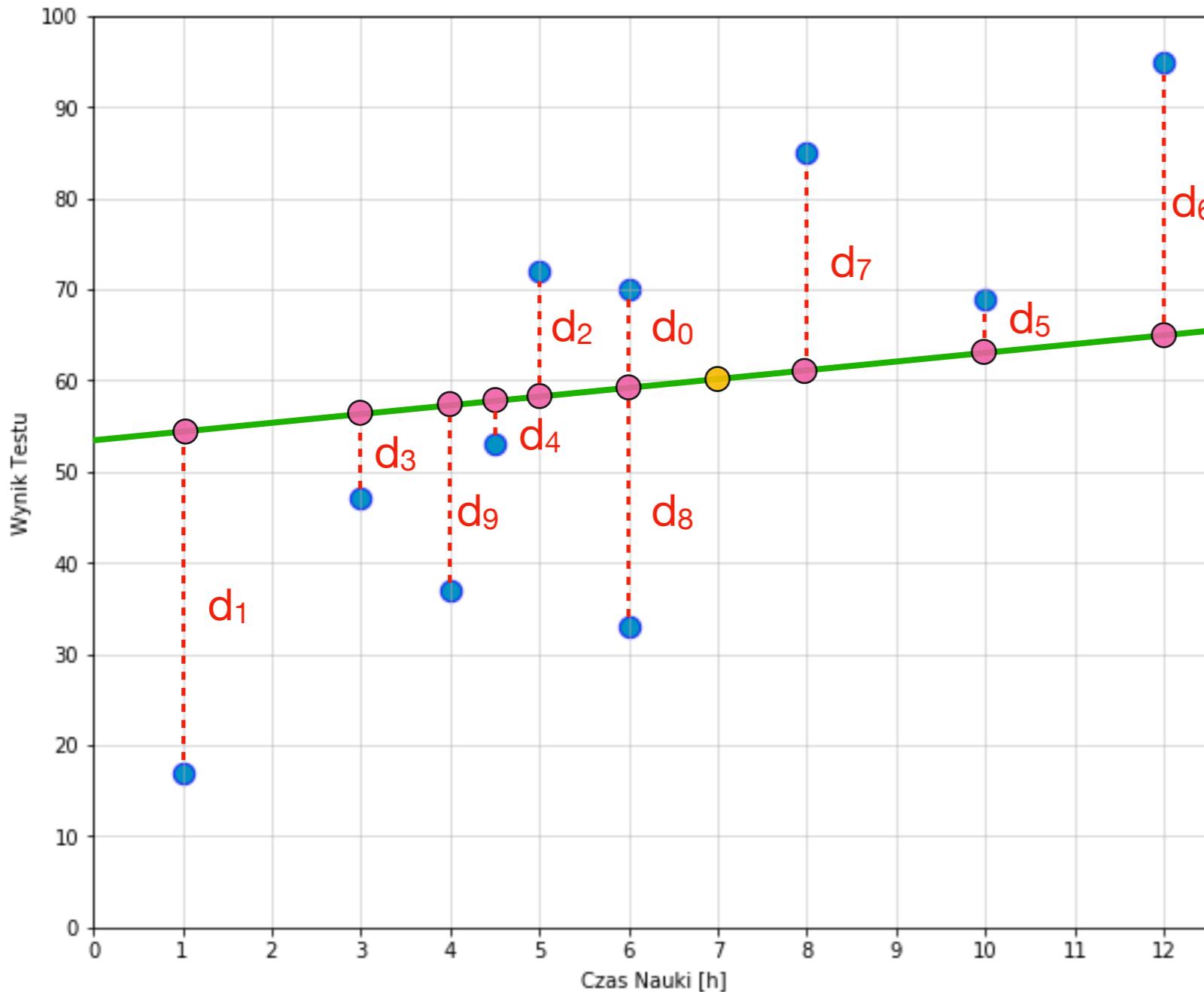
Teoria

$$f(x) : \hat{y} = 1x + 57$$



Teoria

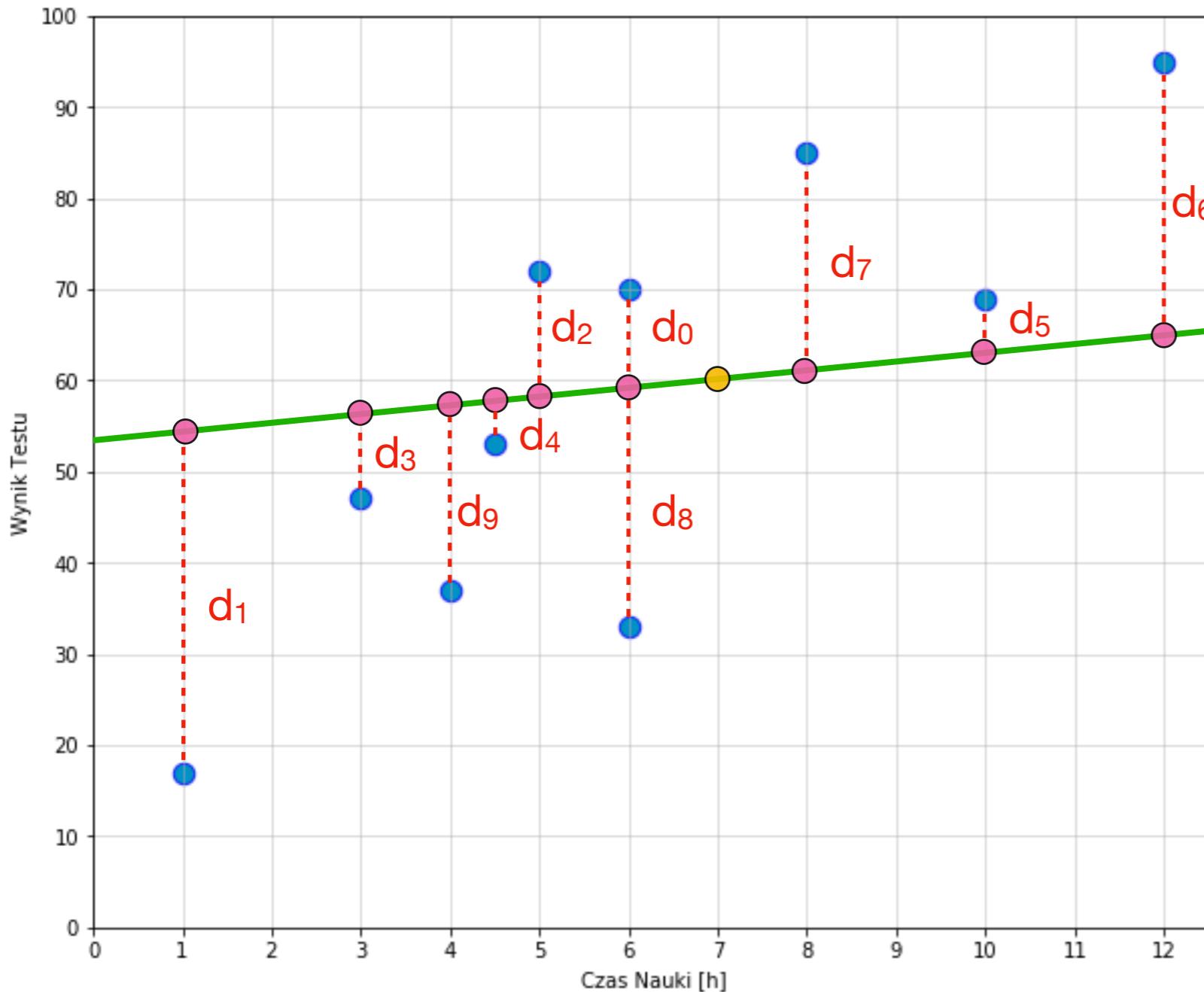
$$f(x) : \hat{y} = 1x + 57$$



$$\begin{aligned} J = & d_0 + d_1 + d_2 + d_3 + d_4 \\ & + d_5 + d_6 + d_7 + d_8 + d_9 \end{aligned}$$

Teoria

$$f(x) : \hat{y} = 1x + 57$$

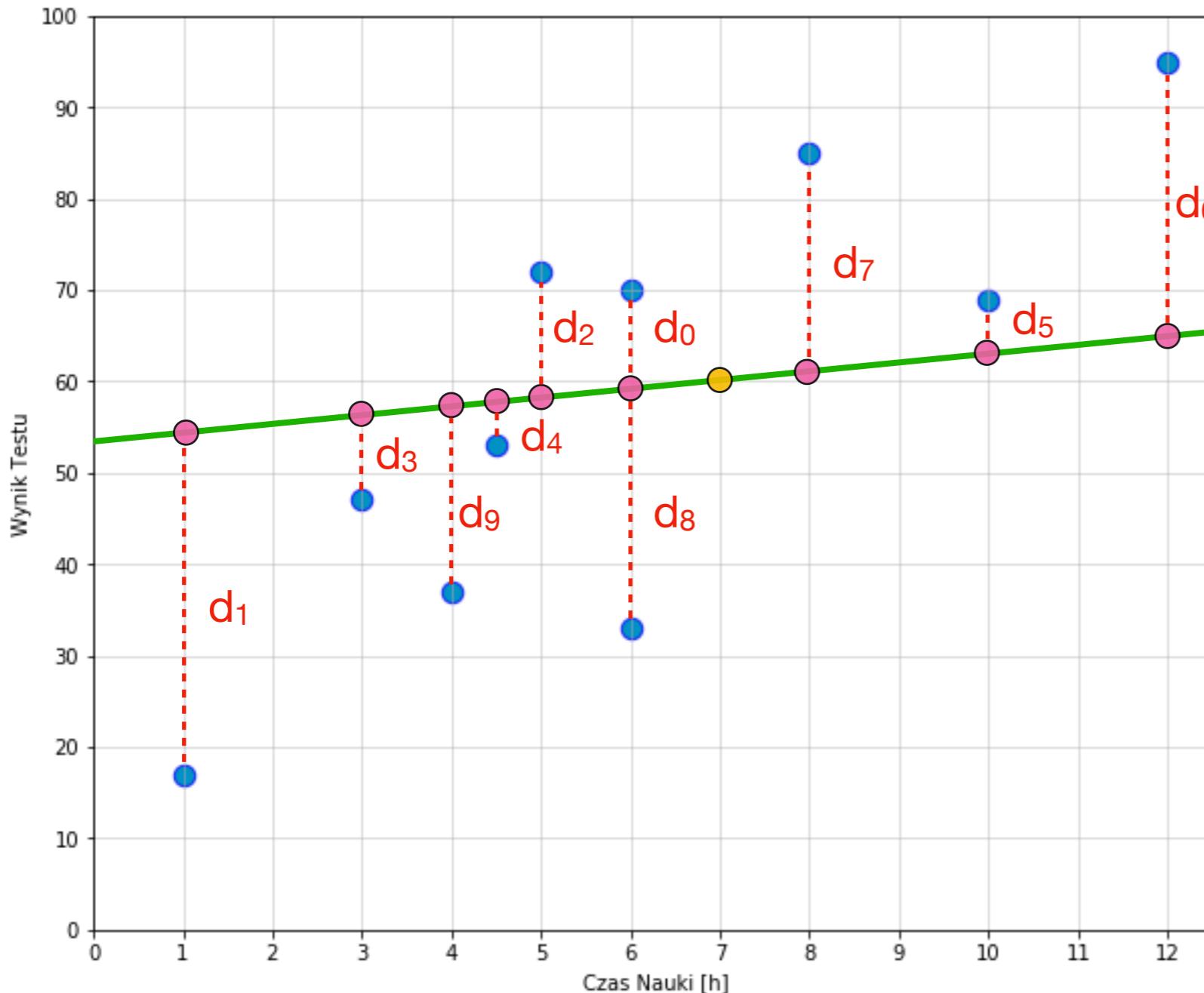


$$\begin{aligned} J = & d_0 + d_1 + d_2 + d_3 + d_4 \\ & + d_5 + d_6 + d_7 + d_8 + d_9 \end{aligned}$$

$$J = 191,5$$

Teoria

$$f(x) : \hat{y} = 1x + 57$$



$$\begin{aligned} J &= d_0 + d_1 + d_2 + d_3 + d_4 \\ &+ d_5 + d_6 + d_7 + d_8 + d_9 \end{aligned}$$

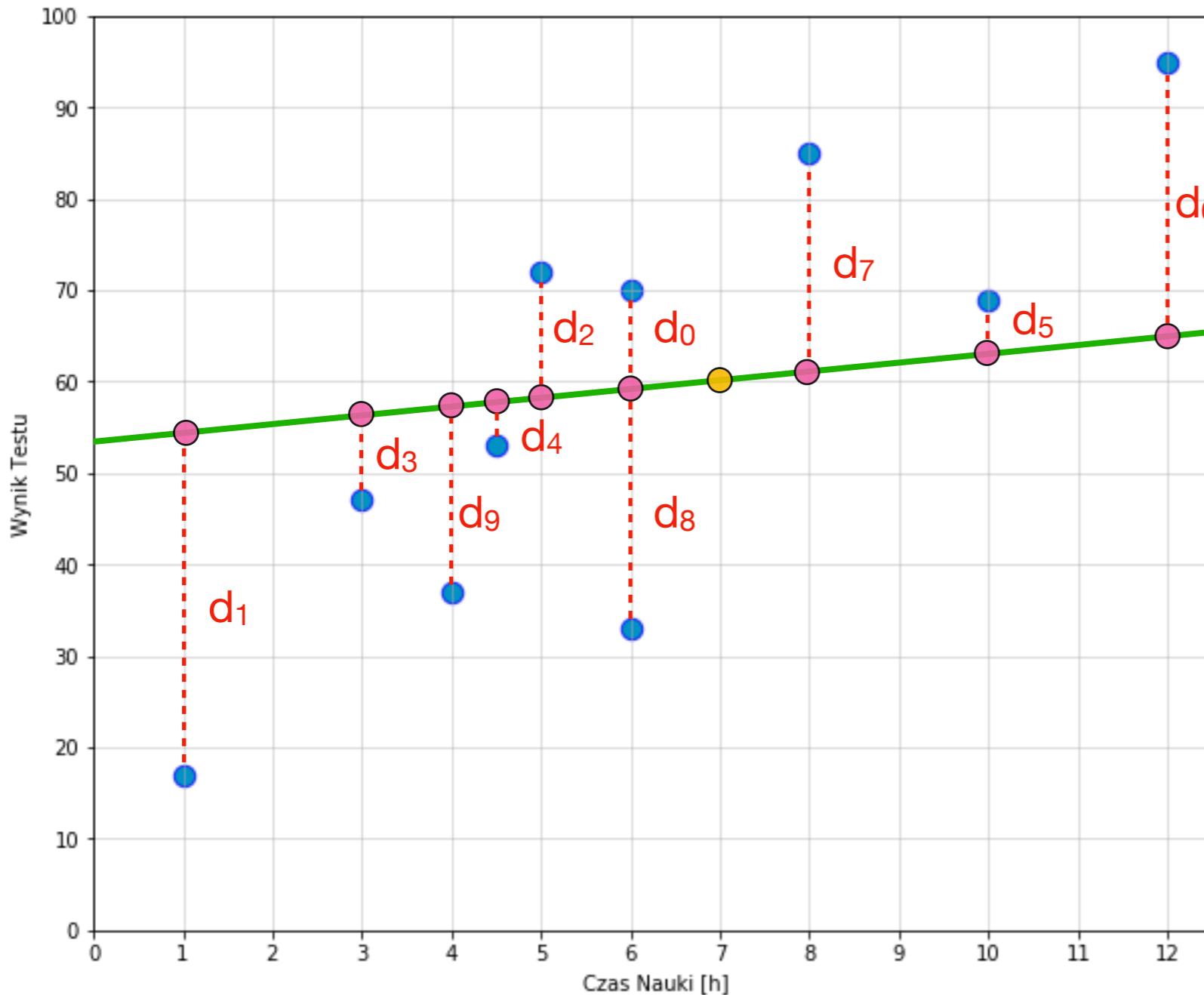
$$J = 191,5$$

$$J_{\text{średnie}} = 191,5 / 10$$

↑
ilość
danych

Teoria

$$f(x) : \hat{y} = 1x + 57$$



$$\begin{aligned} J &= d_0 + d_1 + d_2 + d_3 + d_4 \\ &+ d_5 + d_6 + d_7 + d_8 + d_9 \end{aligned}$$

$$J = 191,5$$

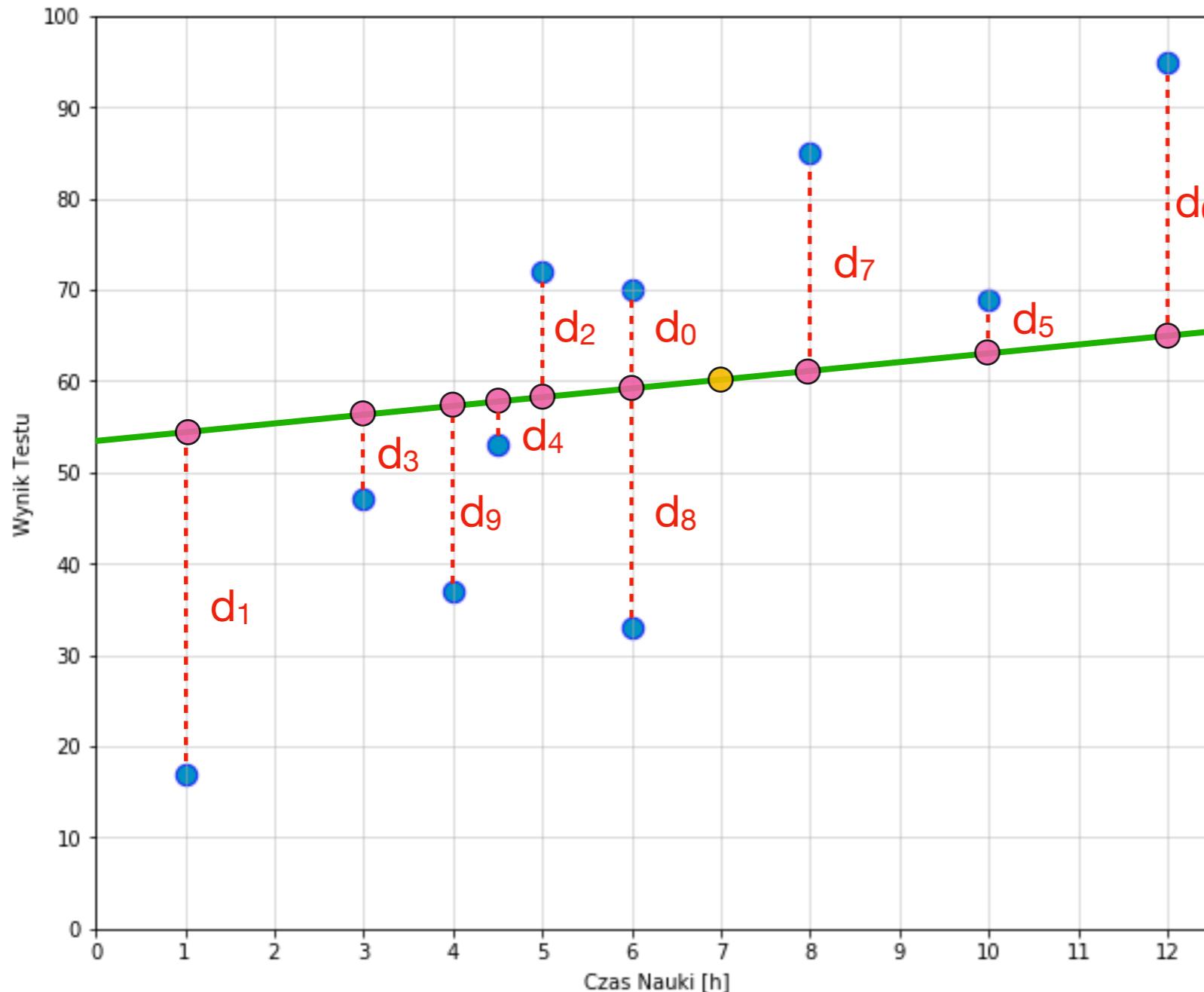
$$J_{\text{średnie}} = 191,5 / 10$$

↑
ilość
danych

$$J_{\text{średnie}} = 19,15$$

Teoria

$$f(x) : \hat{y} = 1x + 57$$



$$\begin{aligned} J &= d_0 + d_1 + d_2 + d_3 + d_4 \\ &+ d_5 + d_6 + d_7 + d_8 + d_9 \end{aligned}$$

$$J = 191,5$$

$$J_{\text{średnie}} = 191,5 / 10$$

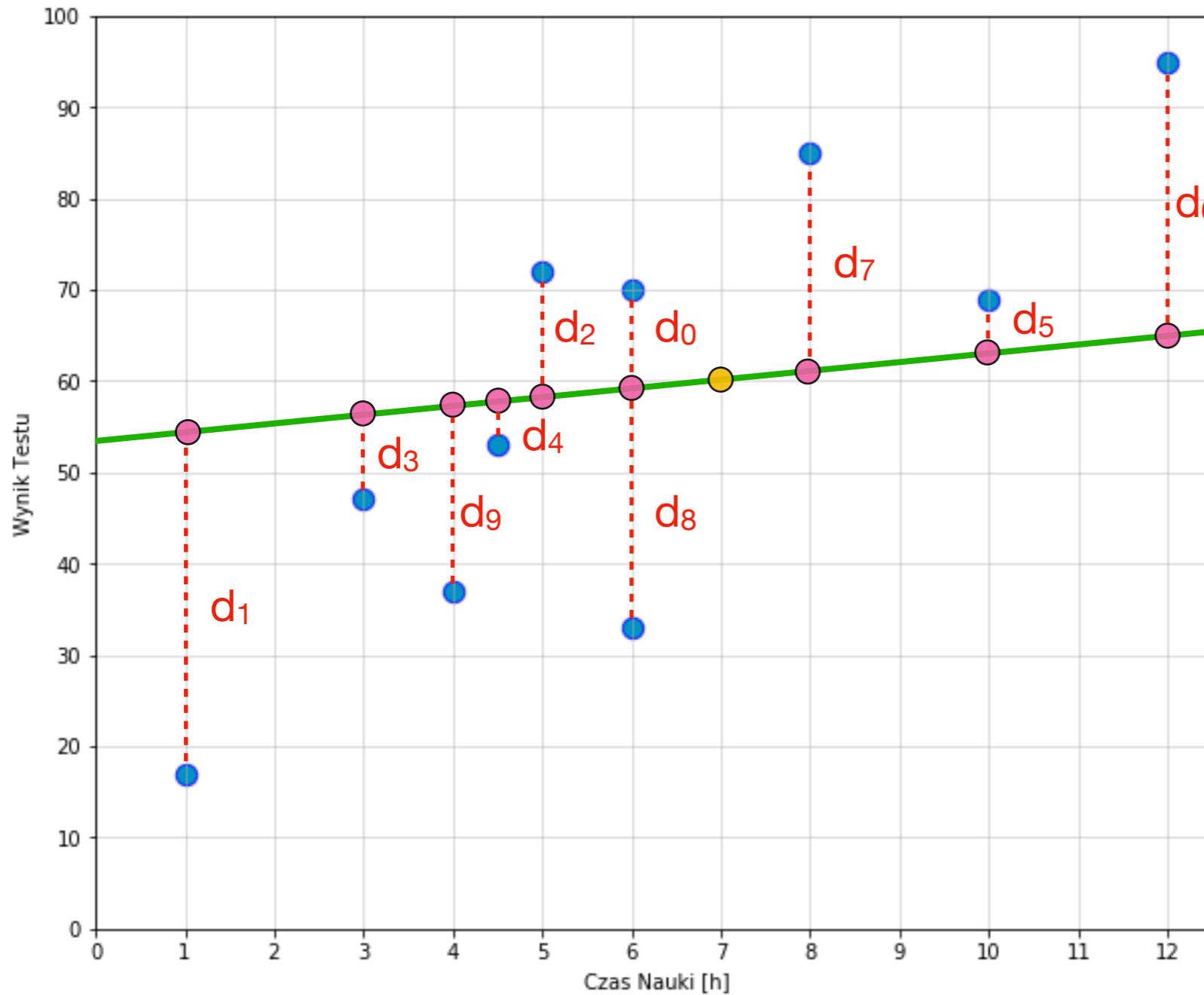
↑
ilość
danych

$$J_{\text{średnie}} = 19,15$$

Średni błąd na punkt wynosi
19.15, a chcemy by było 0.

Teoria

$$f(x) : \hat{y} = 1x + 57$$



$$\begin{aligned} J &= d_0 + d_1 + d_2 + d_3 + d_4 \\ &+ d_5 + d_6 + d_7 + d_8 + d_9 \end{aligned}$$

$$J = 191,5$$

$$J_{\text{średnie}} = 191,5 / 10$$

↑
ilość
danych

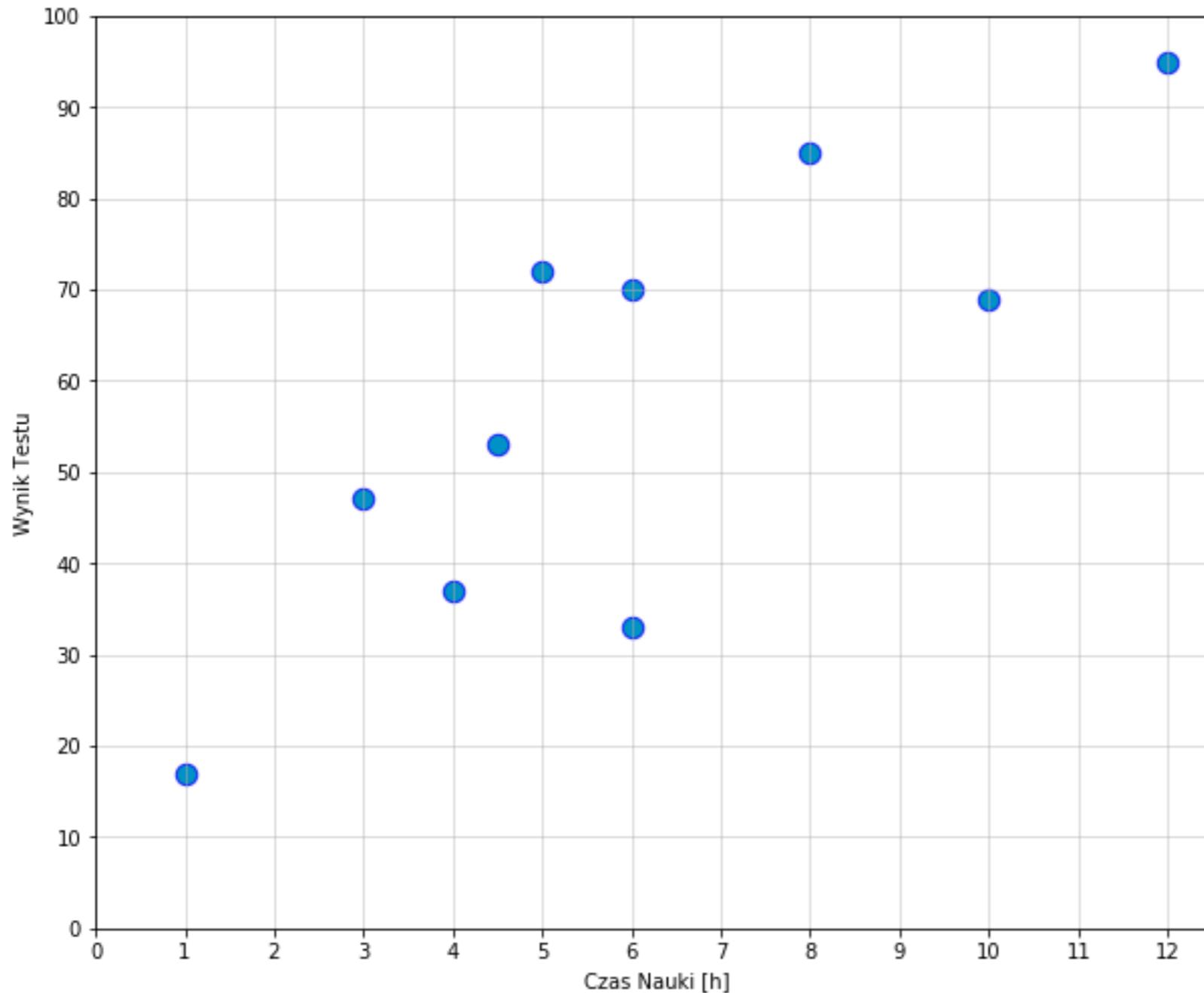
$$J_{\text{średnie}} = 19,15$$

Średni błąd na punkt wynosi
19.15, a chcemy by było 0.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Teoria

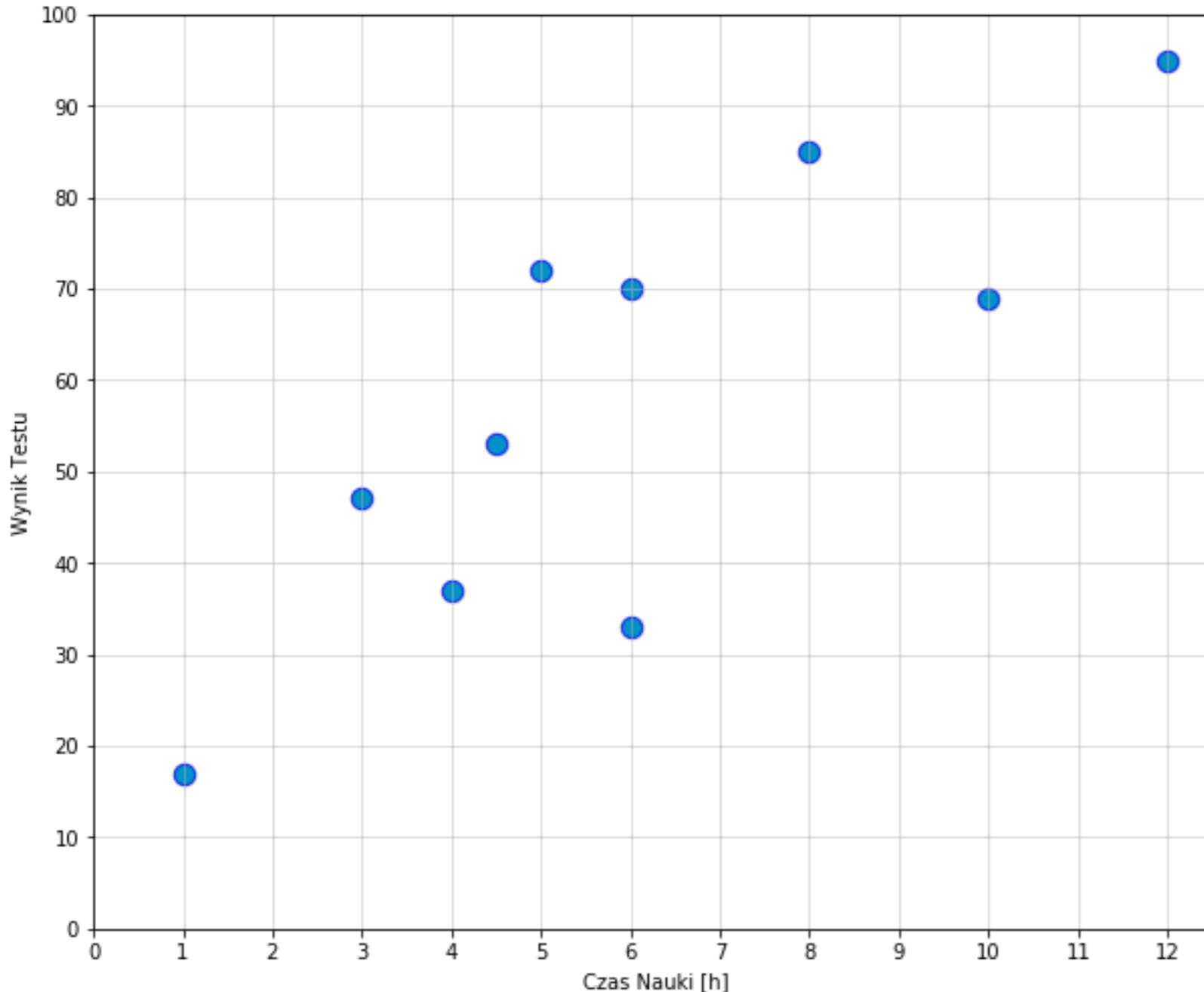
$$f(x) : \hat{y} = wx + b$$



Chcemy znaleźć takie w i b aby
funkcja kosztu była jak
najmniejsza!

Teoria

$$f(x) : \hat{y} = wx + b$$

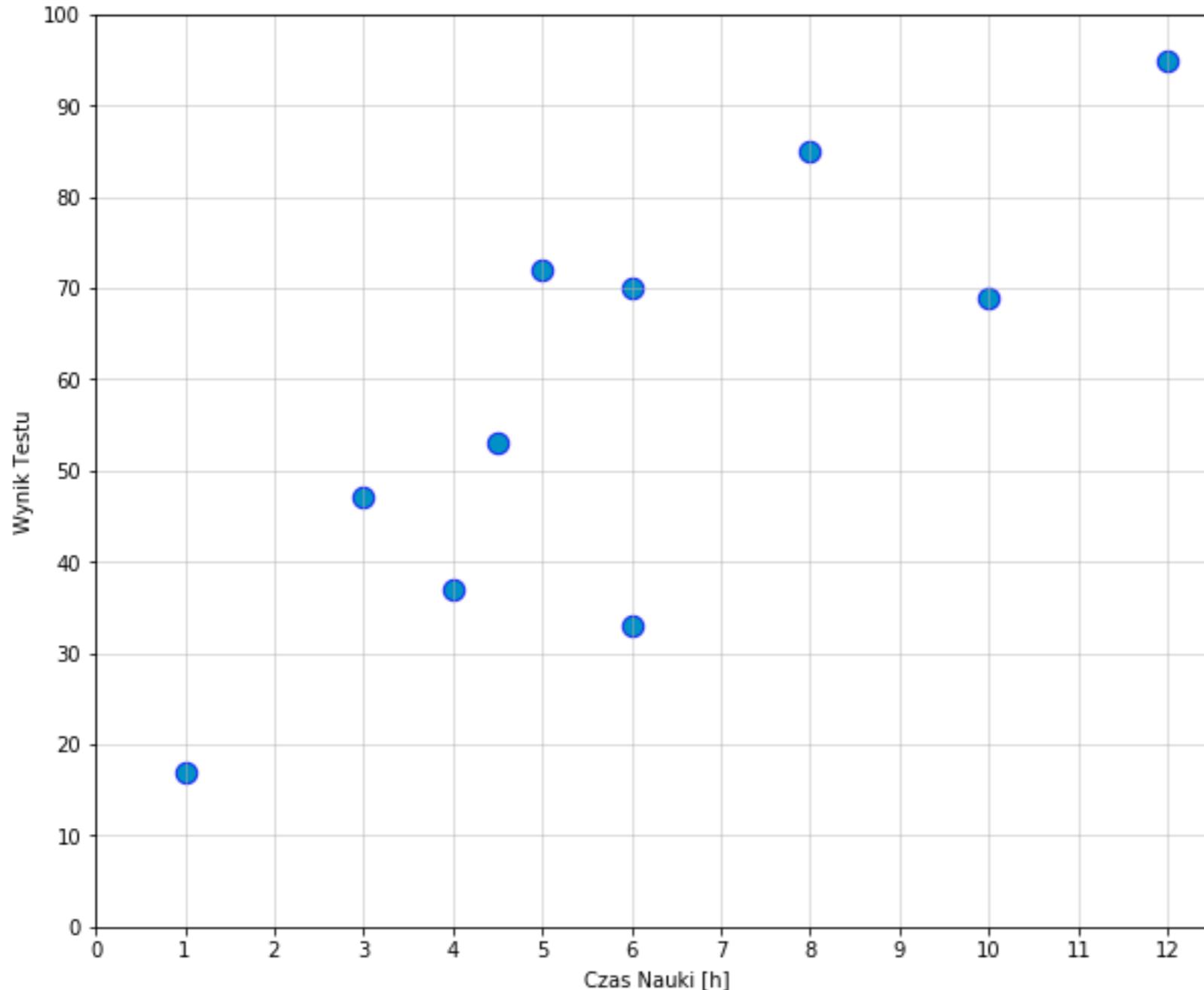


Chcemy znaleźć takie w i b aby
funkcja kosztu była jak
najmniejsza!

Nie uda się znaleźć funkcji
liniowej, która przetnie
każdy punkt więc **błąd nigdy**
nie będzie wynosić 0.

Teoria

$$f(x) : \hat{y} = wx + b$$



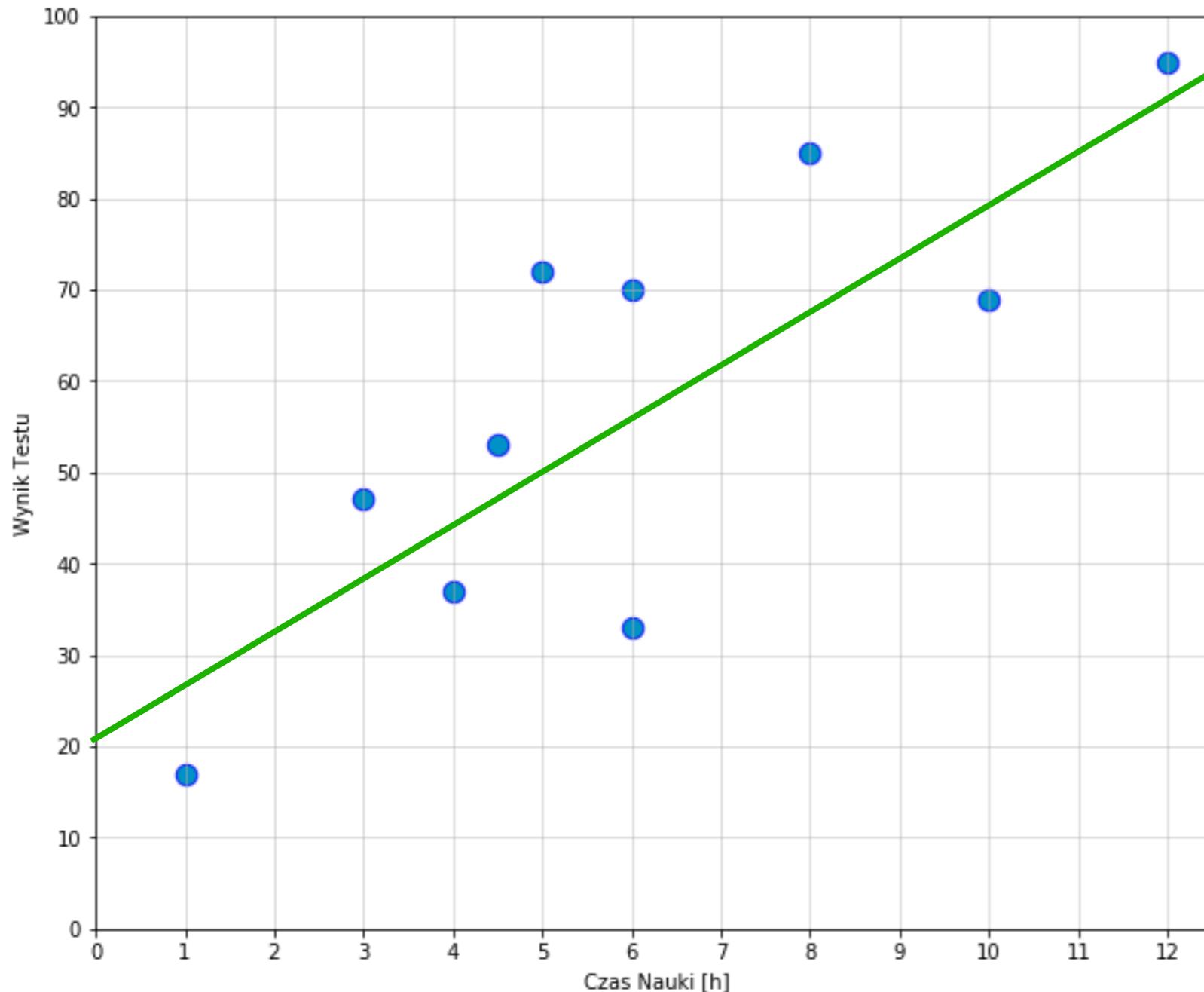
Chcemy znaleźć takie w i b aby
funkcja kosztu była jak
najmniejsza!

Nie uda się znaleźć funkcji
liniowej, która przetnie
każdy punkt więc **błąd nigdy**
nie będzie wynosić 0.

Trzeba iść na kompromis
znać prostą, która jest jak
najbliżej każdego punktu.

Teoria

$$f(x) : \hat{y} = 6.1x + 21.4 \quad MAE = 11.62$$



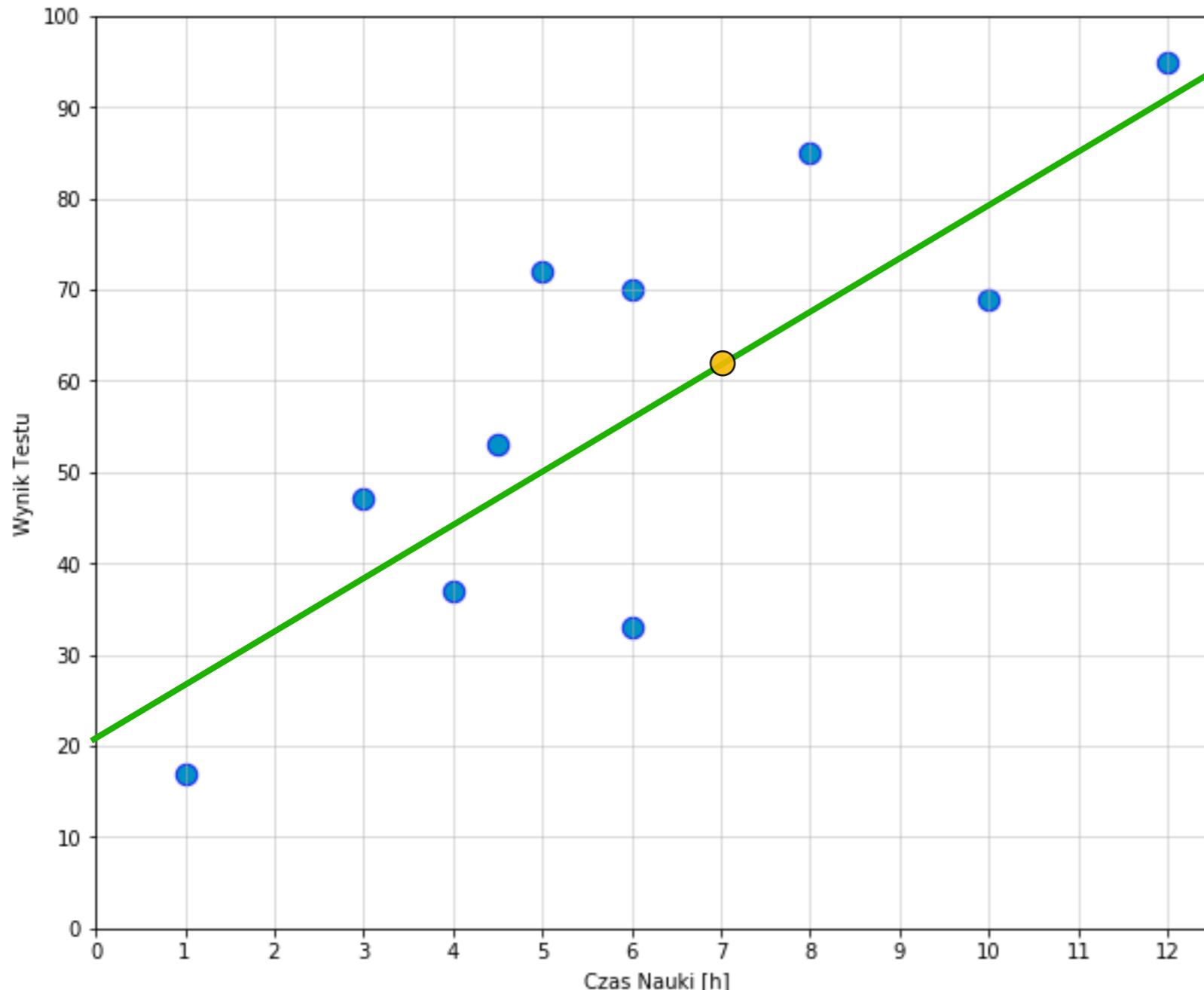
Chcemy znaleźć takie w i b aby funkcja kosztu była jak najmniejsza!

Nie uda się znaleźć funkcji liniowej, która przetnie każdy punkt więc błąd nigdy nie będzie wynosić 0.

Trzeba iść na kompromis znaleźć prostą, która jest jak najbliżej każdego punktu.

Teoria

$$f(x) : \hat{y} = 6.1x + 21.4 \quad MAE = 11.62$$



Chcemy znaleźć takie w i b aby
funkcja kosztu była jak
najmniejsza!

Nie uda się znaleźć funkcji
liniowej, która przetnie
każdy punkt więc **błąd nigdy**
nie będzie wynosić 0.

Trzeba iść na kompromis
znać prostą, która jest jak
najbliżej każdego punktu.

$$f(7) = 61.1$$

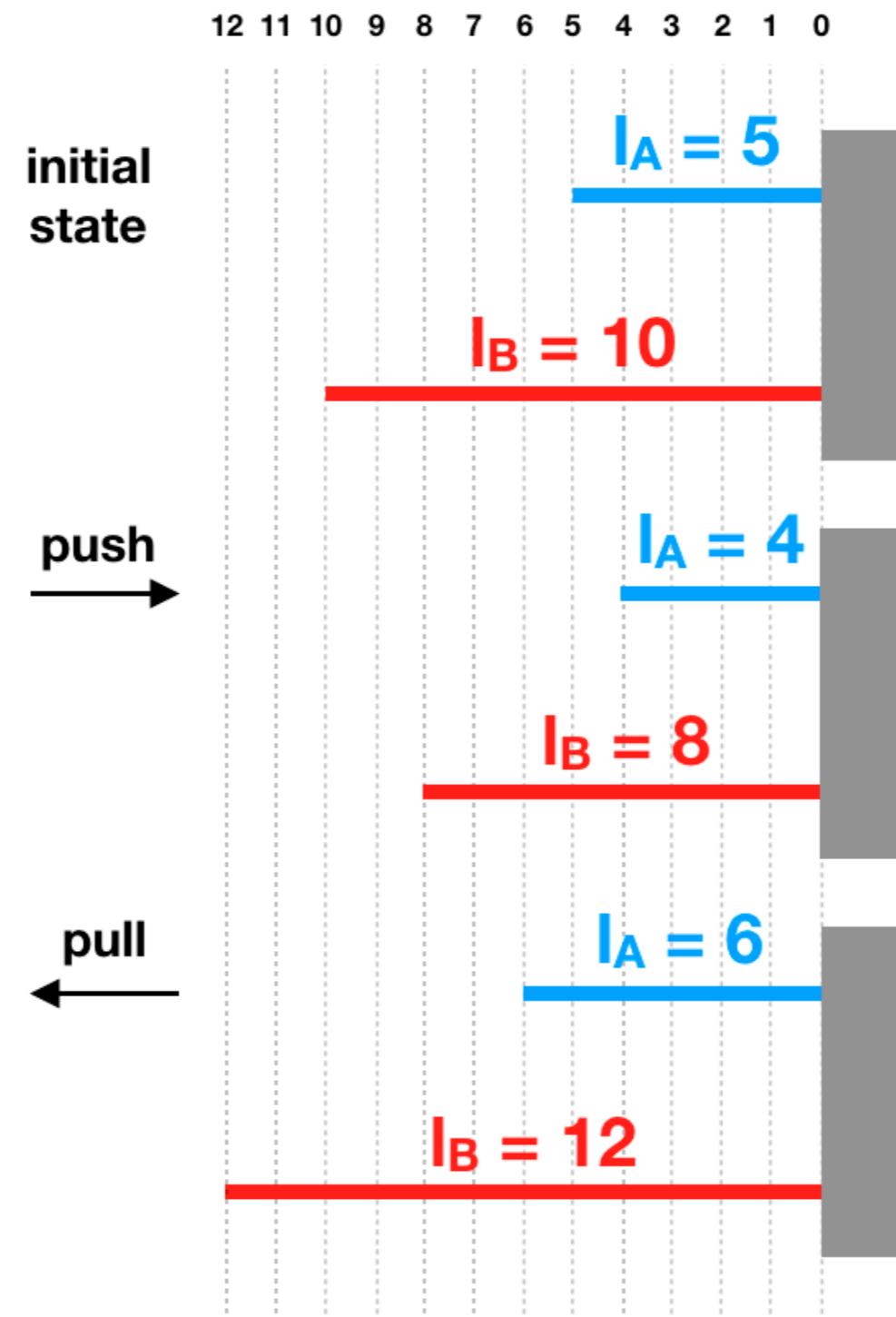
Teoria

Skąd wziąć najlepsze wartości dla parametrów funkcji (w** oraz **b**)?**

Teoria

Skąd wziąć najlepsze wartości dla parametrów funkcji (**w** oraz **b**)?

Algorytm Spadku Gradientu



$$I_B = 2 \cdot I_A$$

$$I_B = 2 \cdot I_A$$



POCHODNA

$$I_B = 2 \cdot I_A$$


POCHODNA

OPISUJE ZALEŻNOŚĆ POMIĘDZY
DOWOLNYMI DWOMA WARTOŚCIAMI WE WZORZE

$$I_B = 2 \cdot I_A$$

↑

POCHODNA

OPISUJE ZALEŻNOŚĆ POMIĘDZY
DOWOLNYMI DWOMA WARTOŚCIAMI WE WZORZE

OPISUJE O ILE ZMIENIA SIĘ DANA WARTOŚĆ
KIEDY MODYFIKOWANA JEST INNA

$$\frac{dI_B}{dI_A} = 2$$

$$I_B = 2 \cdot I_A$$



POCHODNA

OPISUJE ZALEŻNOŚĆ POMIĘDZY
DOWOLNYMI DWOMA WARTOŚCIAMI WE WZORZE

OPISUJE O ILE ZMIENIA SIĘ DANA WARTOŚĆ
KIEDY MODYFIKOWANA JEST INNA

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Gdzie:

n - ilość danych

i - indeks danej

y - oczekiwana wartość

ŷ - przewidziana wartość

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Gdzie:

n - ilość danych

i - indeks danej

y - oczekiwana wartość

\hat{y} - przewidziana wartość

Id	Czas Nauki	Wynik Testu
0	6	70
1	1	17
2	5	72
3	3	47
4	4.5	53
5	10	69
6	12	95
7	8	85
8	6	33
9	4	37

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Gdzie:

n - ilość danych

i - indeks danej

y - oczekiwana wartość

\hat{y} - przewidziana wartość

$$\hat{y} = \mathbf{w}\mathbf{x} + \mathbf{b}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Gdzie:

n - ilość danych

i - indeks danej

y - oczekiwana wartość

\hat{y} - przewidziana wartość

$$\hat{y} = w\mathbf{x} + b$$

$$MAE(w, b) = \frac{1}{n} \sum_{i=1}^n |y_i - (wx_i + b)|$$

Gdzie:

n - ilość danych

i - indeks danej

y - oczekiwana wartość

ŷ - przewidziana wartość

$$MAE(w, b) = \frac{1}{n} \sum_{i=1}^n |y_i - (wx_i + b)|$$

Gdzie:

n - ilość danych

i - indeks danej

y - oczekiwana wartość

\hat{y} - przewidziana wartość

Id	Czas Nauki	Wynik Testu
0	6	70
1	1	17
2	5	72
3	3	47
4	4.5	53
5	10	69
6	12	95
7	8	85
8	6	33
9	4	37

$$MAE(w, b) = \frac{1}{n} \sum_{i=1}^n |y_i - (wx_i + b)|$$

Gdzie:

n - ilość danych

i - indeks danej

y - oczekiwana wartość

ŷ - przewidziana wartość

Jeżeli policzymy:

$$MAE(w, b) = \frac{1}{n} \sum_{i=1}^n |y_i - (wx_i + b)|$$

Gdzie:

n - ilość danych

i - indeks danej

y - oczekiwana wartość

ŷ - przewidziana wartość

Jeżeli policzymy:

$\frac{\partial MAE(w, b)}{\partial w}$ - wiemy jak zmieniać **w** aby **minimalizować MAE**

$$MAE(w, b) = \frac{1}{n} \sum_{i=1}^n |y_i - (wx_i + b)|$$

Gdzie:

n - ilość danych

i - indeks danej

y - oczekiwana wartość

ŷ - przewidziana wartość

Jeżeli policzymy:

$$\frac{\partial MAE(w, b)}{\partial w}$$
 - wiemy jak zmieniać **w** aby **minimalizować MAE**

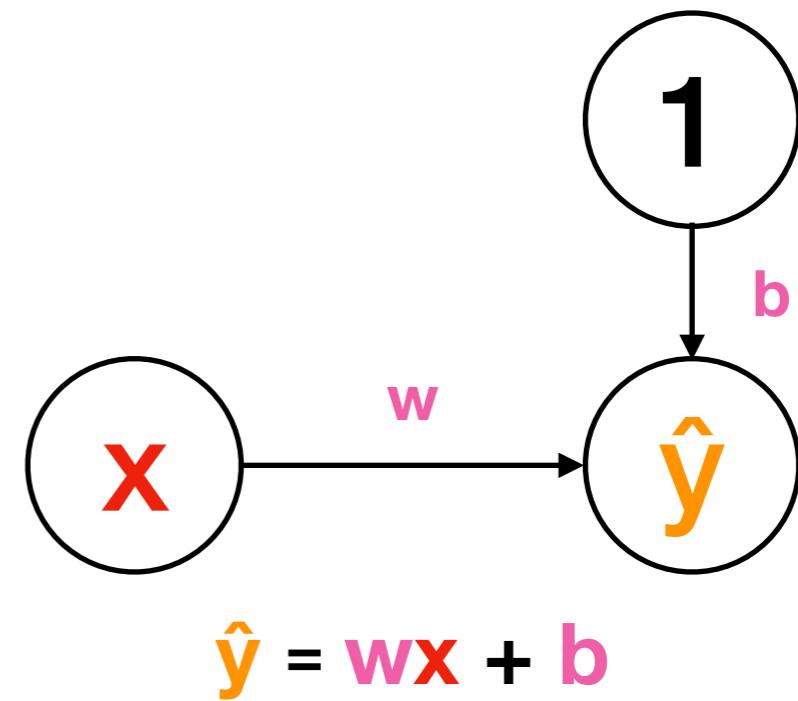
$$\frac{\partial MAE(w, b)}{\partial b}$$
 - wiemy jak zmieniać **b** aby **minimalizować MAE**

Uczenie Nadzorowane

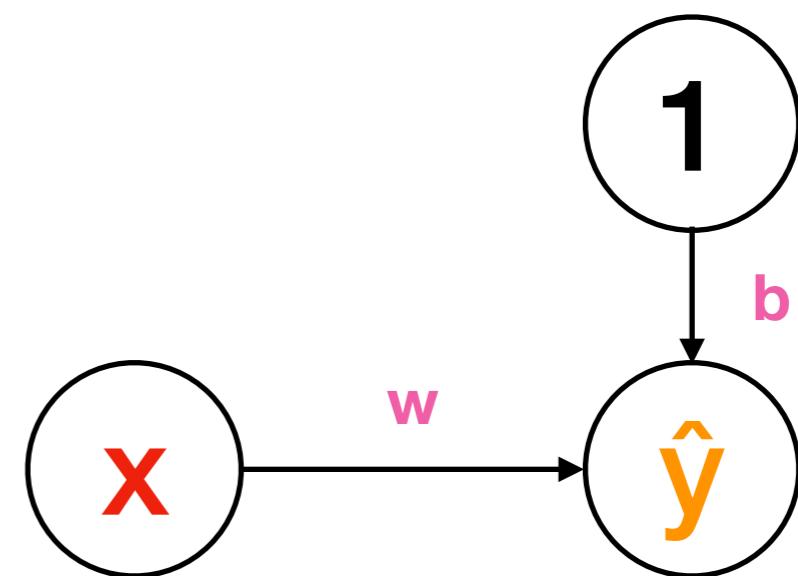
Uczenie Nadzorowane

$$\hat{y} = \textcolor{magenta}{w}\textcolor{red}{x} + \textcolor{magenta}{b}$$

Uczenie Nadzorowane

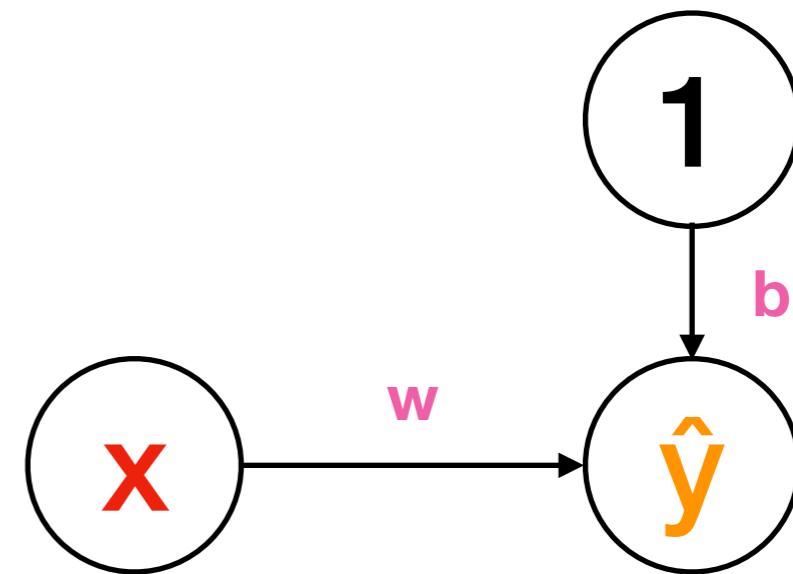


Uczenie Nadzorowane



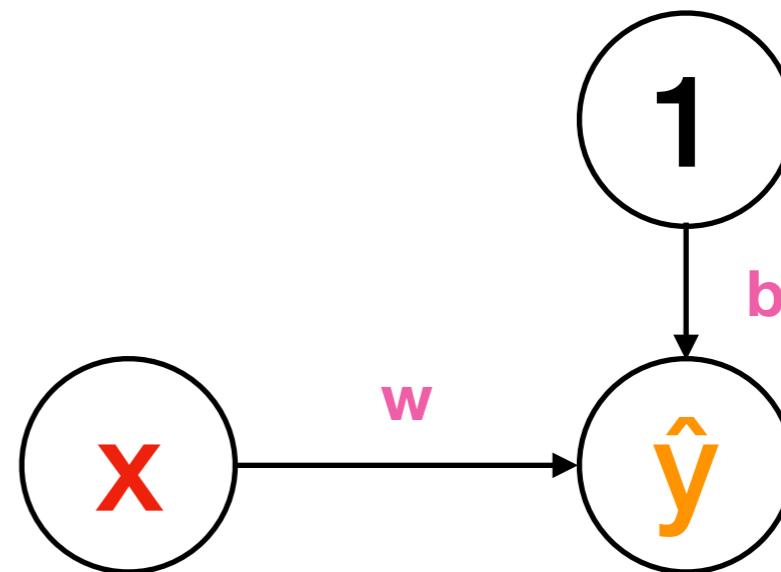
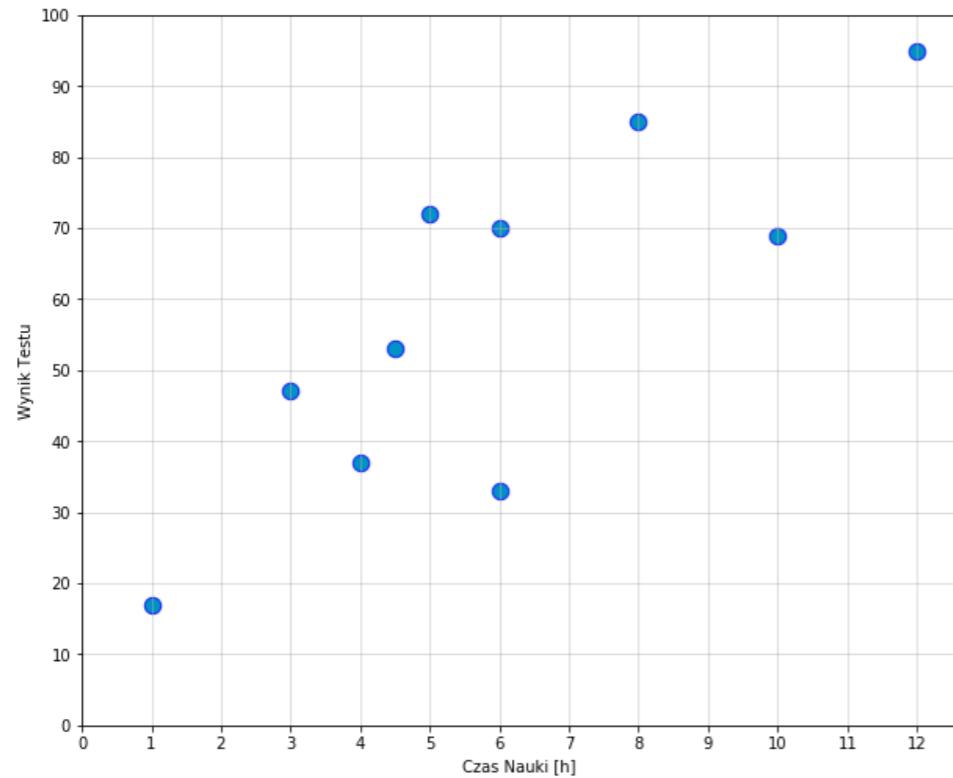
Uczenie Nadzorowane

Id	Czas Nauki	Wynik Testu
0	6	70
1	1	17
2	5	72
3	3	47
4	4.5	53
5	10	69
6	12	95
7	8	85
8	6	33
9	4	37



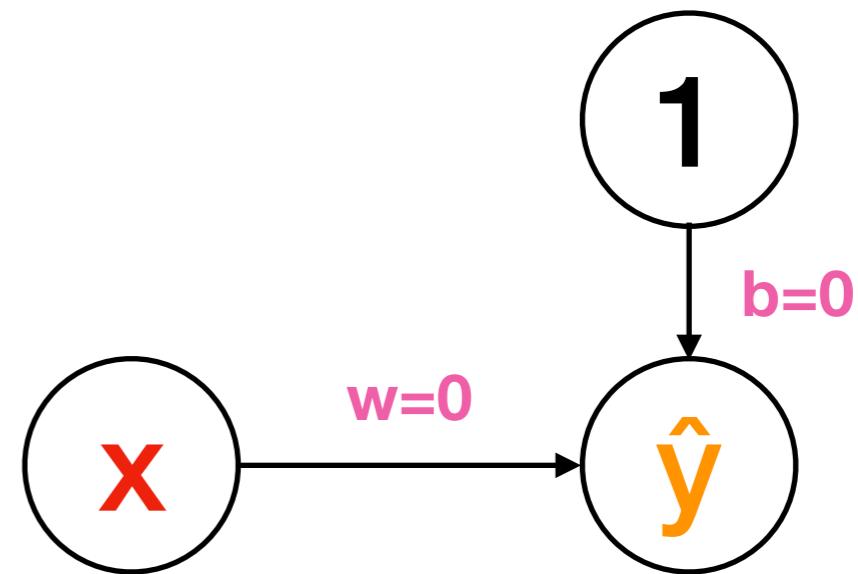
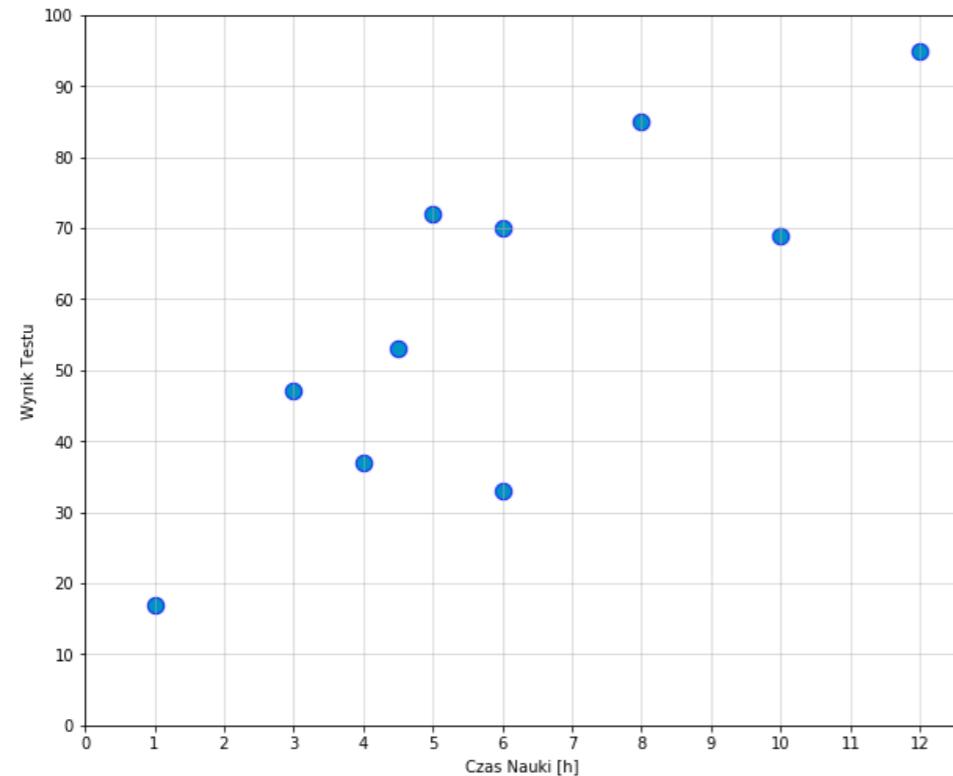
Uczenie Nadzorowane

Id	Czas Nauki	Wynik Testu
0	6	70
1	1	17
2	5	72
3	3	47
4	4.5	53
5	10	69
6	12	95
7	8	85
8	6	33
9	4	37



Uczenie Nadzorowane

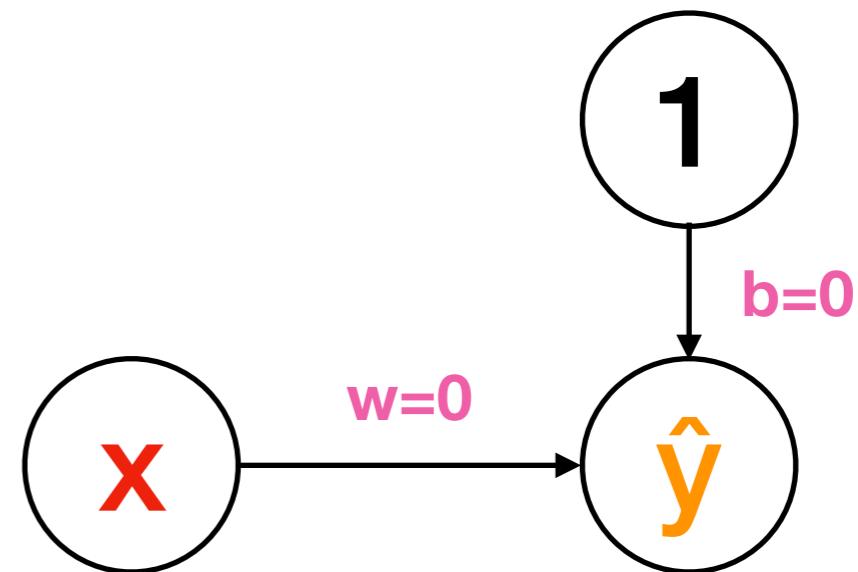
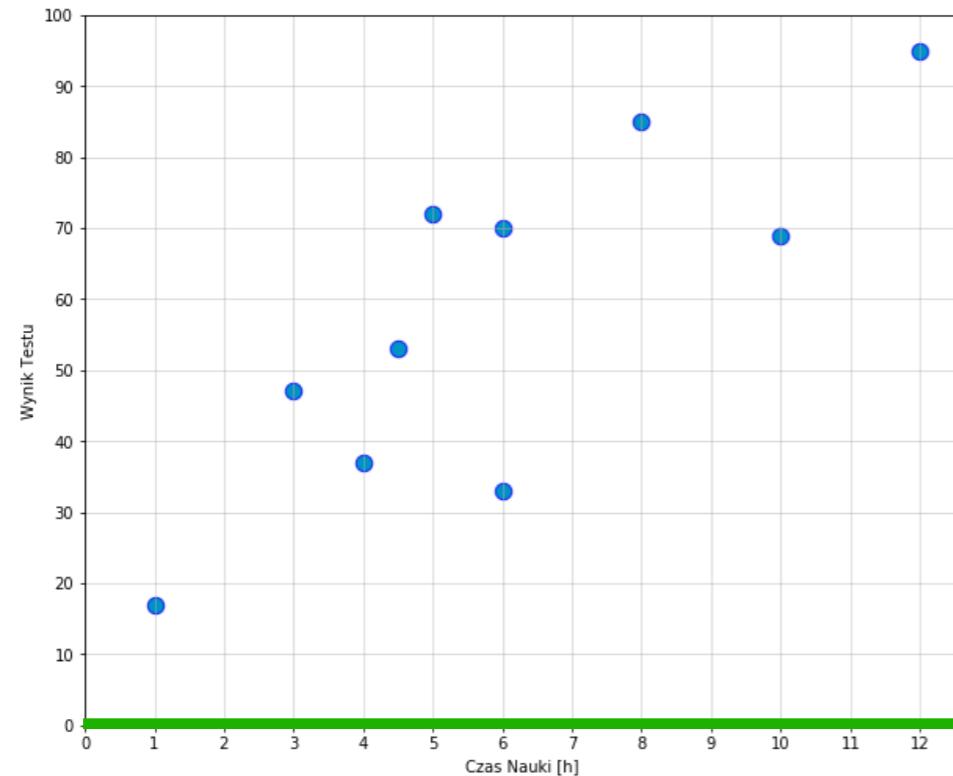
Id	Czas Nauki	Wynik Testu
0	6	70
1	1	17
2	5	72
3	3	47
4	4.5	53
5	10	69
6	12	95
7	8	85
8	6	33
9	4	37



**Inicjalizujemy model poprzez
nadanie wartości początkowych
dla w oraz b .**

Uczenie Nadzorowane

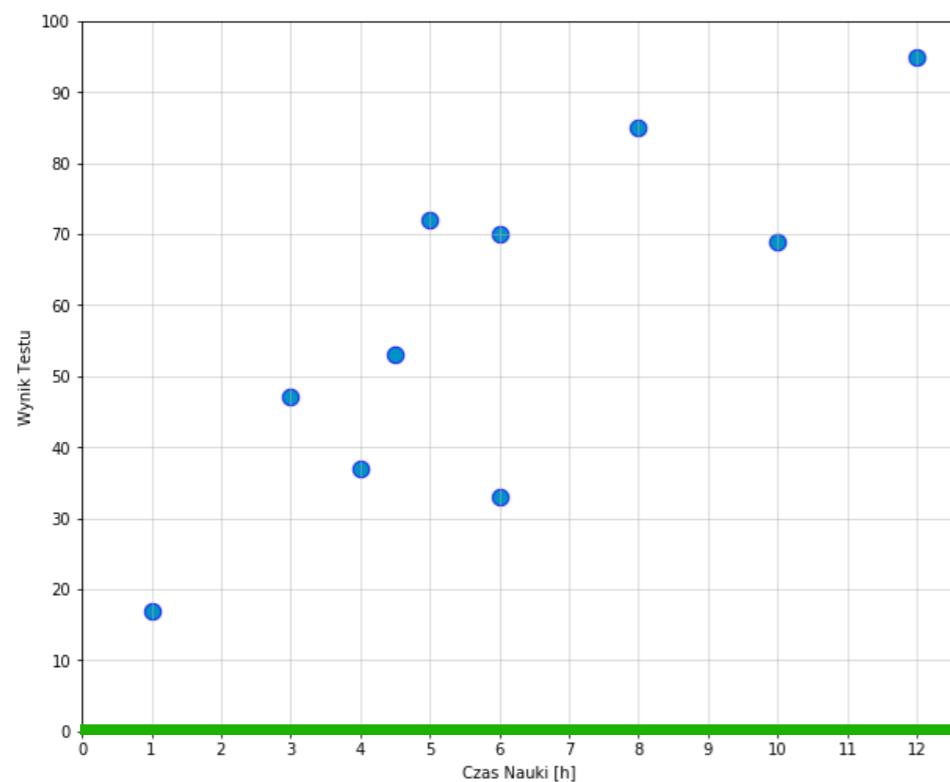
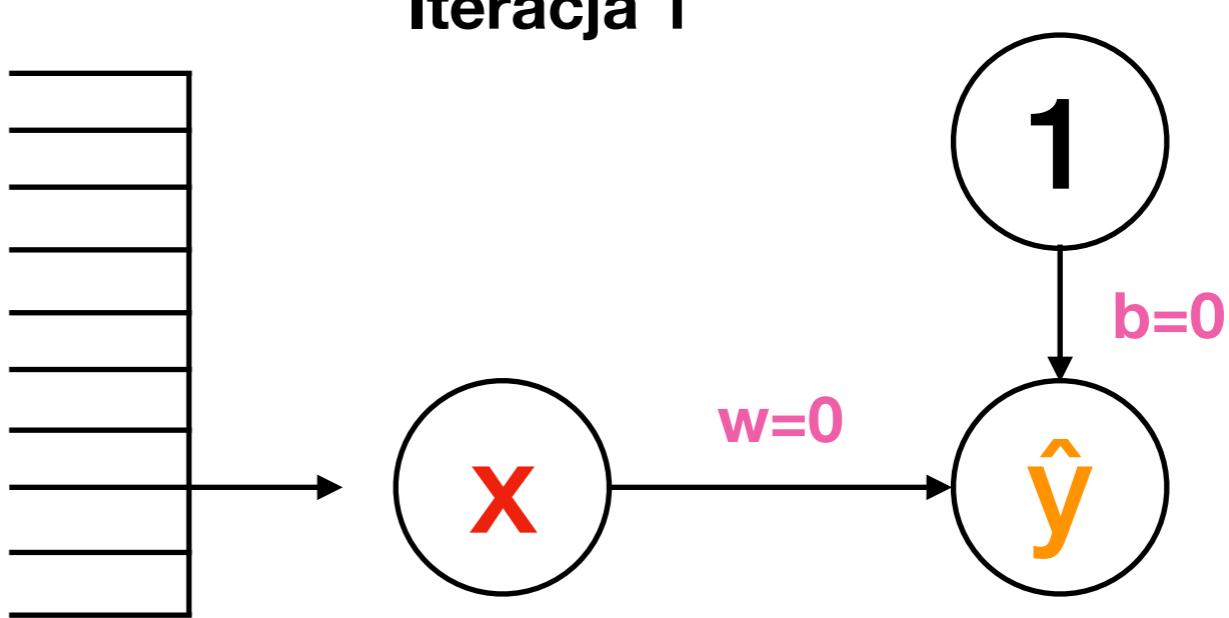
Id	Czas Nauki	Wynik Testu
0	6	70
1	1	17
2	5	72
3	3	47
4	4.5	53
5	10	69
6	12	95
7	8	85
8	6	33
9	4	37



Inicjalizujemy model poprzez nadanie wartości początkowych dla w oraz b .

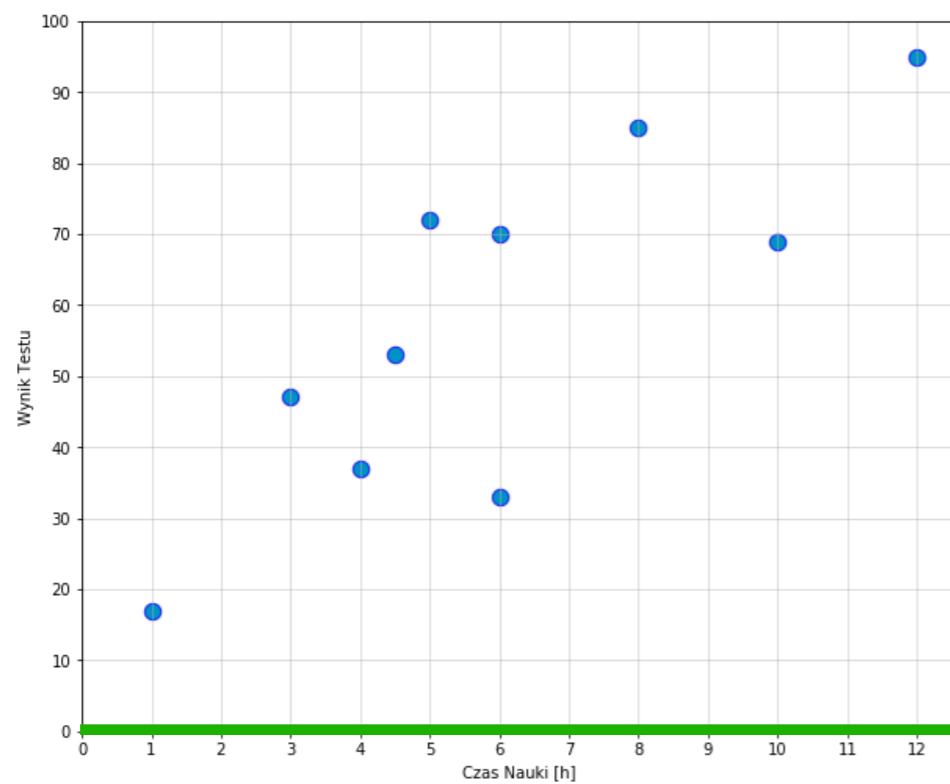
Uczenie Nadzorowane

Id	Czas Nauki	Wynik Testu
0	6	70
1	1	17
2	5	72
3	3	47
4	4.5	53
5	10	69
6	12	95
7	8	85
8	6	33
9	4	37

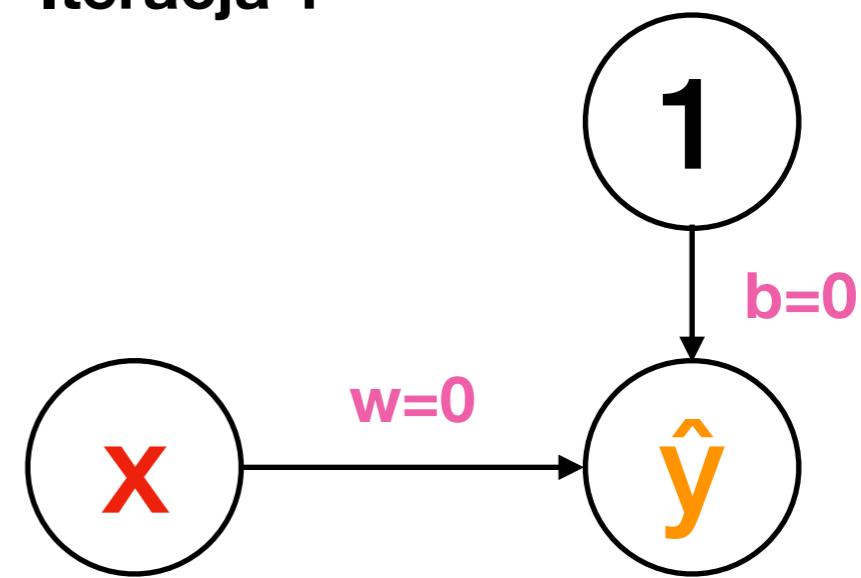


Uczenie Nadzorowane

Id	Czas Nauki	Wynik Testu
0	6	70
1	1	17
2	5	72
3	3	47
4	4.5	53
5	10	69
6	12	95
7	8	85
8	6	33
9	4	37



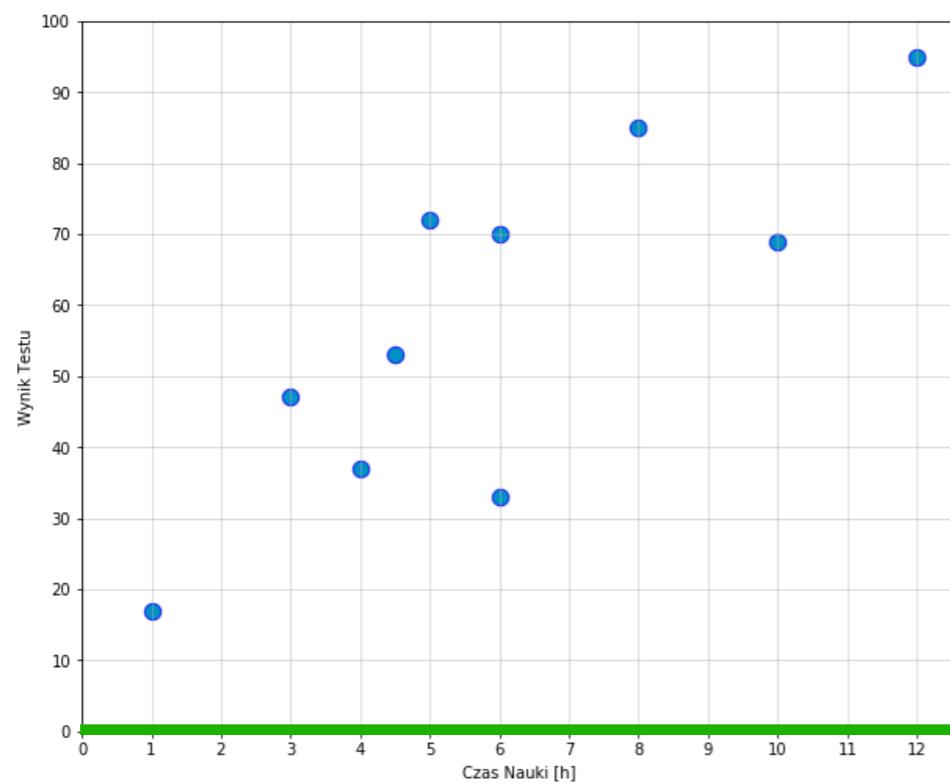
Iteracja 1



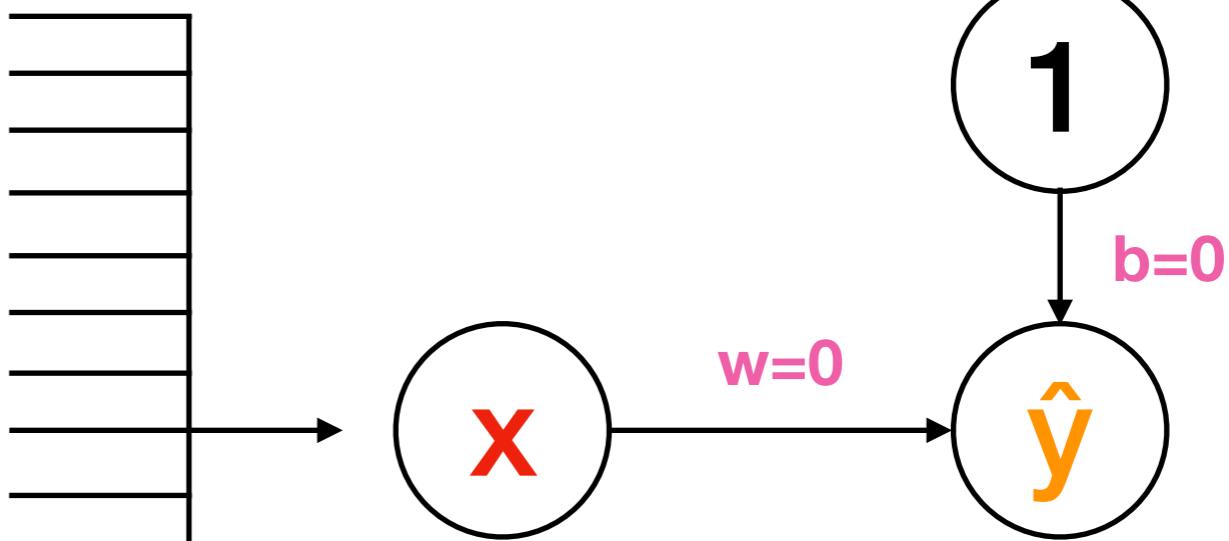
MAE(w, b) 57,8

Uczenie Nadzorowane

Id	Czas Nauki	Wynik Testu
0	6	70
1	1	17
2	5	72
3	3	47
4	4.5	53
5	10	69
6	12	95
7	8	85
8	6	33
9	4	37



Iteracja 1



$MAE(w, b)$

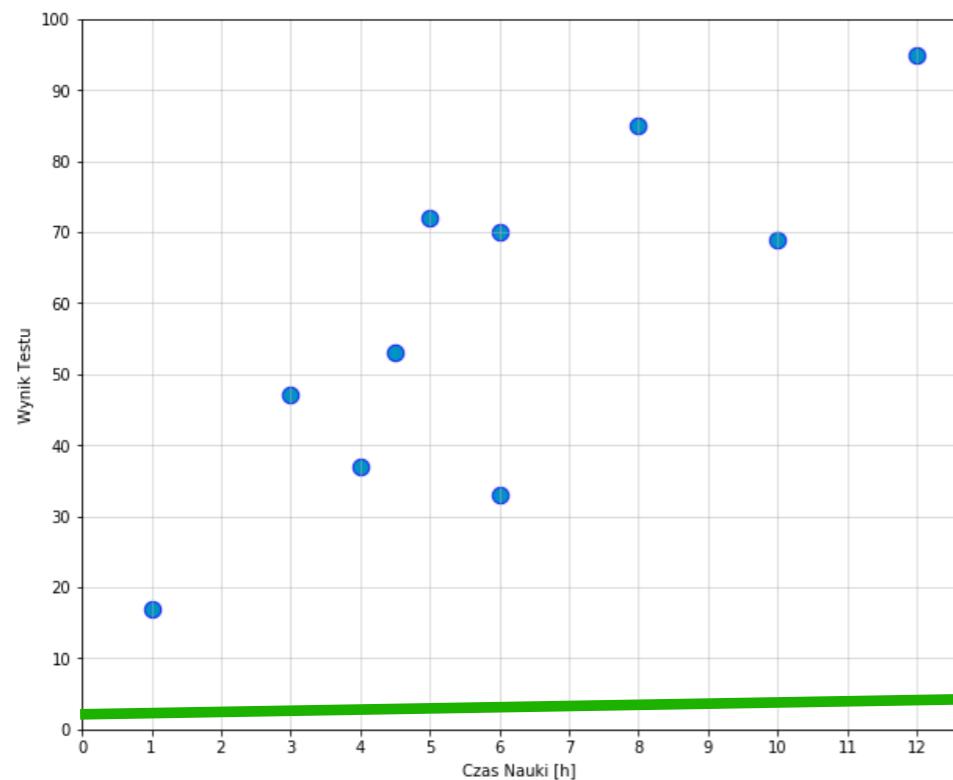
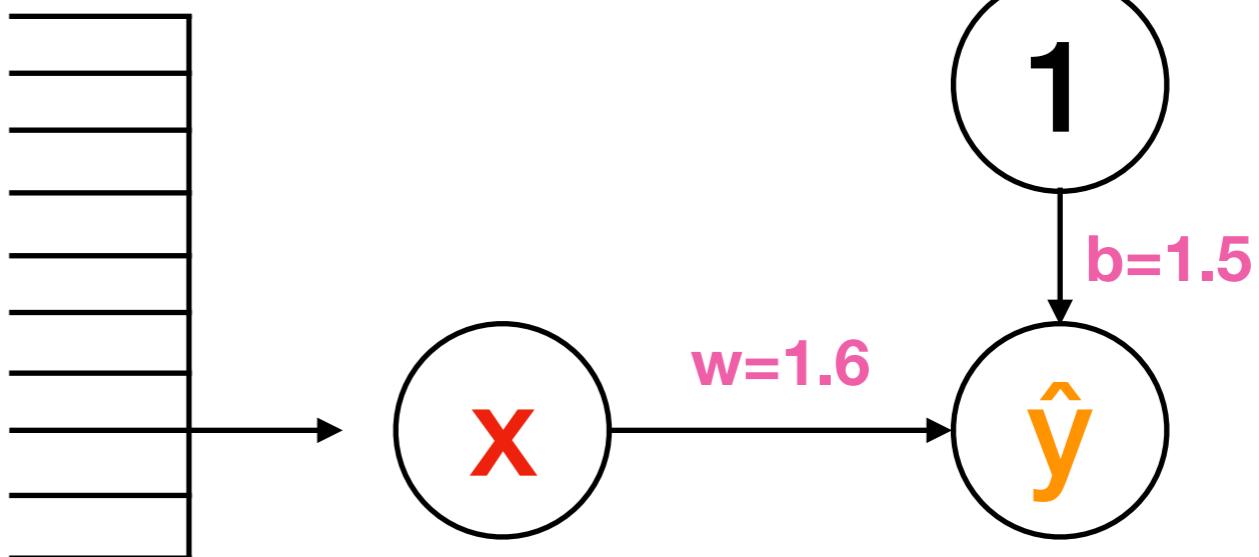
$$w' = w - \text{learning_rate} \cdot \frac{\partial MAE(w, b)}{\partial w}$$

$$b' = b - \text{learning_rate} \cdot \frac{\partial MAE(w, b)}{\partial b}$$

Uczenie Nadzorowane

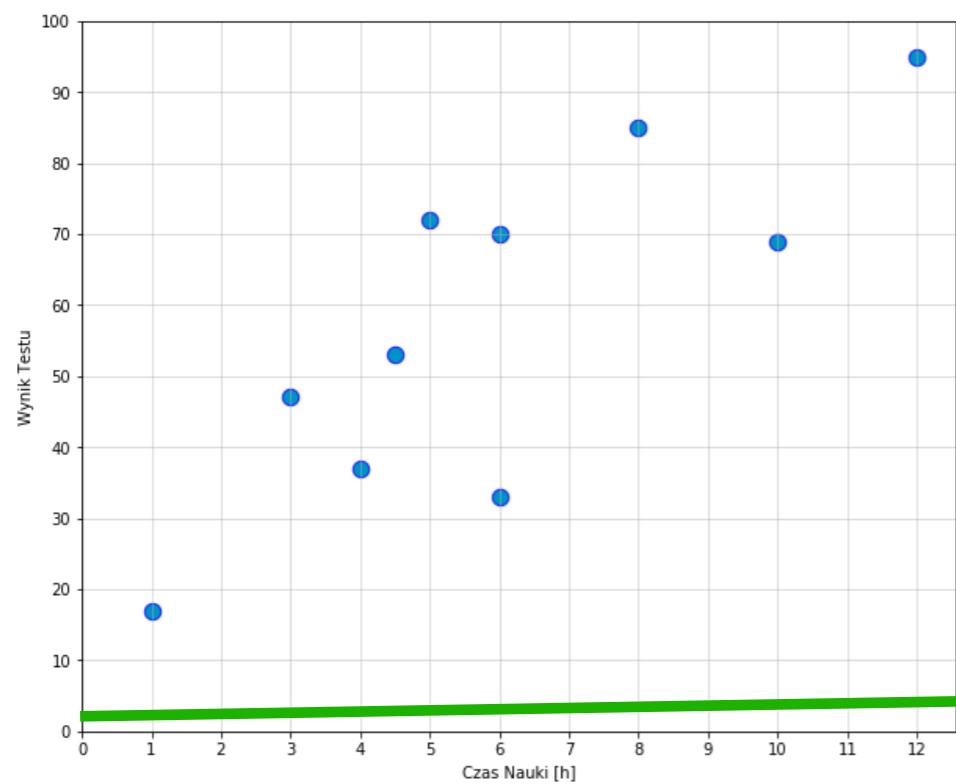
Id	Czas Nauki	Wynik Testu
0	6	70
1	1	17
2	5	72
3	3	47
4	4.5	53
5	10	69
6	12	95
7	8	85
8	6	33
9	4	37

Iteracja 1

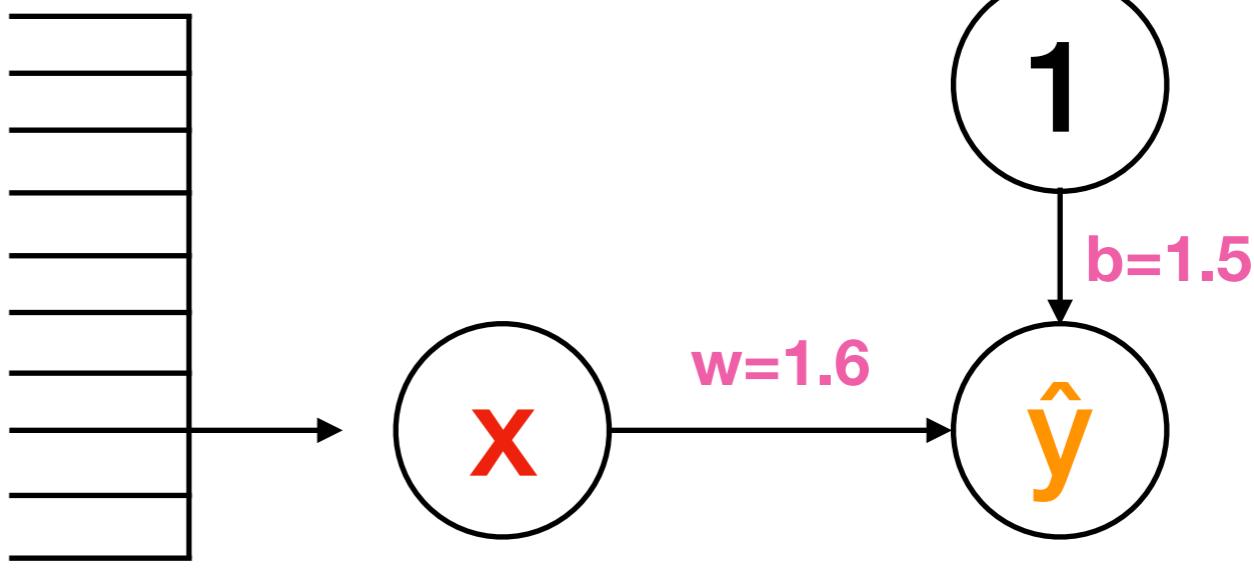


Uczenie Nadzorowane

Id	Czas Nauki	Wynik Testu
0	6	70
1	1	17
2	5	72
3	3	47
4	4.5	53
5	10	69
6	12	95
7	8	85
8	6	33
9	4	37



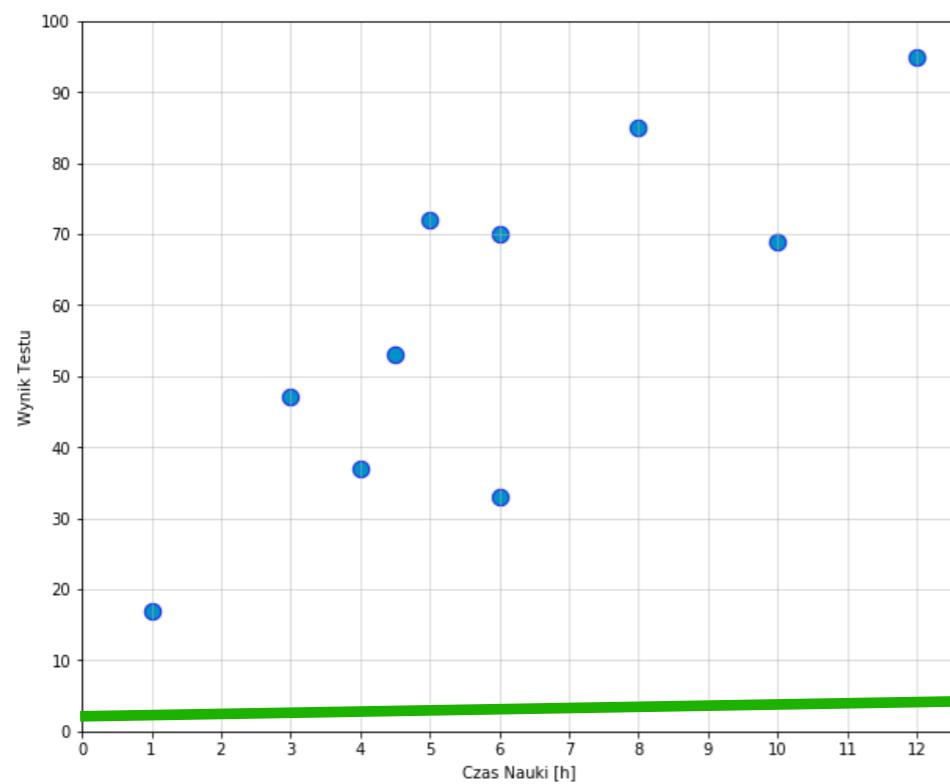
Iteracja 2



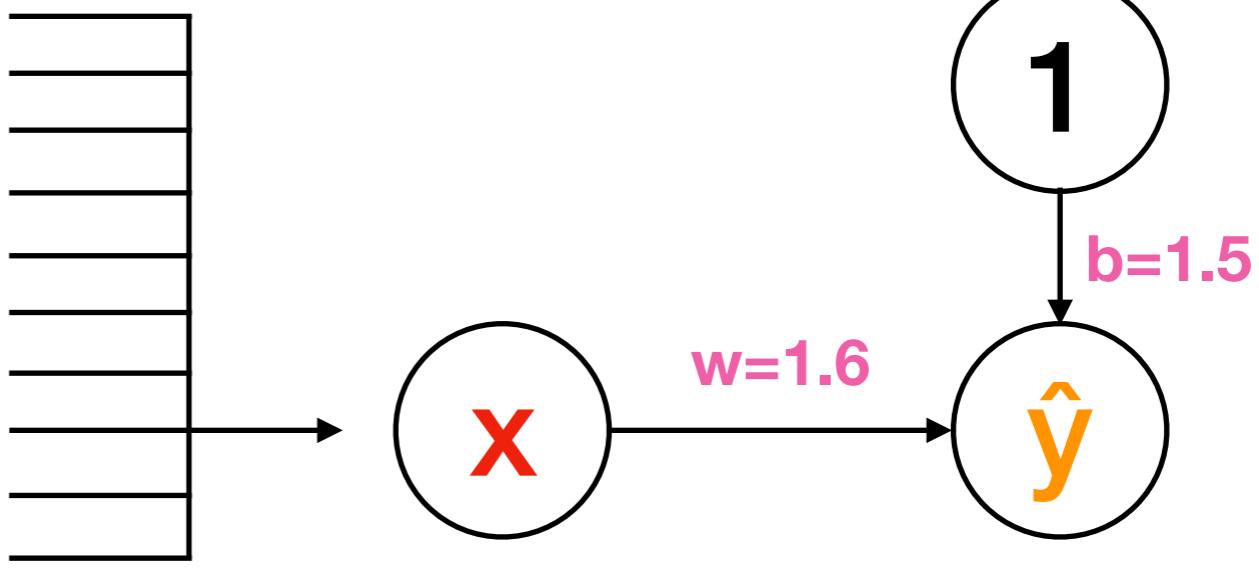
$MAE(w, b)$ 46.8

Uczenie Nadzorowane

Id	Czas Nauki	Wynik Testu
0	6	70
1	1	17
2	5	72
3	3	47
4	4.5	53
5	10	69
6	12	95
7	8	85
8	6	33
9	4	37



Iteracja 2



MAE(w, b)

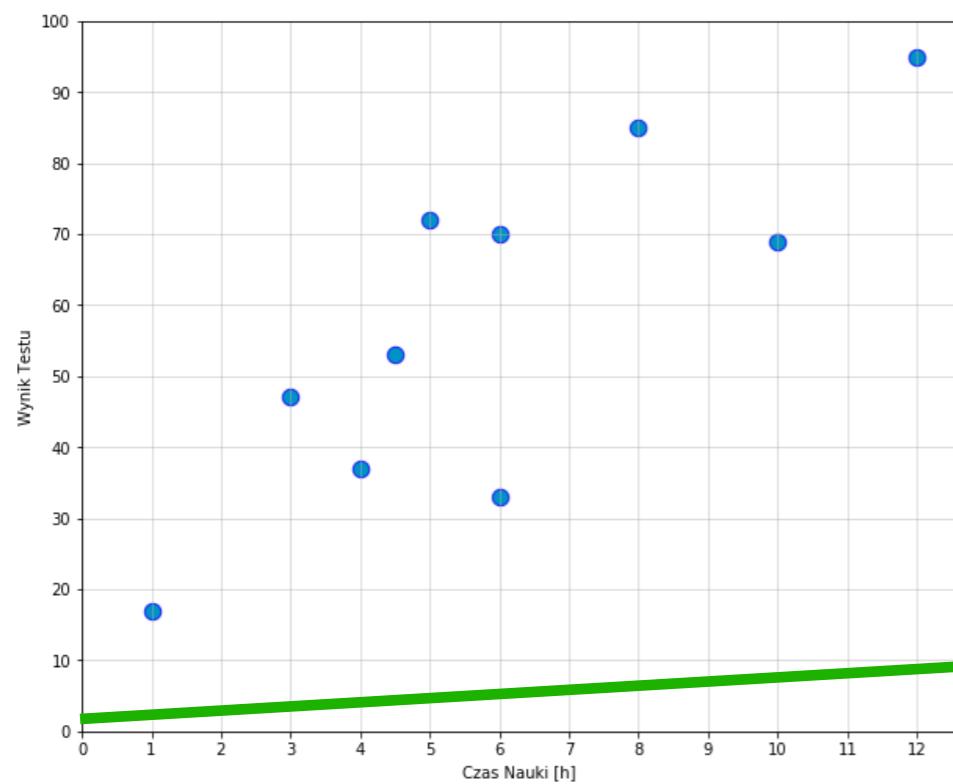
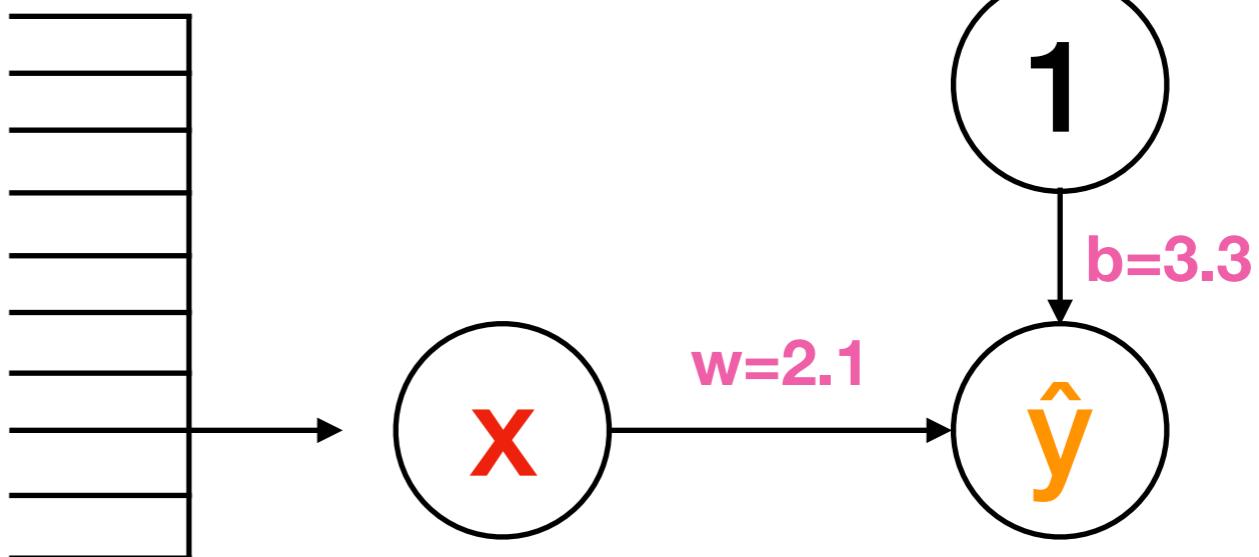
$$w' = w + \text{learning_rate} \cdot \frac{\partial \text{MAE}(w, b)}{\partial w}$$

$$b' = b + \text{learning_rate} \cdot \frac{\partial \text{MAE}(w, b)}{\partial b}$$

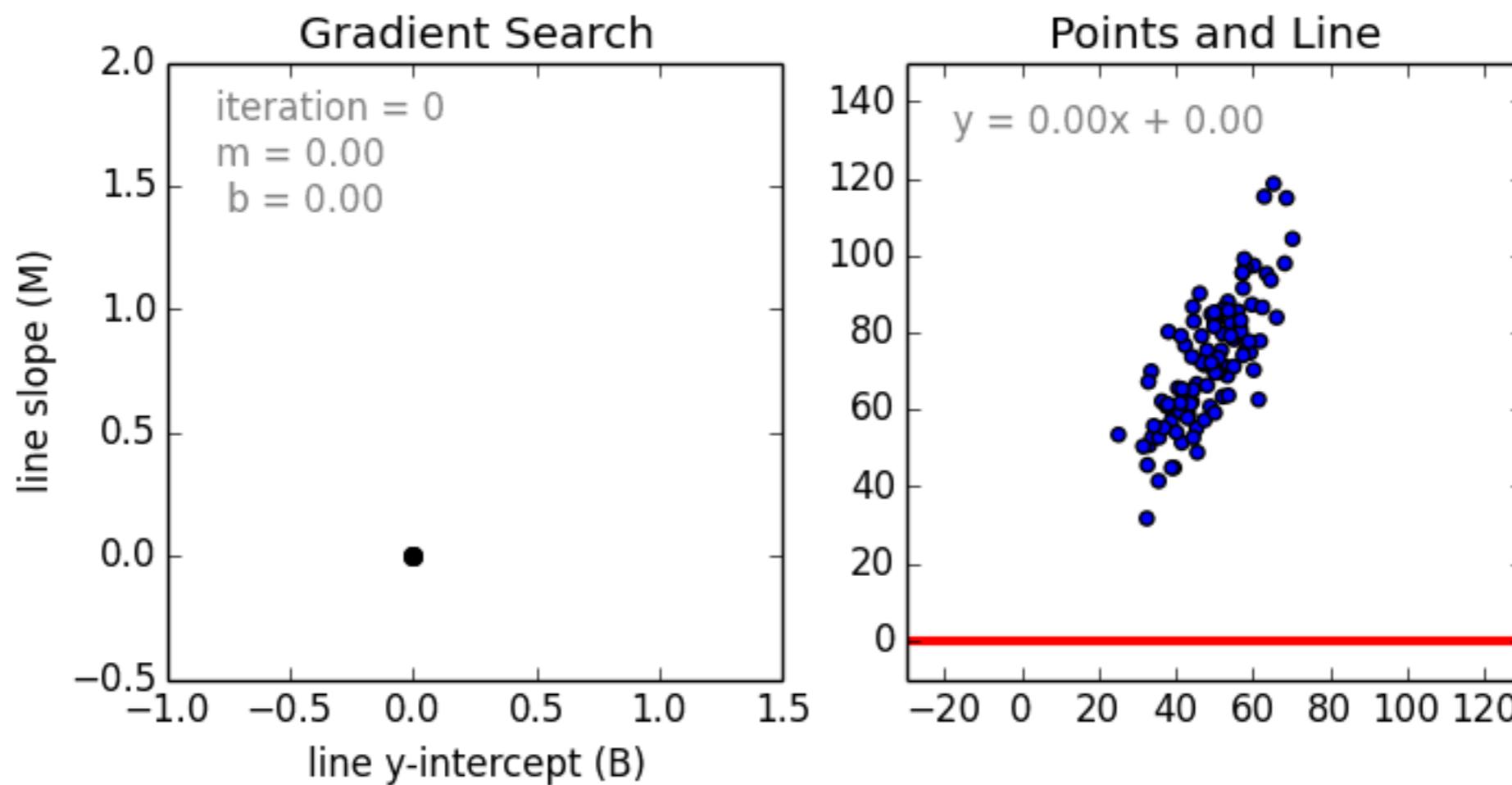
Uczenie Nadzorowane

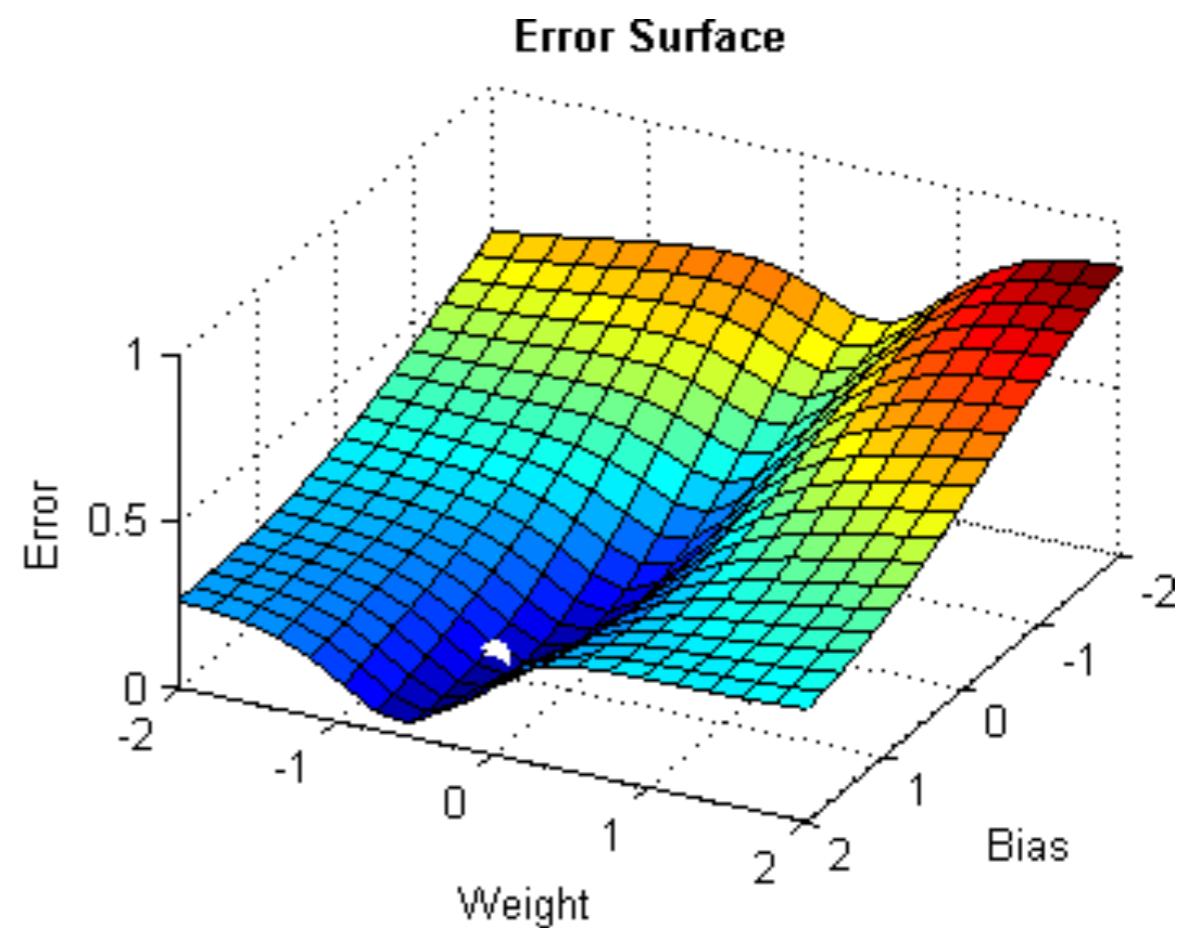
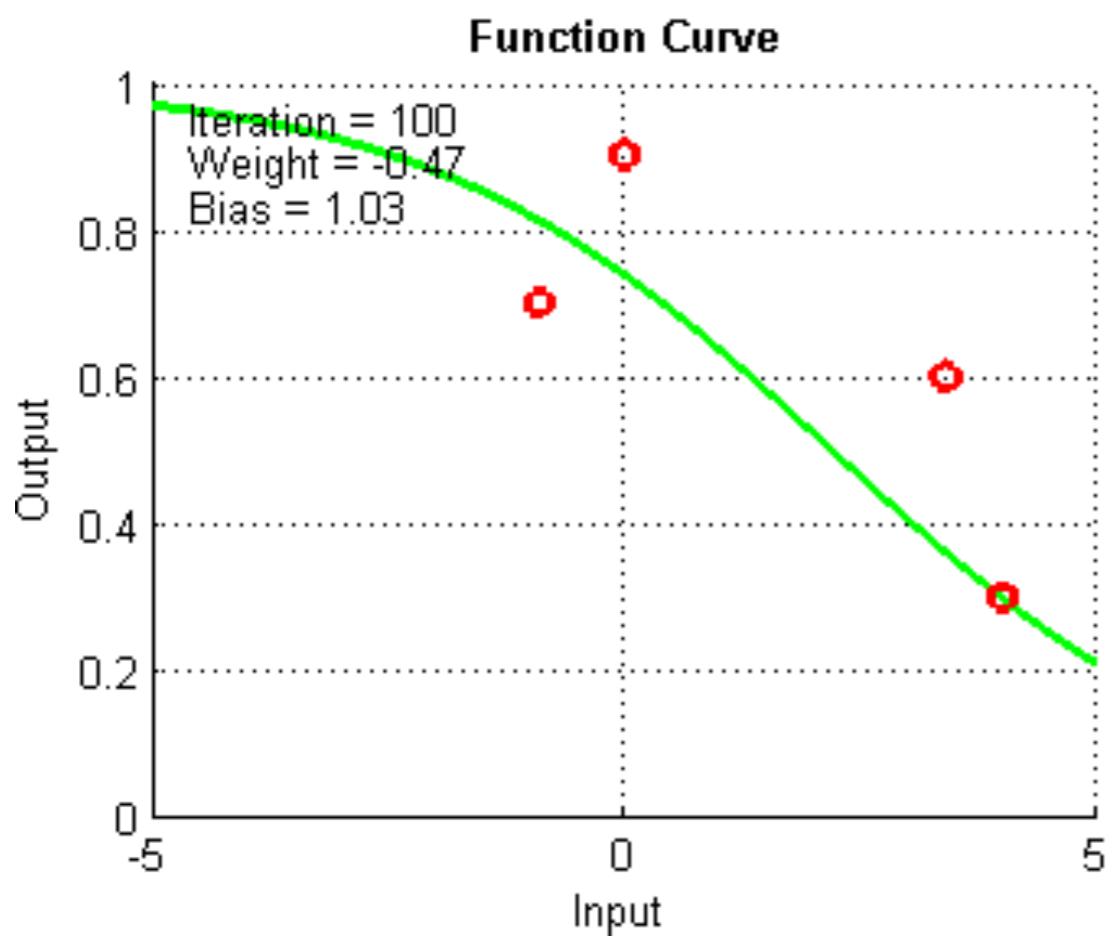
Id	Czas Nauki	Wynik Testu
0	6	70
1	1	17
2	5	72
3	3	47
4	4.5	53
5	10	69
6	12	95
7	8	85
8	6	33
9	4	37

Iteracja 2

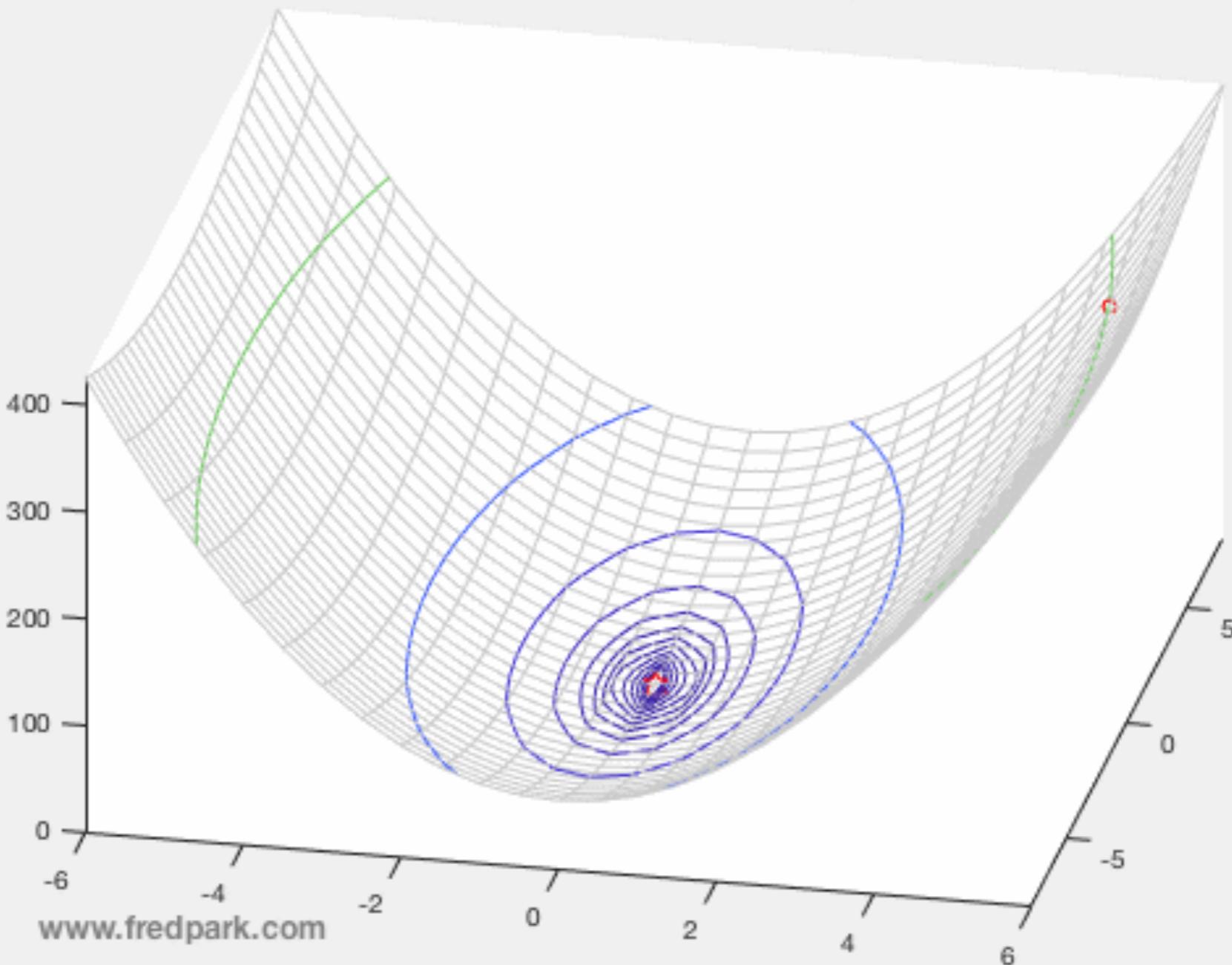


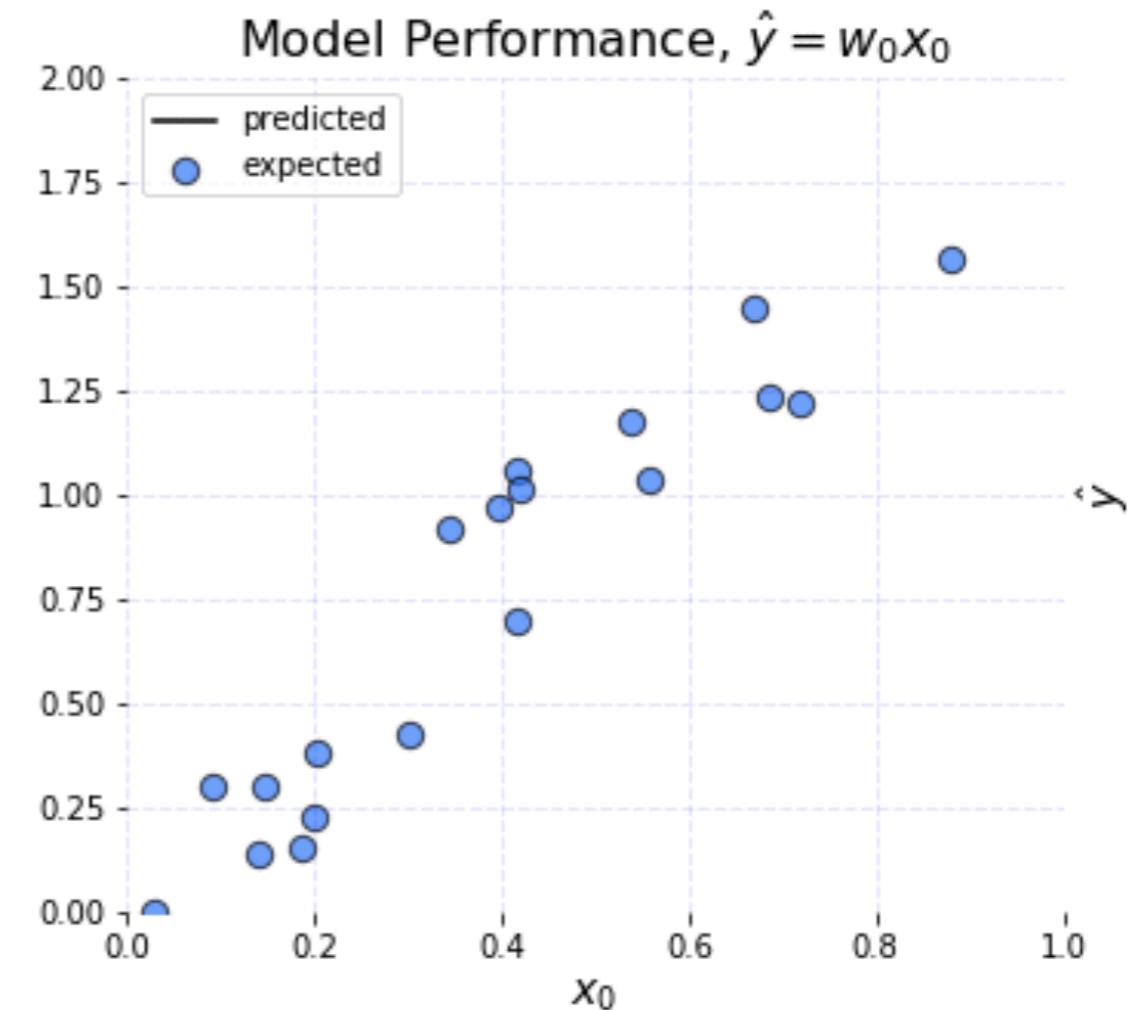
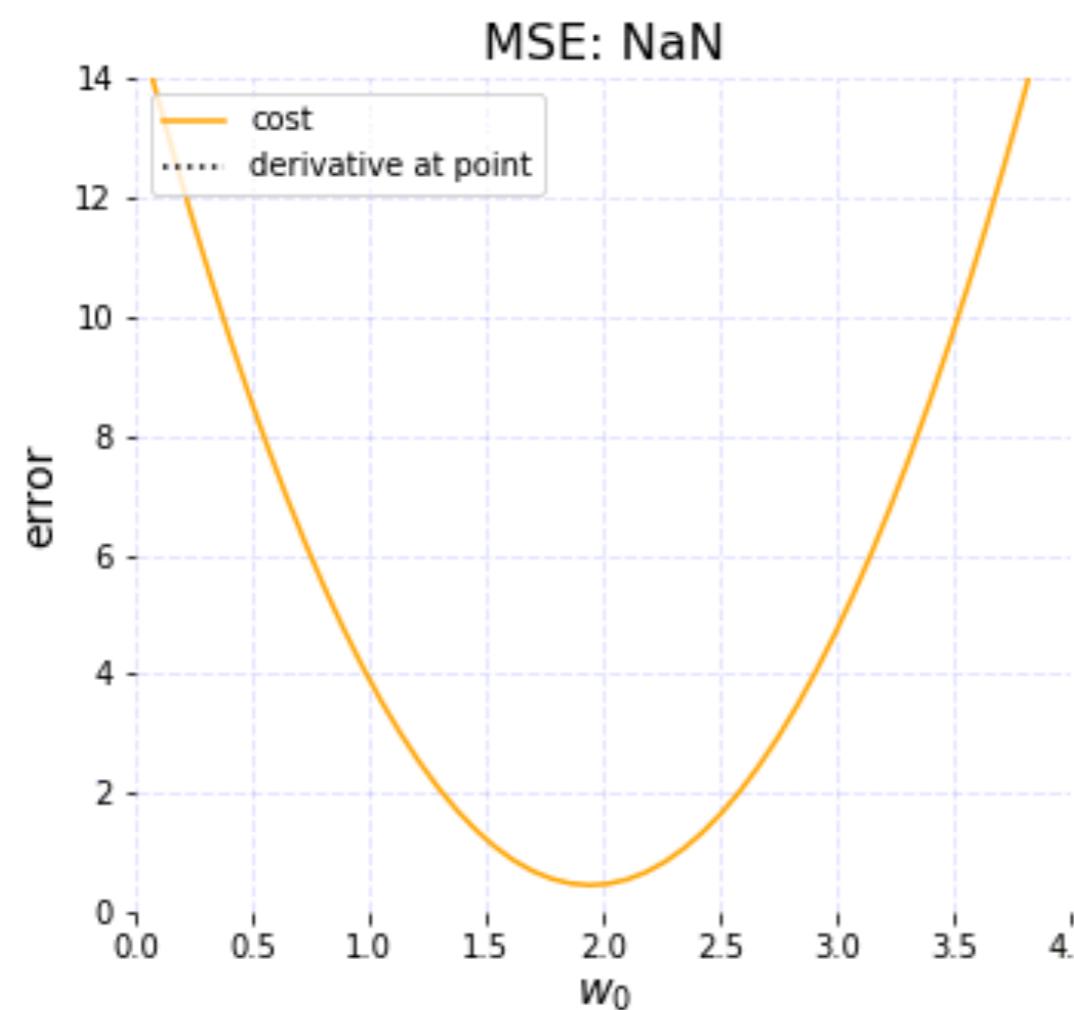
Uczenie Nadzorowane

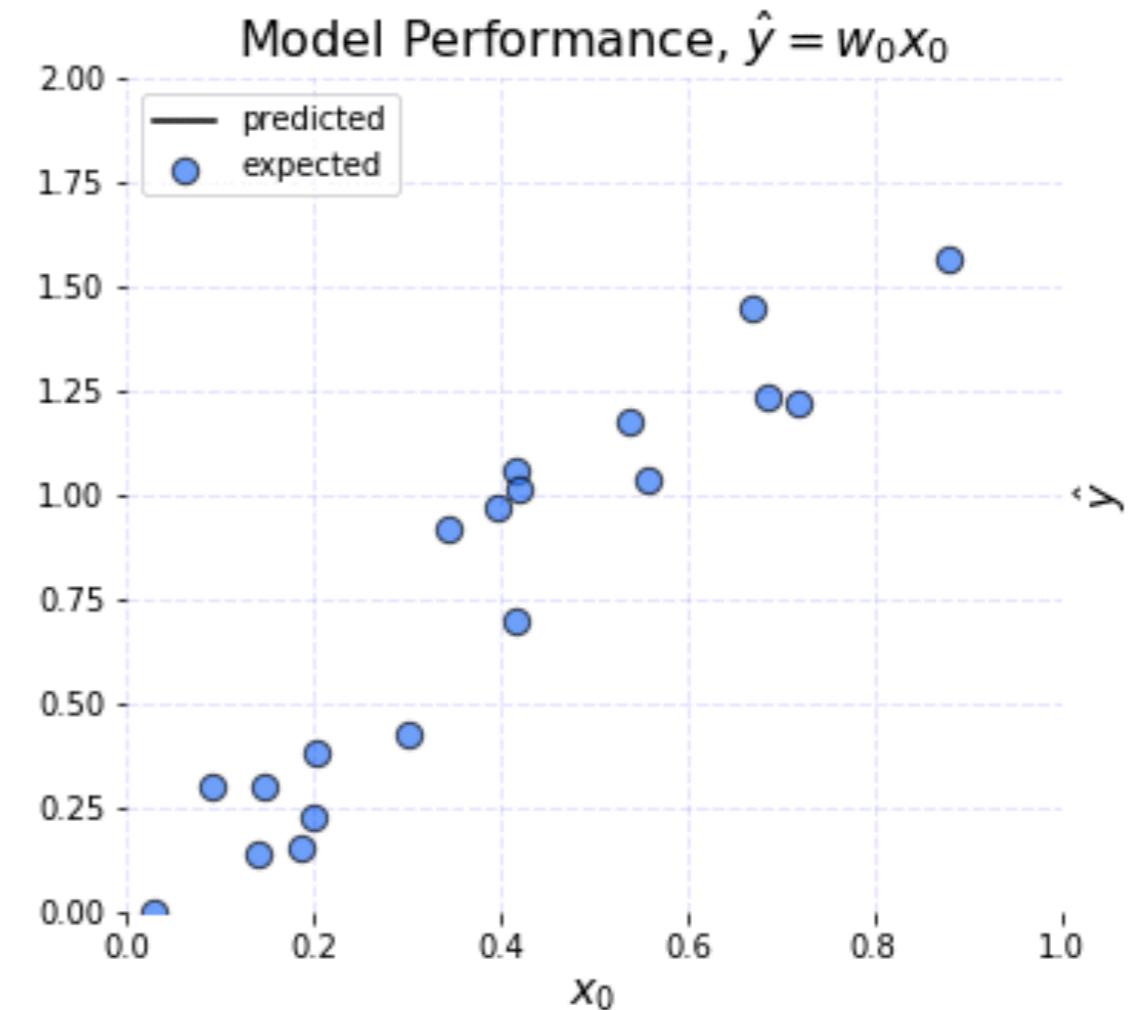
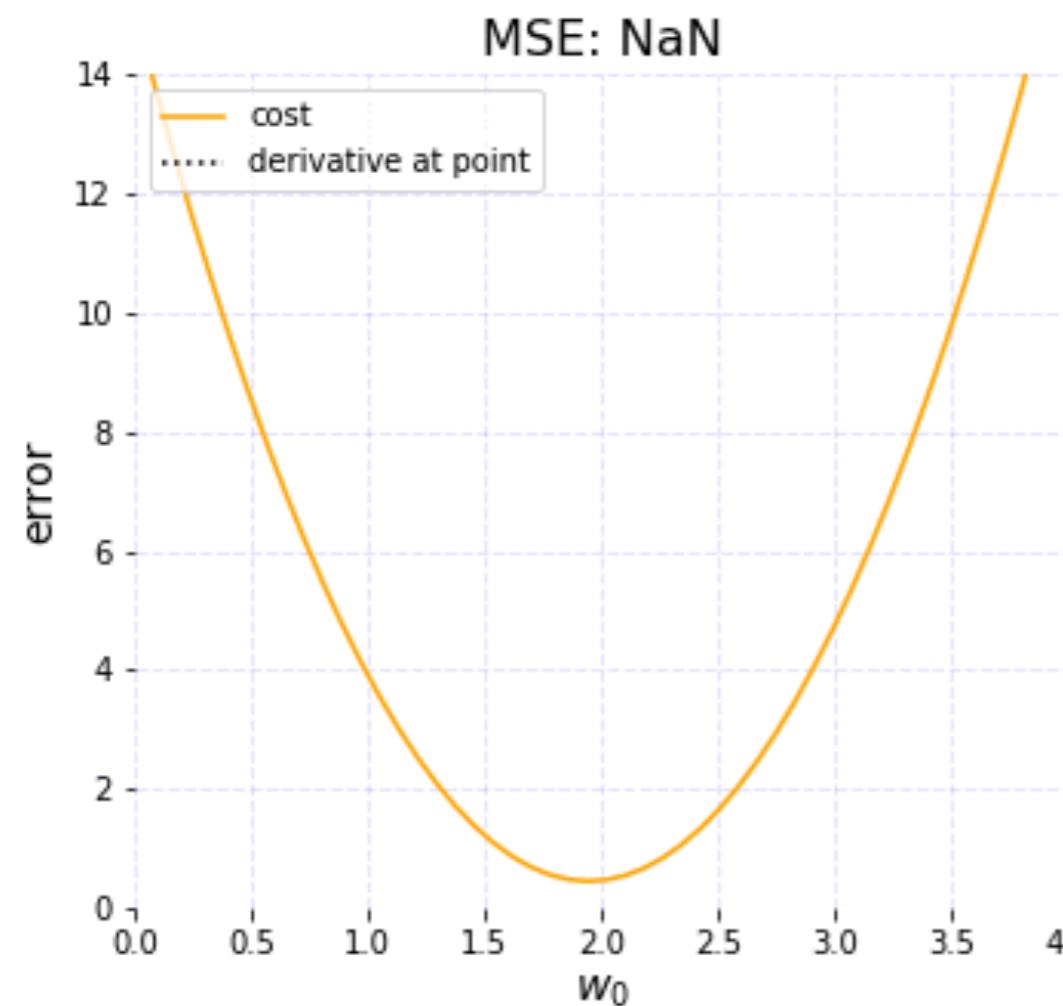




Gradient Descent on $z = 10x^2 + y^2$



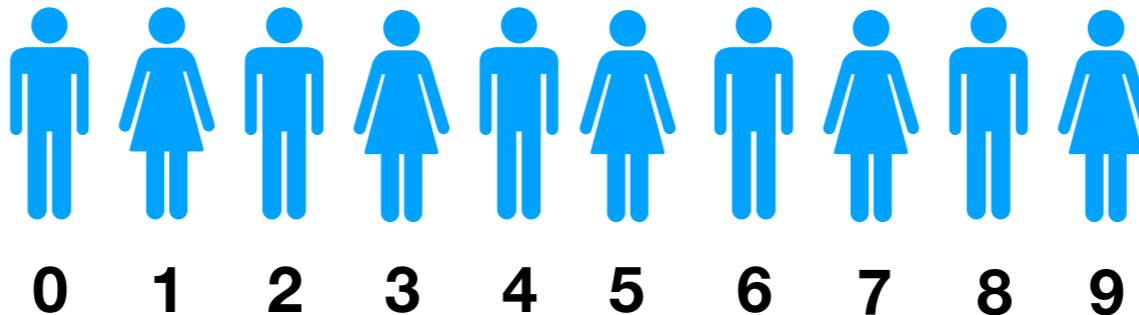




Zbyt duży learning_rate powoduje, że model się nie będzie uczyć!

Teoria

Studenci:

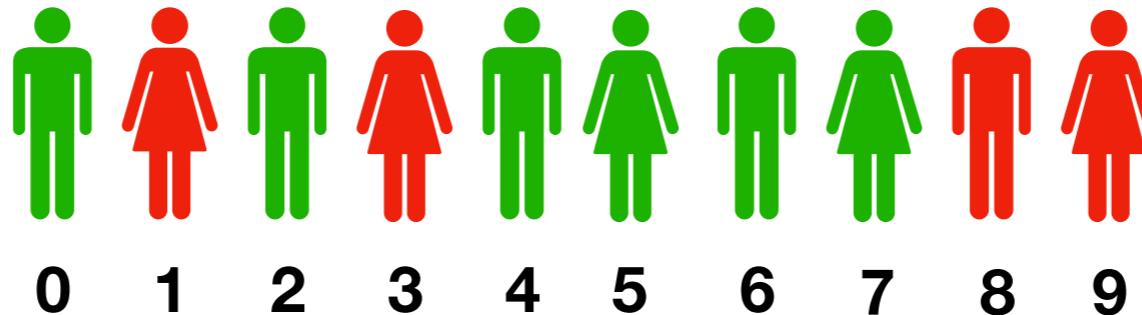


Dane:

Id	Czas Nauki [h]	Czas Snu [h]	Wynik Testu
0	6	8	70
1	1	3	17
2	5	7.5	72
3	3	6	47
4	4.5	6.5	53
5	10	4	69
6	12	8	95
7	8	8	85
8	6	2	33
9	4	5	37

Teoria

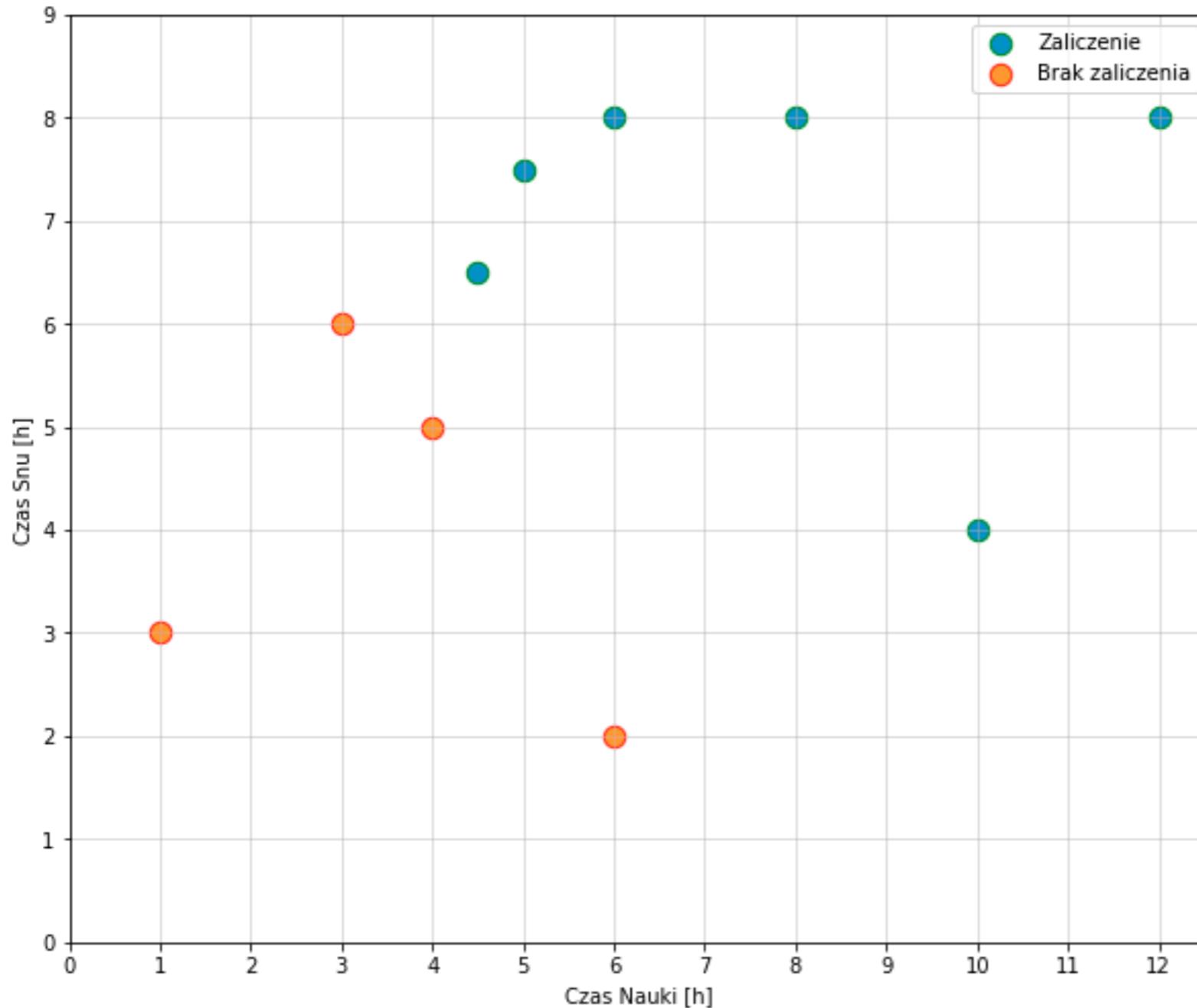
Studenci:



Dane:

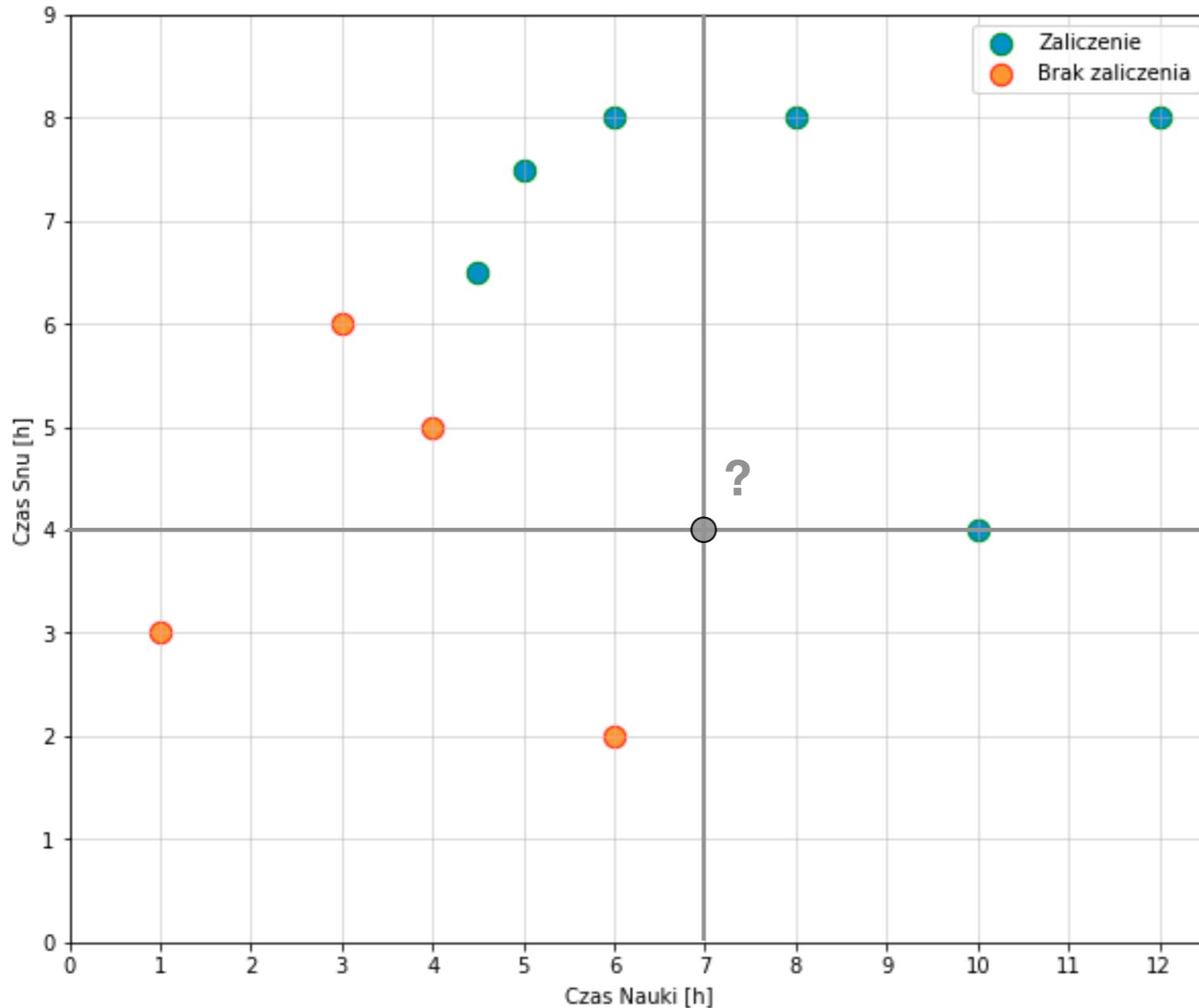
Id	Czas Nauki	Czas Snu [h]	Wynik Testu	Zaliczenie
0	6	8	70	1
1	1	3	17	0
2	5	7.5	72	1
3	3	6	47	0
4	4.5	6.5	53	1
5	10	4	69	1
6	12	8	95	1
7	8	8	85	1
8	6	2	33	0
9	4	5	37	0

Teoria



Czy zdam egzamin, jeżeli będę się uczyć przez 7 godzin i będę spać 4 godziny.

Teoria

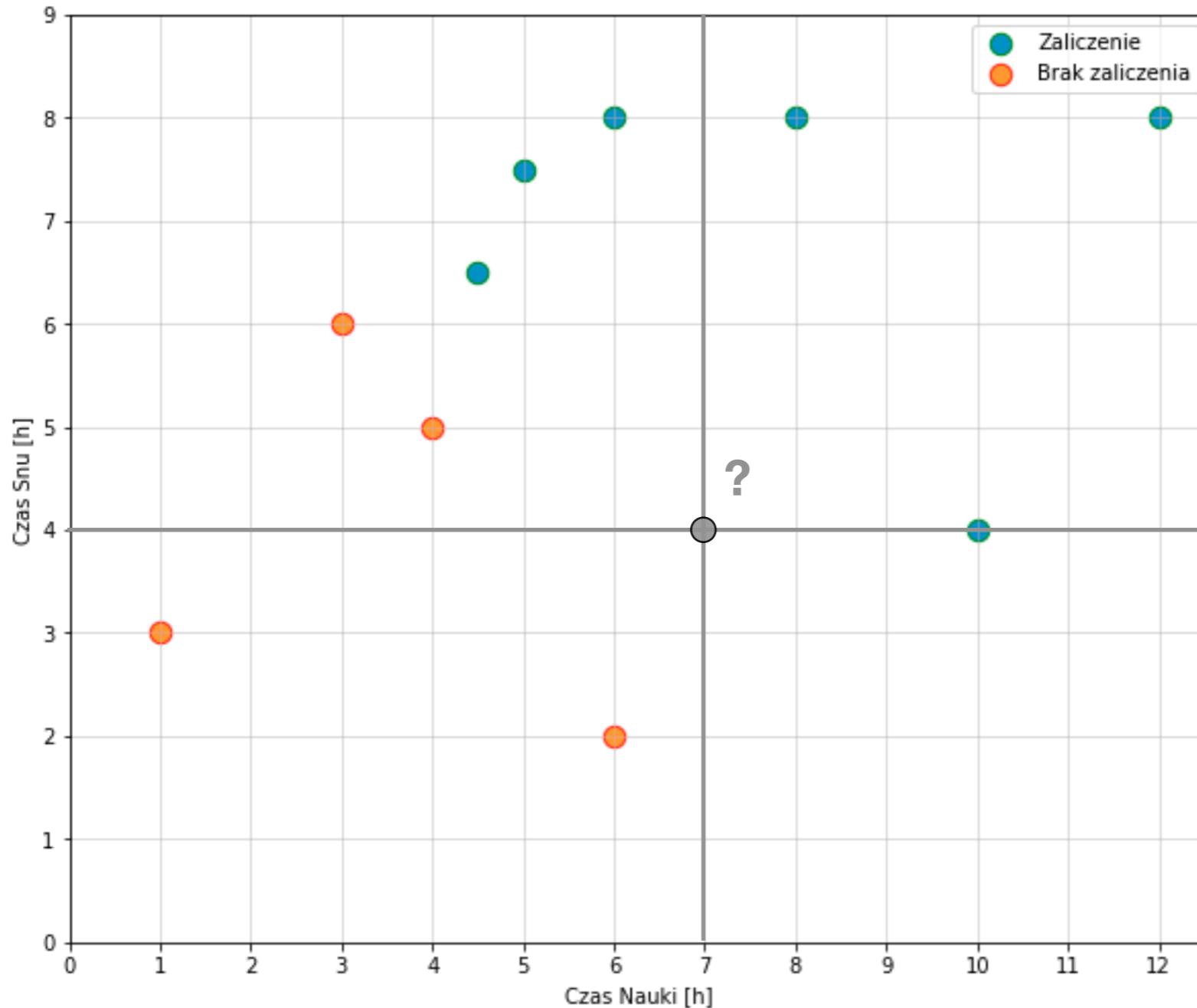


Czy zdam egzamin, jeżeli będę się uczyć przez 7 godzin i będę spać 4 godziny.

$$x_0 = 7$$

$$x_1 = 4$$

Teoria



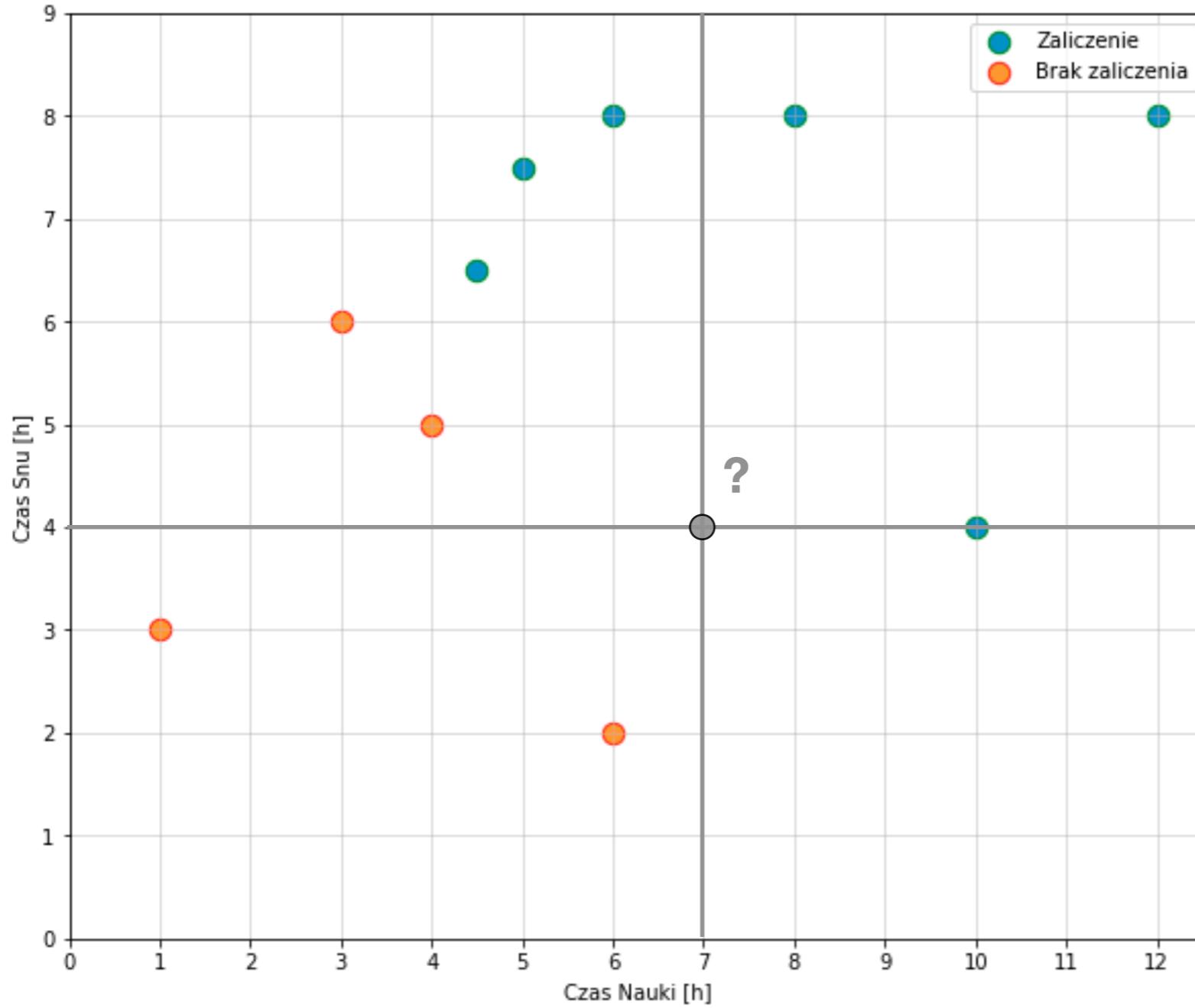
Czy zdam egzamin, jeżeli będę się uczyć przez 7 godzin i będę spać 4 godziny.

$$x_0 = 7$$

$$x_1 = 4$$

$$f(x), \quad x = [x_0, x_1]$$

Teoria



Czy zdam egzamin, jeżeli będę się uczyć przez 7 godzin i będę spać 4 godziny.

$$x_0 = 7 \quad x_1 = 4$$

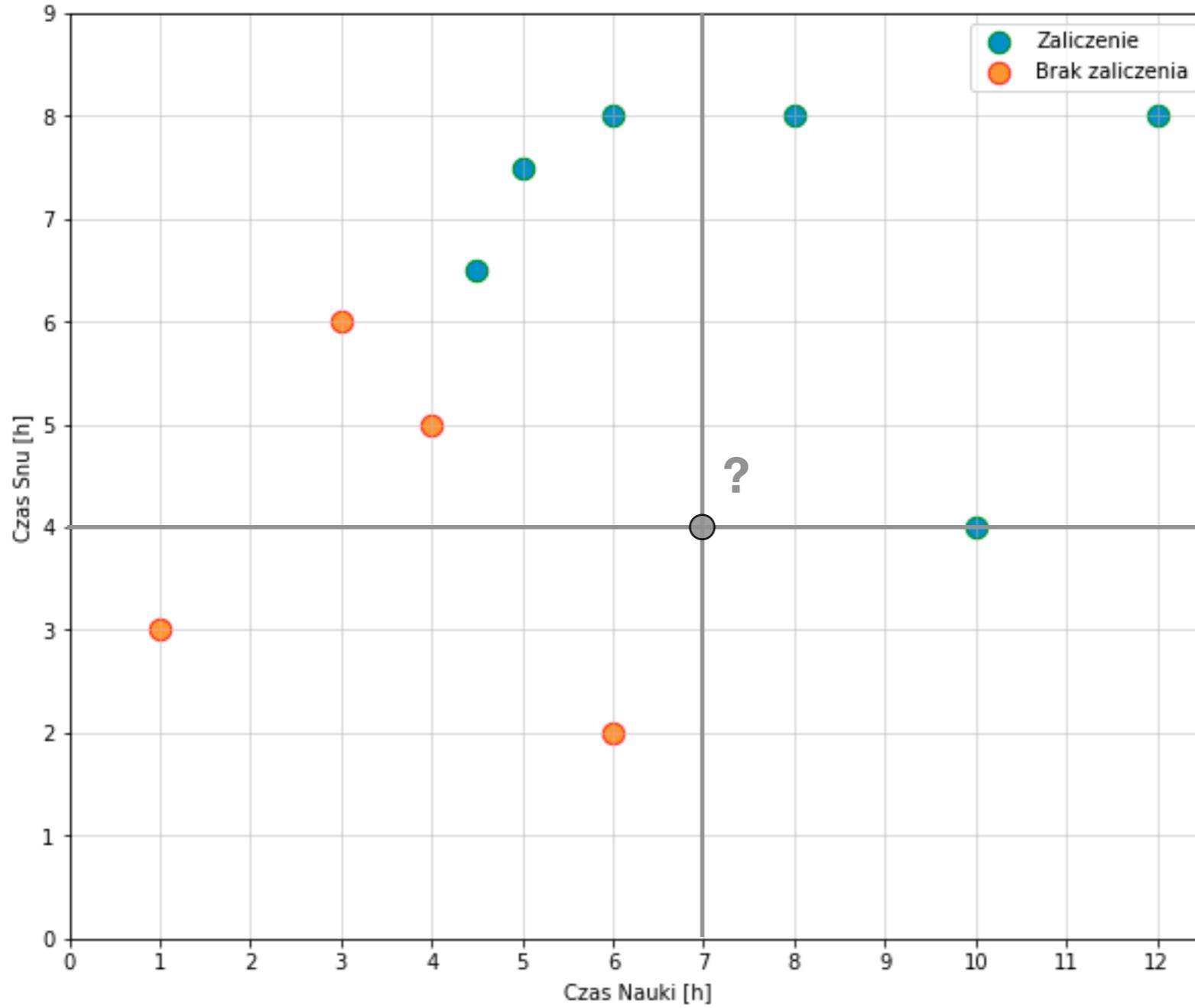
$$f(x), \quad x = [x_0, x_1]$$

$\hat{y} = \text{Status Zaliczenia}$

1 - zaliczone

0 - nie zaliczone

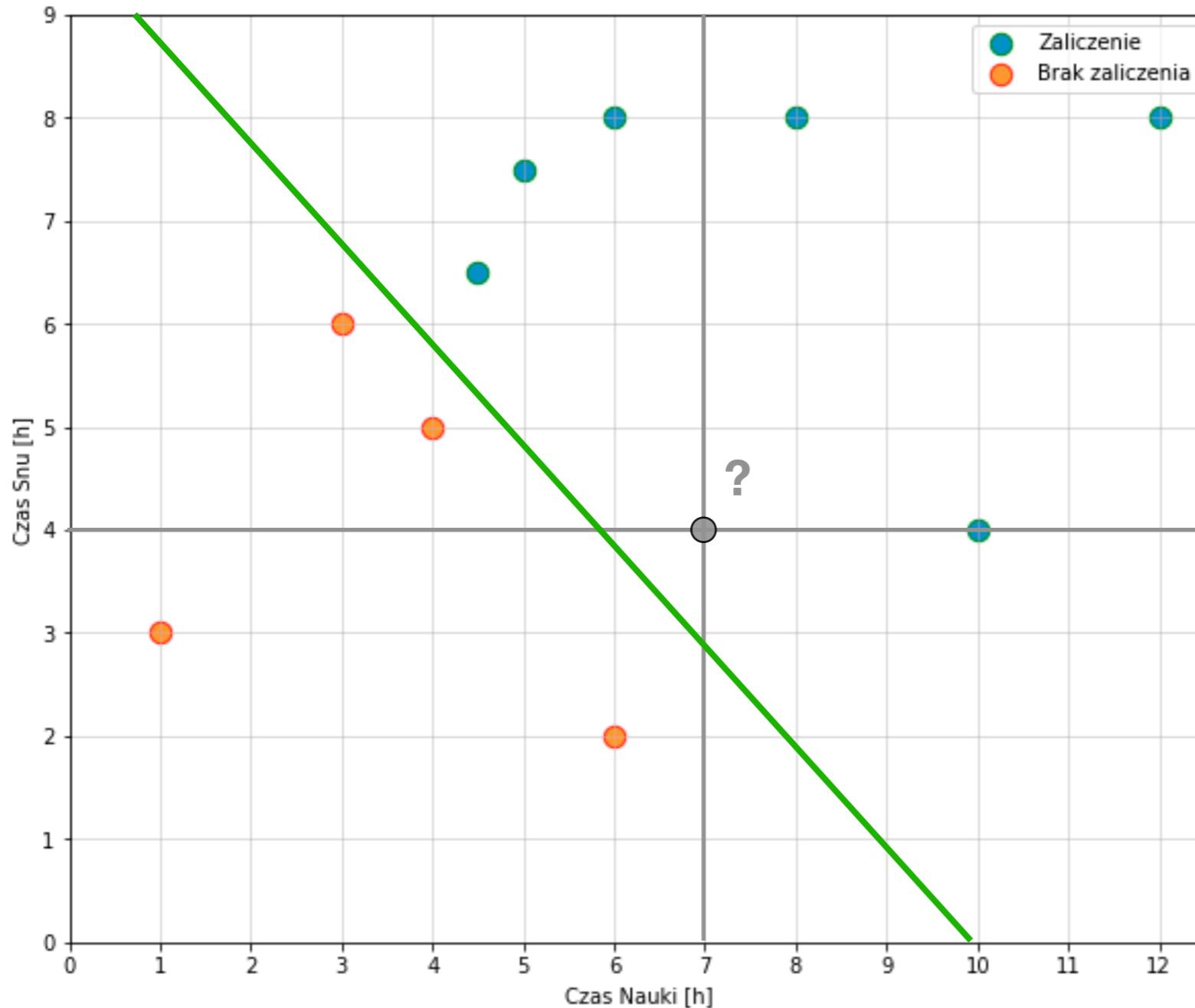
Teoria



Znowu użyjmy funkcji liniowej

$$f(x): \hat{y} = w_0x_0 + w_1x_1 + b$$

Teoria

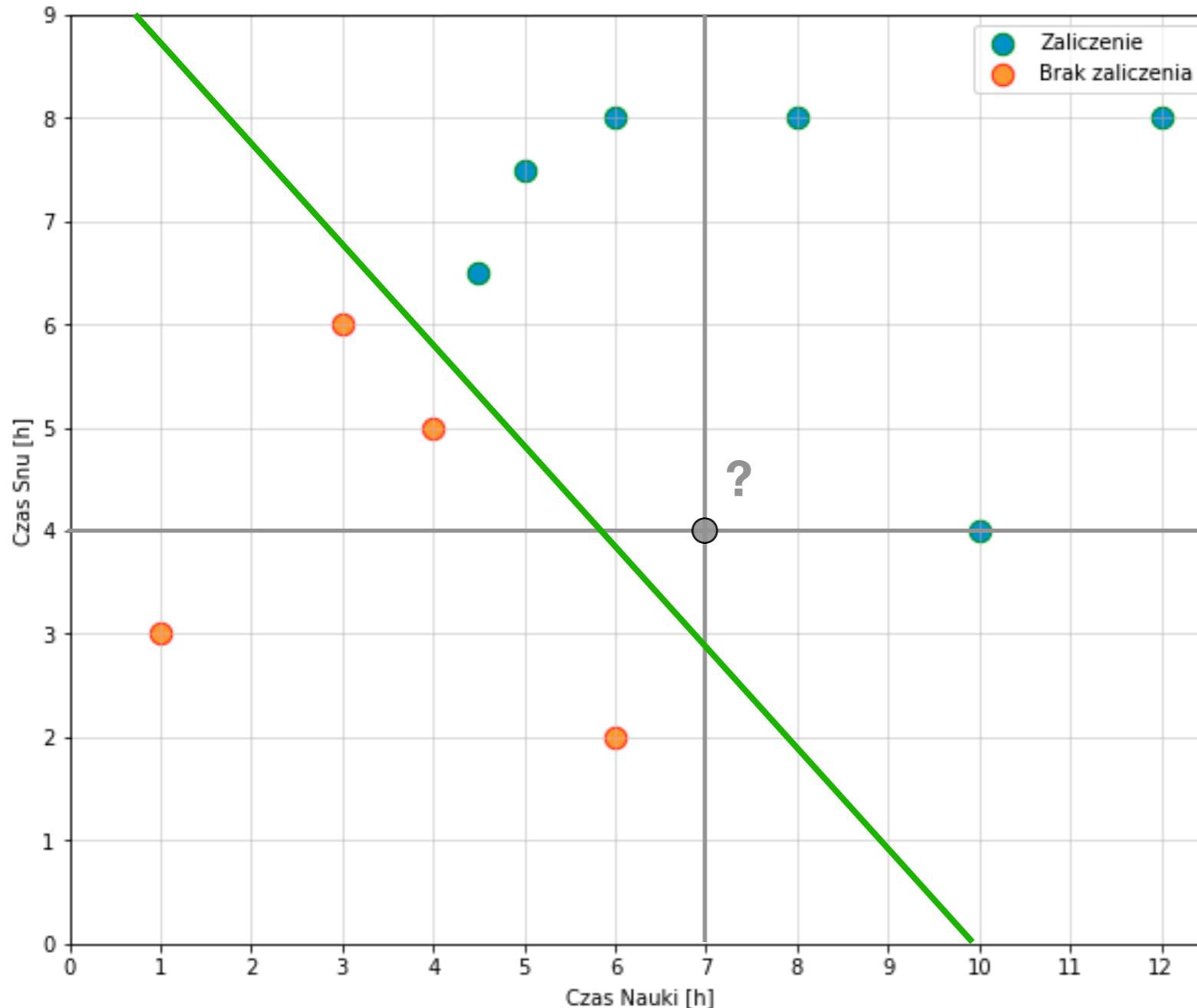


Znowu użyjmy funkcji liniowej

$$f(x): \hat{y} = w_0x_0 + w_1x_1 + b$$

Ale tym razem do oddzielania punktów!

Teoria



Znowu użyjmy funkcji liniowej

$$f(x): \hat{y} = w_0x_0 + w_1x_1 + b$$

Ale tym razem do oddzielania punktów!

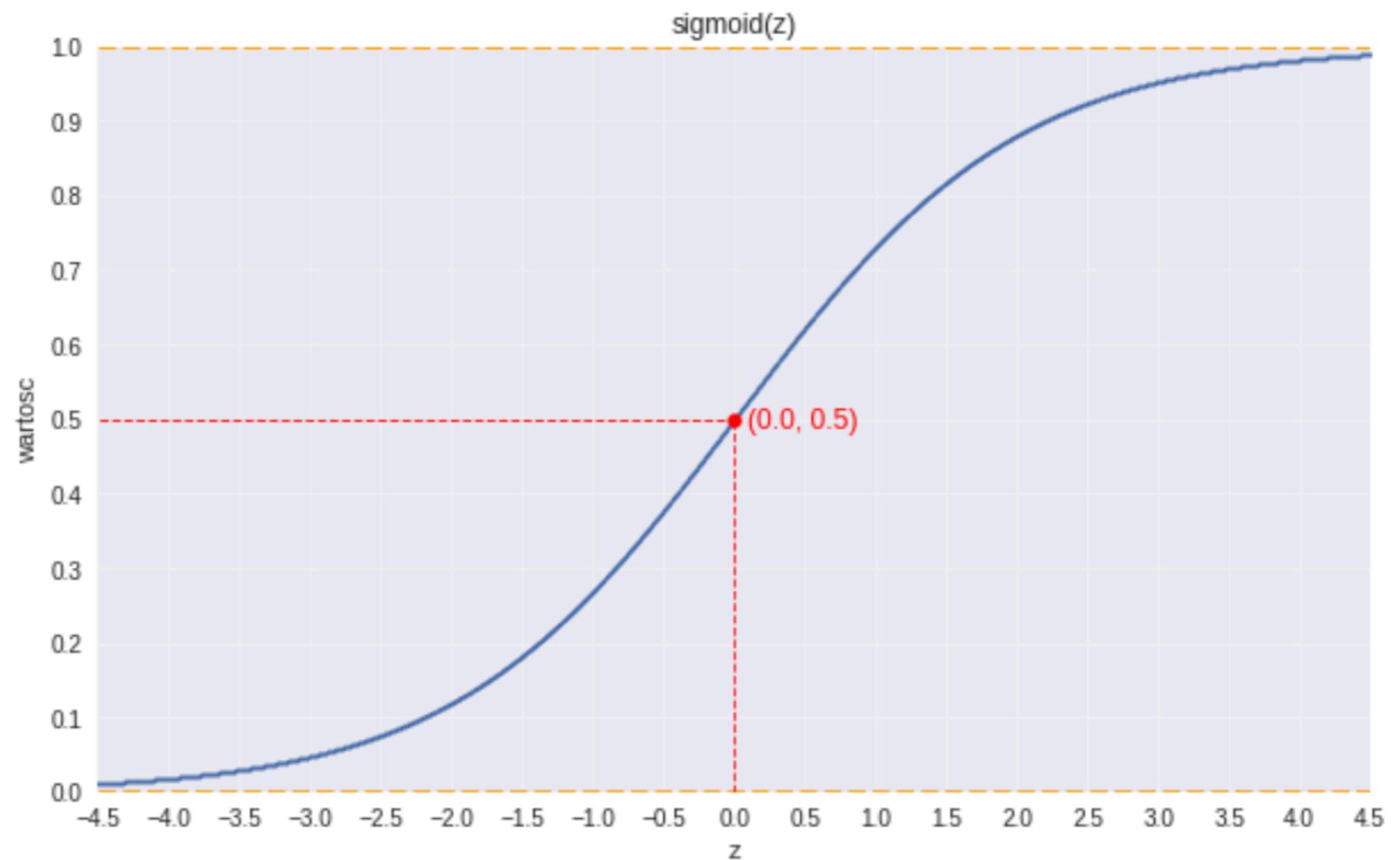
Jak tego dokonać?

Funkcja Aktywacji

Teoria

Funkcja Aktywacji - Sigmoid

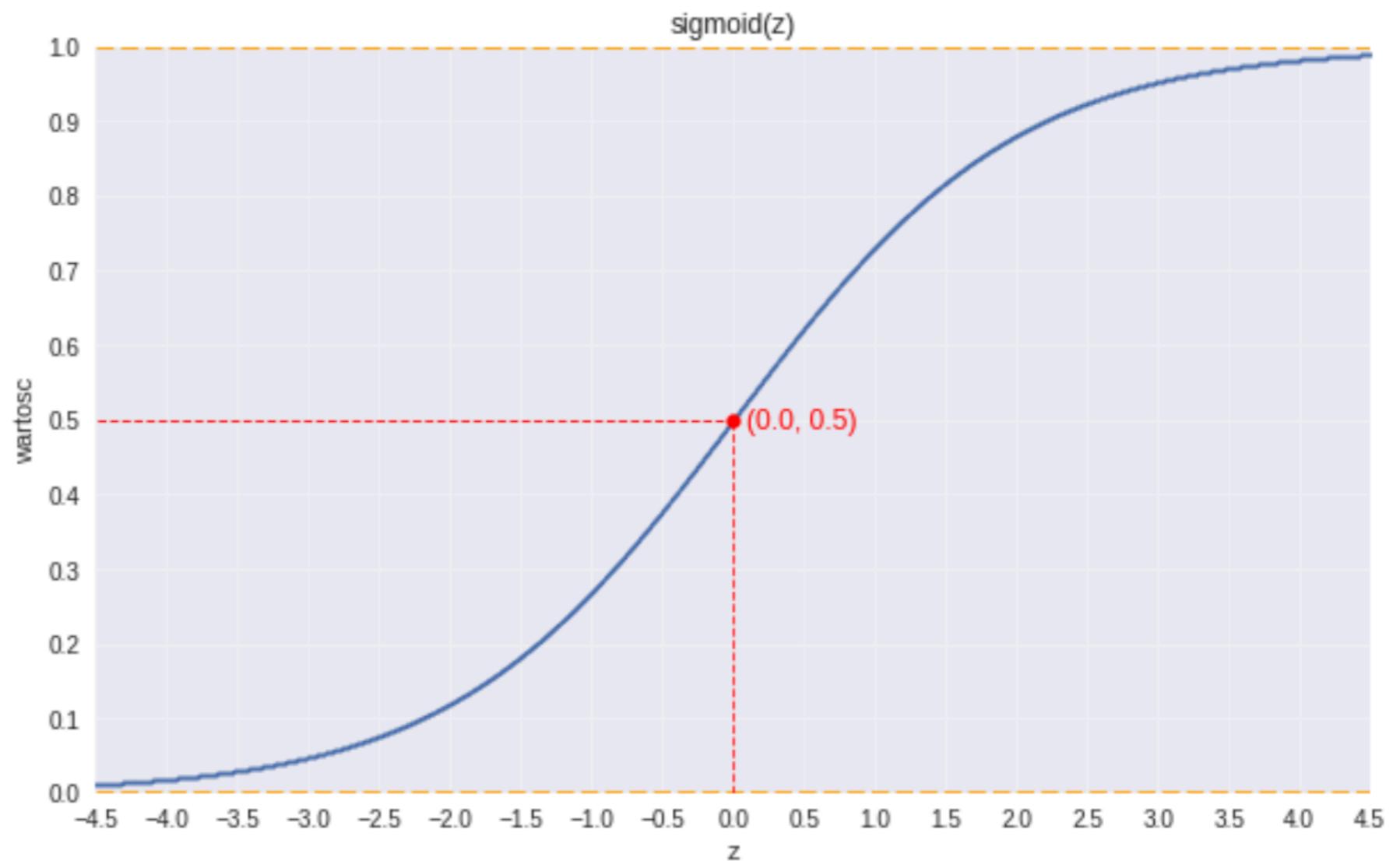
$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



Teoria

Funkcja Aktywacji - Sigmoid

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



przedział wartości (0, 1)

Teoria

$$f(x) : \hat{y} = w_0x_0 + w_1x_1 + b \quad (-\infty, +\infty)$$

Teoria

$$f(x) : z = w_0x_0 + w_1x_1 + b \quad (-\infty, +\infty)$$

Teoria

$$z = w_0x_0 + w_1x_1 + b$$

(-inf, +inf)



$$f(x) : \text{sigmoid}(z)$$

(0, 1)

Teoria

$$z = w_0x_0 + w_1x_1 + b \quad (-\infty, +\infty)$$



$$f(x) : \text{sigmoid}(z) \quad (0, 1)$$



$\hat{y} =$ Prawdopodobieństwo,
że ktoś zaliczy egzamin.

Teoria

$$z = w_0x_0 + w_1x_1 + b \quad (-\infty, +\infty)$$



$$f(x) : \text{sigmoid}(z) \quad (0, 1)$$



$\hat{y} =$ Prawdopodobieństwo,
że ktoś zaliczy egzamin.

0.7 -> pewnie zda
0.15 -> mała szansa

Teoria

$$z = w_0x_0 + w_1x_1 + b \quad (-\infty, +\infty)$$



$$f(x) : \text{sigmoid}(z) \quad (0, 1)$$



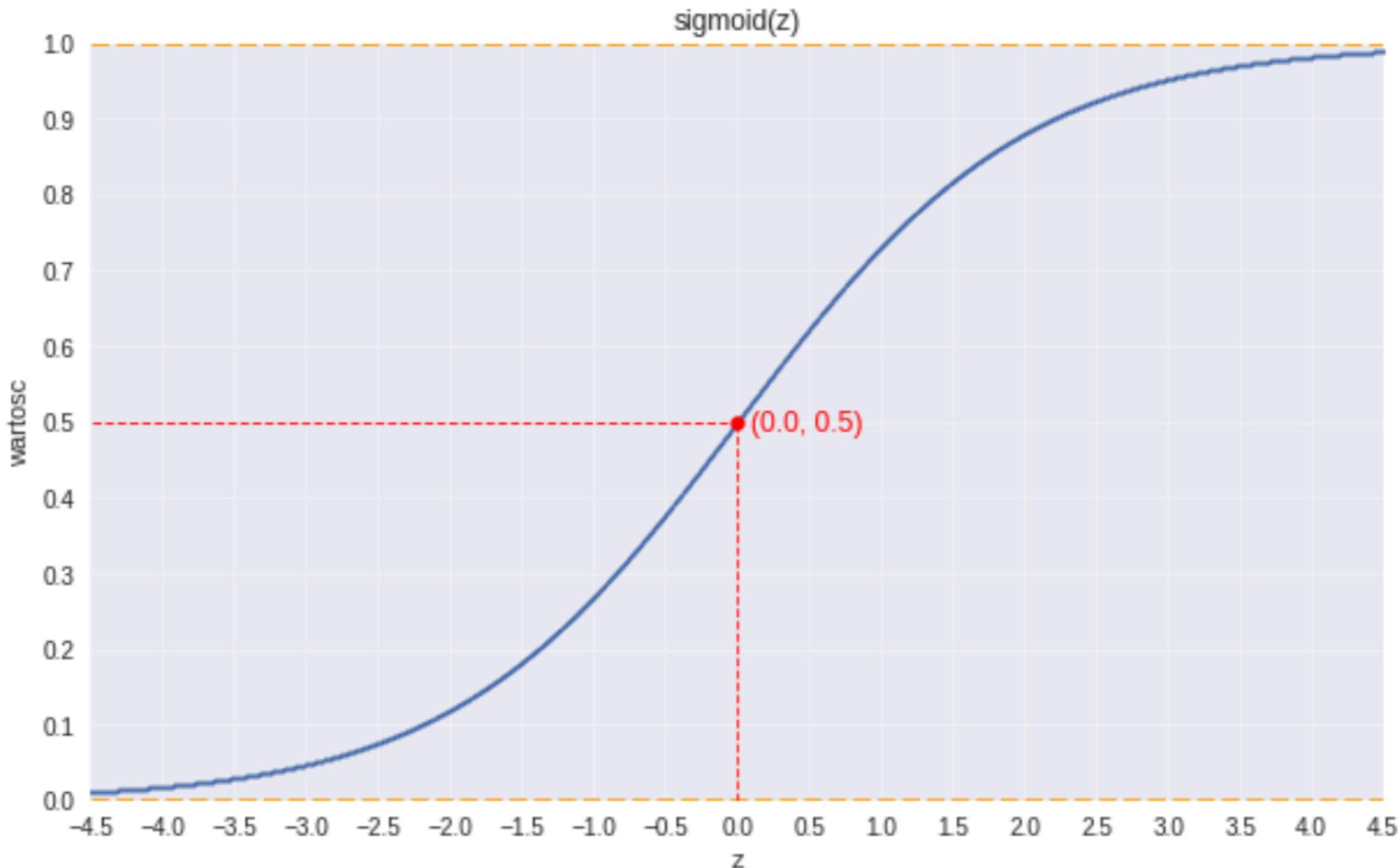
$\hat{y} =$ Prawdopodobieństwo,
że ktoś zaliczy egzamin.

Jeżeli $\hat{y} \geq 0.5$ zwróć 1.

{0, 1}

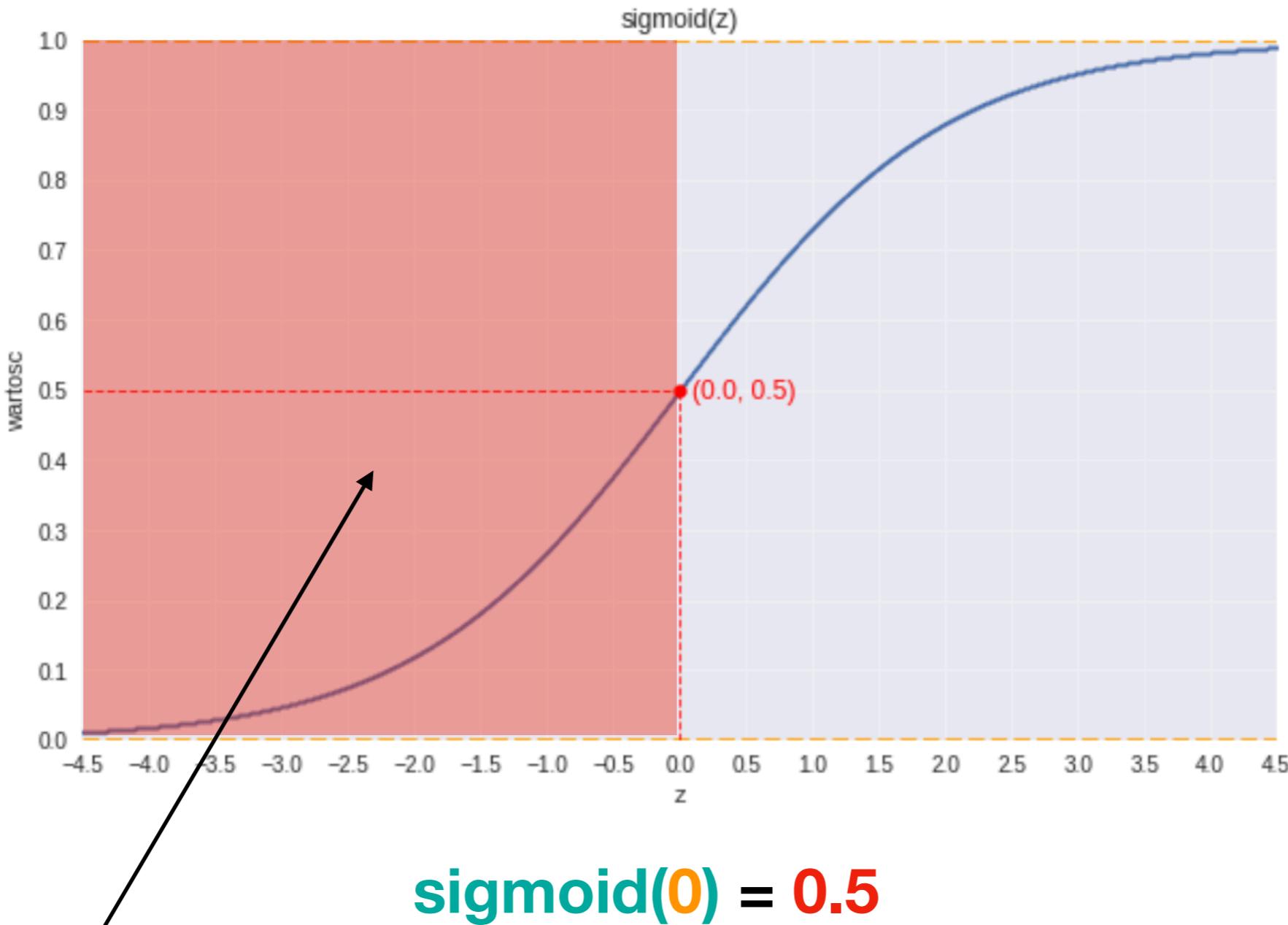
Jeżeli $\hat{y} < 0.5$ zwróć 0.

Teoria



$$\text{sigmoid}(0) = 0.5$$

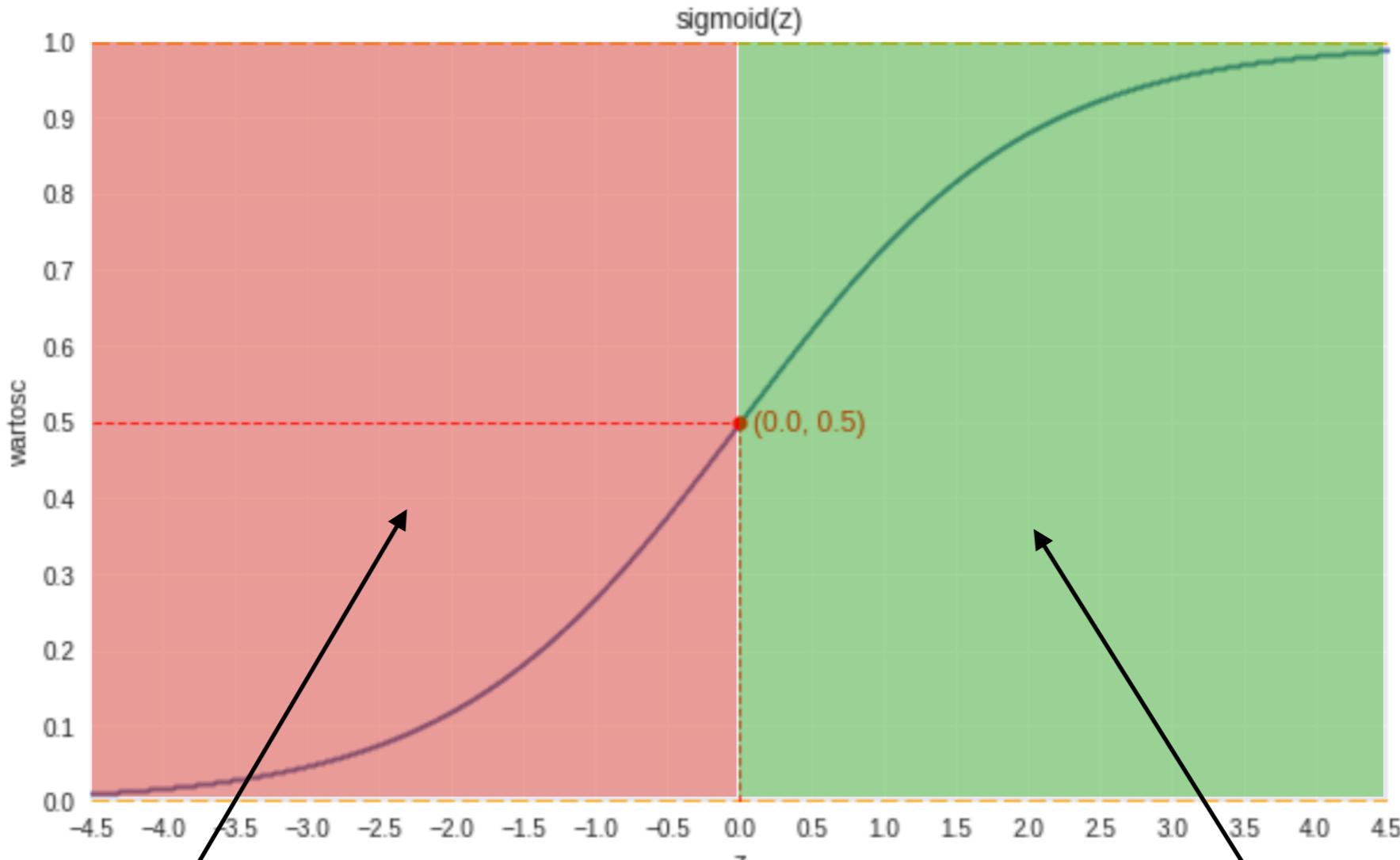
Teoria



Zwróć 0 dla:

$$w_0x_0 + w_1x_1 + b < 0$$

Teoria



$$\text{sigmoid}(0) = 0.5$$

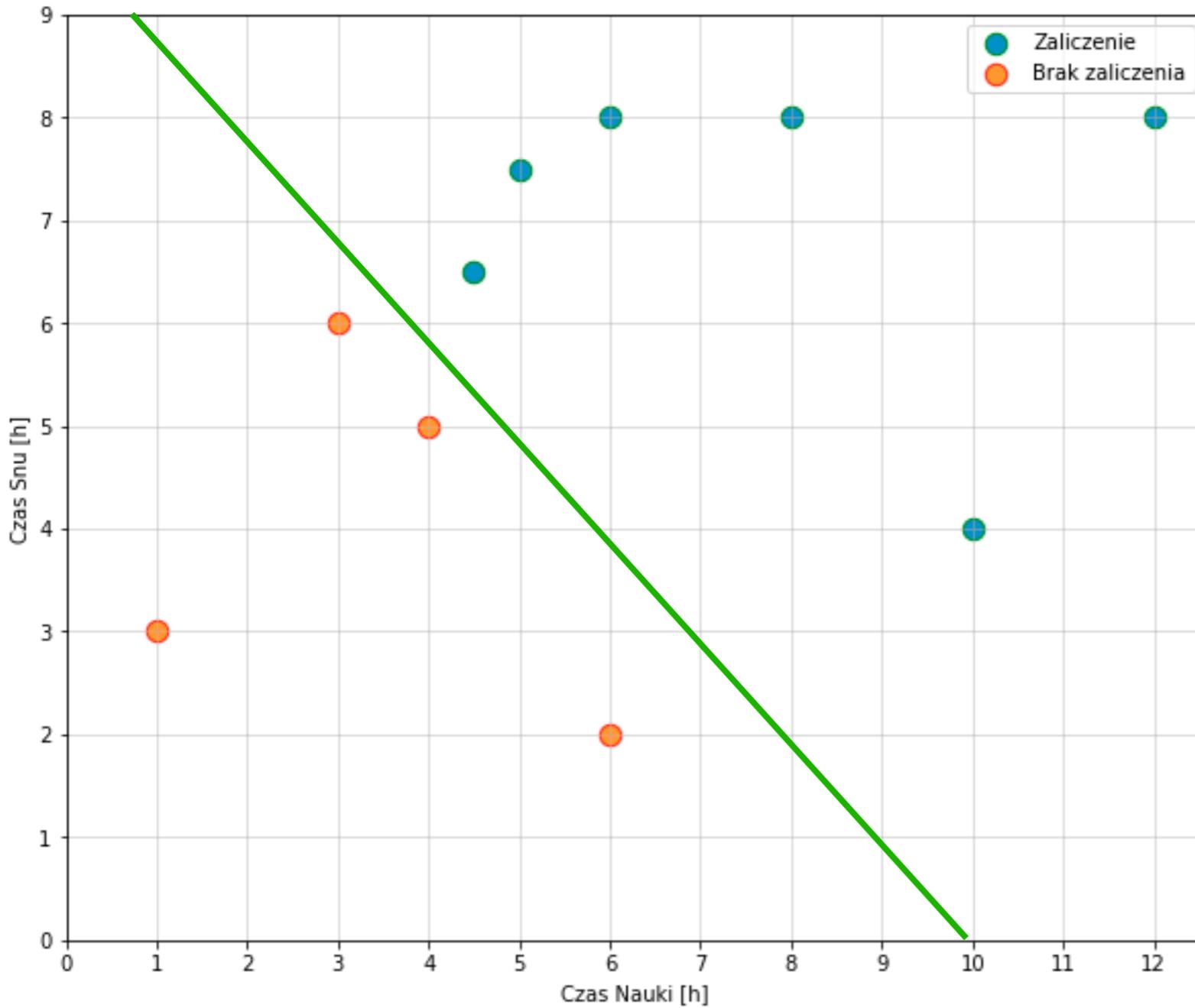
Zwróć 0 dla:

$$w_0x_0 + w_1x_1 + b < 0$$

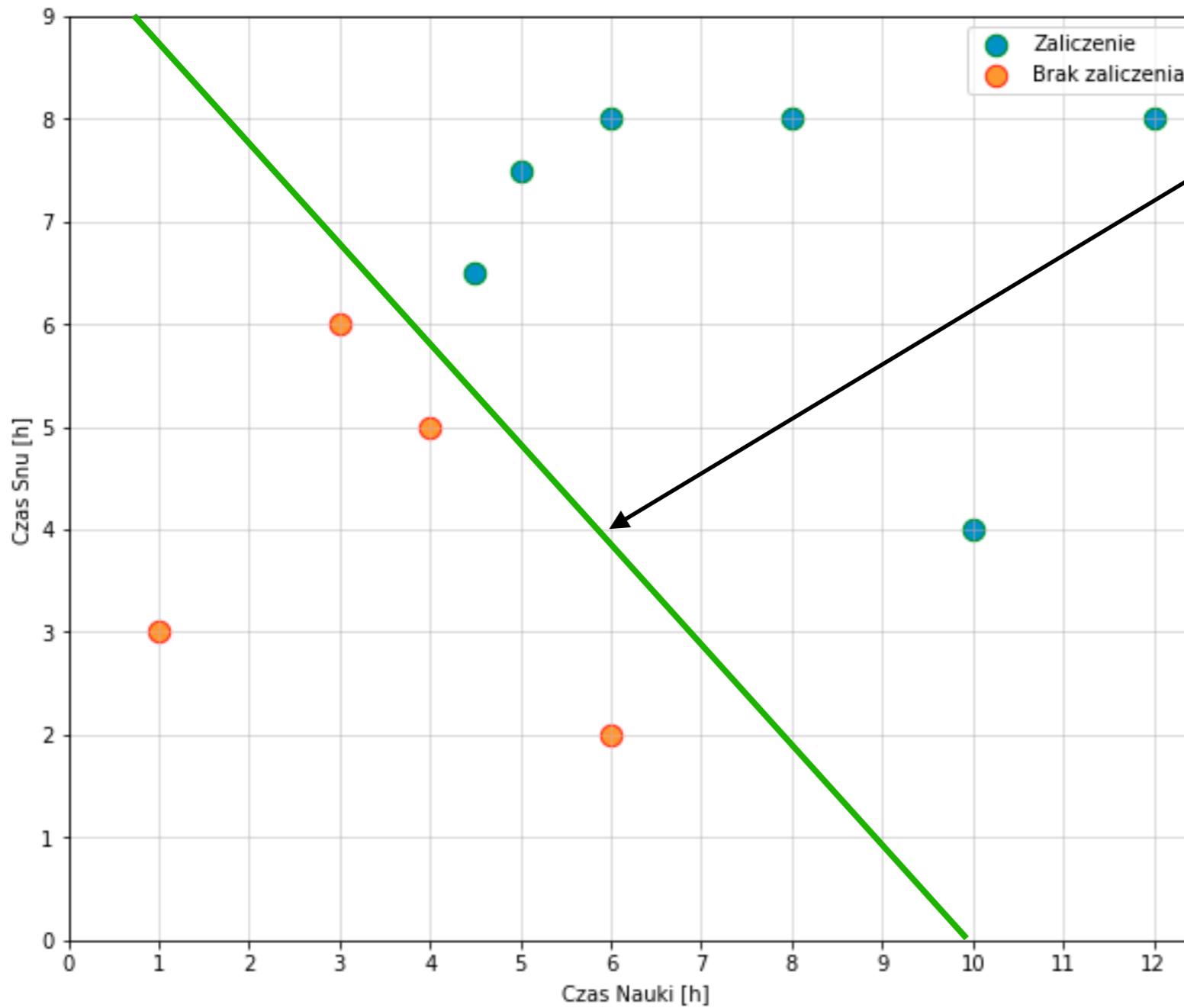
Zwróć 1 dla:

$$w_0x_0 + w_1x_1 + b \geq 0$$

Teoria

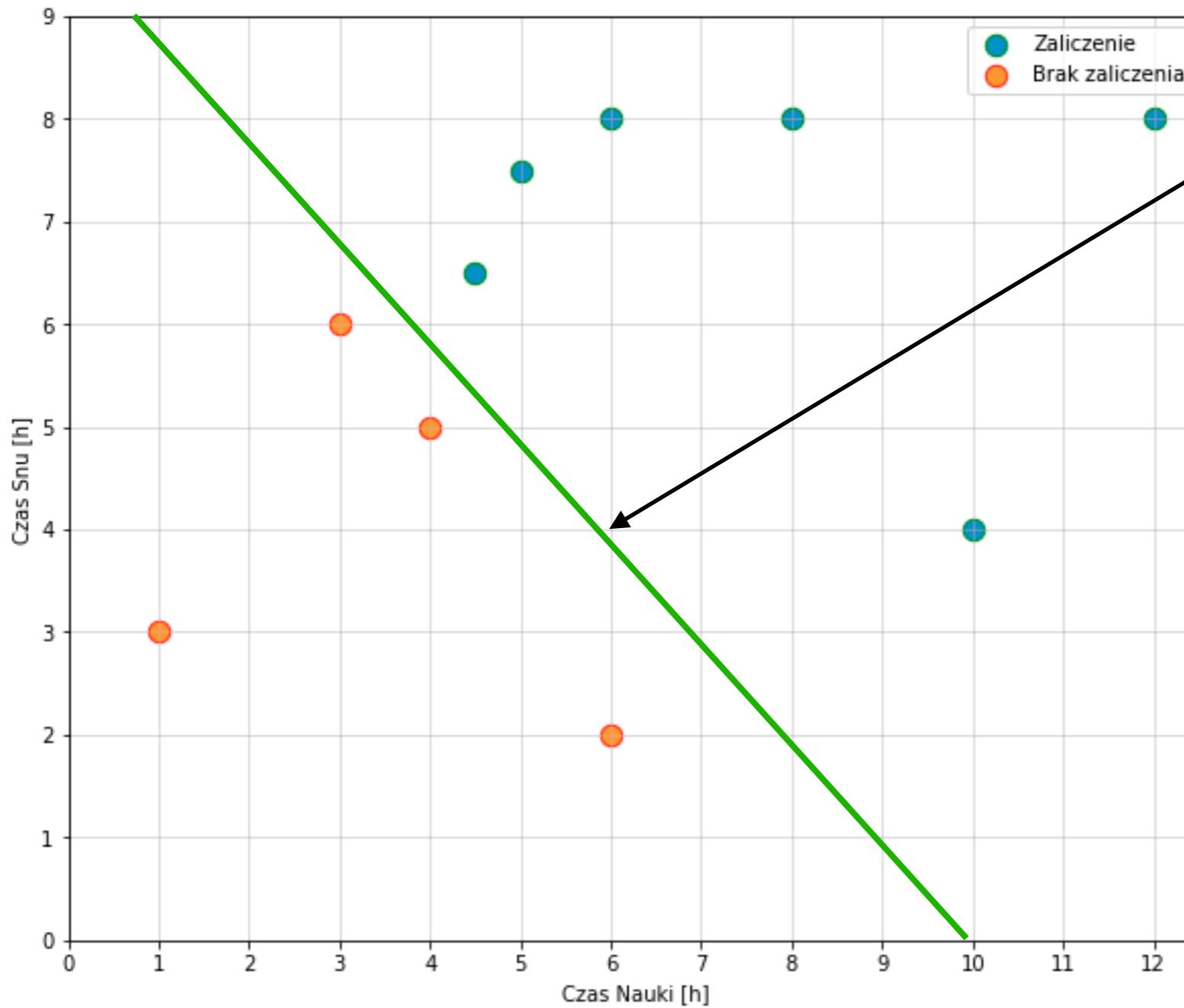


Teoria



$$w_0x_0 + w_1x_1 + b = 0$$

Teoria



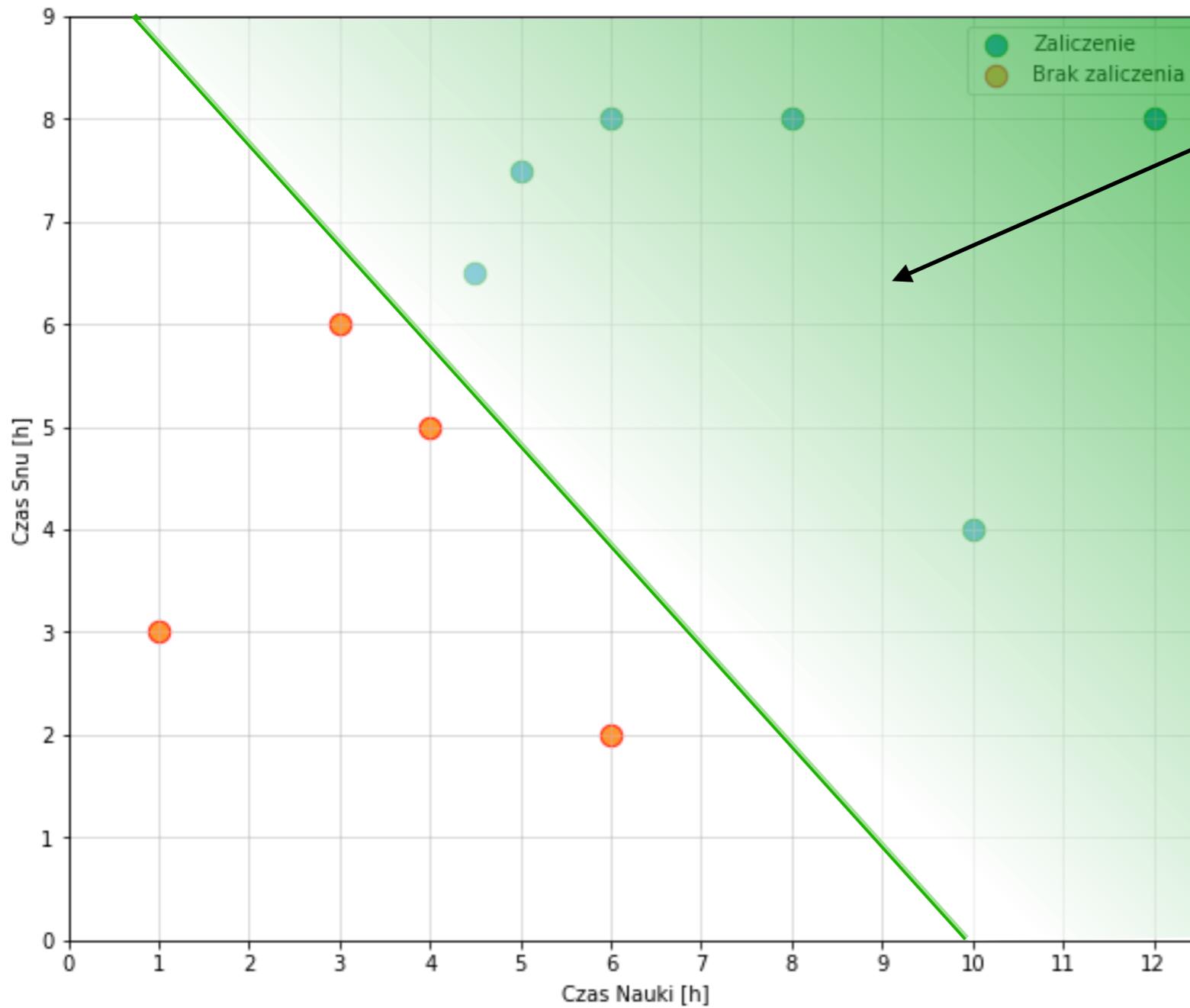
$$w_0x_0 + w_1x_1 + b = 0$$

bo

$$\text{sigmoid}(0) = 0.5$$

a to wartość graniczna

Teoria



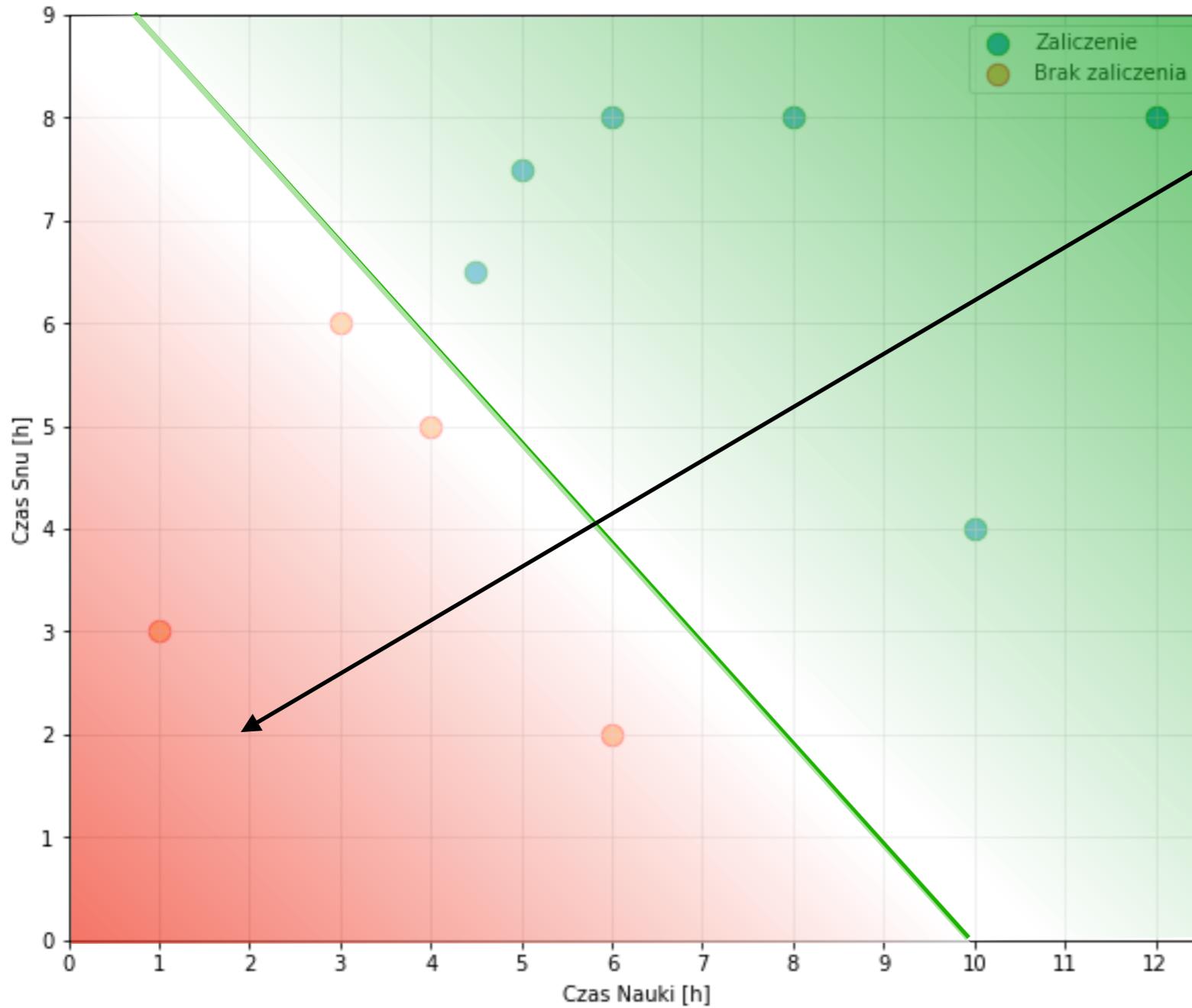
$$w_0x_0 + w_1x_1 + b \geq 0$$

bo

$$\text{sigmoid}(\geq 0)$$

dąży do 1

Teoria



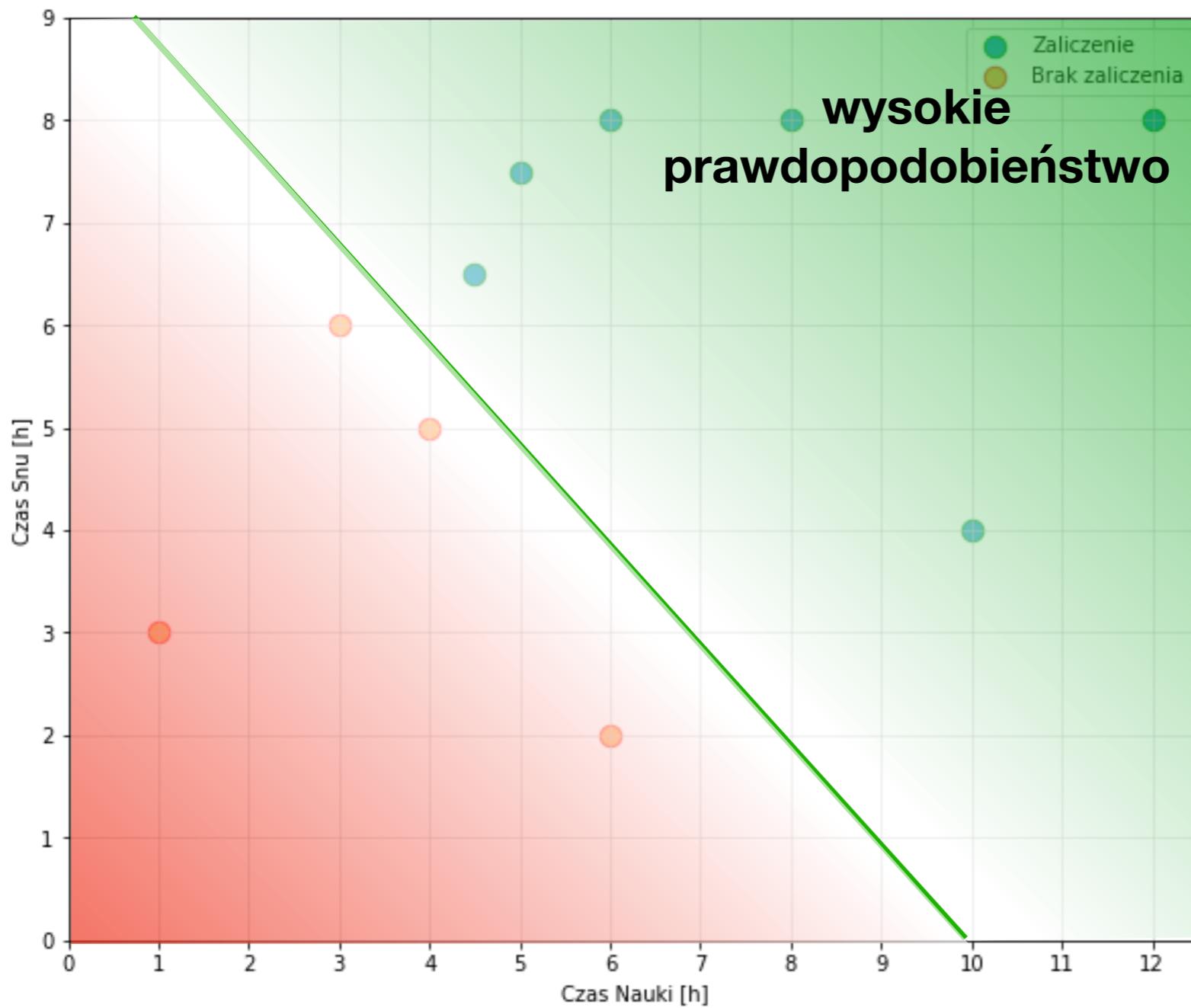
$$w_0x_0 + w_1x_1 + b < 0$$

bo

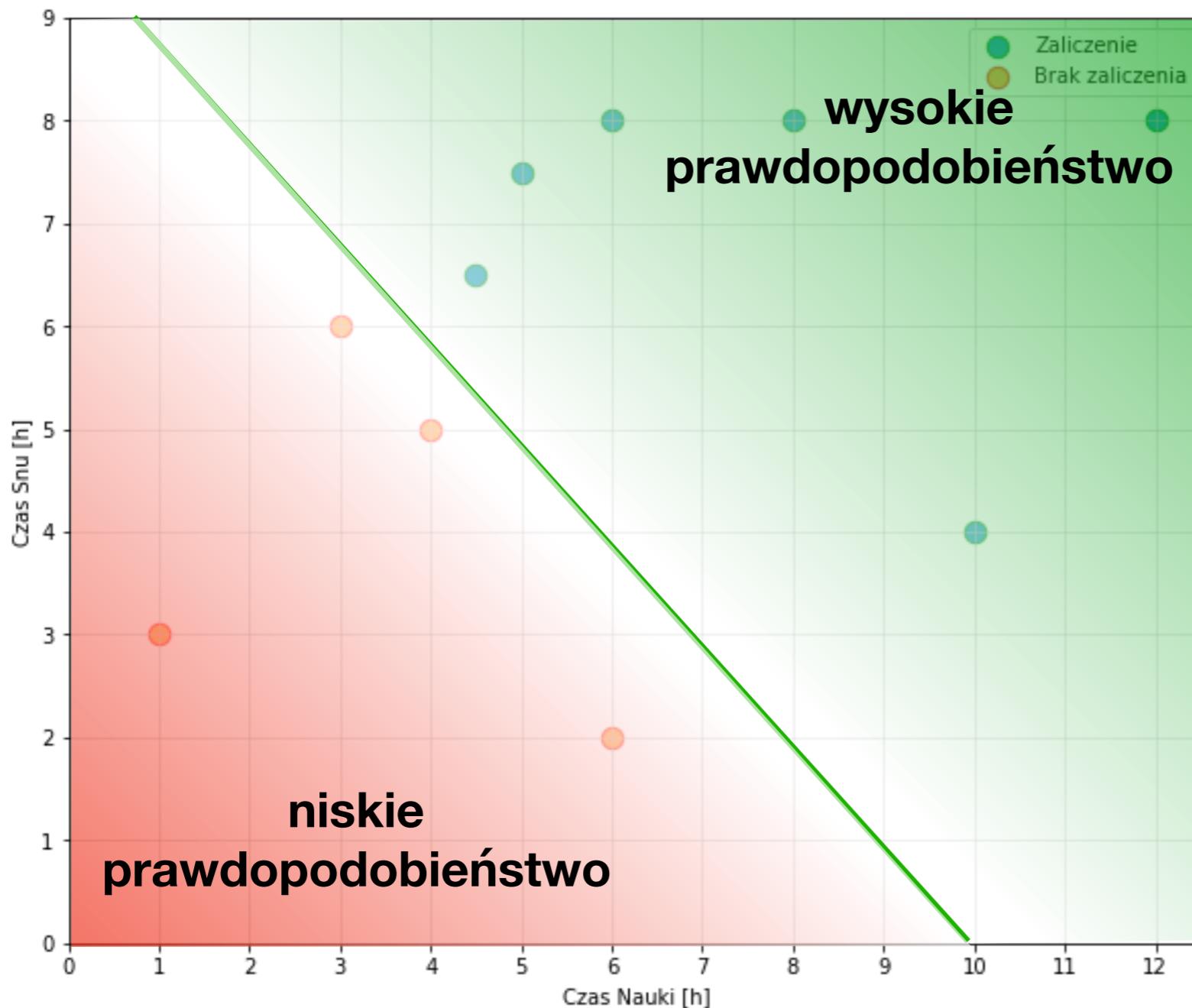
sigmoid(< 0)

dąży do 0

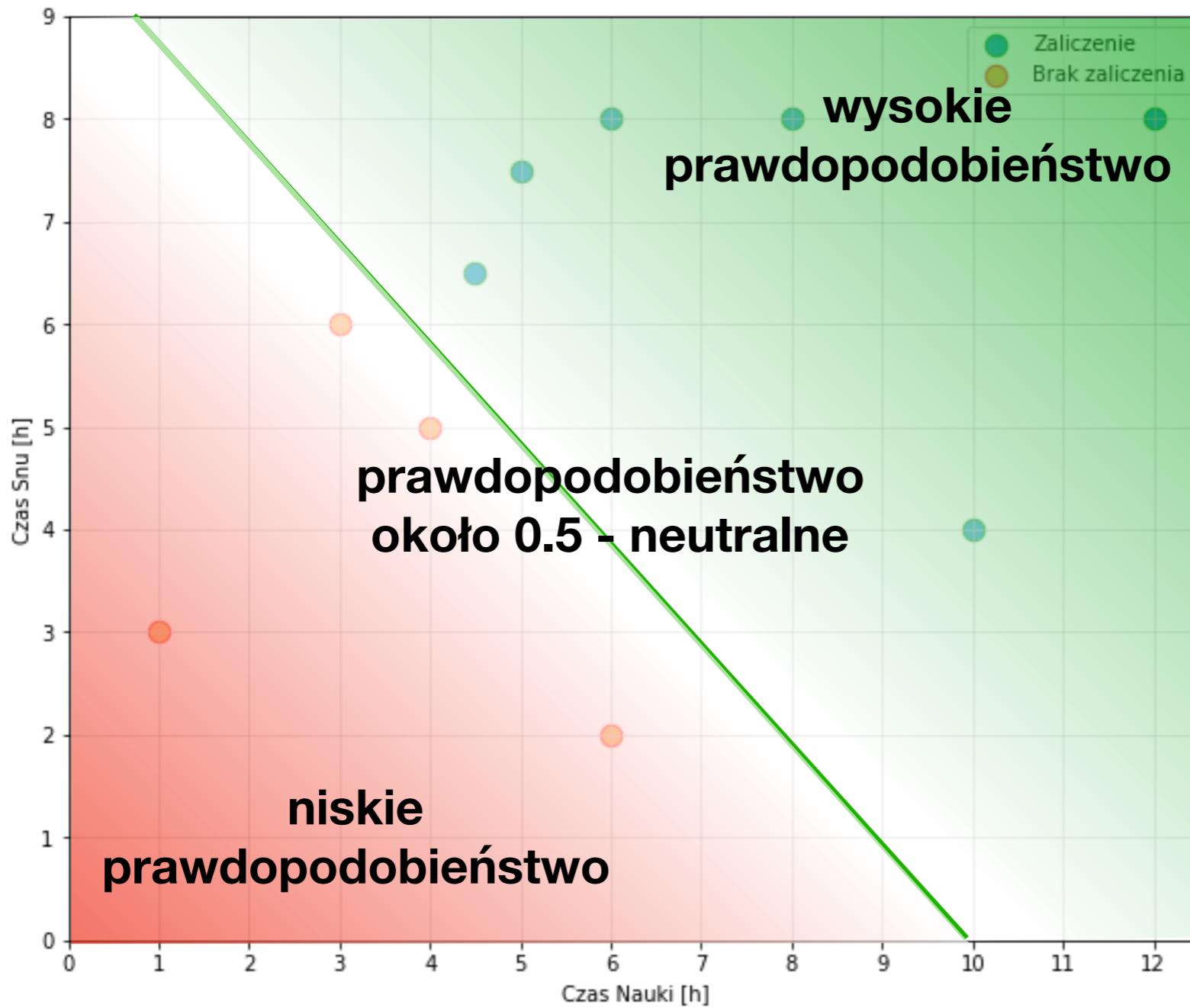
Teoria



Teoria



Teoria



Teoria

f(x) : sigmoid(z)

gdzie

z = w₀x₀ + w₁x₁ + b

Teoria

f(x): sigmoid(**z**)

gdzie

z = **w₀x₀ + w₁x₁ + b**

Id	Czas Nauki [h]	Czas Snu [h]	Zaliczenie
0	6	8	1
1	1	3	0
2	5	7.5	1
3	3	6	0
4	4.5	6.5	1
5	10	4	1
6	12	8	1
7	8	8	1
8	6	2	0
9	4	5	0

Teoria

f(x) : sigmoid(z)

gdzie

z = w₀x₀ + w₁x₁ + b



Id	Czas Nauki [h]	Czas Snu [h]	Zaliczenie
0	6	8	1
1	1	3	0
2	5	7.5	1
3	3	6	0
4	4.5	6.5	1
5	10	4	1
6	12	8	1
7	8	8	1
8	6	2	0
9	4	5	0

Teoria

$f(x)$: sigmoid(z)

gdzie

$$z = w_0x_0 + w_1x_1 + b$$



$$x_0 = 6 \quad x_1 = 8$$



Id	Czas Nauki [h]	Czas Snu [h]	Zaliczenie
0	6	8	1
1	1	3	0
2	5	7.5	1
3	3	6	0
4	4.5	6.5	1
5	10	4	1
6	12	8	1
7	8	8	1
8	6	2	0
9	4	5	0

Teoria

$f(x)$: sigmoid(z)

gdzie

$$z = w_0x_0 + w_1x_1 + b$$



$$x_0 = 6 \quad x_1 = 8$$

$$y = 1$$

A red arrow points to the first row of the table, highlighting the data point for Id 0.

Id	Czas Nauki [h]	Czas Snu [h]	Zaliczenie
0	6	8	1
1	1	3	0
2	5	7.5	1
3	3	6	0
4	4.5	6.5	1
5	10	4	1
6	12	8	1
7	8	8	1
8	6	2	0
9	4	5	0

Teoria

$f(x)$: sigmoid(z)

gdzie

$$z = w_0x_0 + w_1x_1 + b$$



$$\hat{y} = 0.8$$

$$x_0 = 6 \quad x_1 = 8$$

$$y = 1$$



Id	Czas Nauki [h]	Czas Snu [h]	Zaliczenie
0	6	8	1
1	1	3	0
2	5	7.5	1
3	3	6	0
4	4.5	6.5	1
5	10	4	1
6	12	8	1
7	8	8	1
8	6	2	0
9	4	5	0

Teoria

$f(x)$: sigmoid(z)



gdzie

$$z = w_0x_0 + w_1x_1 + b$$

$$x_0 = 6 \quad x_1 = 8$$

$$\hat{y} = 1$$



Powinno być 1 a jest 0.8.

Teoria

$f(x)$: sigmoid(z)

gdzie

$z = w_0x_0 + w_1x_1 + b$



$\hat{y} = 0.8$



$x_0 = 6$

$x_1 = 8$

$y = 1$

Powinno być 1 a jest 0.8.

Do klasyfikacji używa się zwykle
innej funkcji kosztu.

Teoria

$f(x)$: sigmoid(z)



gdzie

$z = w_0x_0 + w_1x_1 + b$

$x_0 = 6 \quad x_1 = 8$

$y = 1$



Cross Entropy

$$\mathcal{L} = y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

Teoria

$f(x)$: sigmoid(z)



gdzie

$z = w_0x_0 + w_1x_1 + b$

$x_0 = 6 \quad x_1 = 8$

$y = 1$



Cross Entropy

$$\mathcal{L} = y \log(\hat{y})$$

Dla $y = 1$

Teoria

$f(x)$: sigmoid(z)



gdzie

$$z = w_0x_0 + w_1x_1 + b$$

$$x_0 = 6 \quad x_1 = 8$$

$$y = 1$$



$$\hat{y} = 0.8$$

Cross Entropy

$$\mathcal{L} = - (1 - y) \log(1 - \hat{y}) \quad \text{Dla } y = 0$$

Teoria

$f(x)$: sigmoid(z)



gdzie

$$z = w_0x_0 + w_1x_1 + b$$

$$x_0 = 6 \quad x_1 = 8$$

$$y = 1$$



Cross Entropy

$$\mathcal{L} = y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

w₀, w₁, b

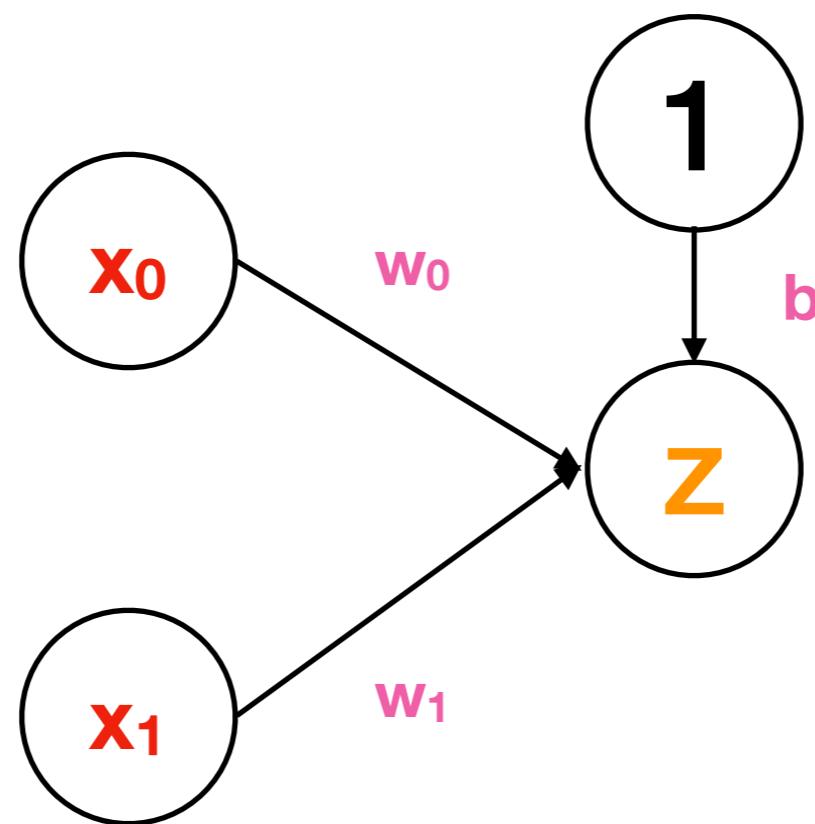


Spadek Gradientu

Teoria

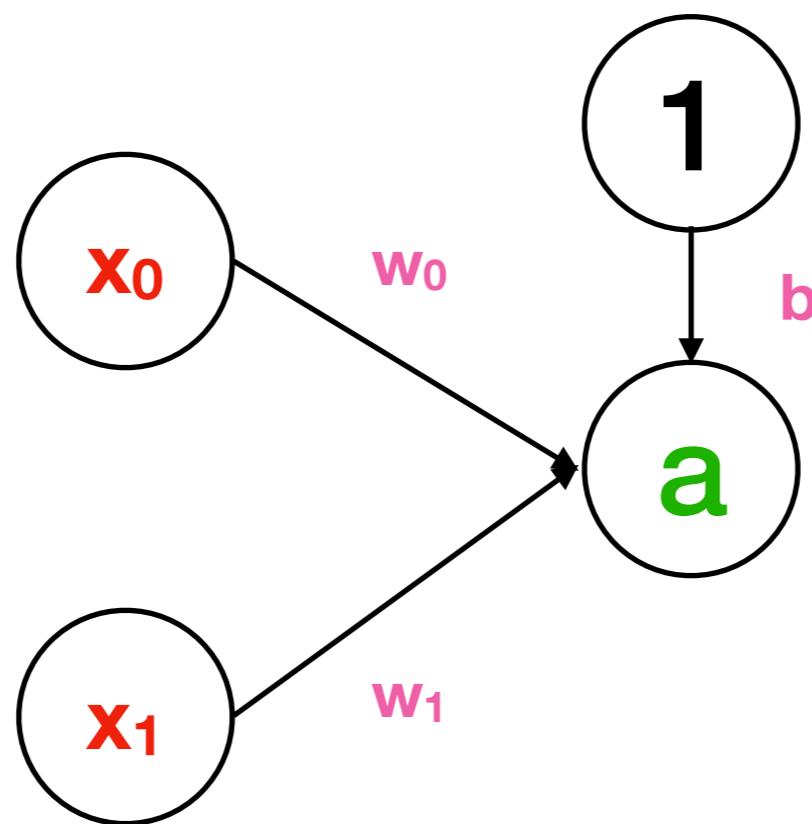
$$z = w_0x_0 + w_1x_1 + b$$

Teoria



$$z = w_0x_0 + w_1x_1 + b$$

Teoria



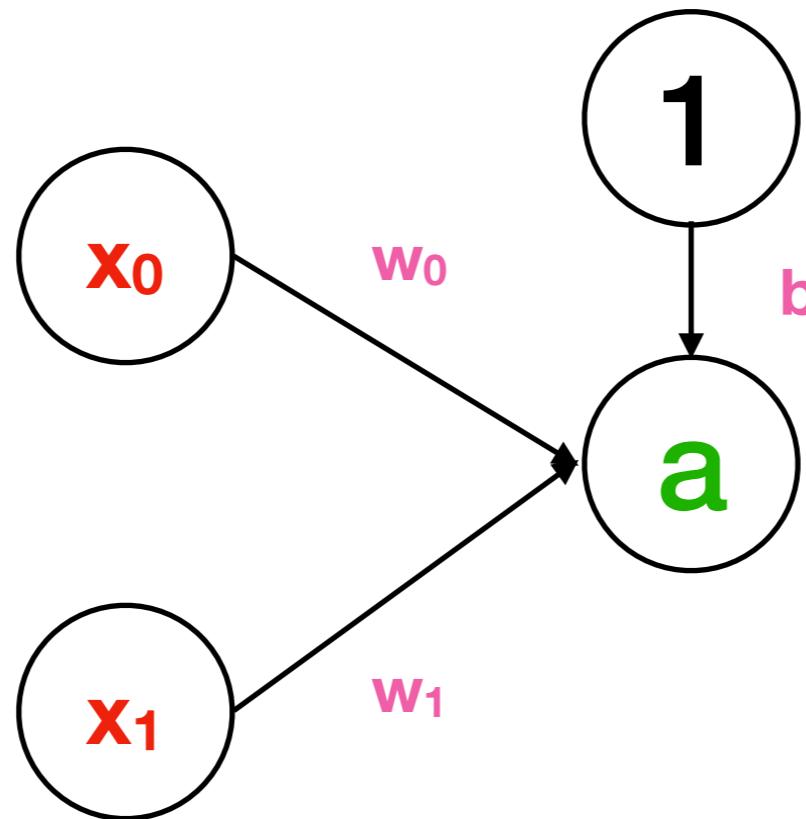
$$a = \text{sigmoid}(z)$$

$$z = w_0x_0 + w_1x_1 + b$$

Teoria

a - aktywacja

z - kombinacja liniowa



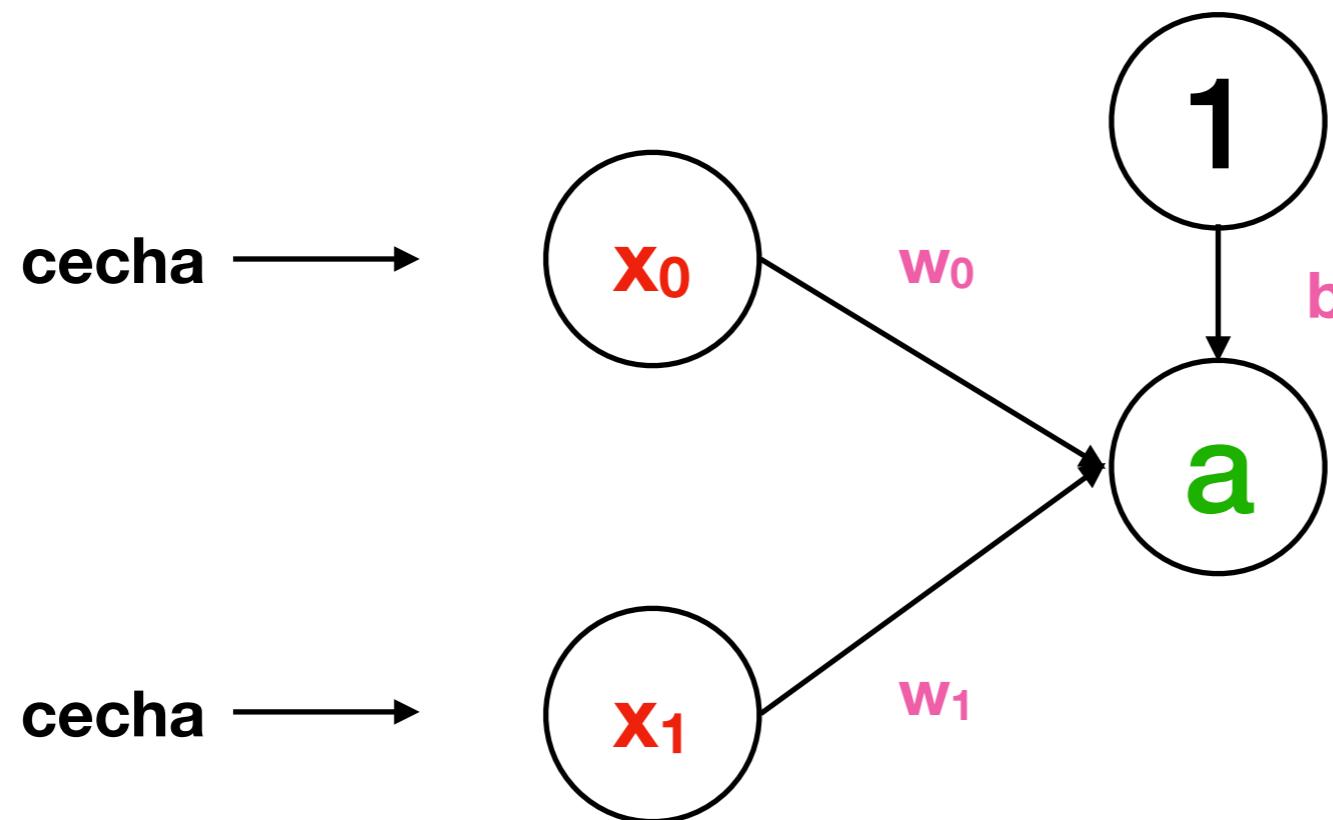
$$a = \text{sigmoid}(z)$$

$$z = w_0x_0 + w_1x_1 + b$$

Teoria

a - aktywacja

z - kombinacja liniowa

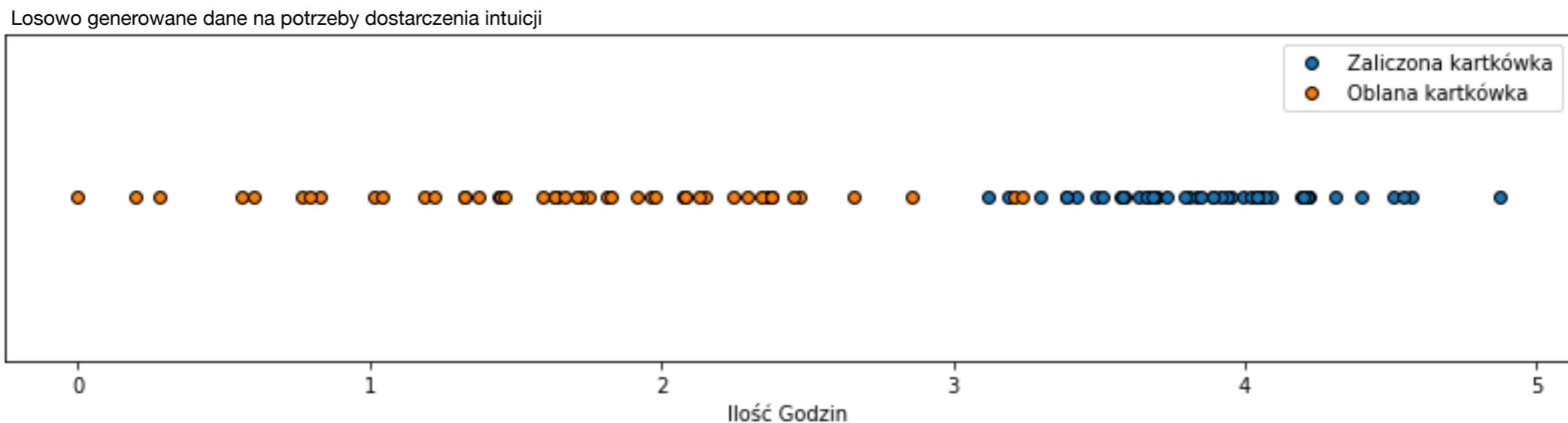


$$a = \text{sigmoid}(z)$$

$$z = w_0x_0 + w_1x_1 + b$$

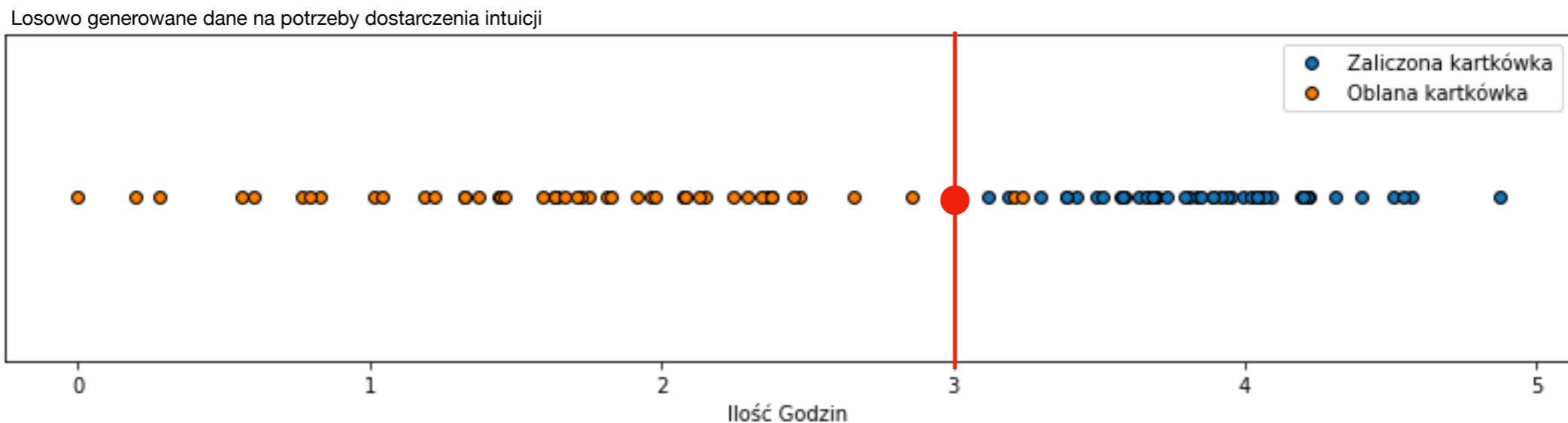
Teoria

**Gdybyśmy chcieli przewidywać przy pomocy tylko jednej cechy:
- ile godzin student spędził na nauce**



Teoria

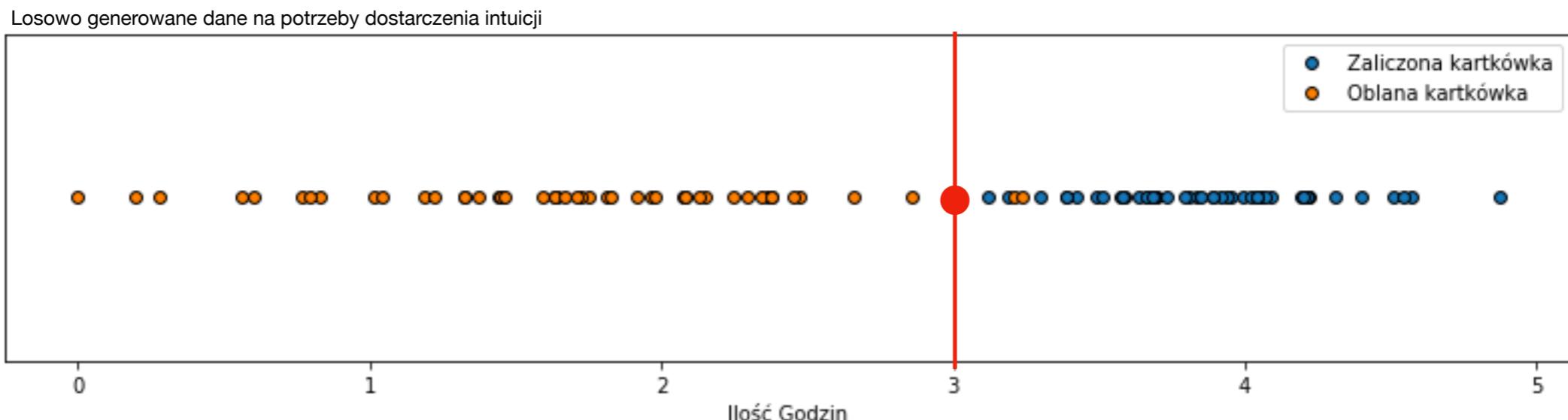
Gdybyśmy chcieli przewidywać przy pomocy tylko **jednej cechy**:
- ile godzin student spędził na nauce



Ponieważ dane są **1-wymiarowe**, potrzebny jest **jeden punkt** by je rozdzielić np. 3.0

Teoria

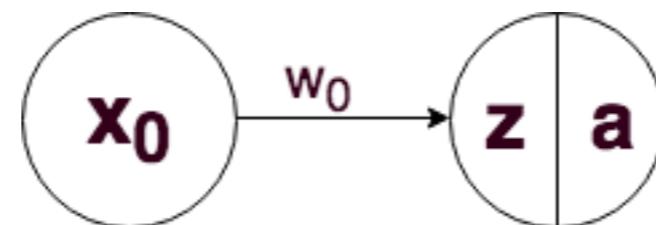
Gdybyśmy chcieli przewidywać przy pomocy tylko jednej cechy:
- ile godzin student spędził na nauce



Ponieważ dane są 1-wymiarowe, potrzebny jest jeden punkt by je rozdzielić np. 3.0

Do modelu Regresji Logistycznej wchodziła by tylko jedna cecha:

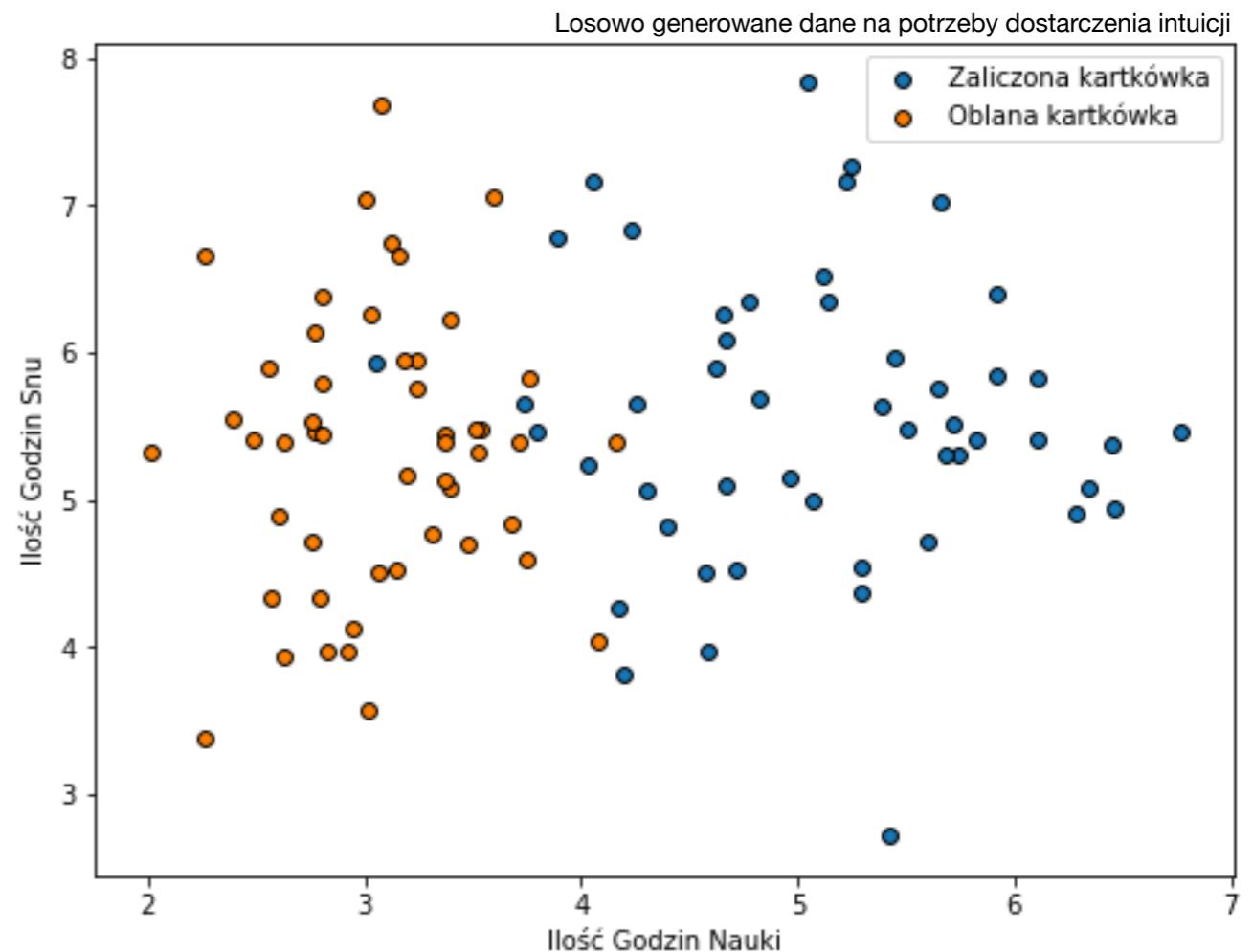
Ille godzin student
spędził na nauce



Teoria

Gdybyśmy chcieli przewidywać przy pomocy tylko **dwoch cech**:

- ile godzin student spędził na nauce
- ile godzin student spał

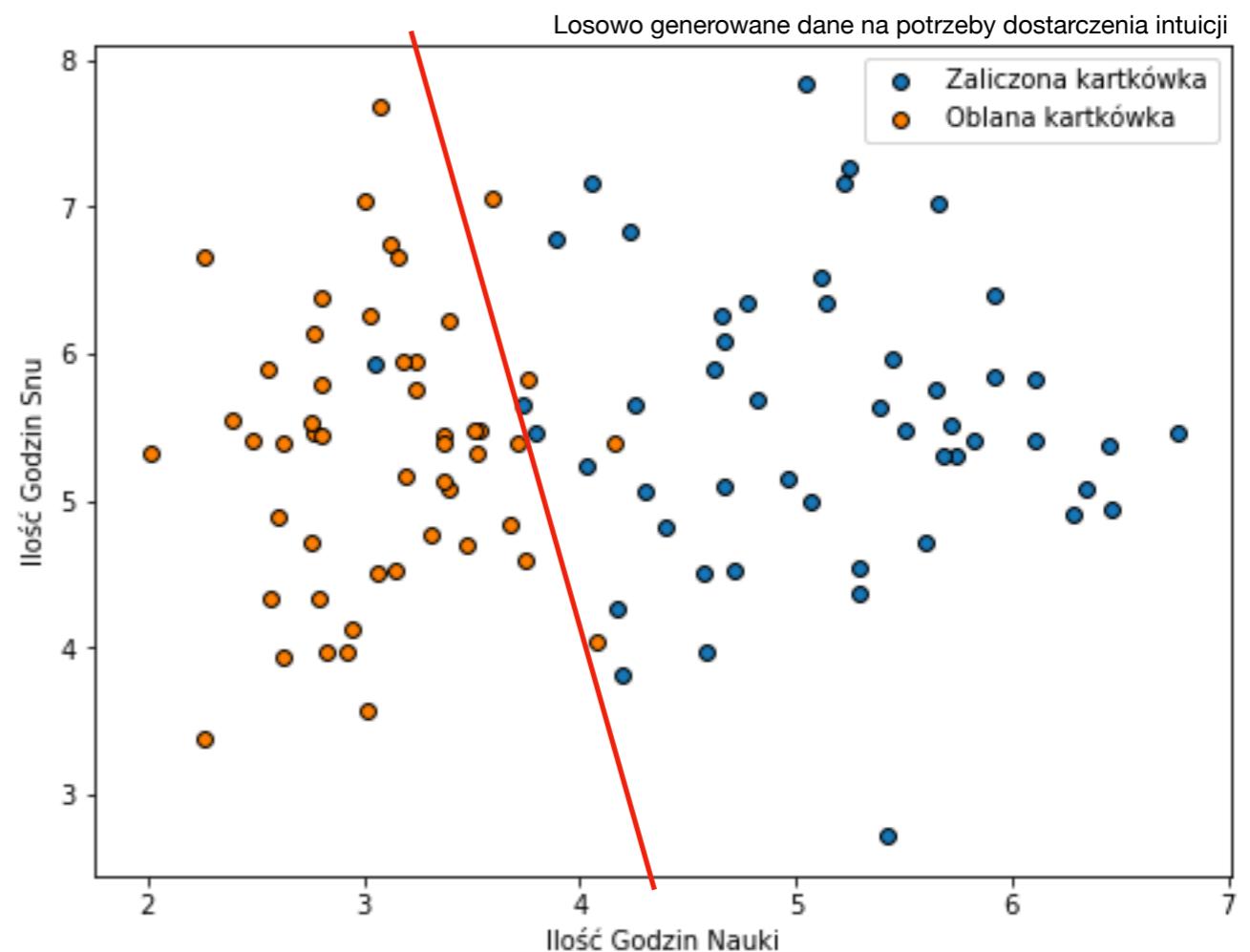


Teoria

Gdybyśmy chcieli przewidywać przy pomocy tylko **dwoch cech**:

- ile godzin student spędził na nauce
- ile godzin student spał

Ponieważ dane są **2-wymiarowe**,
potrzebna jest **linia** by je rozdzielić.



Teoria

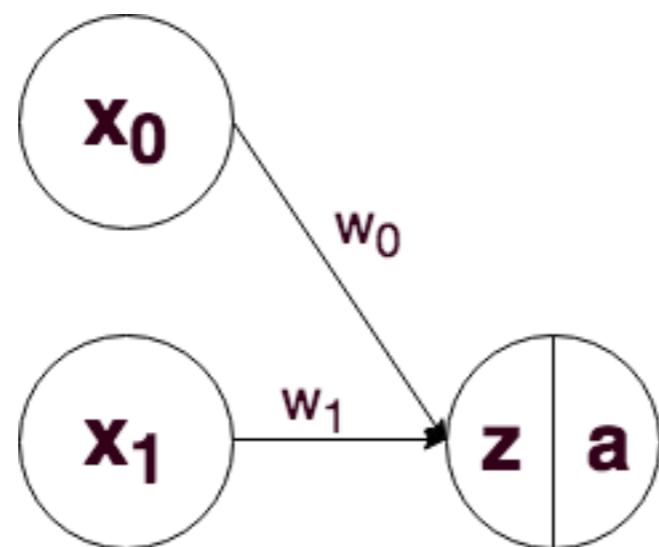
Gdybyśmy chcieli przewidywać przy pomocy tylko **dwoch cech**:

- ile godzin student spędził na nauce
- ile godzin student spał

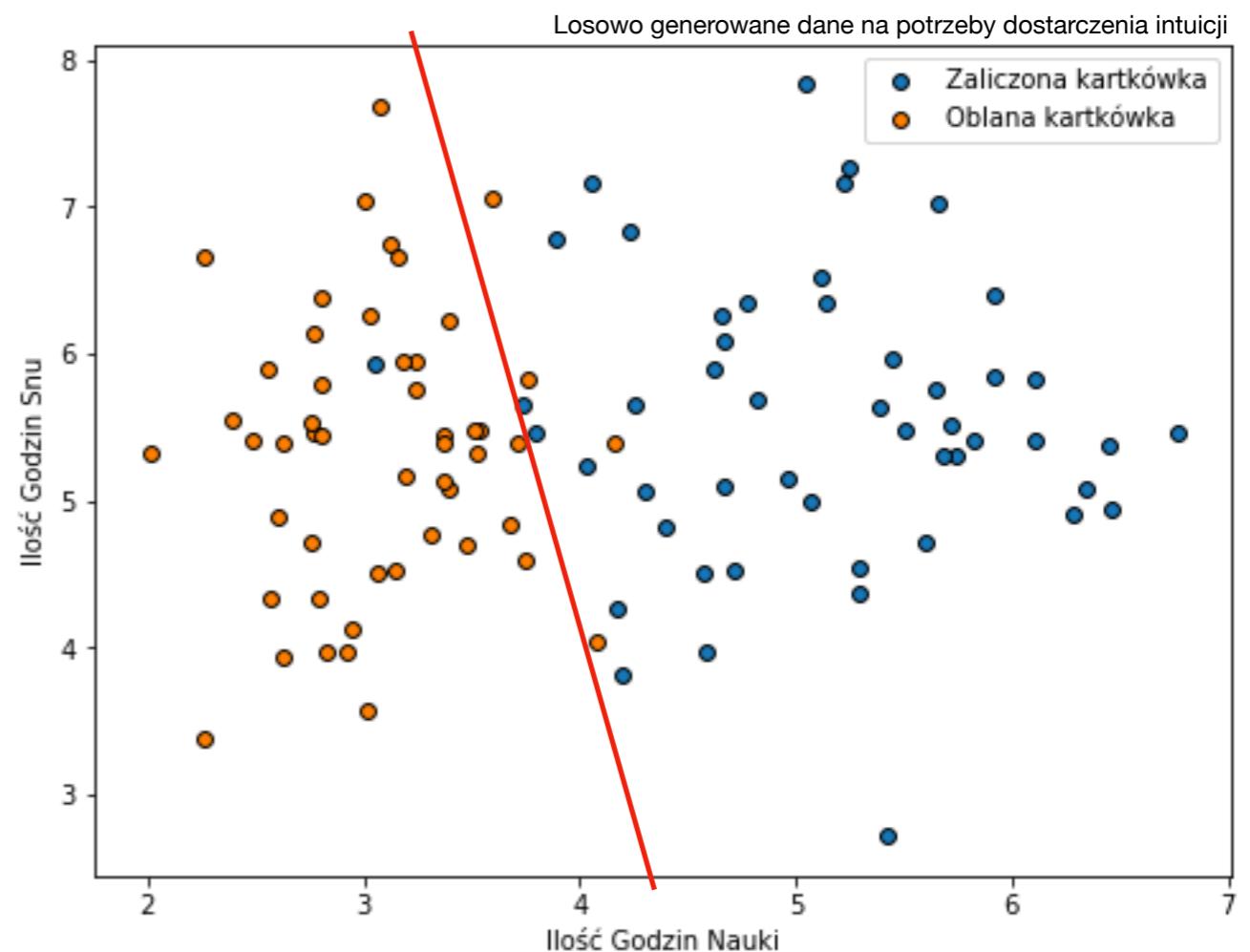
Ponieważ dane są **2-wymiarowe**, potrzebna jest **linia** by je rozdzielić.

Do modelu Regresji Logistycznej:
wchodziły by 2 cechy:

Ile godzin student
spędził na nauce



Ile godzin student
spał

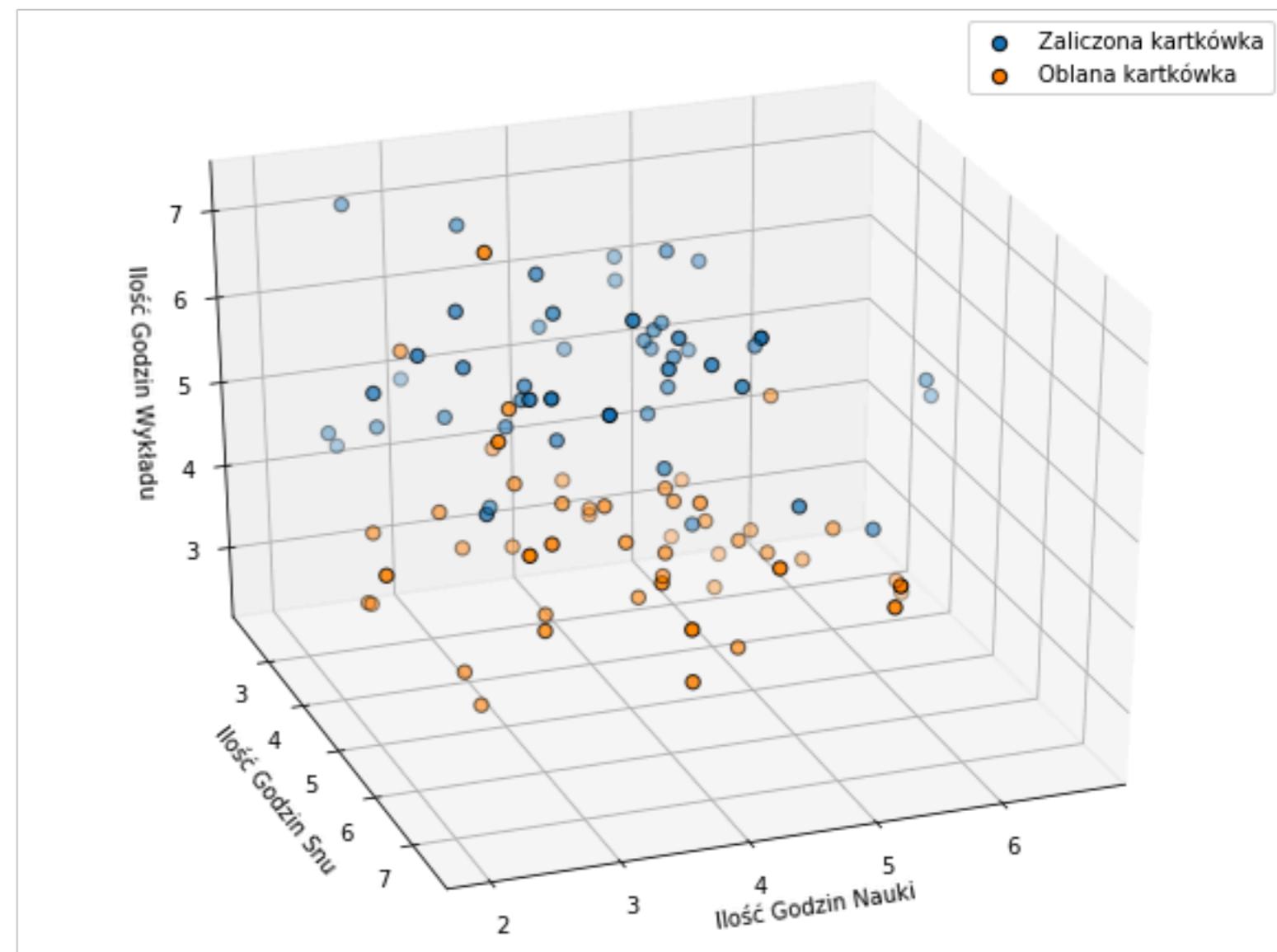


Teoria

Gdybyśmy chcieli przewidywać przy pomocy tylko **trzech cech**:

- ile godzin student spędził na nauce
- ile godzin student spał
- ilość godzin spędzona na wykładach

Losowo generowane dane na potrzeby dostarczenia intuicji



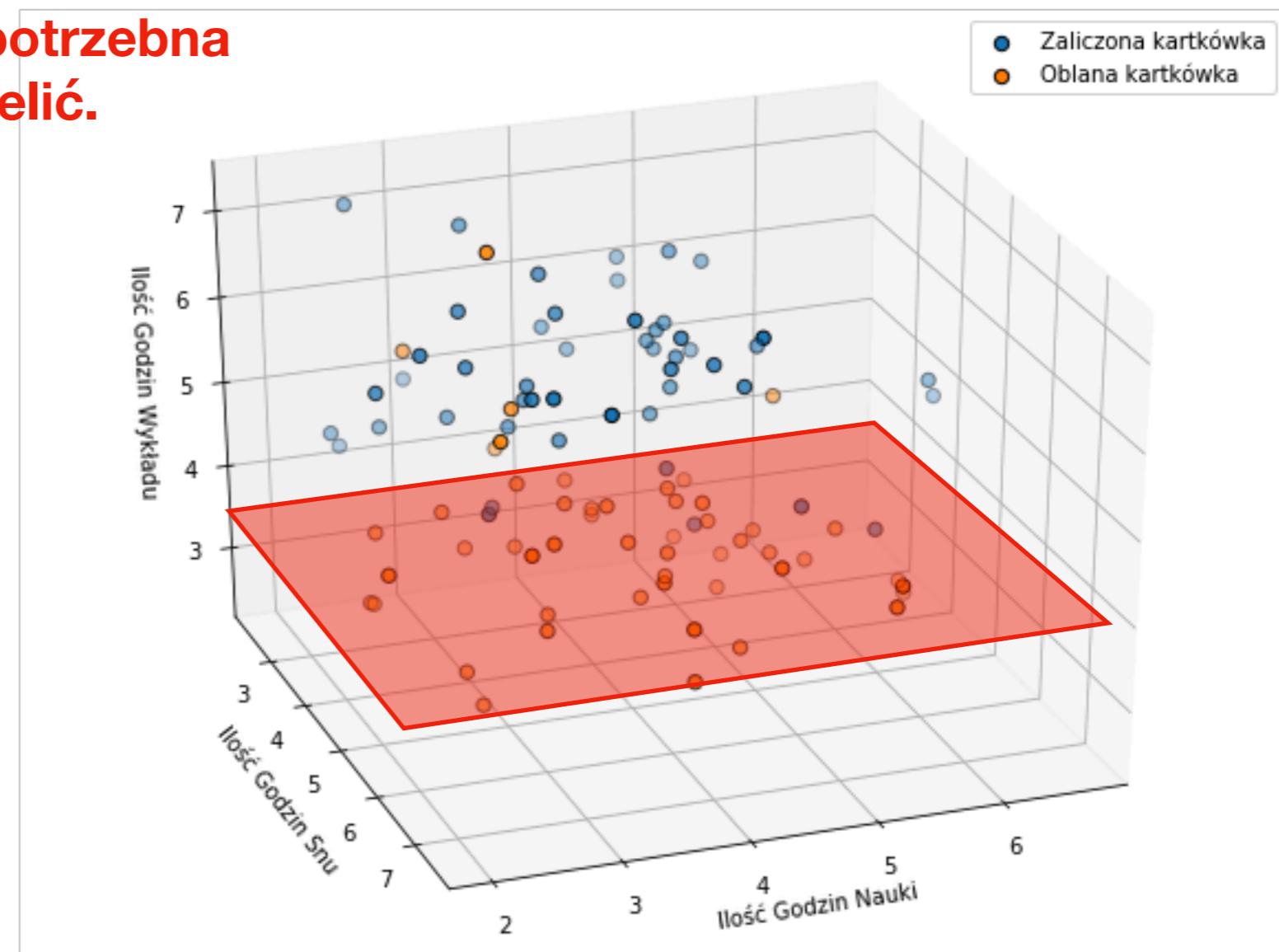
Teoria

Gdybyśmy chcieli przewidywać przy pomocy tylko **trzech cech**:

- ile godzin student spędził na nauce
- ile godzin student spał
- ilość godzin spędzona na wykładach

Losowo generowane dane na potrzeby dostarczenia intuicji

Ponieważ dane są **3-wymiarowe, potrzebna jest płaszczyzna by je rozdzielić.**



Teoria

Gdybyśmy chcieli przewidywać przy pomocy tylko **trzech cech**:

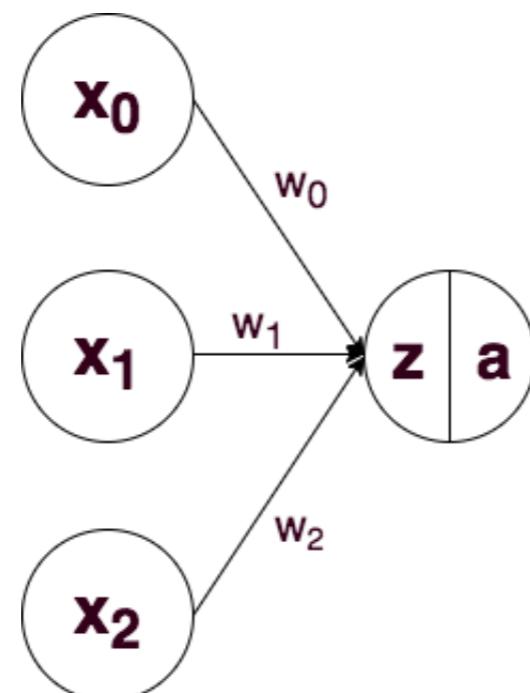
- ile godzin student spędził na nauce
- ile godzin student spał
- ilość godzin spędzona na wykładach

Losowo generowane dane na potrzeby dostarczenia intuicji

Ponieważ dane są **3-wymiarowe, potrzebna jest płaszczyzna by je rozdzielić.**

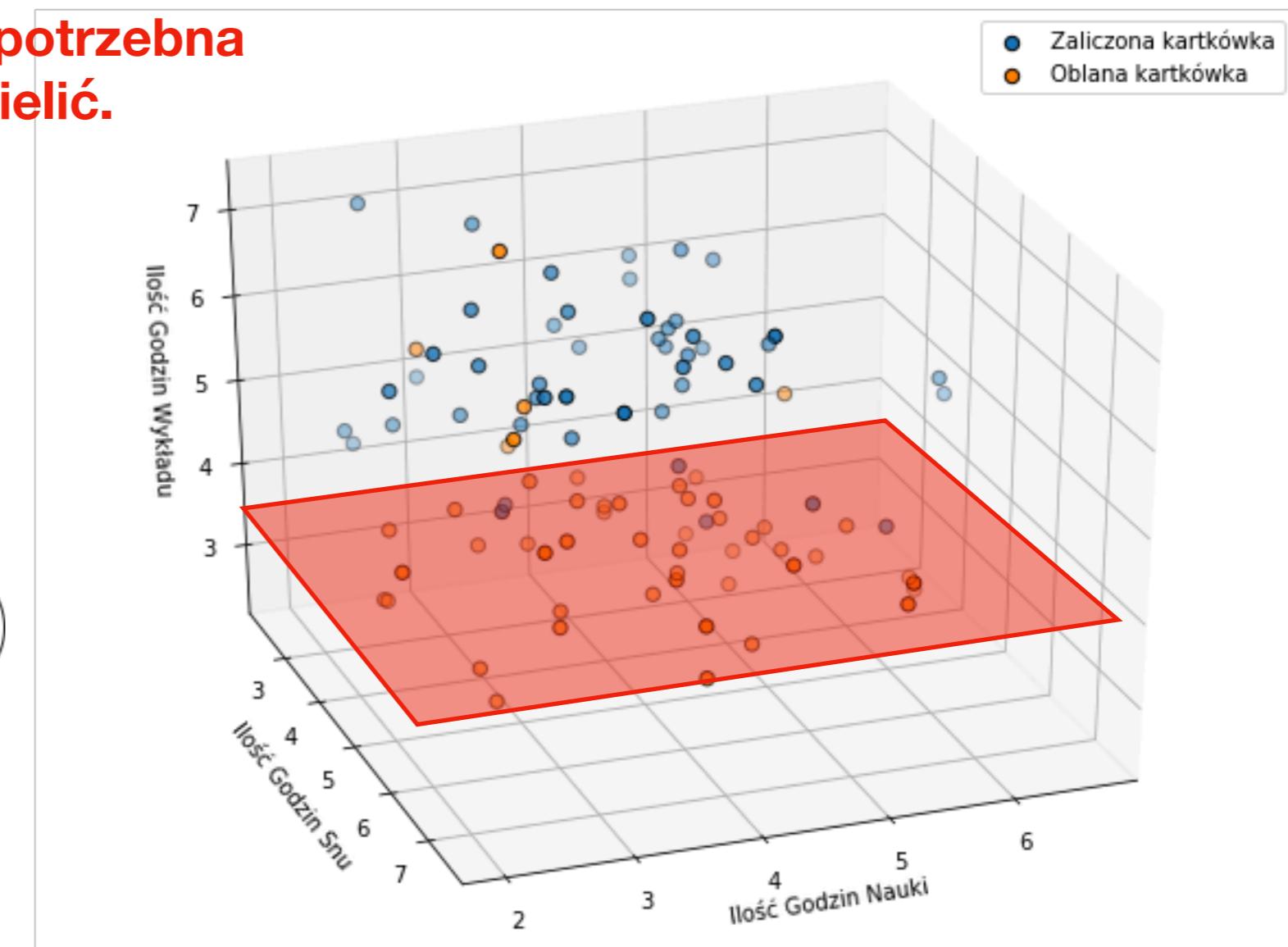
Do modelu Regresji Logistycznej wchodziły by 3 cechy:

Ile godzin student spędził na nauce



Ile godzin student spał

Ile godzin student spędził na wykładzie



Teoria

Gdybyśmy chcieli przewidywać przy pomocy tylko czterech lub więcej cech:

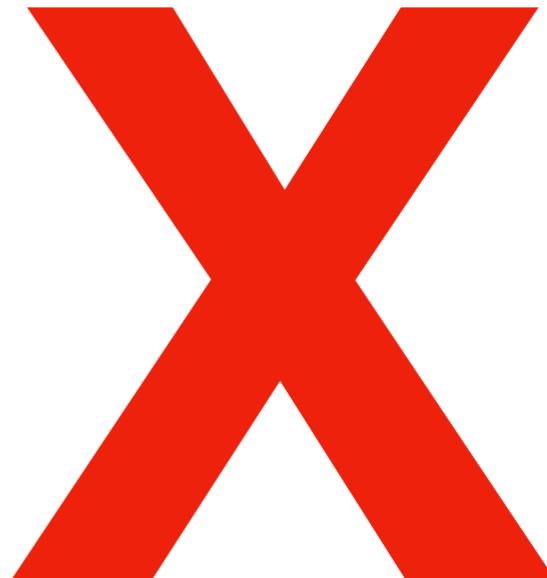
- ile godzin student spędził na nauce
- ile godzin student spał
- ilość godzin spędzona na wykładach
- stan zdrowia
- ...
- n

Teoria

Gdybyśmy chcieli przewidywać przy pomocy tylko czterech lub więcej cech:

- ile godzin student spędził na nauce
- ile godzin student spał
- ilość godzin spędzona na wykładach
- stan zdrowia
- ...
- n

**Dane wtedy są n-wymiarowe.
Dzielimy je n-wymiarową
hiperprzestrzenią.**



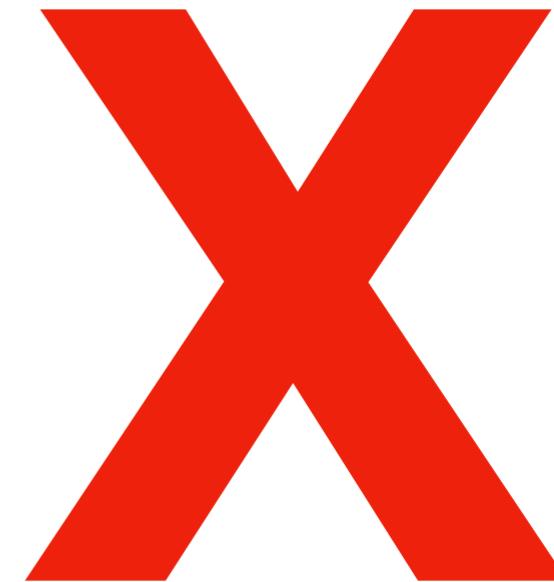
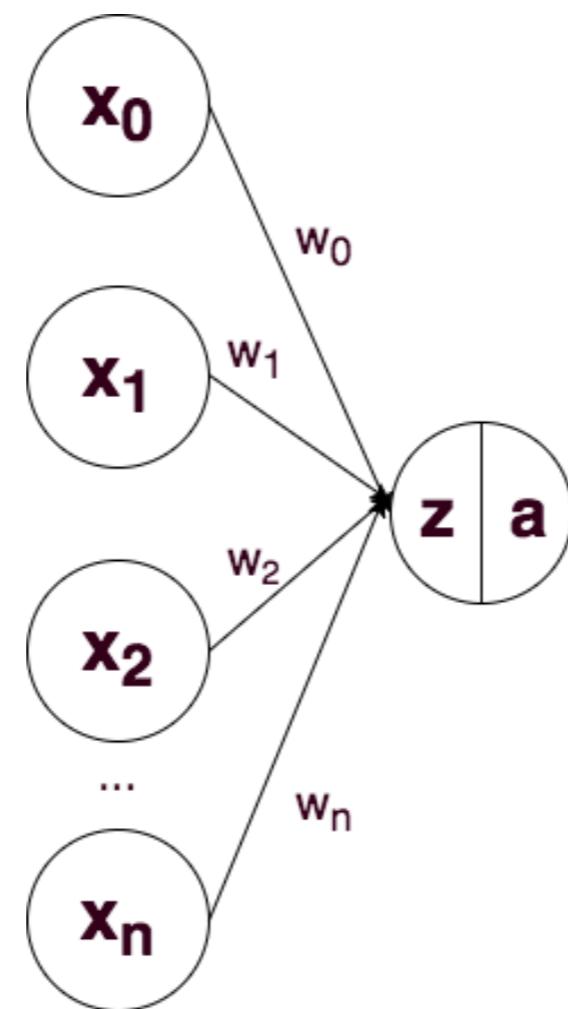
**Nie możemy wyświetlić więcej niż
3 wymiarów na wykresie.**

Teoria

Gdybyśmy chcieli przewidywać przy pomocy tylko **czterech lub więcej cech**:

- ile godzin student spędził na nauce
- ile godzin student spał
- ilość godzin spędzona na wykładach
- stan zdrowia
- ...
- n

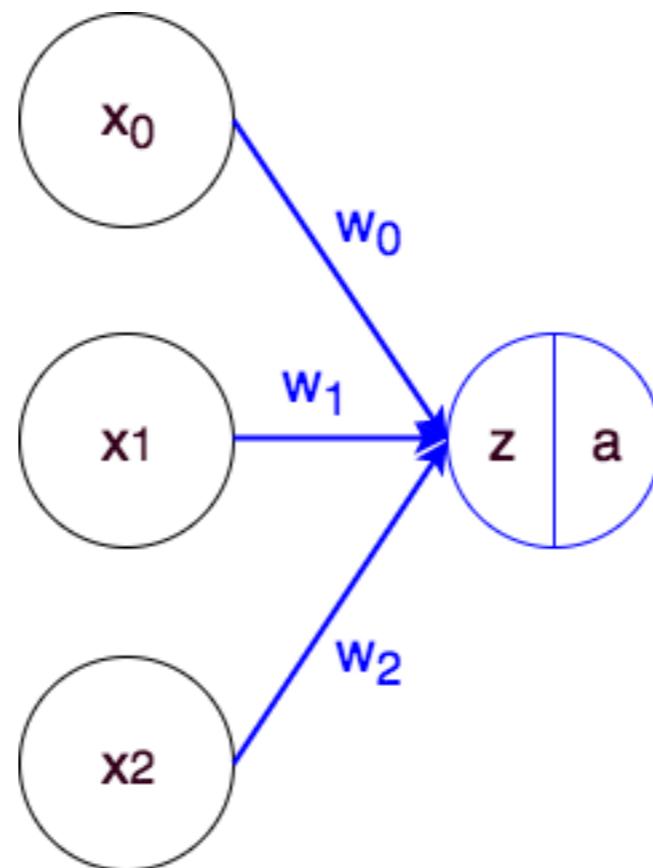
Dane wtedy są **n-wymiarowe**.
Dzielimy je **n-wymiarową hiperprzestrzenią**.



Nie możemy wyświetlić więcej niż 3 wymiarów na wykresie.

Teoria

Czyli **JEDEN** taki zestaw parametrów: w_0 , w_1 , w_2 , b ,



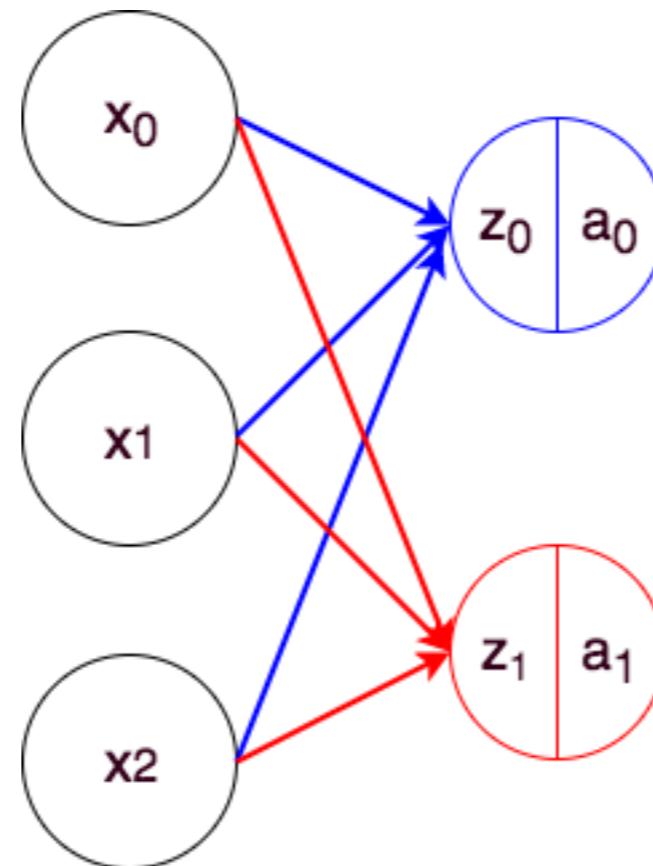
Daje nam odpowiednio **JEDEN** punkt/linie/płaszczyznę/hiperprzestrzeń do rozdzielenia danych, w zależności od ich **wymiarowości n (ilości cech)**.

Teoria

A teraz mamy **DWA** zestawy parametrów:

w₀₀, w₁₀, w₂₀, b₀

w₀₁, w₁₁, w₂₁, b₁



Daje nam odpowiednio **DWA** punkty/linie/płaszczyzny/hiperprzestrzenie do rozdzielenia danych, w zależności od ich **wymiarowości n (ilości cech)**.

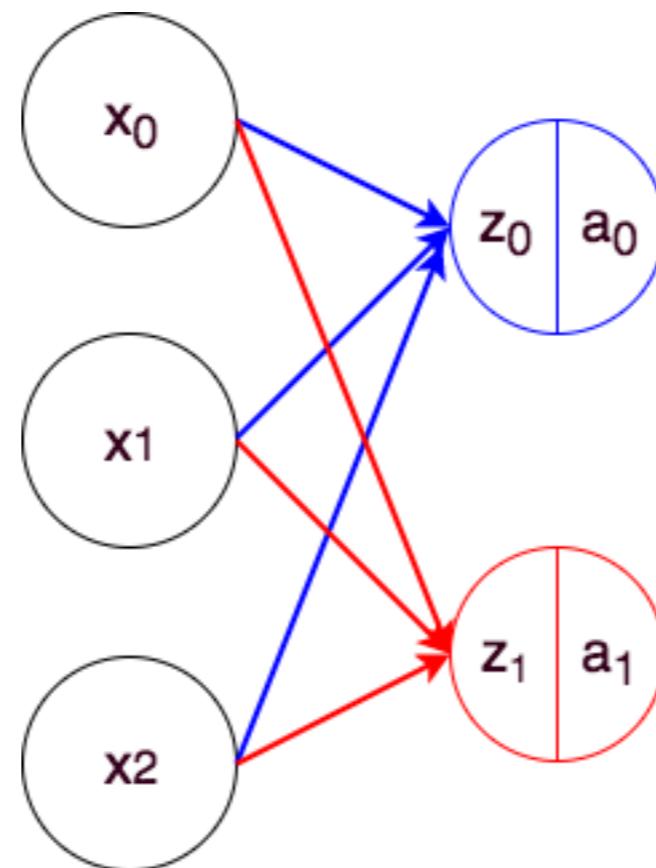
Teoria

A teraz mamy **DWA** zestawy parametrów:

w₀₀, w₁₀, w₂₀, b₀

w₀₁, w₁₁, w₂₁, b₁

**Ale co jest na wyjściu
z naszego modelu w
tej chwili?**



Daje nam odpowiednio **DWA** punkty/linie/płaszczyzny/hiperprzestrzenie do rozdzielenia danych, w zależności od ich **wymiarowości n (ilości cech)**.

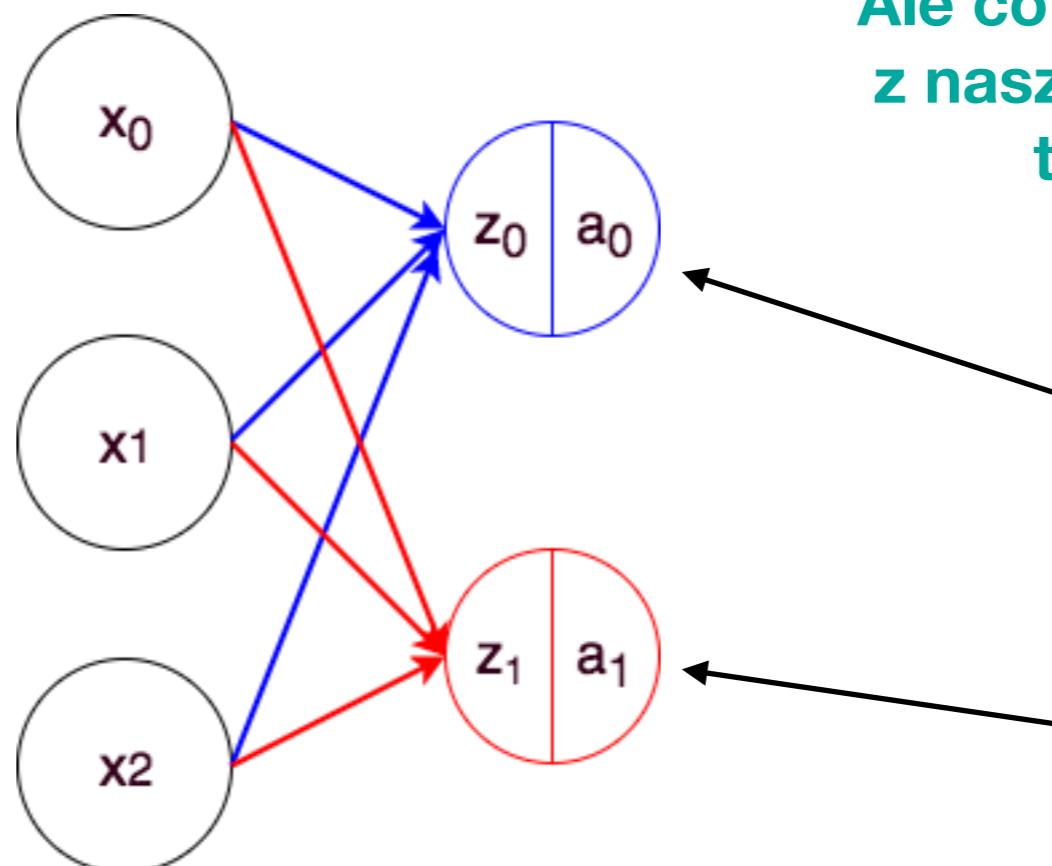
Teoria

A teraz mamy **DWA** zestawy parametrów:

w₀₀, w₁₀, w₂₀, b₀

w₀₁, w₁₁, w₂₁, b₁

Ale co jest na wyjściu
z naszego modelu w
tej chwili?



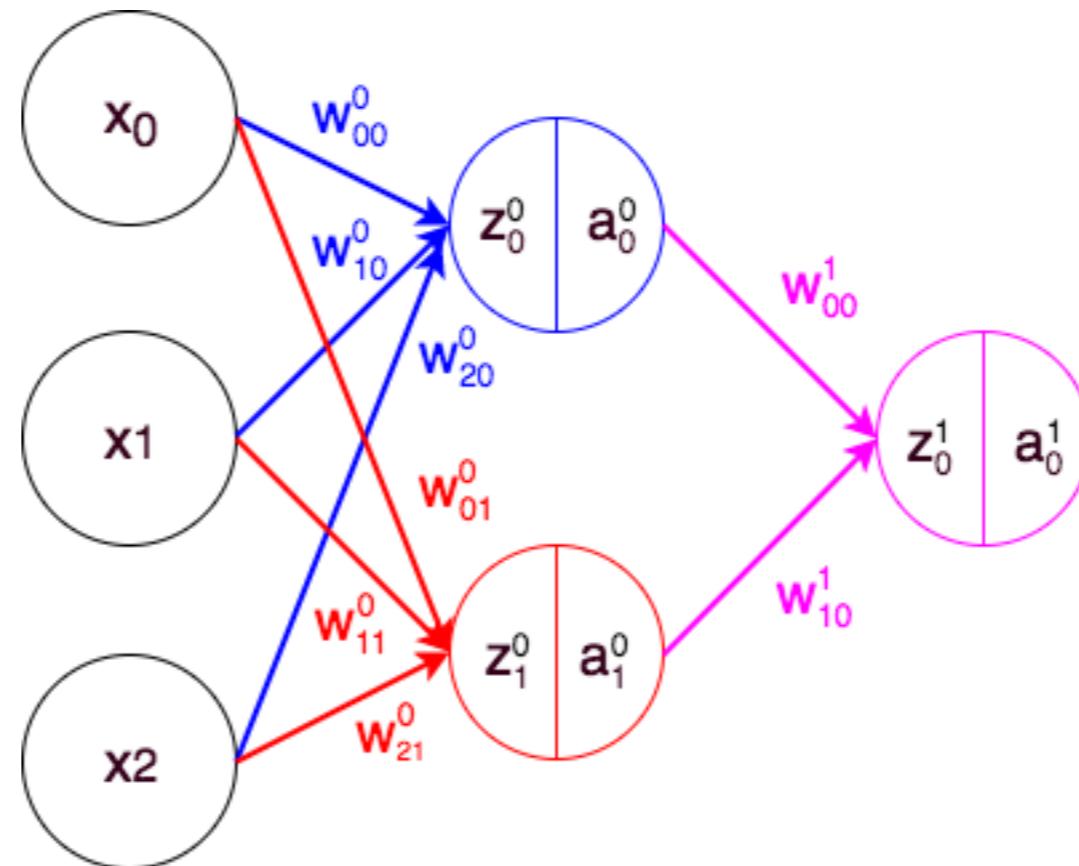
Jedna opinia
o wyniku

Druga opinia
o wyniku

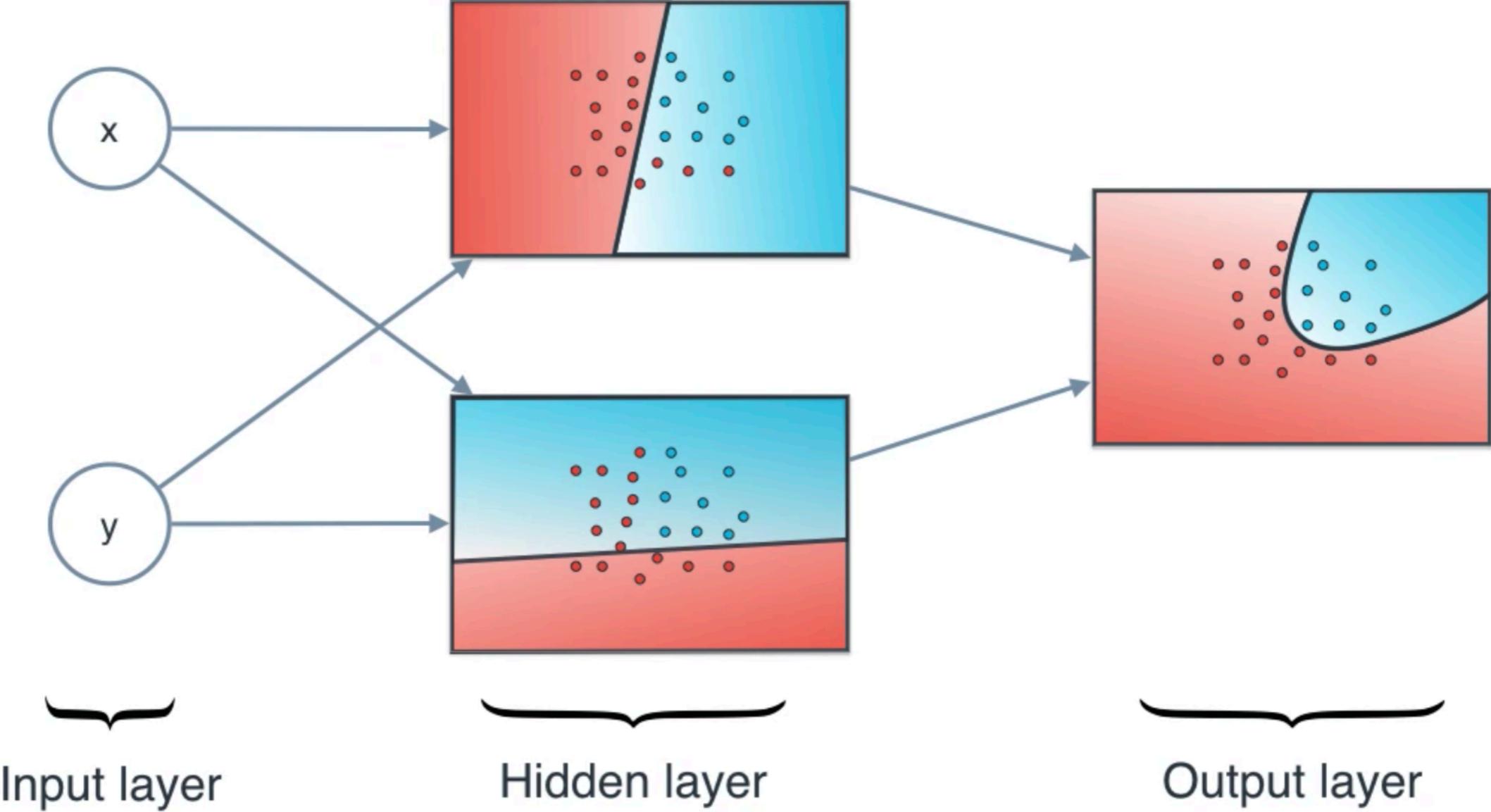
Daje nam odpowiednio **DWA** punkty/linie/płaszczyzny/hiperprzestrzenie
do rozdzielenia danych, w zależności od ich **wymiarowości n (ilości cech)**.

Teoria

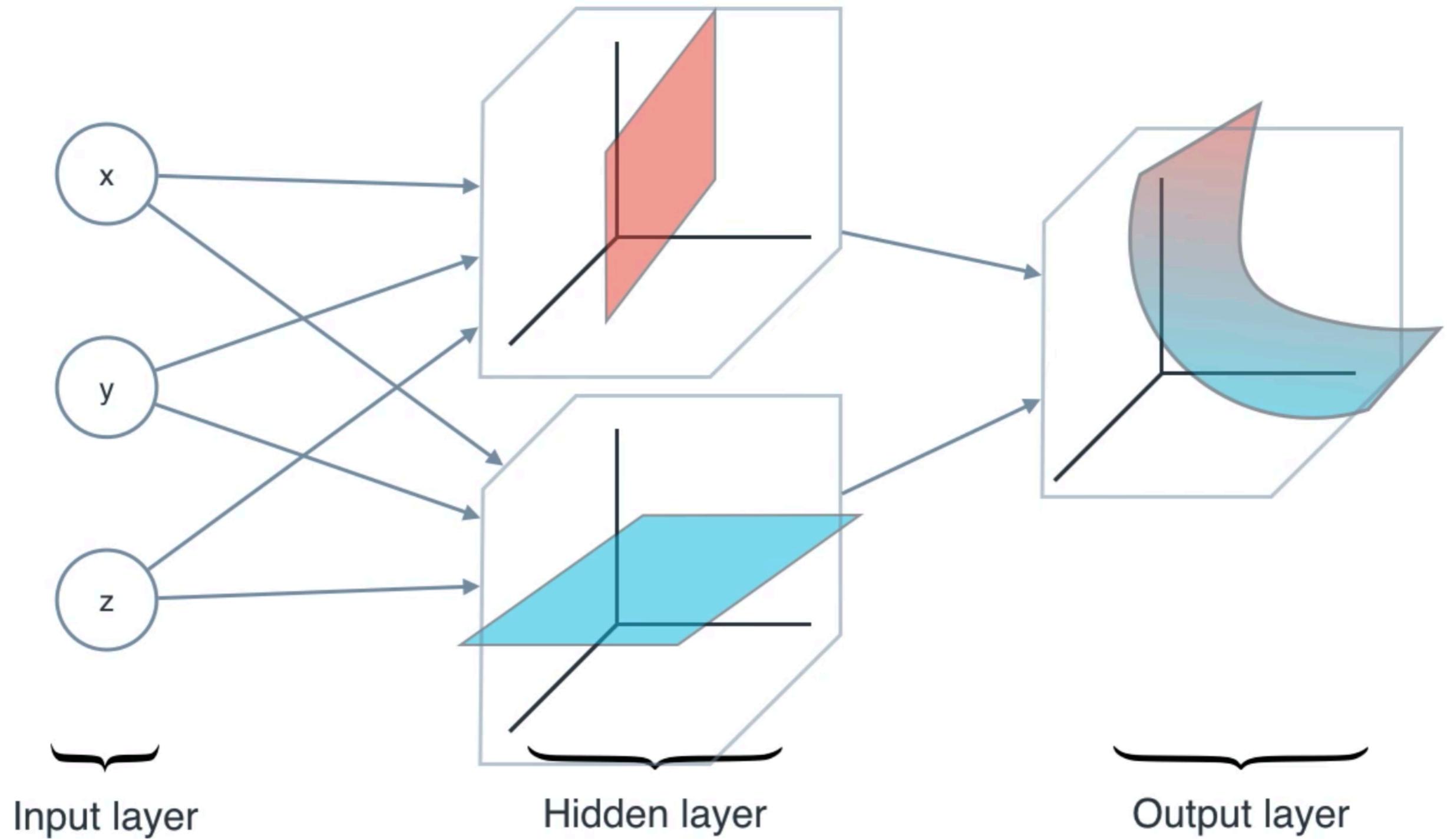
A teraz mamy **TRZY**
zestawy parametrów:
 $w_{00}, w_{10}, w_{20}, b_0$
 $w_{01}, w_{11}, w_{21}, b_1$
 w_{00}, w_{10}, b_0



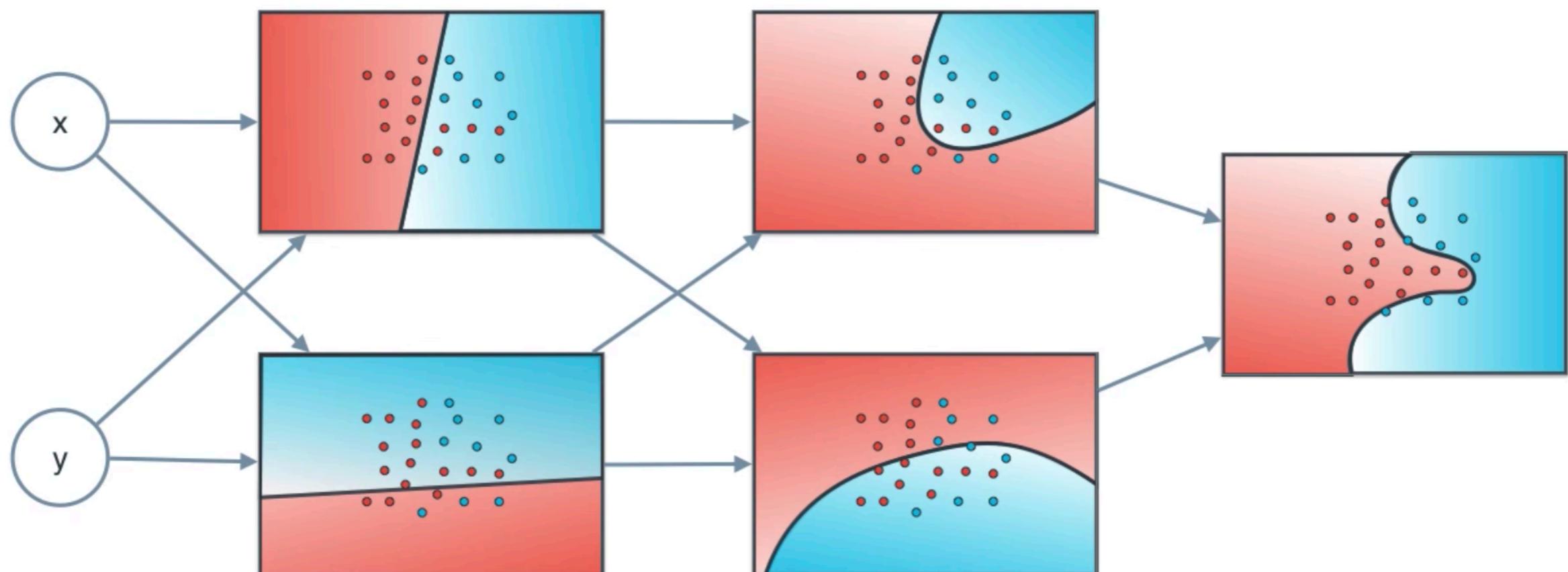
Teoria



Teoria

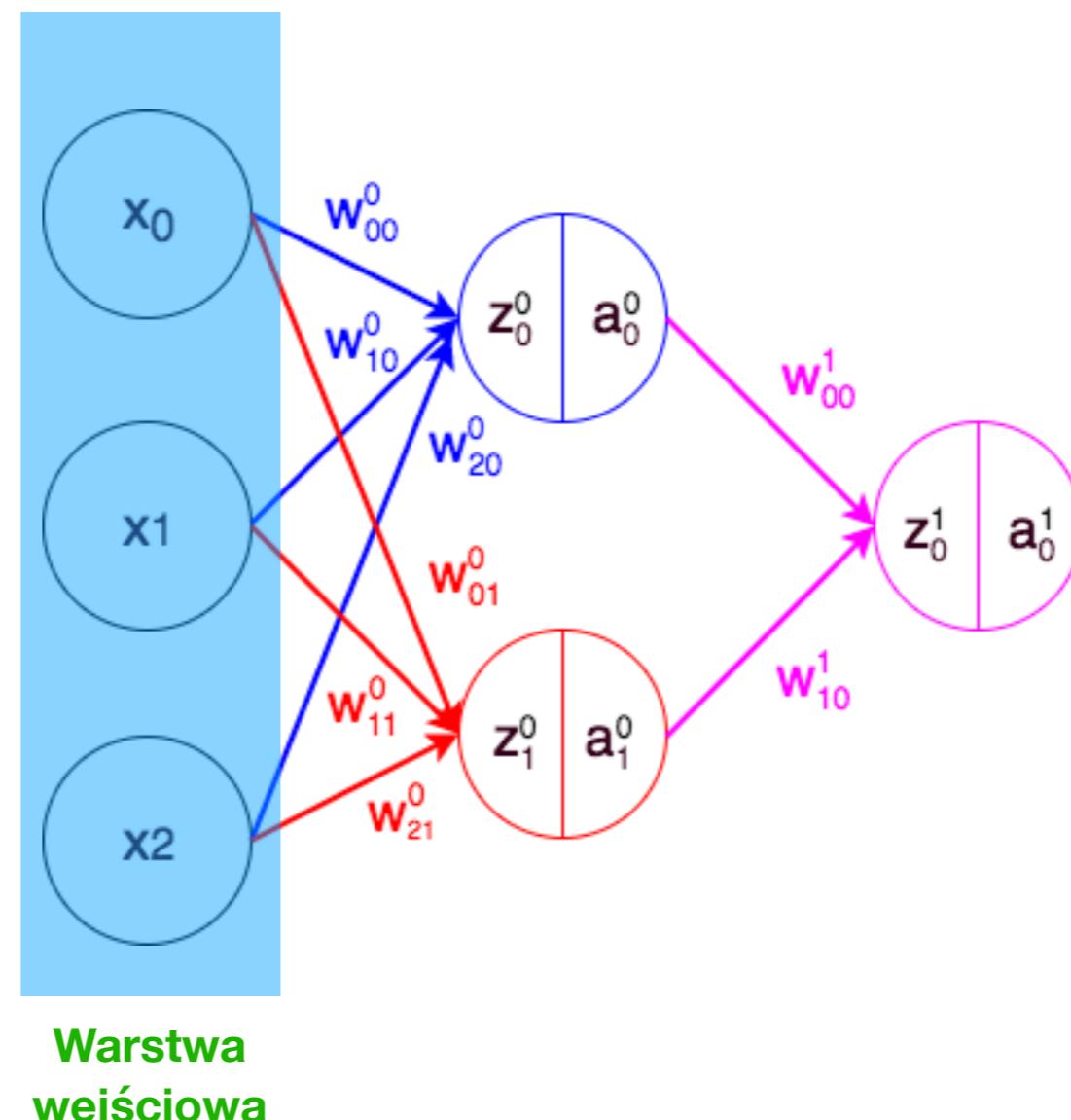


Teoria



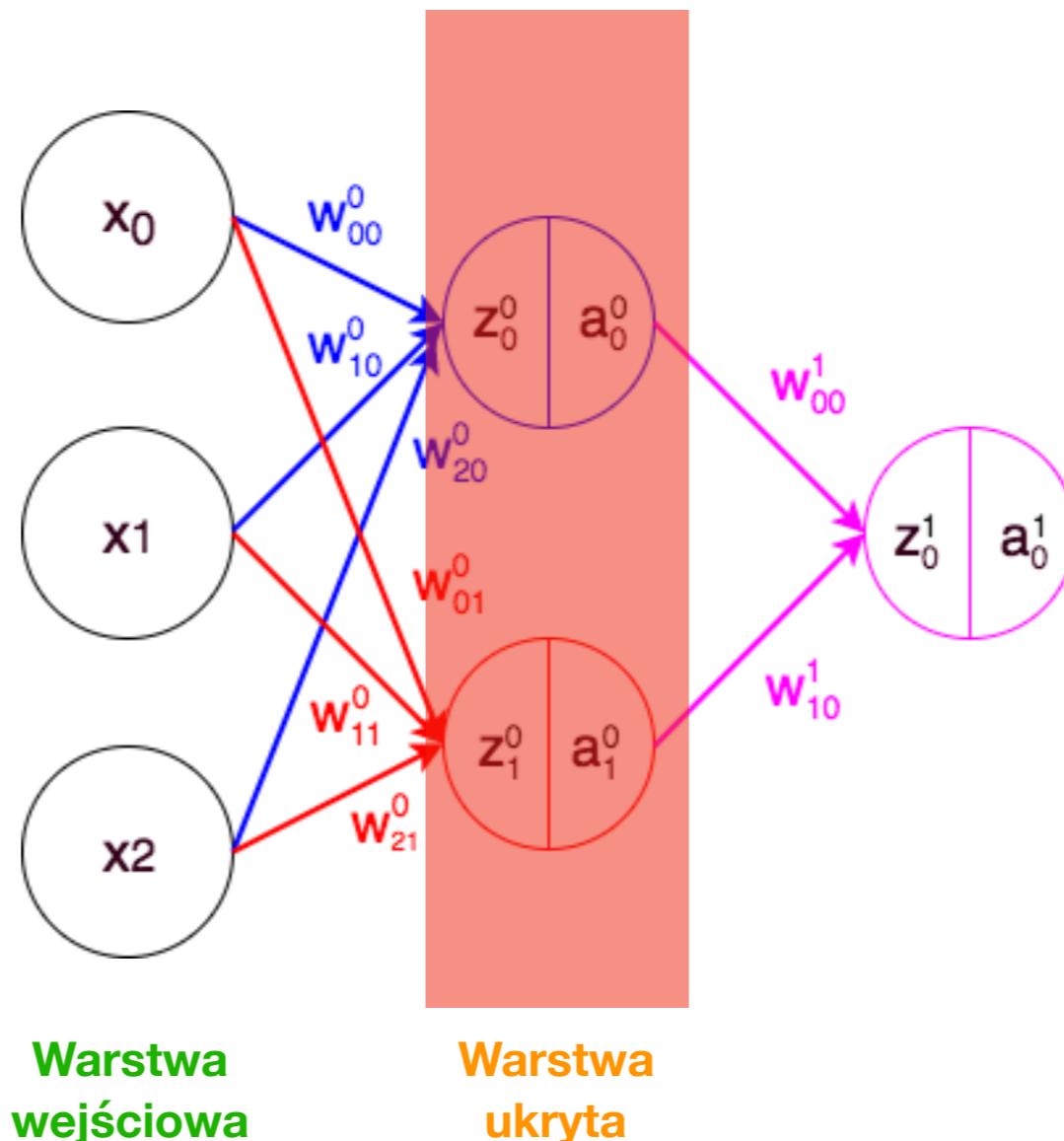
Sieć Neuronowa

A teraz mamy **TRZY**
zestawy parametrów:
w₀₀, w₁₀, w₂₀, b₀
w₀₁, w₁₁, w₂₁, b₁
w₀₀, w₁₀, b₀



Sieć Neuronowa

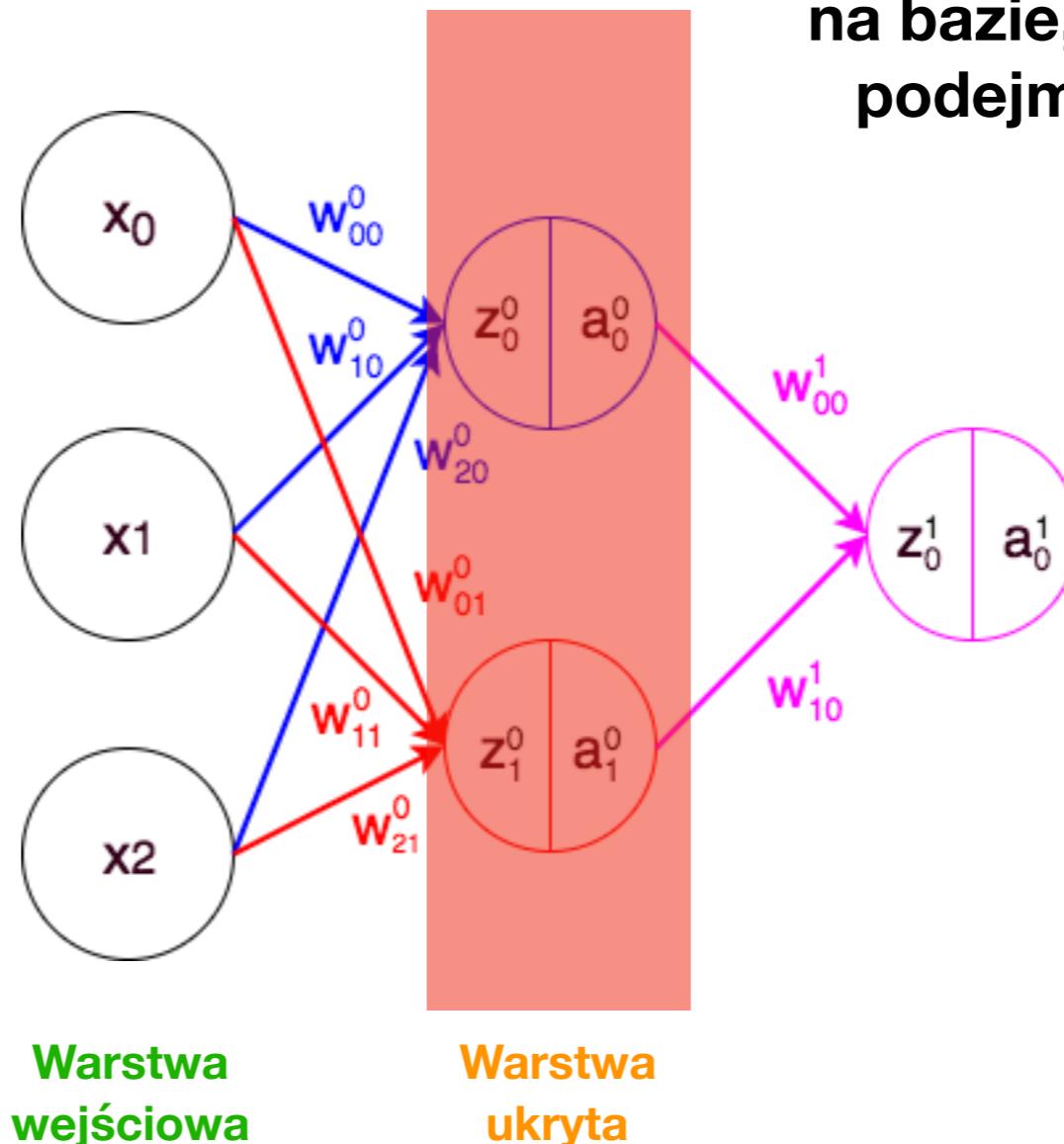
A teraz mamy **TRZY**
zestawy parametrów:
w₀₀, w₁₀, w₂₀, b₀
w₀₁, w₁₁, w₂₁, b₁
w₀₀, w₁₀, b₀



Sieć Neuronowa

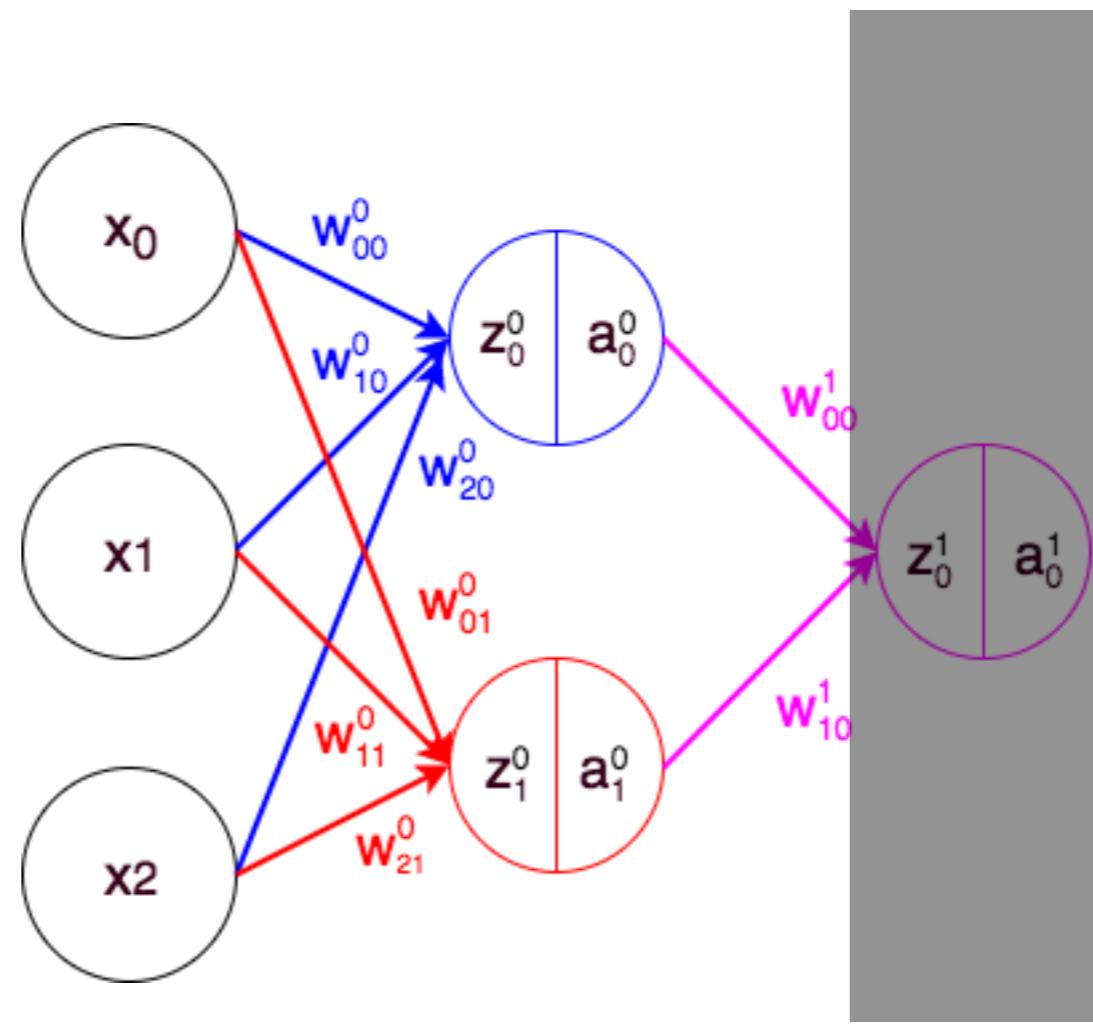
A teraz mamy **TRZY**
zestawy parametrów:
w₀₀, w₁₀, w₂₀, b₀
w₀₁, w₁₁, w₂₁, b₁
w₀₀, w₁₀, b₀

**Nowe cechy
SAMODZIELNIE
stworzone przez model
na bazie, których będzie
podejmował decyzje!**



Sieć Neuronowa

A teraz mamy **TRZY**
zestawy parametrów:
w₀₀, w₁₀, w₂₀, b₀
w₀₁, w₁₁, w₂₁, b₁
w₀₀, w₁₀, b₀



Warstwa
wejściowa

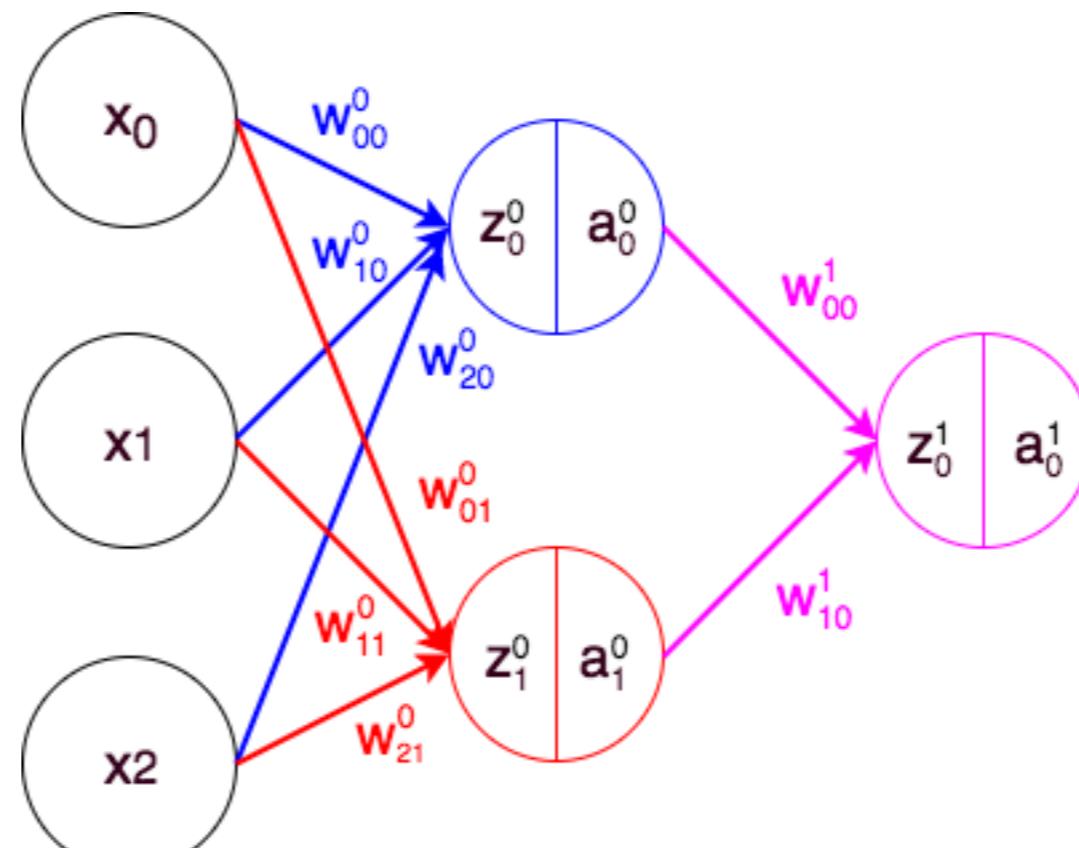
Warstwa
ukryta

Warstwa
wyjściowa

Sieć Neuronowa

A teraz mamy **TRZY**
zestawy parametrów:
 $w_{00}, w_{10}, w_{20}, b_0$
 $w_{01}, w_{11}, w_{21}, b_1$
 w_{00}, w_{10}, b_0

Głębokość sieci = ilość warstw
Ilość warstw = ilość macierzy wag
Ilość warstw = 2

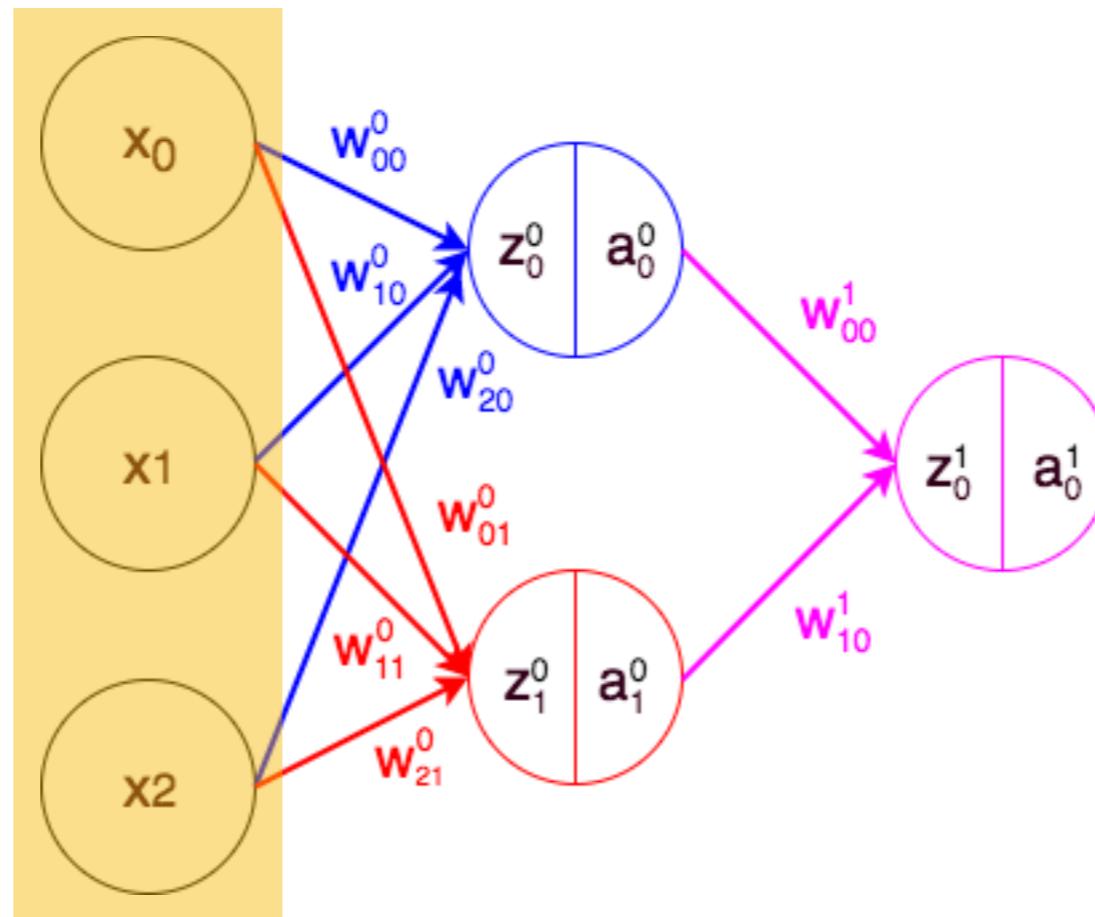


**Warstwa
wejściowa**

**Warstwa
ukryta**

**Warstwa
wyjściowa**

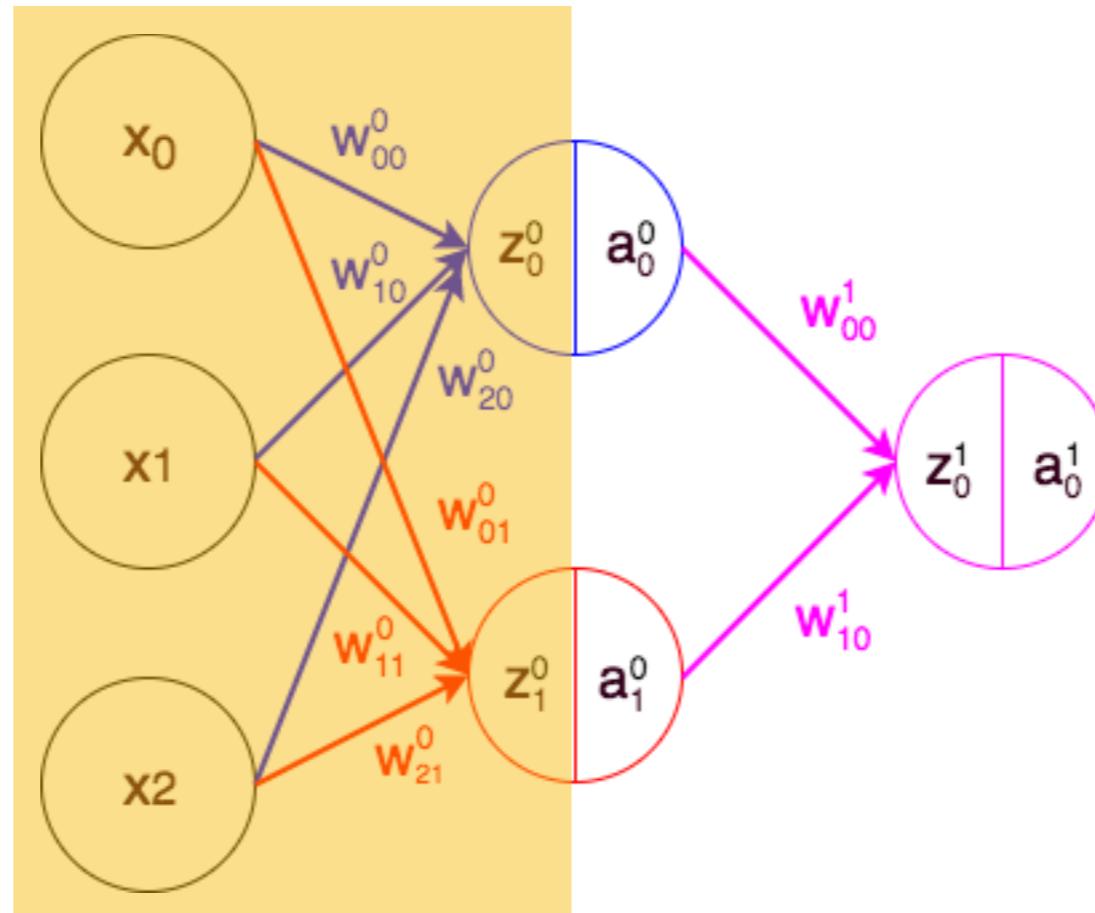
Sieć Neuronowa - matematycznie



$$\begin{bmatrix} x_0 & x_1 & x_2 \end{bmatrix}$$

(1,3)

Sieć Neuronowa - matematycznie



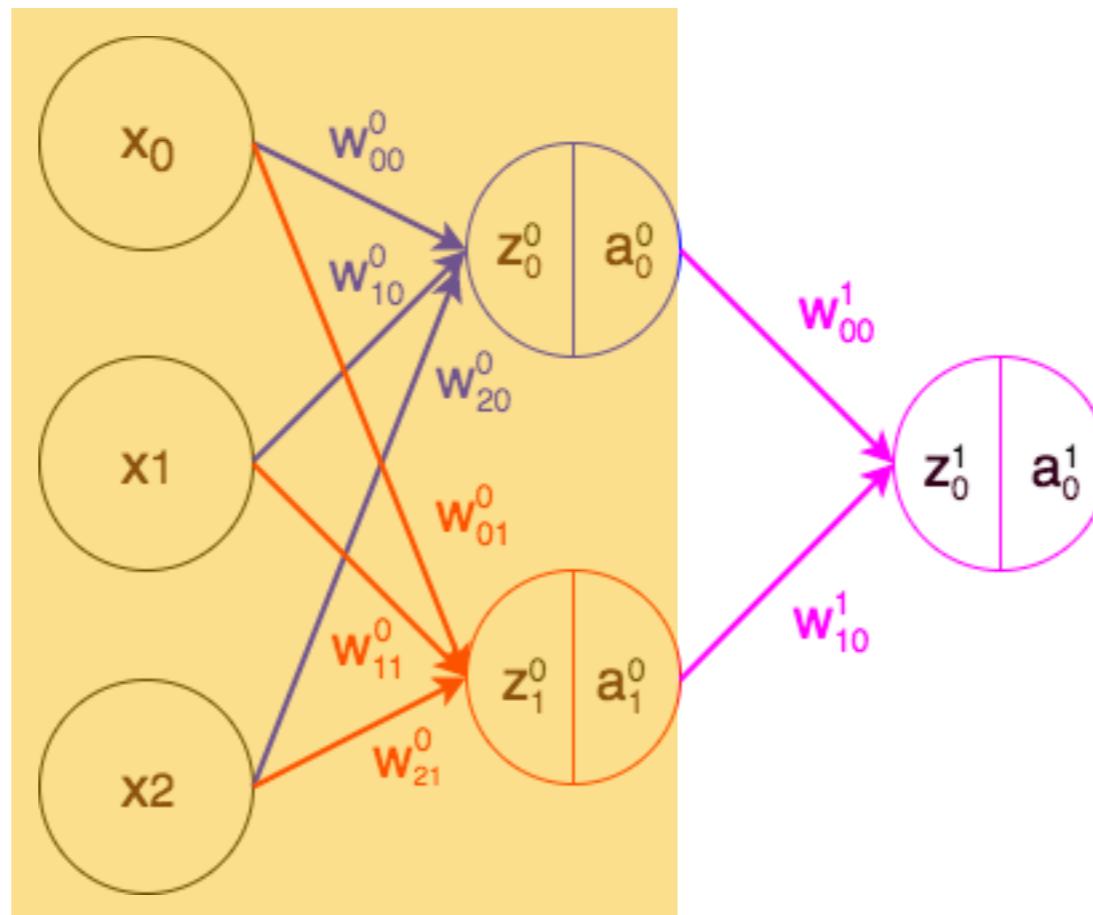
$$\begin{bmatrix} x_0 & x_1 & x_2 \end{bmatrix} \begin{bmatrix} w_{00}^0 & w_{01}^0 \\ w_{10}^0 & w_{11}^0 \\ w_{20}^0 & w_{21}^0 \end{bmatrix} = \begin{bmatrix} w_{00}^0 x_0 + w_{10}^0 x_1 + w_{20}^0 x_2 + b_0^0 & w_{01}^0 x_0 + w_{11}^0 x_1 + w_{21}^0 x_2 + b_1^0 \end{bmatrix}$$

(1,3)

(3,2)

(1, 2)

Sieć Neuronowa - matematycznie



funkcja
aktywacji

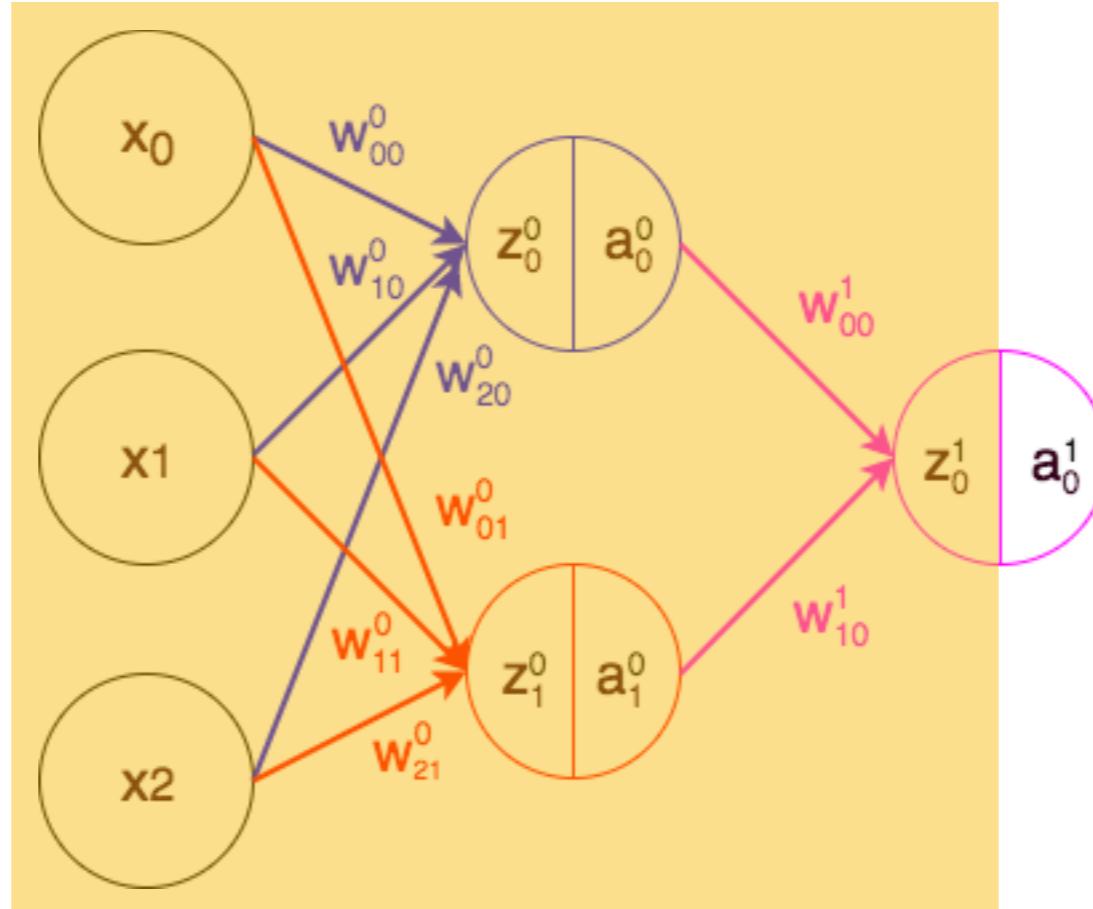
$$\begin{bmatrix} x_0 & x_1 & x_2 \end{bmatrix} \begin{bmatrix} w_{00}^0 & w_{01}^0 \\ w_{10}^0 & w_{11}^0 \\ w_{20}^0 & w_{21}^0 \end{bmatrix} = \begin{bmatrix} a_0(w_{00}^0 x_0 + w_{10}^0 x_1 + w_{20}^0 x_2 + b_0^0) \\ a_0(w_{01}^0 x_0 + w_{11}^0 x_1 + w_{21}^0 x_2 + b_1^0) \end{bmatrix}$$

(1,3)

(3,2)

(1, 2)

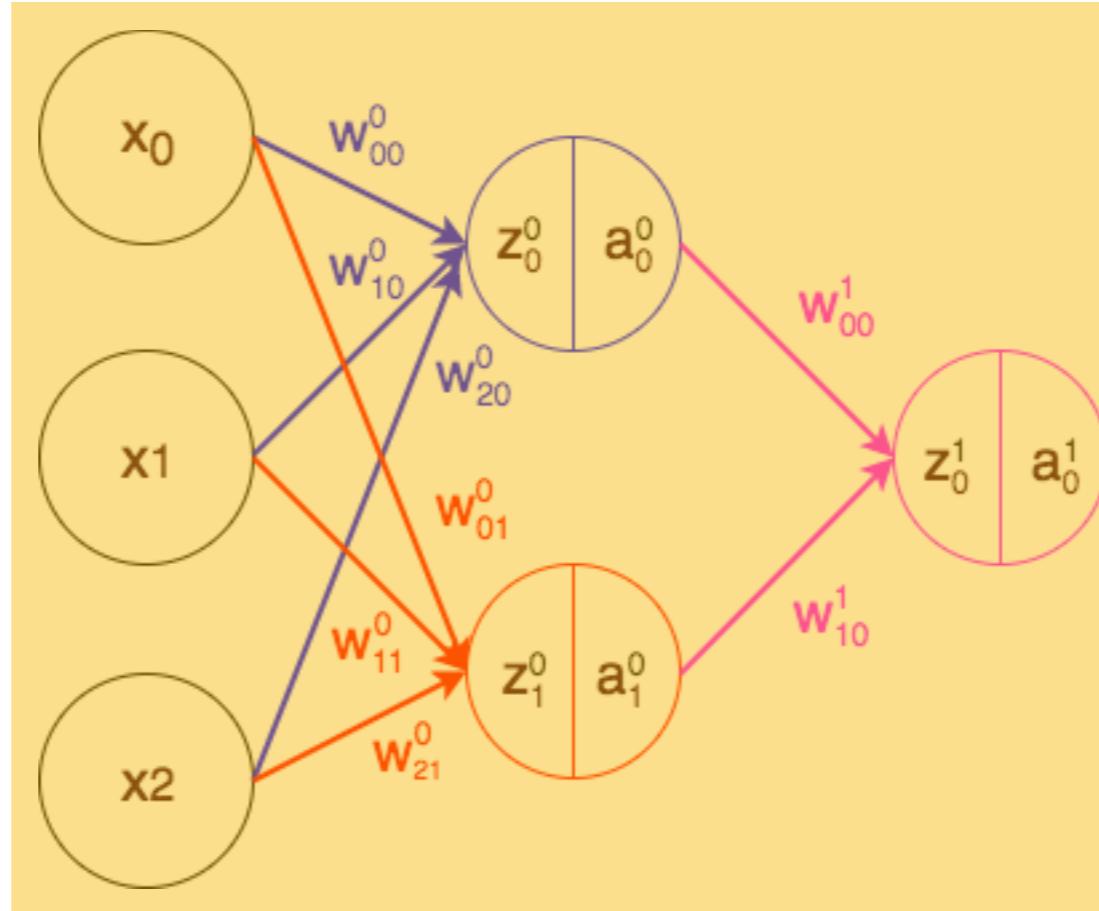
Sieć Neuronowa - matematycznie



$$\begin{bmatrix} x_0 & x_1 & x_2 \end{bmatrix} \begin{bmatrix} w_{00}^0 & w_{01}^0 \\ w_{10}^0 & w_{11}^0 \\ w_{20}^0 & w_{21}^0 \end{bmatrix} \begin{bmatrix} w_{00}^1 \\ w_{10}^1 \end{bmatrix} = [w_{00}^1 a(w_{00}^0 x_0 + w_{10}^0 x_1 + w_{20}^0 x_2 + b_0^0) + w_{10}^1 a(w_{01}^0 x_0 + w_{11}^0 x_1 + w_{21}^0 x_2 + b_1^0) + b_0^1]$$

(1,3) (3,2) (2,1) (1,1)

Sieć Neuronowa - matematycznie

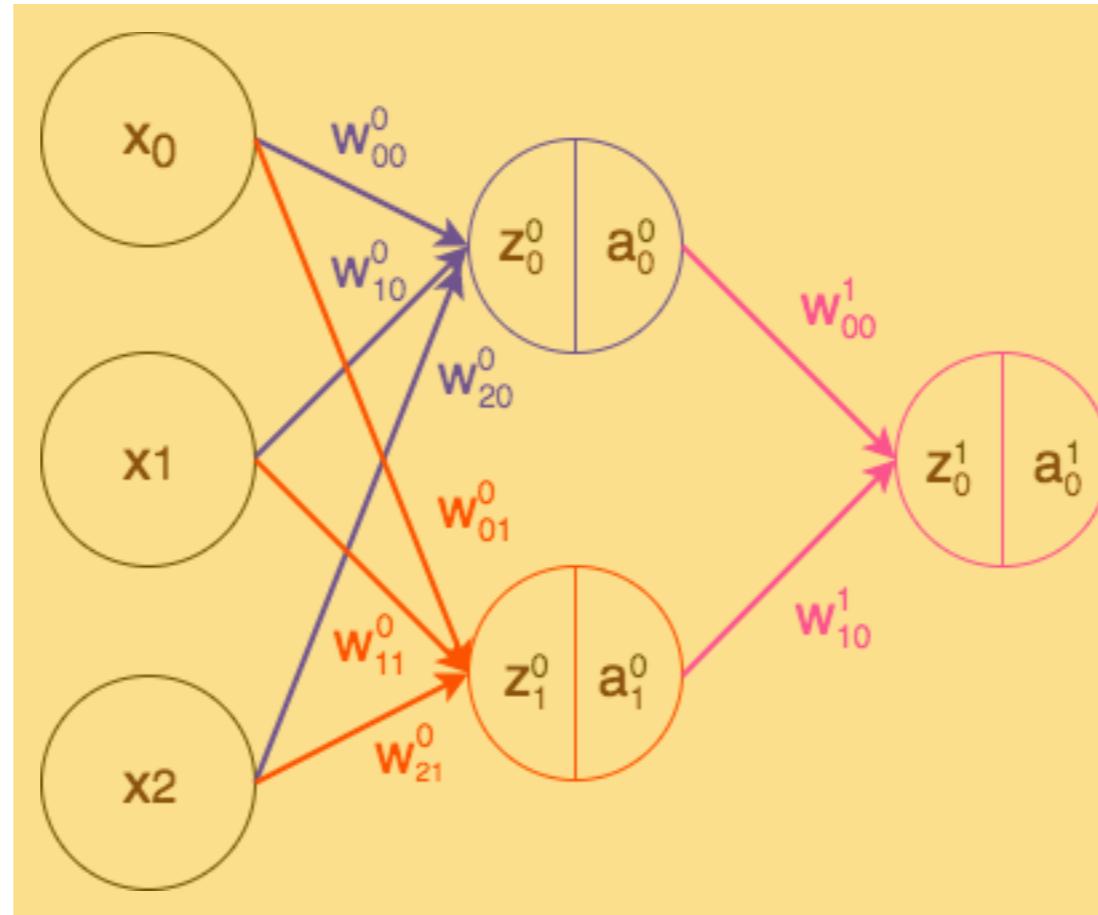


$$\begin{bmatrix} x_0 & x_1 & x_2 \end{bmatrix} \begin{bmatrix} w_{00}^0 & w_{01}^0 \\ w_{10}^0 & w_{11}^0 \\ w_{20}^0 & w_{21}^0 \end{bmatrix} \begin{bmatrix} w_{00}^1 \\ w_{10}^1 \end{bmatrix} = [a_1(w_{00}^1 a_0(w_{00}^0 x_0 + w_{10}^0 x_1 + w_{20}^0 x_2 + b_0^0) + w_{10}^1 a_0(w_{01}^0 x_0 + w_{11}^0 x_1 + w_{21}^0 x_2 + b_1^0) + b_0^1)]$$

(1,3) (3,2) (2,1) (1,1)

Sieć Neuronowa - matematycznie

Wszystko sprowadza się do wymnożenia macierzy.



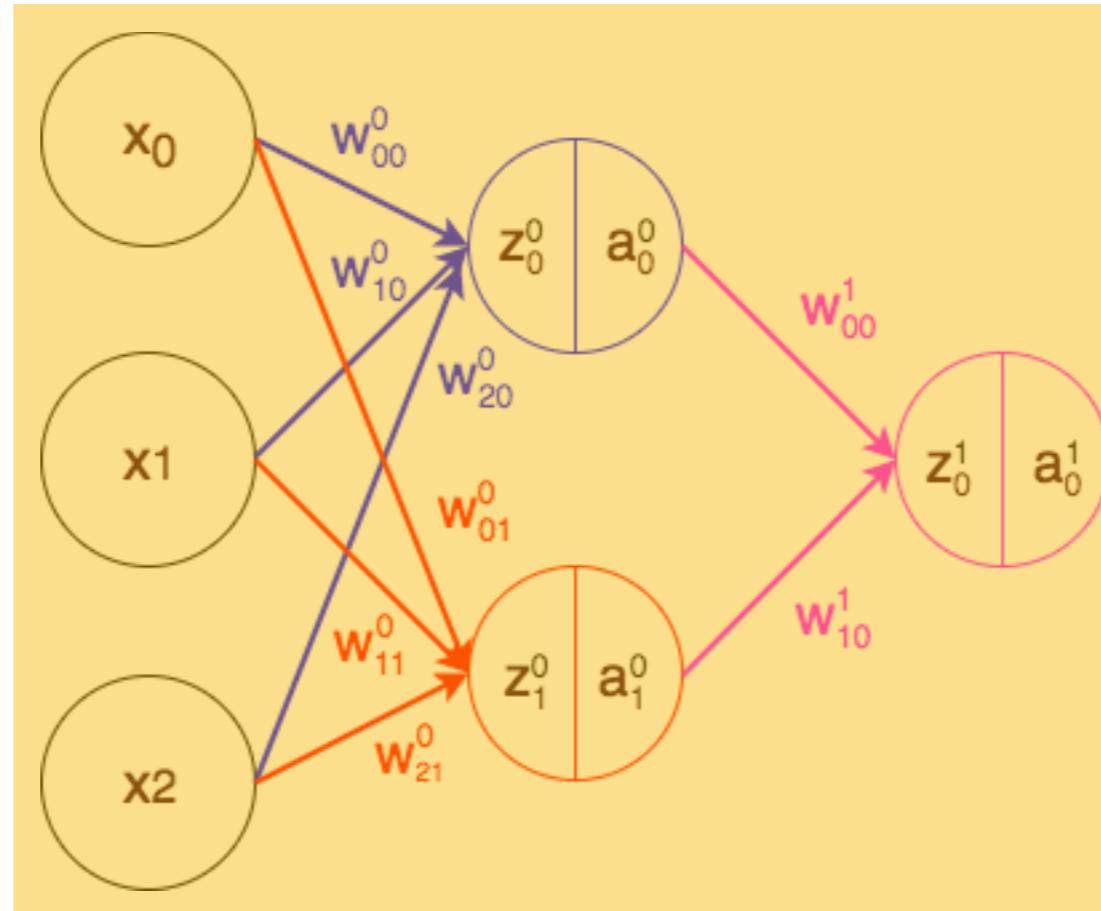
$$\begin{bmatrix} x_0 & x_1 & x_2 \end{bmatrix} \begin{bmatrix} w_{00}^0 & w_{01}^0 \\ w_{10}^0 & w_{11}^0 \\ w_{20}^0 & w_{21}^0 \end{bmatrix} \begin{bmatrix} w_{00}^1 \\ w_{10}^1 \end{bmatrix} = [a_1(w_{00}^1 a_0(w_{00}^0 x_0 + w_{10}^0 x_1 + w_{20}^0 x_2 + b_0^0) + w_{10}^1 a_0(w_{01}^0 x_0 + w_{11}^0 x_1 + w_{21}^0 x_2 + b_1^0) + b_1^1)]$$

(1,3) (3,2) (2,1) (1,1)

Sieć Neuronowa - matematycznie

Wszystko sprowadza się do wymnożenia macierzy.

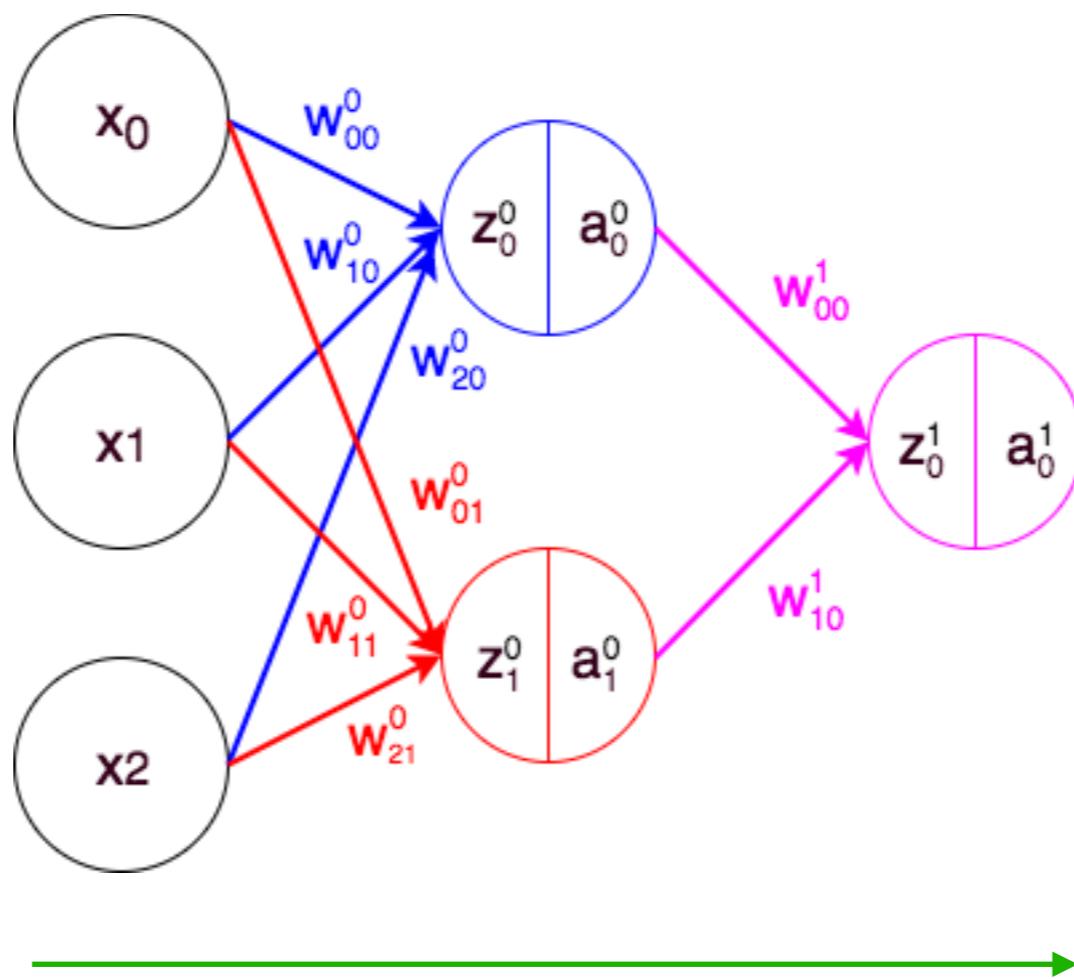
a1, a0 oznacza, że każda warstwa może mieć różne funkcje aktywacji



$$\begin{bmatrix} x_0 & x_1 & x_2 \end{bmatrix} \begin{bmatrix} w_{00}^0 & w_{01}^0 \\ w_{10}^0 & w_{11}^0 \\ w_{20}^0 & w_{21}^0 \end{bmatrix} \begin{bmatrix} w_{00}^1 \\ w_{10}^1 \end{bmatrix} = \begin{bmatrix} a_1(w_{00}^1 a_0(w_{00}^0 x_0 + w_{10}^0 x_1 + w_{20}^0 x_2 + b_0^0) + w_{10}^1 a_0(w_{01}^0 x_0 + w_{11}^0 x_1 + w_{21}^0 x_2 + b_1^0) + b_1^1) \end{bmatrix}$$

(1,3) (3,2) (2,1) (1,1)

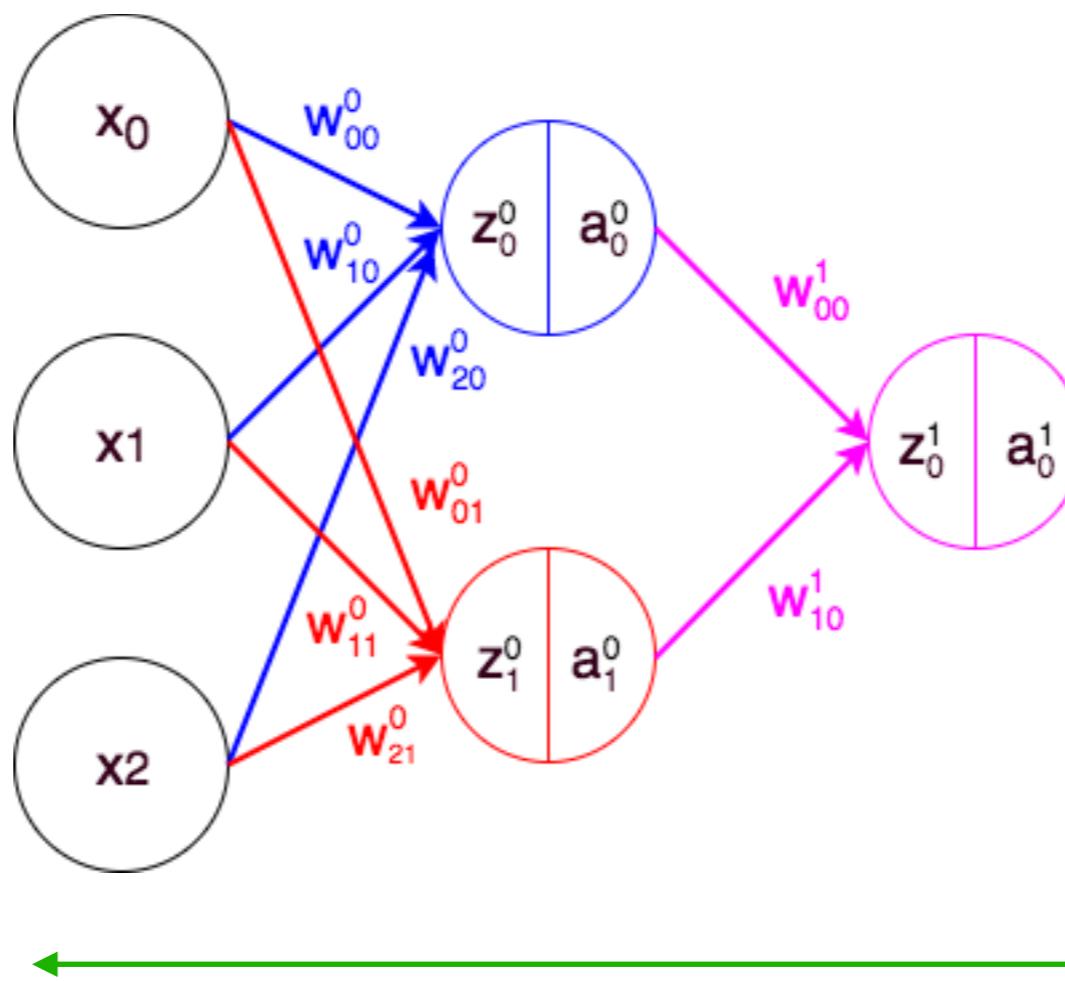
Sieć Neuronowa - Feedforward



Feedforward - dokonanie predykcji

Jest to jednorazowe przejście danych przez sieć od początku do końca.

Sieć Neuronowa - Backpropagation



Backpropagation - nastawienie wag modelu

Proces porównania predykcji z oczekiwania wartością i aktualizacja wag przy pomocy Algorytmu Spadku Gradientu.

Dane a Sieci Neuronowe

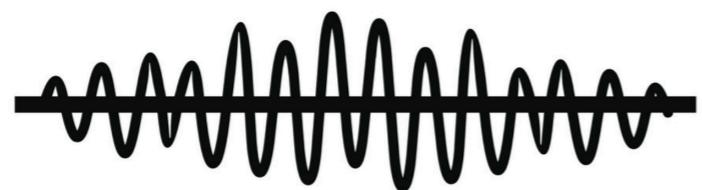
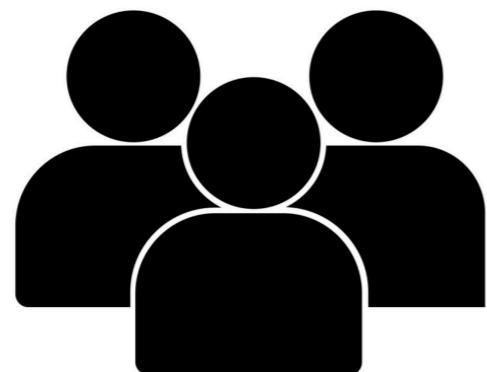
Dane a Sieci Neuronowe



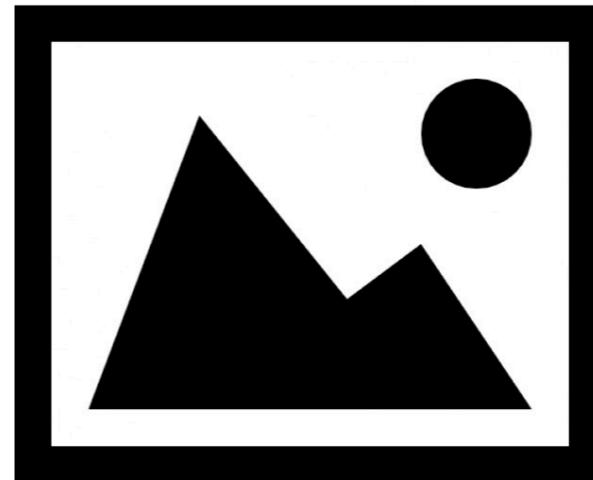
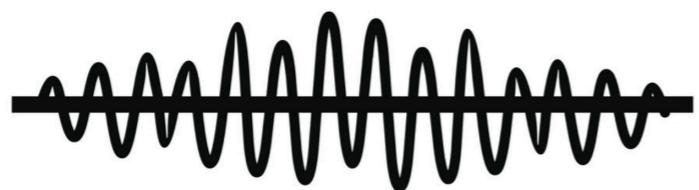
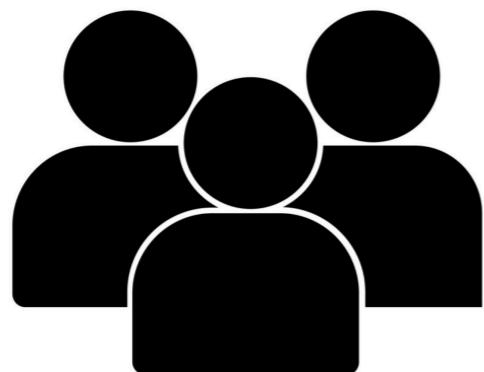
Dane a Sieci Neuronowe



Dane a Sieci Neuronowe



Dane a Sieci Neuronowe



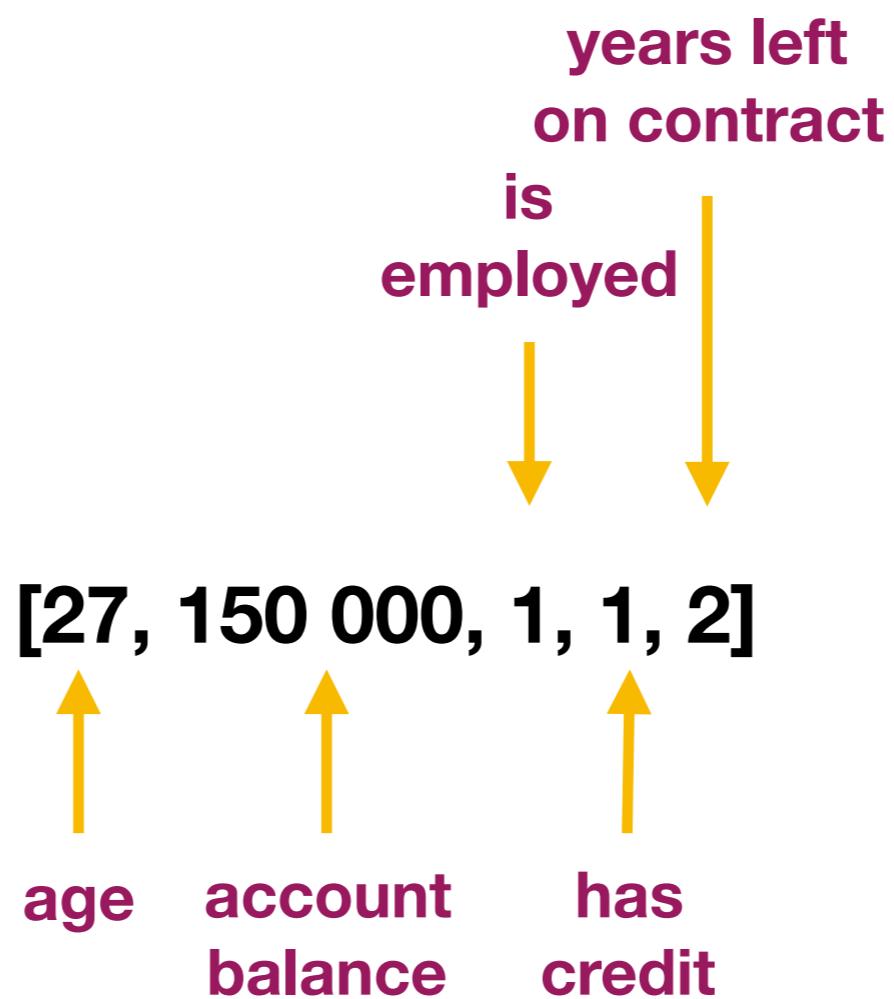
Dane a Sieci Neuronowe

Dane muszą być przedstawione w formie wektora liczb.

[27, 150 000, 1, 1, 2]

Dane a Sieci Neuronowe

Dane muszą być przedstawione w formie wektora liczb.



Dane a Sieci Neuronowe

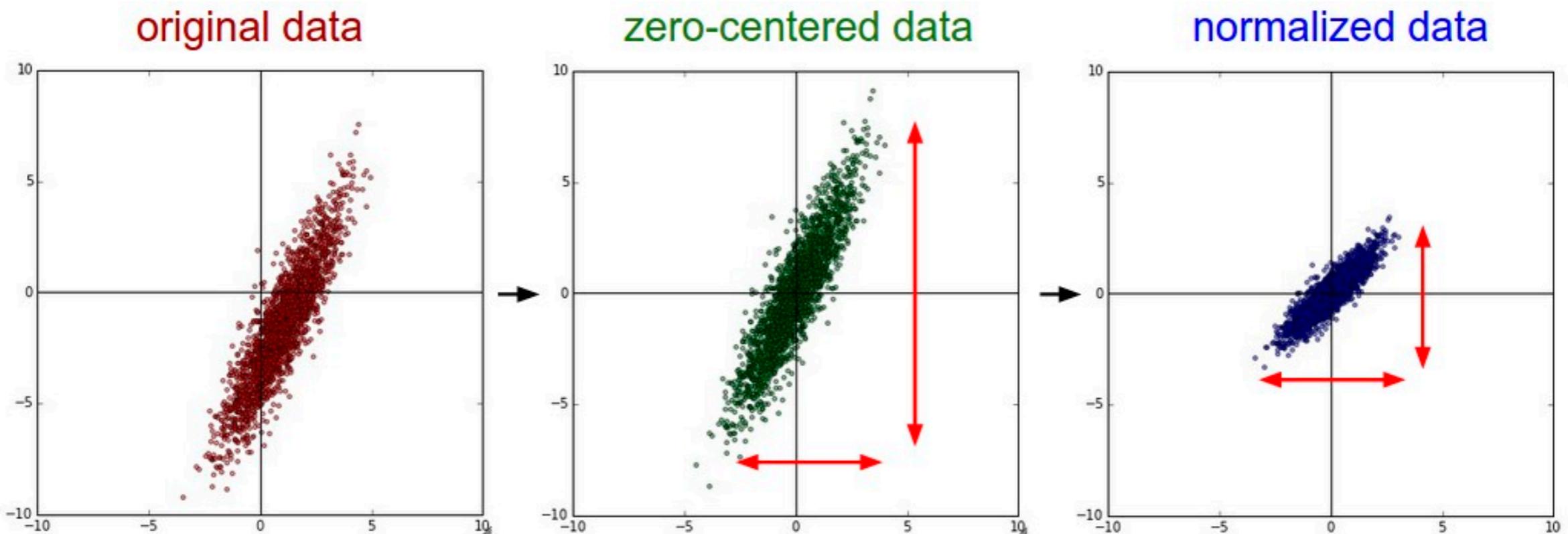
Dane muszą być przedstawione w formie wektora liczb.

Dane powinny być znormalizowane.

Dane a Sieci Neuronowe

Dane muszą być przedstawione w formie wektora liczb.

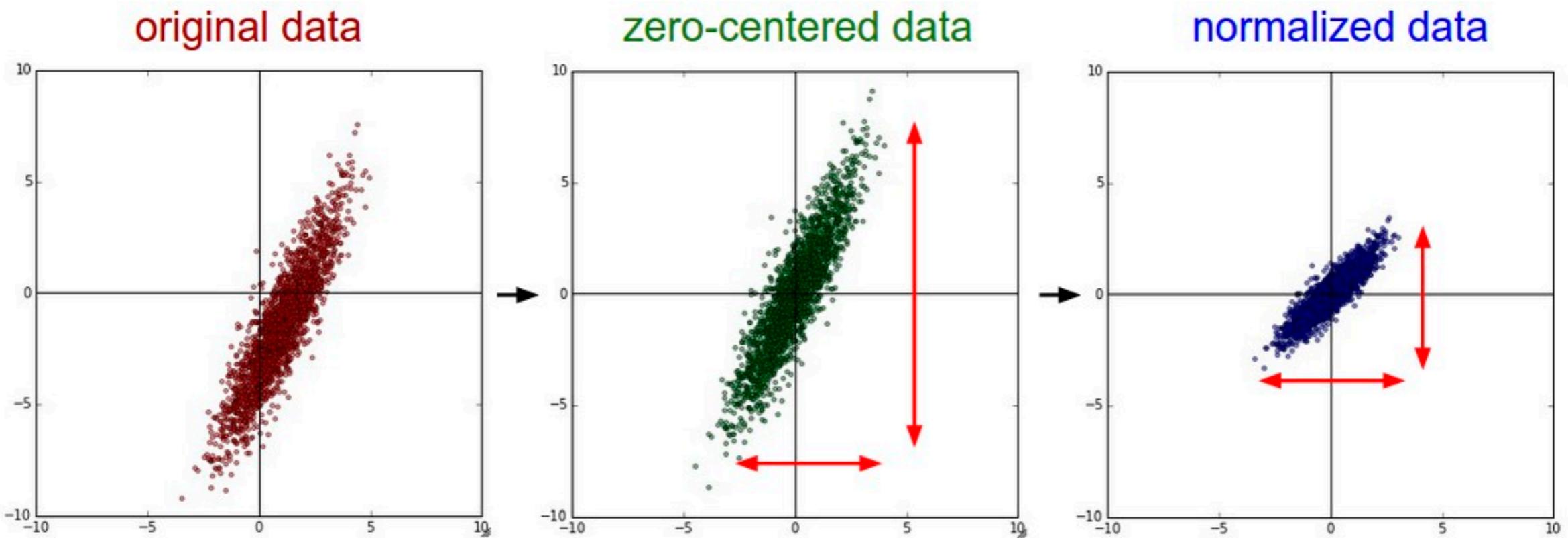
Dane powinny być znormalizowane.



Dane a Sieci Neuronowe

Dane muszą być przedstawione w formie wektora liczb.

Dane powinny być znormalizowane.

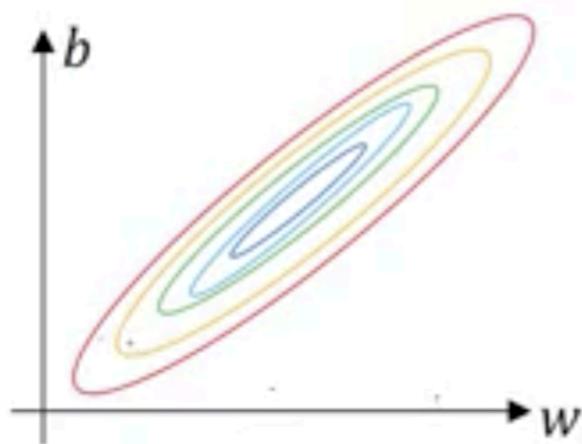
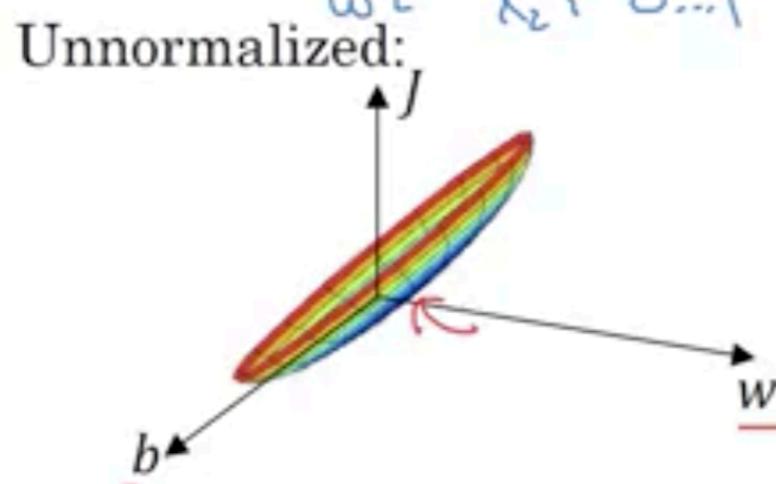


Spadek Gradientu wymaga normalizacji aby zachować stabilność obliczeniową.

Dane a Sieci Neuronowe

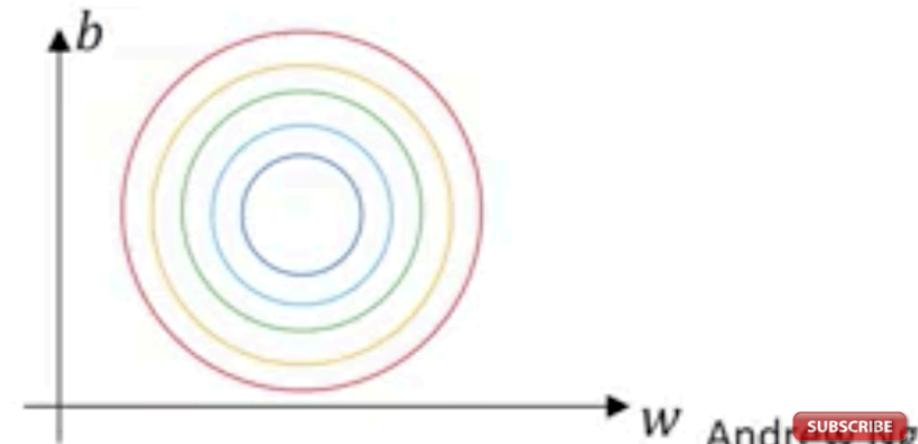
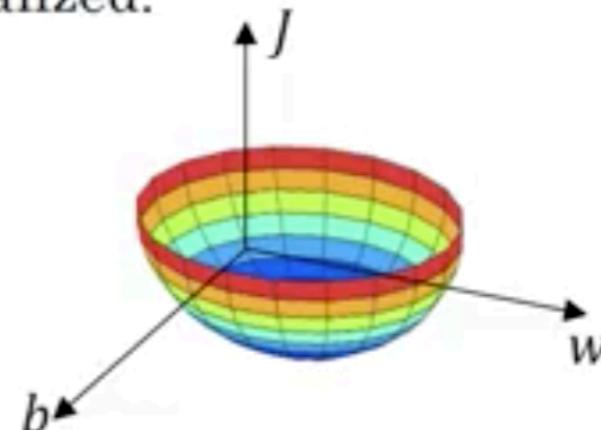
Why normalize inputs?

$$\begin{matrix} \omega_1 & x_1: 1 \dots \infty \\ \omega_2 & x_2: 0 \dots 1 \end{matrix}$$



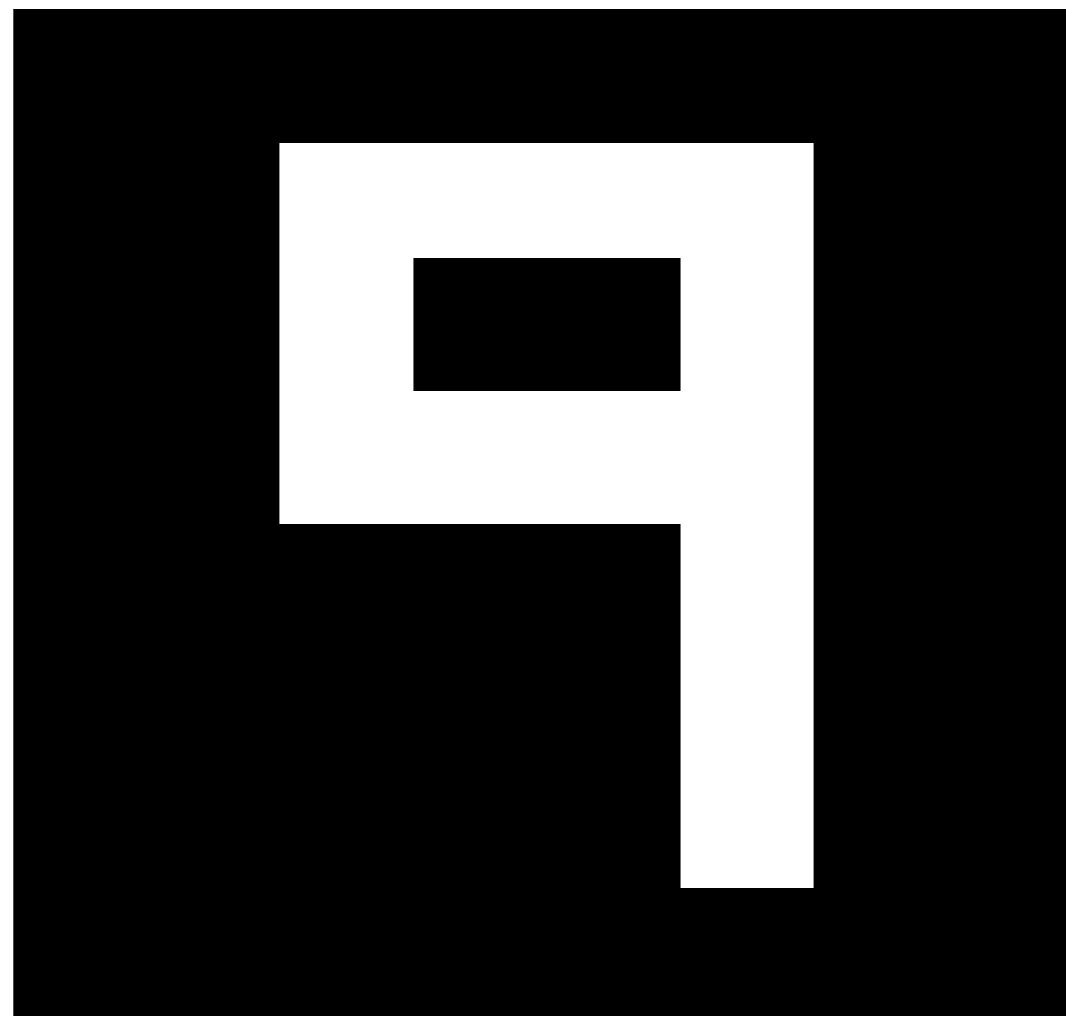
$$J(w, b) = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(\hat{y}^{(i)}, y^{(i)})$$

Normalized:



SUBSCRIBE

Obrazy a Sieci Neuronowe



Obrazy a Sieci Neuronowe

0	0	0	0	0	0	0	0
0	0	255	255	255	255	0	0
0	0	255	0	0	255	0	0
0	0	255	255	255	255	0	0
0	0	0	0	0	255	0	0
0	0	0	0	0	255	0	0
0	0	0	0	0	255	0	0
0	0	0	0	0	0	0	0

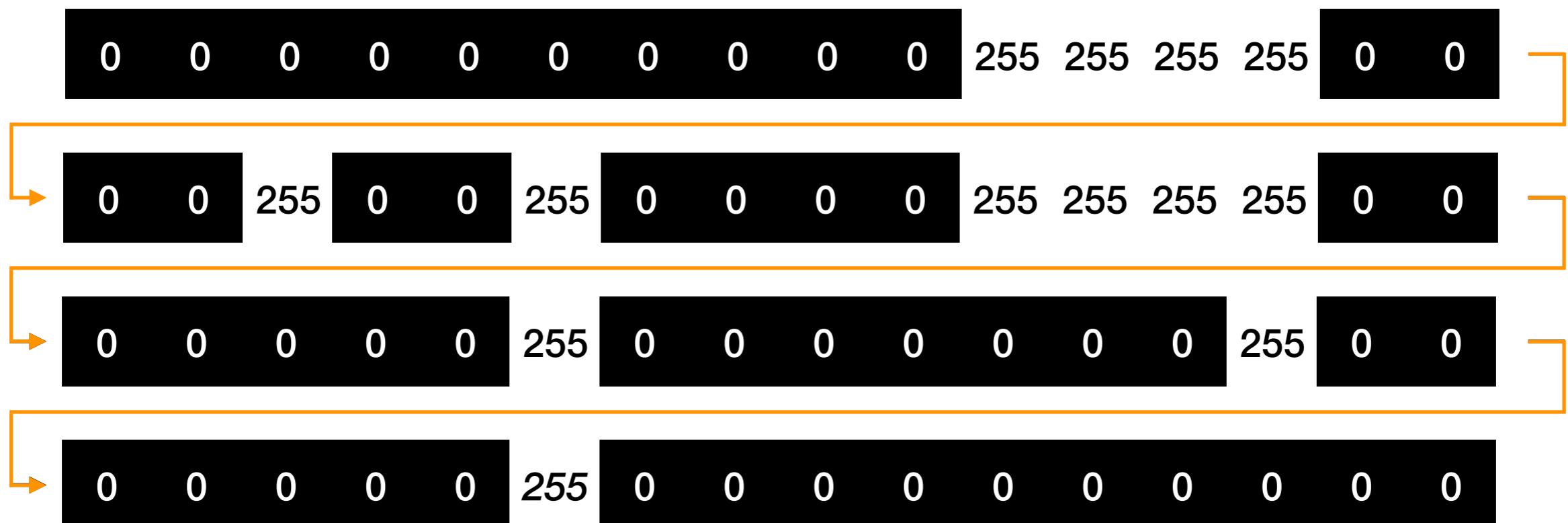
Obrazy a Sieci Neuronowe

8x8x1

0	0	0	0	0	0	0	0
0	0	255	255	255	255	0	0
0	0	255	0	0	255	0	0
0	0	255	255	255	255	0	0
0	0	0	0	0	255	0	0
0	0	0	0	0	255	0	0
0	0	0	0	0	255	0	0
0	0	0	0	0	0	0	0

Obrazy a Sieci Neuronowe

8x8x1 -> wektor (64,)



Obrazy a Sieci Neuronowe

8x8x1 -> wektor (64,)

Obrazy a Sieci Neuronowe

wektor x

pixe_value₀

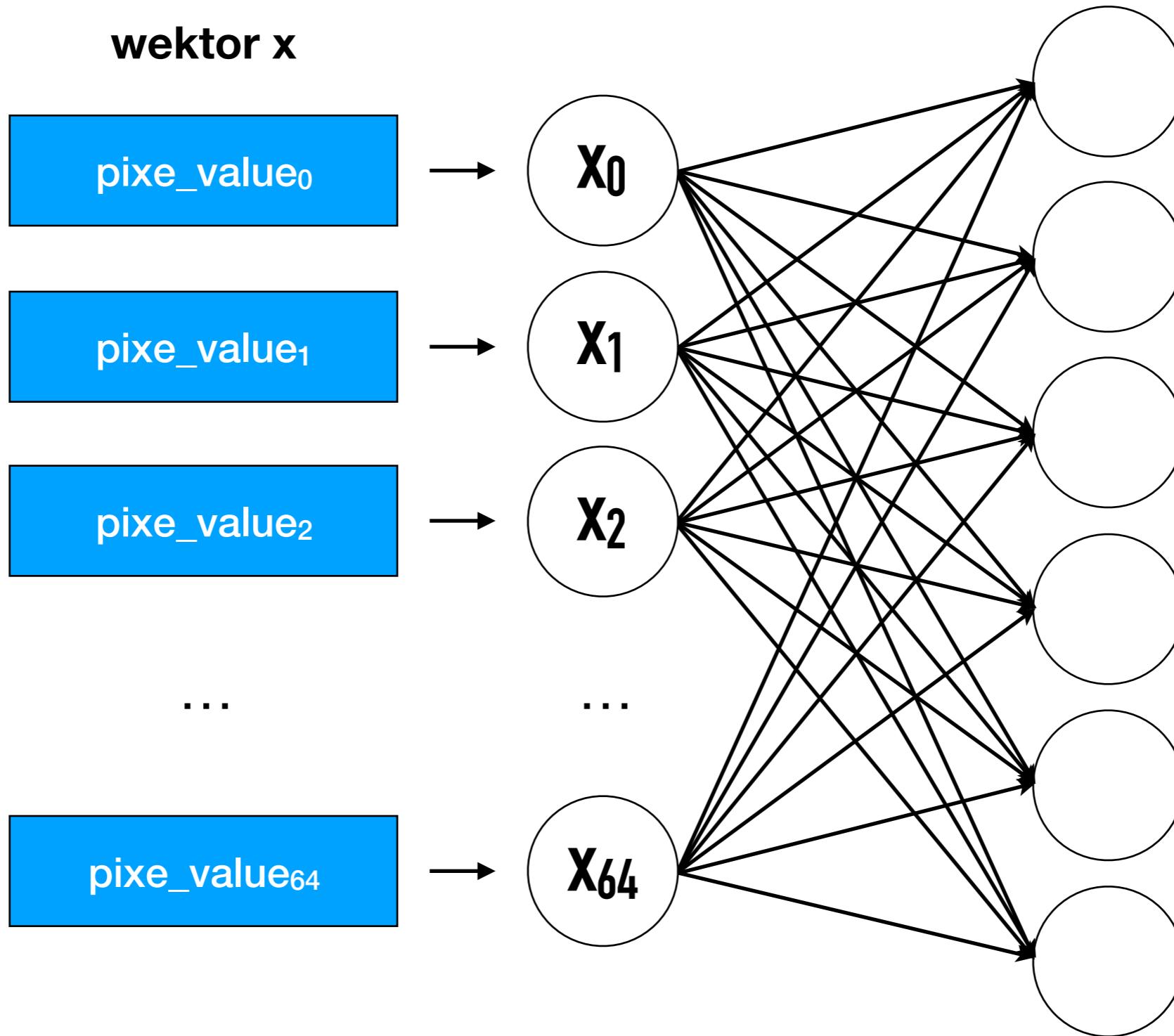
pixe_value₁

pixe_value₂

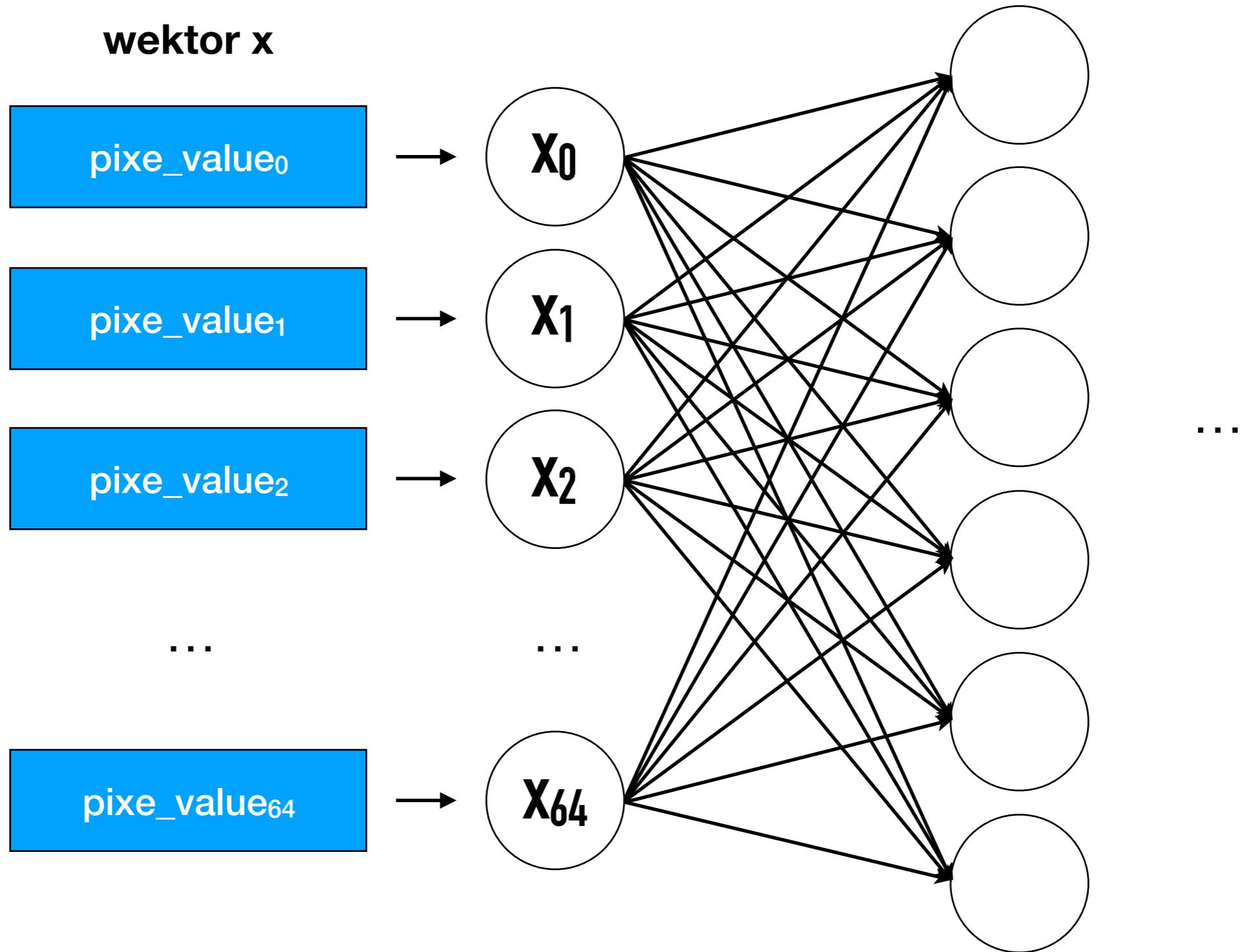
...

pixe_value₆₄

Obrazy a Sieci Neuronowe

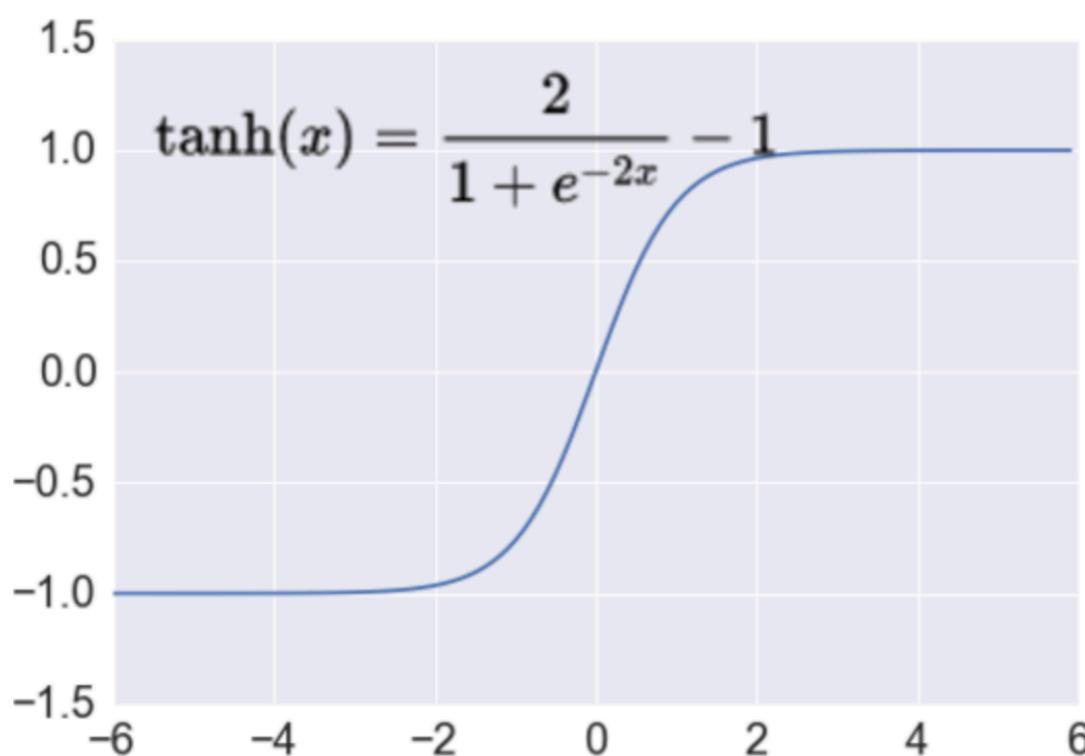


Obrazy a Sieci Neuronowe

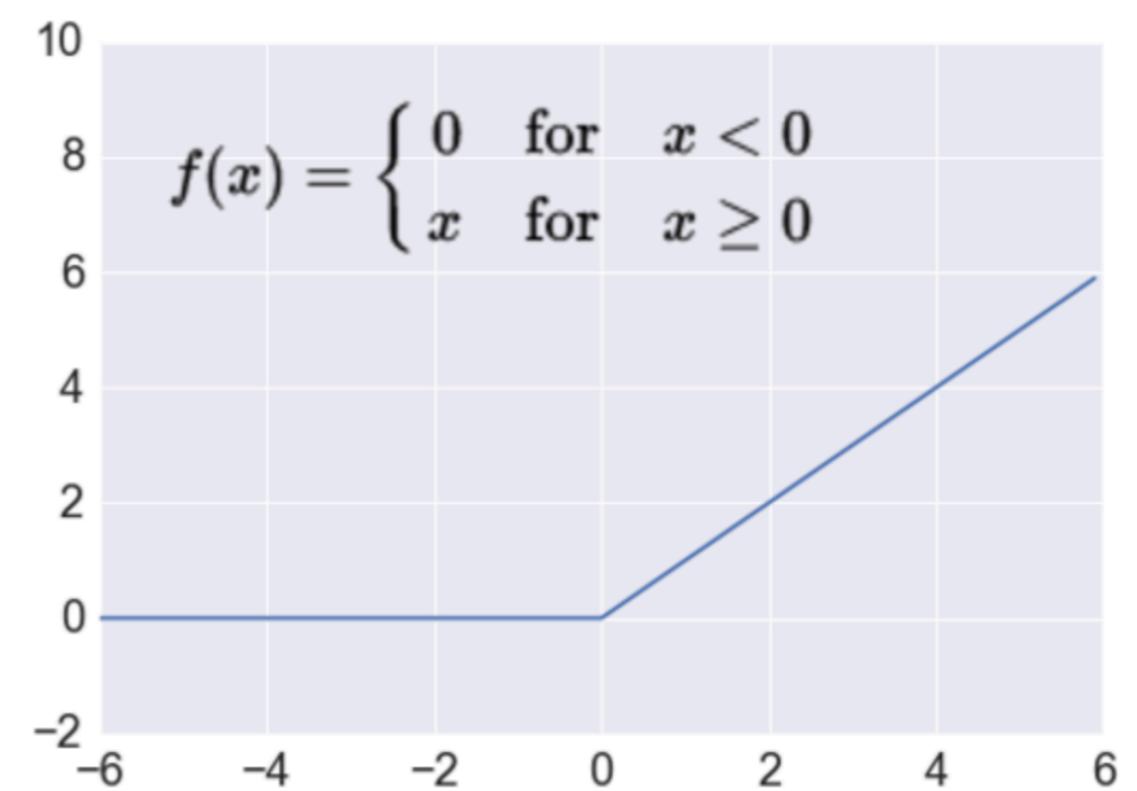


Inne funkcje aktywacji

TanH

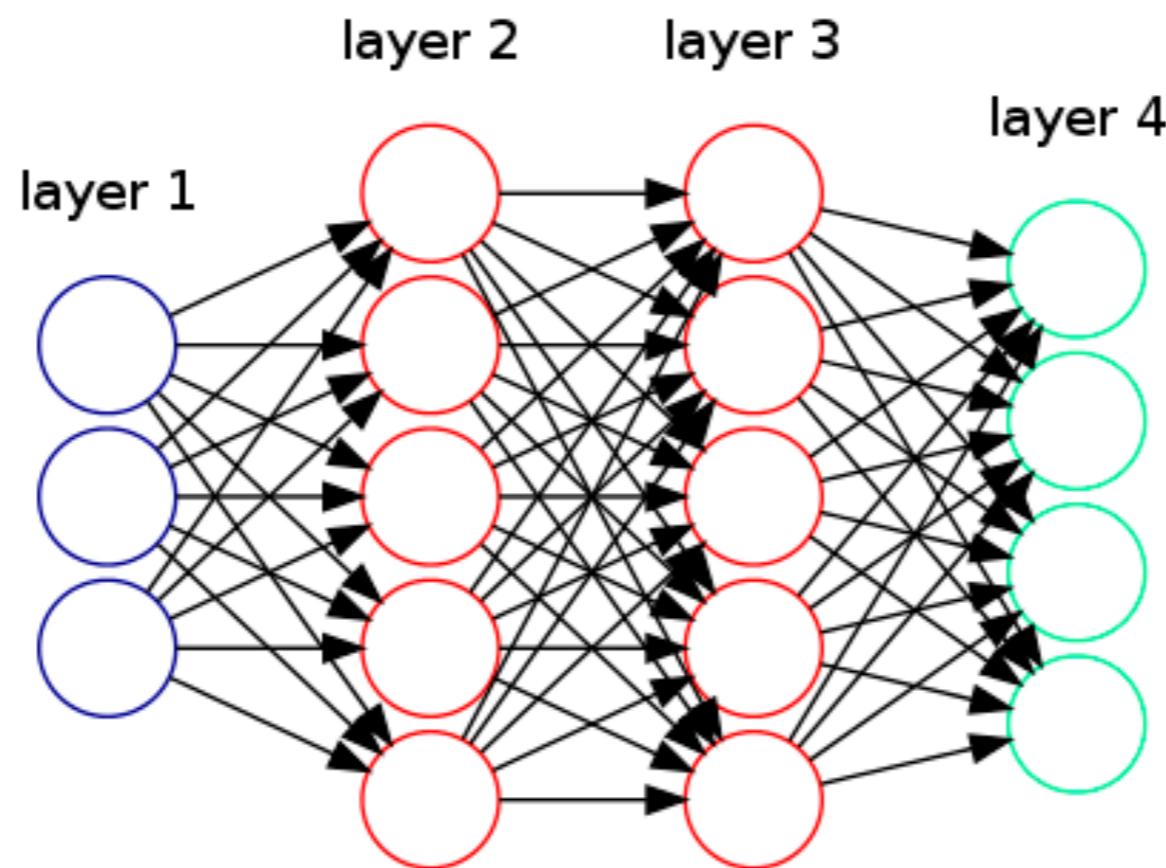


ReLU

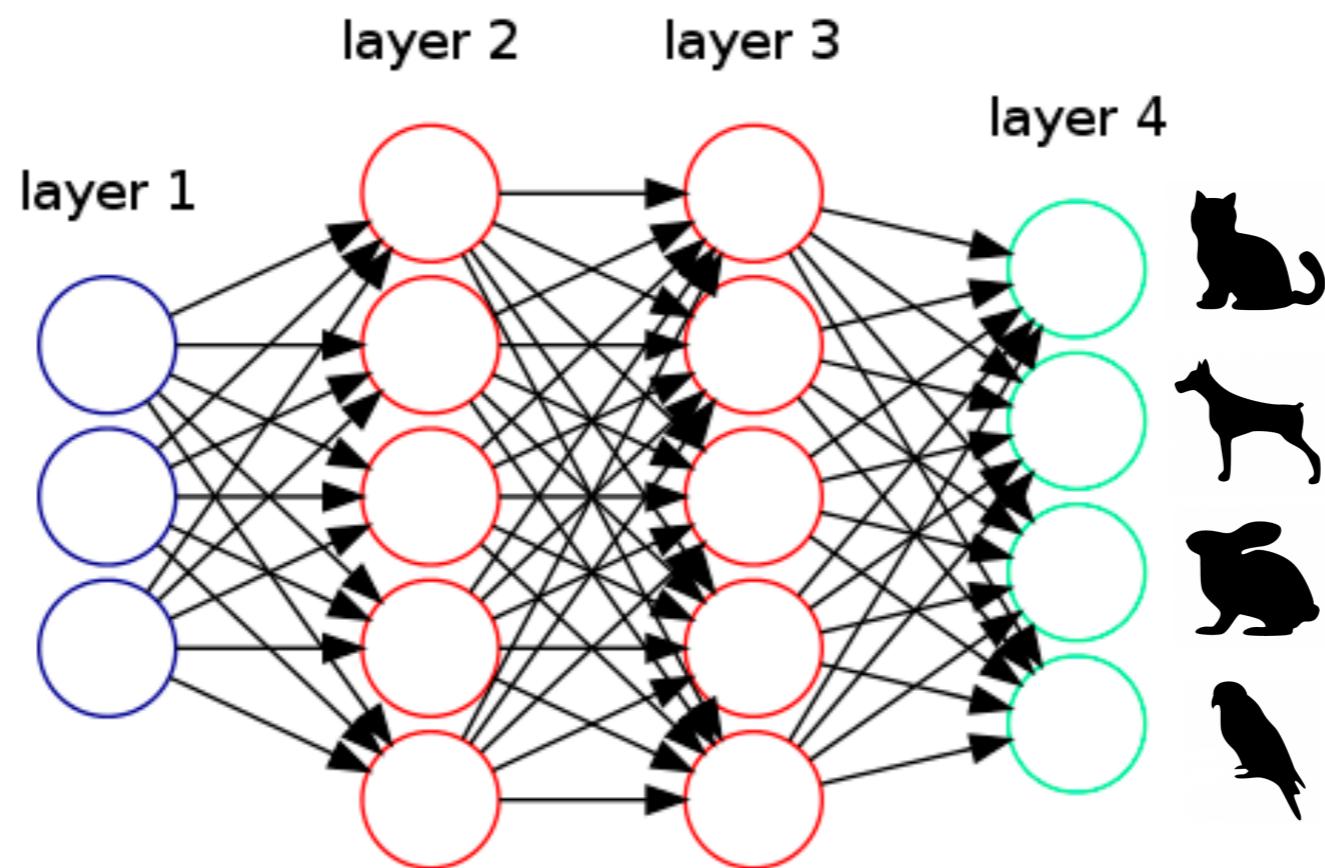


Softmax

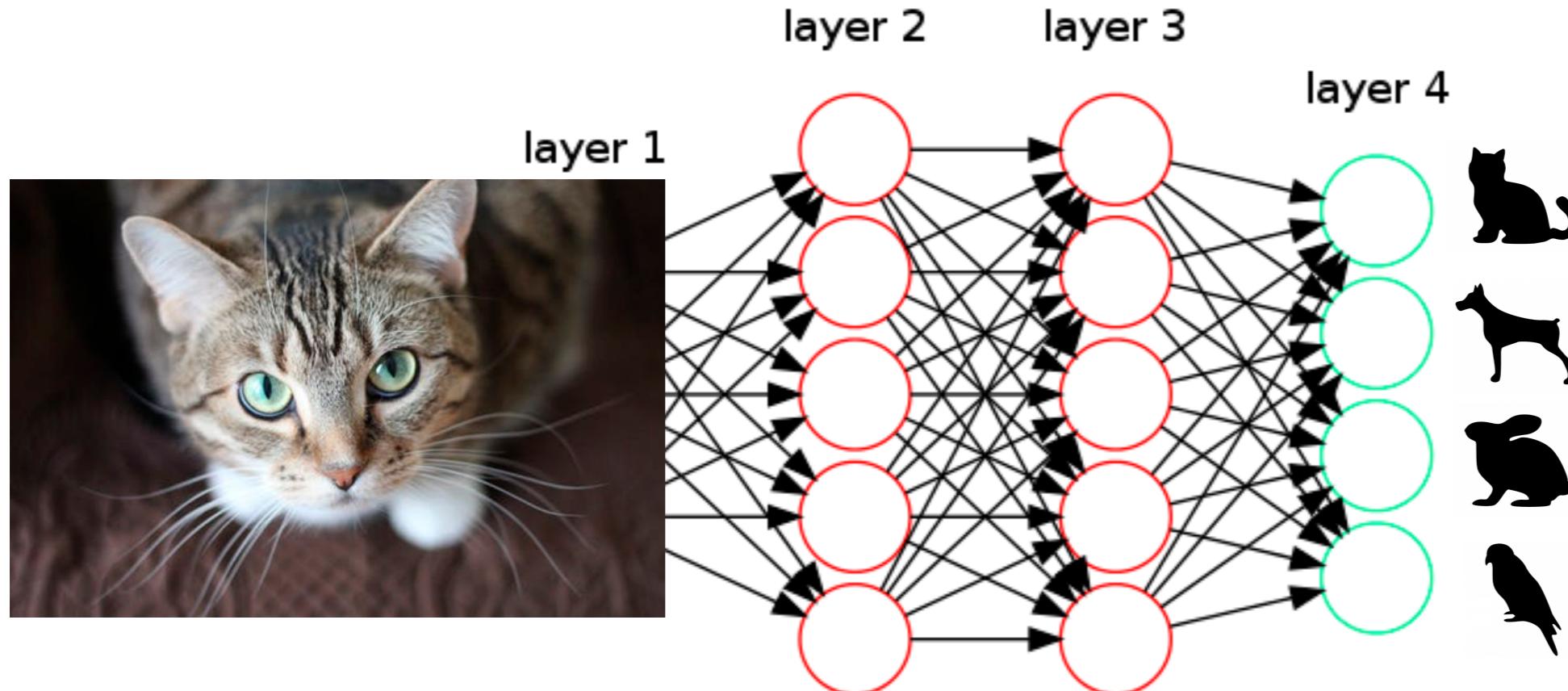
Softmax



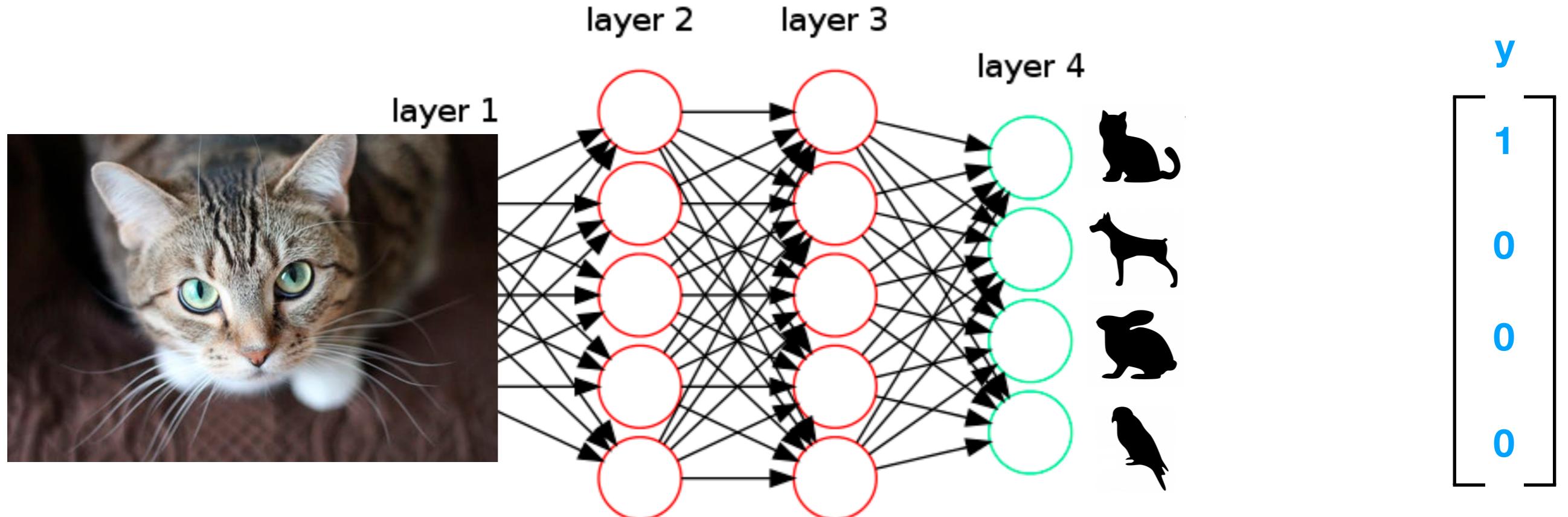
Softmax



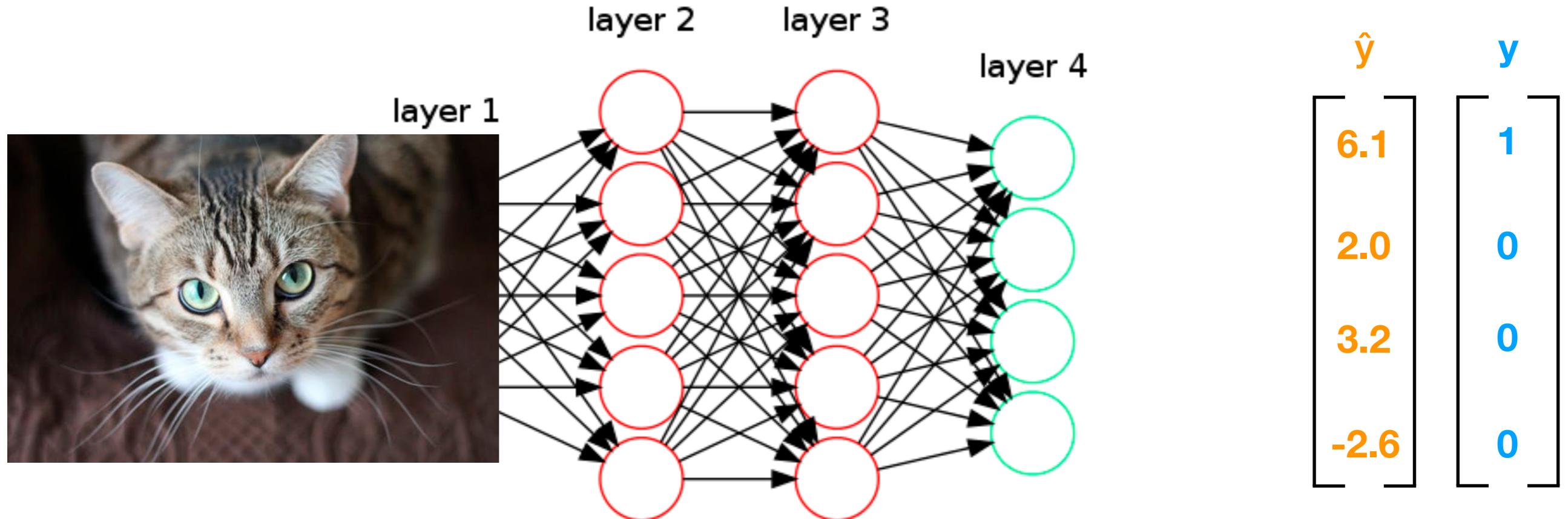
Softmax



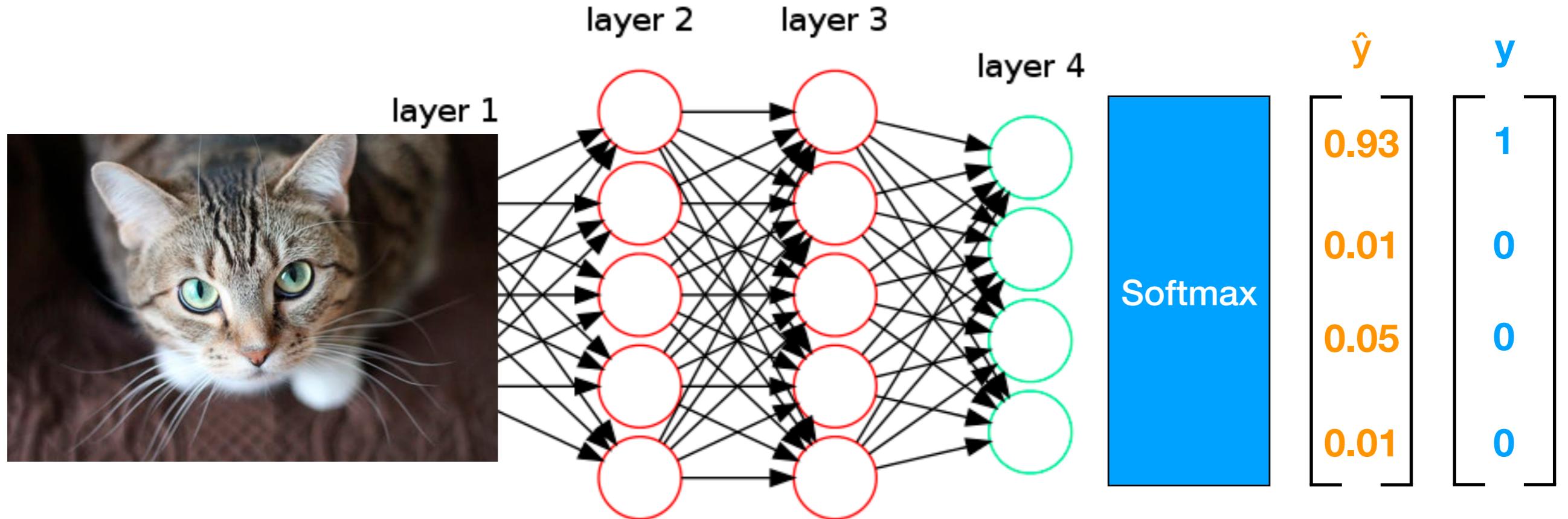
Softmax



Softmax

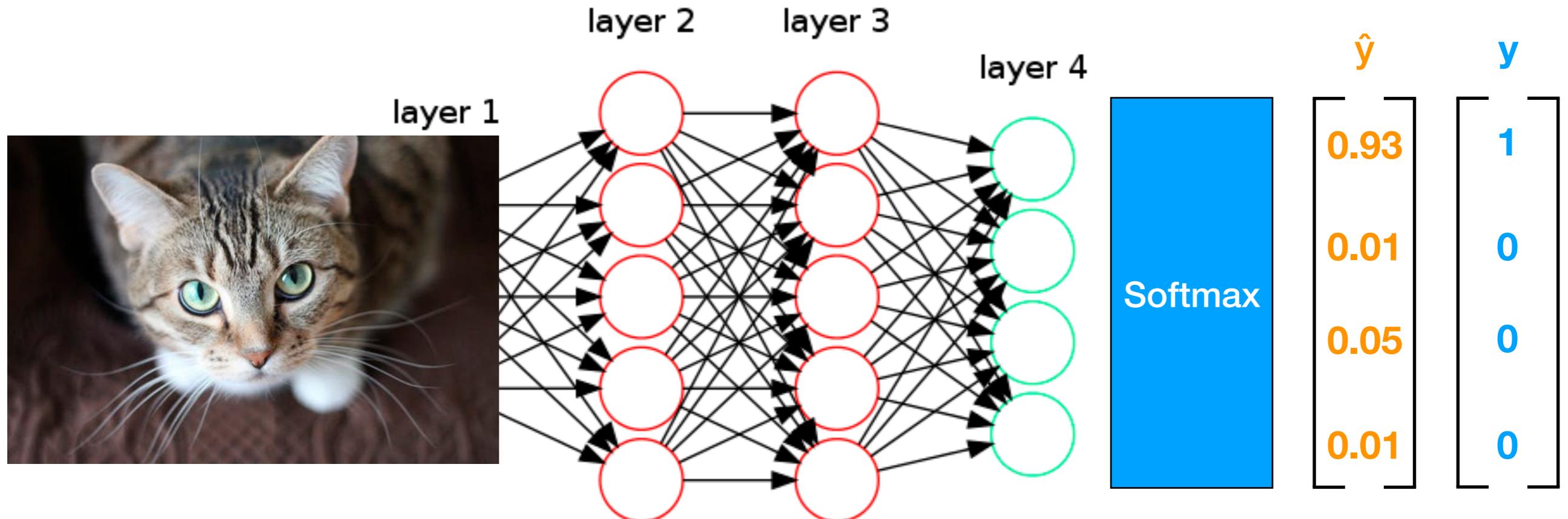


Softmax



Softmax

Używany na wyjściu sieci, która dokonuje klasyfikacji wielu klas.
Zamienia wartości wyjściowe na prawdopodobieństwa.



Więcej na warsztacie :)