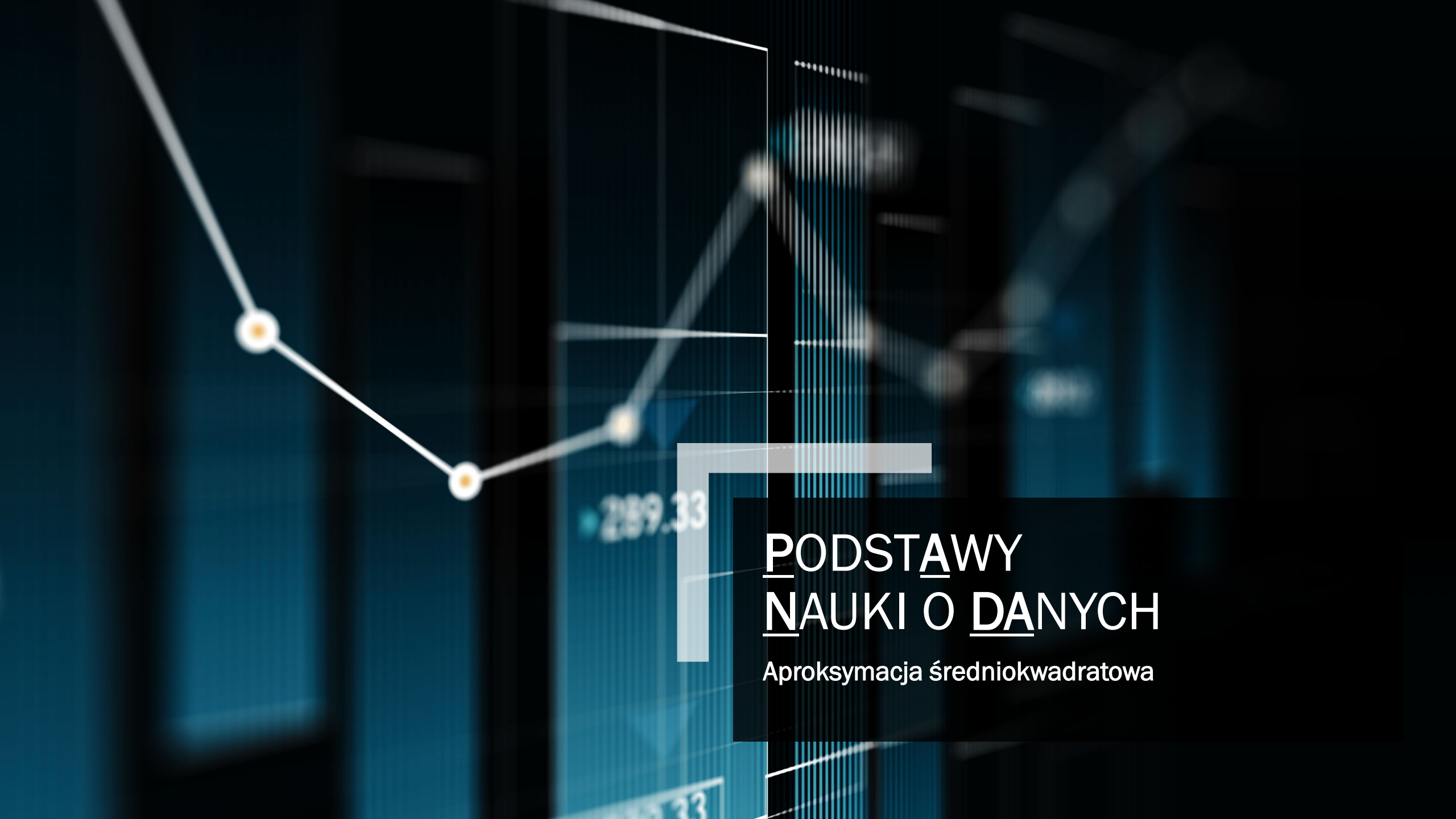




PODSTAWY NAUKI O DANYCH

Aproksymacja średniokwadratowa

Organizacja danych

K – liczba atrybutów/wejść/cech

wejście wyjście
 $\left\{ (\mathbf{x}_i, y_i) \right\}_{i=1}^N$

N – liczba rekordów/par
 wejście-wyjście

	Height	Weight	BMI	Cholesterol LDL	Cholesterol HDL	Gender	Age
306	170	115	39	145	25	Male	70
371	179	108	33	54	29	Male	64
383	160	59	23	183	38	Male	68
611	194	82	21	76	62	Male	63
543	170	99	34	62	30	Male	58
135	156	77	31	99	56	Female	69
644	185	80	23	68	36	Male	68

wiersz/rekord/obserwacja – zmierzona instancja powiązanych informacji

$$\mathbf{x}_i = \begin{bmatrix} x_i^{(1)} \\ x_i^{(2)} \\ \vdots \\ x_i^{(K)} \end{bmatrix} \quad y_i \in \mathbb{R}$$



$$\mathbf{X} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \dots \quad \mathbf{x}_N] = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_N^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_N^{(2)} \\ \dots & \dots & \dots & \dots \\ x_1^{(K)} & x_2^{(K)} & \dots & x_N^{(K)} \end{bmatrix}$$

$$\mathbf{Y} = [y_1 \quad y_2 \quad \dots \quad y_N]$$

Regresja

Dane
pomiarowe



$$\{(\mathbf{x}_i, y_i)\}_{i=1}^N$$

wektor
wartości wejściowych

$\mathbf{x}$

$$x_i \in \mathbb{R}$$

model $\varphi(\mathbf{x}; \theta)$

„predykcja (odgadnięcie)”
wartości wyjściowej

$\bar{y}$

$$y_i \in \mathbb{R}$$

- Przykład 1: model liniowy

$$\bar{y} = \theta^T \mathbf{x}$$

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_K \end{bmatrix}$$

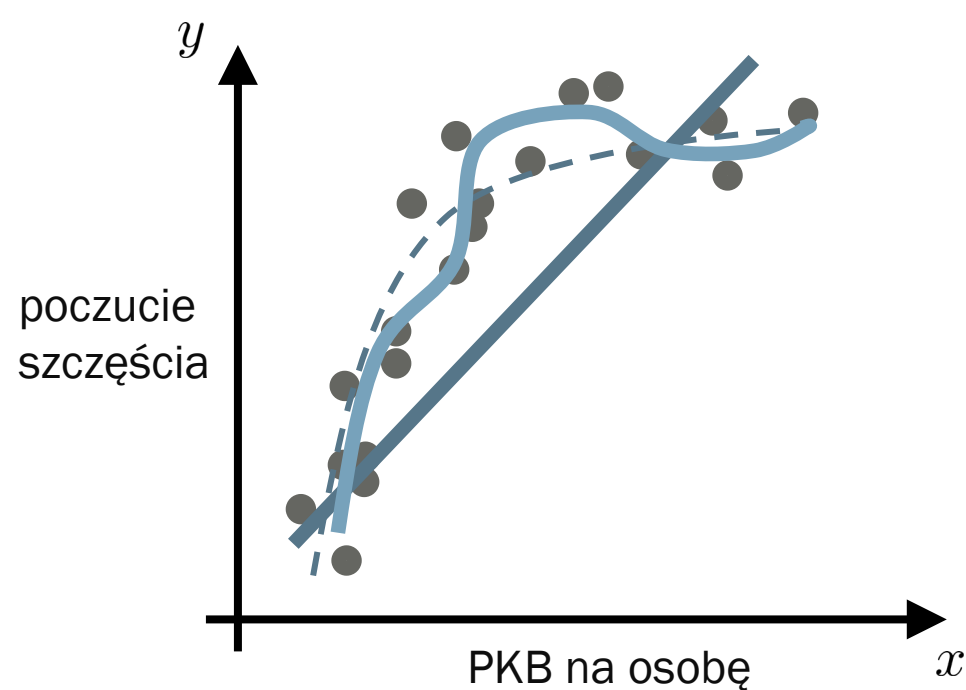
$$\theta_k \in \mathbb{R}$$

- Przykład 2: model „liniowy”

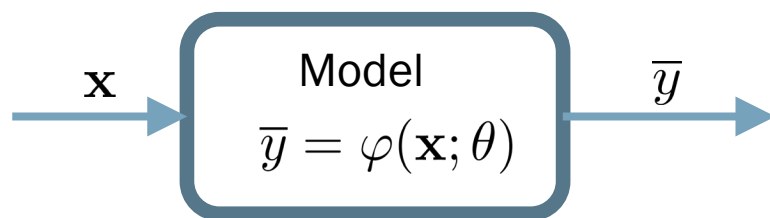
$$\begin{aligned} \bar{y} &= ax + b \\ &= \theta^T \mathbf{x} \end{aligned}$$

$$\mathbf{x} = \begin{bmatrix} x \\ 1 \end{bmatrix}$$

$$\theta = \begin{bmatrix} a \\ b \end{bmatrix}$$



Organizacja wyników obliczeń



θ – wektor parametrów modelu

φ – ogólny opis obliczeń
czyli algorytm predykcji

$$\bar{y}_i = \varphi(\mathbf{x}_i; \theta)$$

błąd predykcji:

$$\begin{aligned} e_i &= y_i - \bar{y}_i \\ &= y_i - \varphi(\mathbf{x}_i; \theta) \end{aligned}$$

„predykcje” modelu

$$\bar{\mathbf{Y}} = [\bar{y}_1 \quad \bar{y}_2 \quad \dots \quad \bar{y}_N]$$

poprawne „odpowiedzi”
ze zbioru danych

$$\mathbf{Y} = [y_1 \quad y_2 \quad \dots \quad y_N]$$

błędy predykcji

$$\mathbf{E} = [e_1 \quad e_2 \quad \dots \quad e_N]$$

Wyznaczenie parametrów modelu

- Dane: $\mathbf{X}, \mathbf{Y}$ - pomiary

$\bar{y} = \varphi(\mathbf{x}; \theta)$ - postać modelu

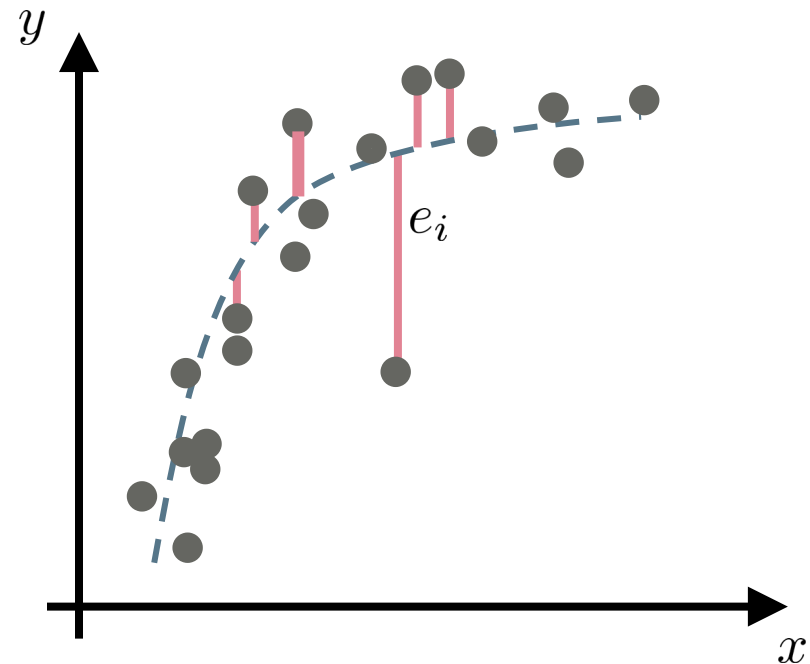
Kryterium jakości modelu:

$$Q(\theta; \mathbf{X}, \mathbf{Y}) \equiv Q(\theta) \quad \text{np.}$$

$$Q_1(\theta) = \sum_{i=1}^N |e_i|$$


- Szukane:

$\theta = ?$ - wartości parametrów modelu



Wyznaczenie parametrów modelu

- Dane: $\mathbf{X}, \mathbf{Y}$ – pomiary

$\bar{y} = \varphi(\mathbf{x}; \theta)$ – postać modelu

Kryterium jakości modelu:

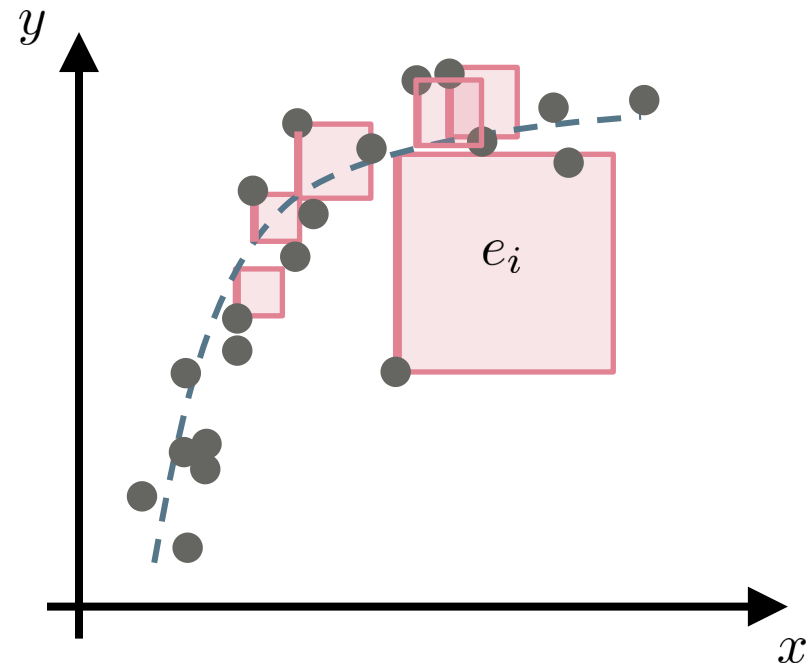
$$Q(\theta; \mathbf{X}, \mathbf{Y}) \equiv Q(\theta) \quad \text{np.}$$

$$Q_1(\theta) = \sum_{i=1}^N |e_i|$$

$$Q_2(\theta) = \sum_{i=1}^N e_i^2 = \mathbf{E}\mathbf{E}^T$$

- Szukane:

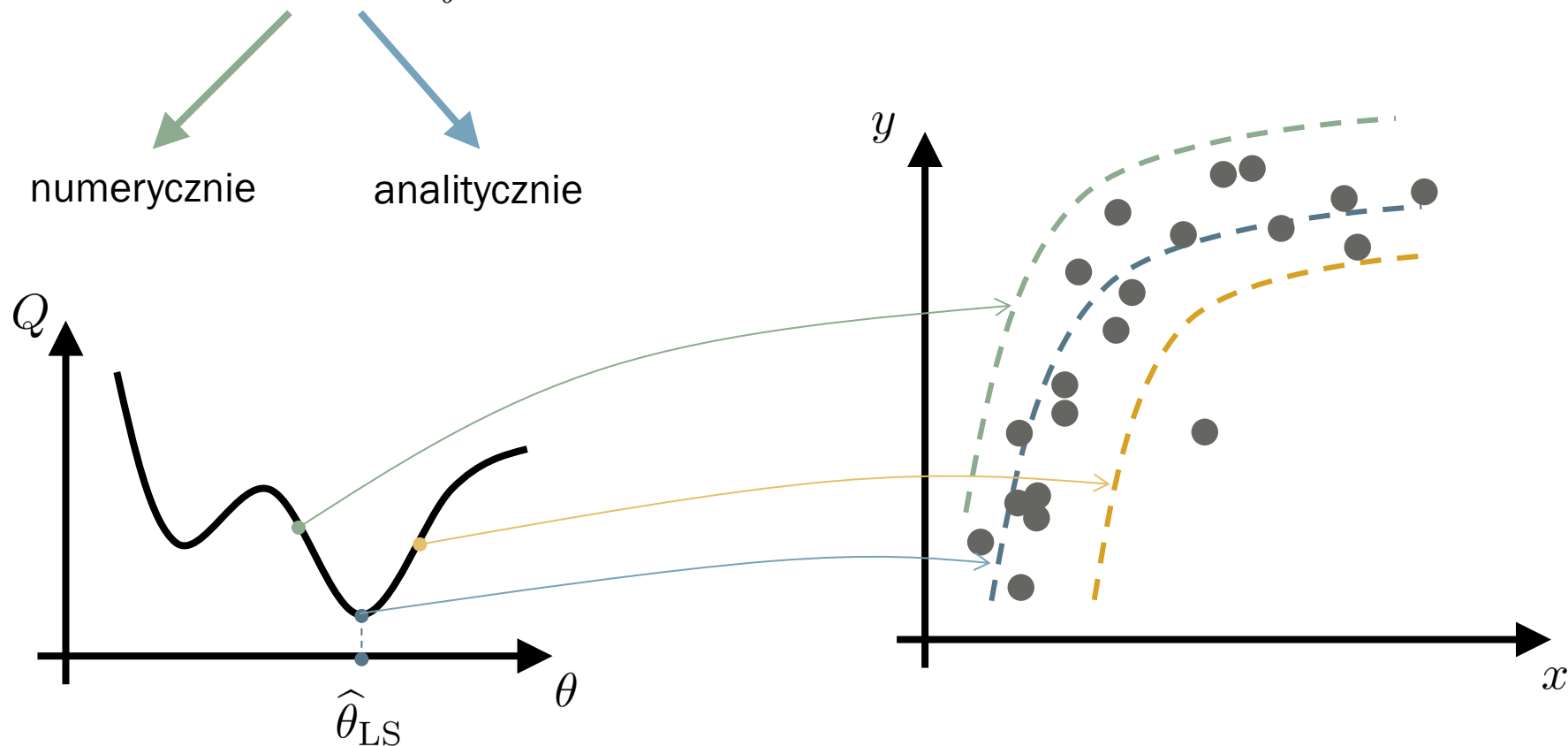
$\theta = ?$ – wartości parametrów modelu



Wyznaczenie parametrów modelu

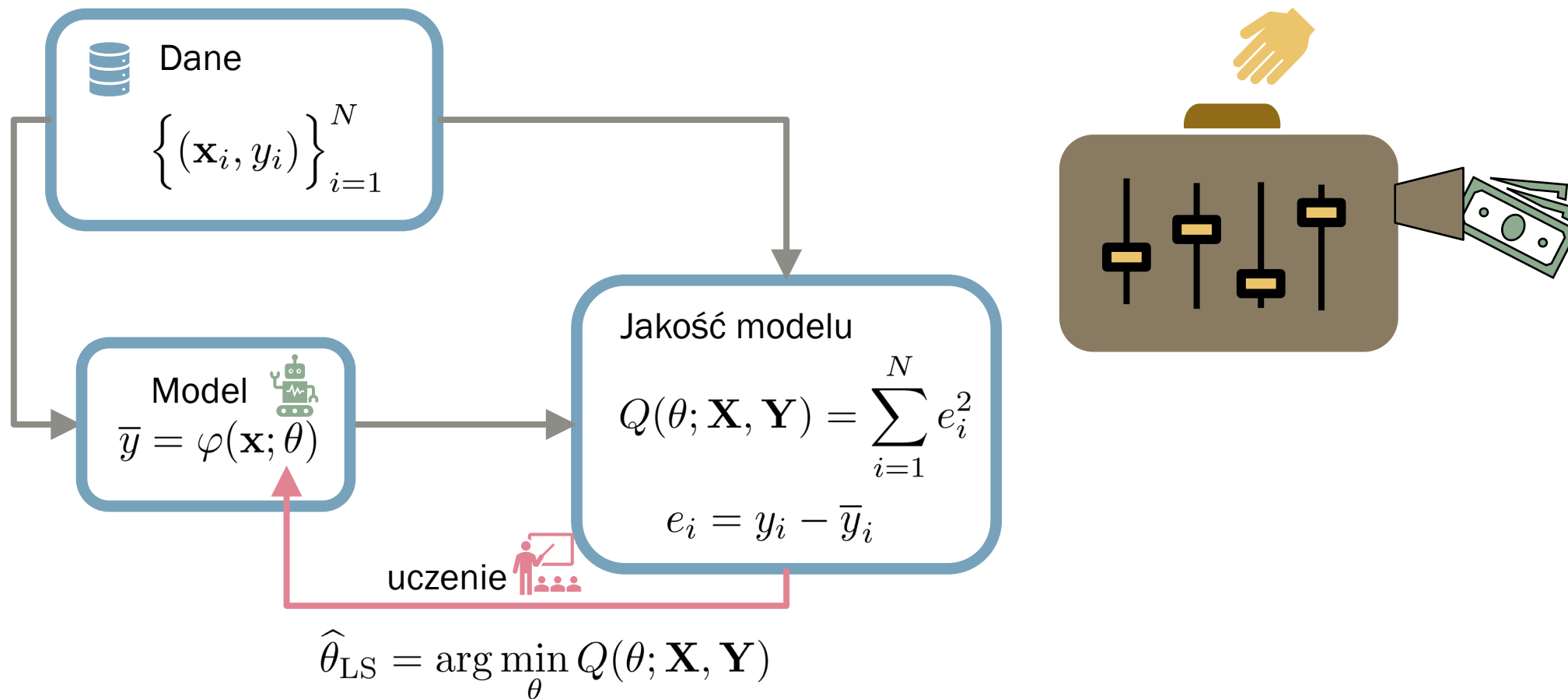
Estymacja, aproksymacja

- Rozwiązanie: $\hat{\theta}_{LS} = \arg \min_{\theta} Q(\theta)$ dla ustalonego zbioru danych $\mathbf{X}, \mathbf{Y}$



Dopasowanie modelu do danych

Metoda najmniejszych kwadratów (ang. *Least Squares method*)



Rozwiązanie analityczne czy numeryczne?

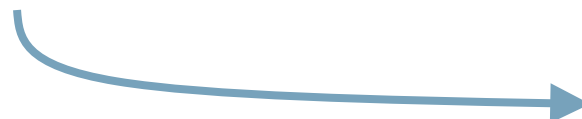
$$\begin{bmatrix} \frac{\partial Q}{\partial \theta_1} \\ \vdots \\ \frac{\partial Q}{\partial \theta_K} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

Jestem w stanie policzyć pochodne
i rozwiązać układ



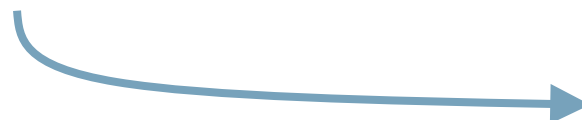
analitycznie

Q nieciągła



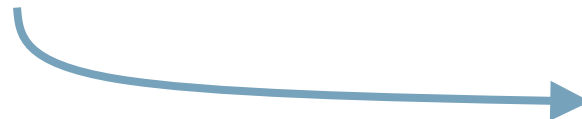
numerycznie optymalizacja

Q nieróżniczkowalna



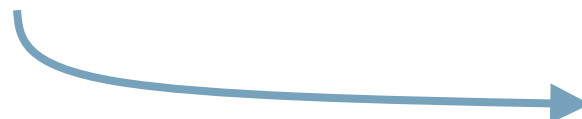
numerycznie optymalizacja

Q zbyt skomplikowana



numerycznie optymalizacja

układ równań zbyt skomplikowany



numeryczne rozwiązanie
układu równań

Reguły „gradientowania”

$$\frac{\partial}{\partial \theta} (\mathbf{A} \mathbf{A}^{-1}) = \frac{\partial \mathbf{A}}{\partial \theta} \mathbf{A}^{-1} + \mathbf{A} \frac{\partial}{\partial \theta} \mathbf{A}^{-1} \quad \Rightarrow \quad \frac{\partial}{\partial \theta} \mathbf{A}^{-1} = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial \theta} \mathbf{A}^{-1}$$

$$\frac{\partial}{\partial \theta} \theta^{\mathrm{T}} \mathbf{A} = \mathbf{A}$$

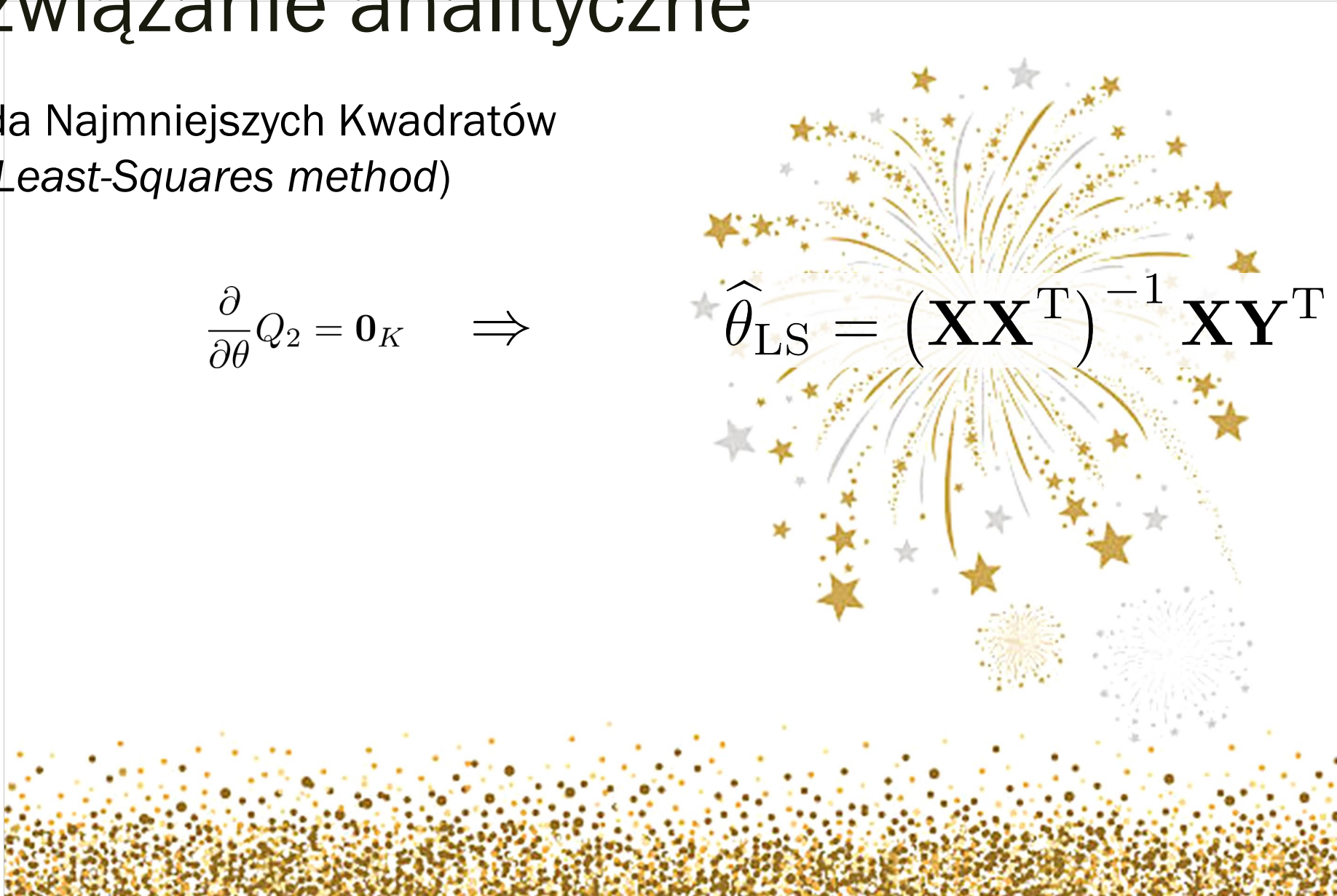
$$\frac{\partial}{\partial \theta} \theta^{\mathrm{T}} \mathbf{A} \theta = (\mathbf{A} + \mathbf{A}^{\mathrm{T}}) \theta$$

Rozwiązanie analityczne

Metoda Najmniejszych Kwadratów
(ang. *Least-Squares method*)

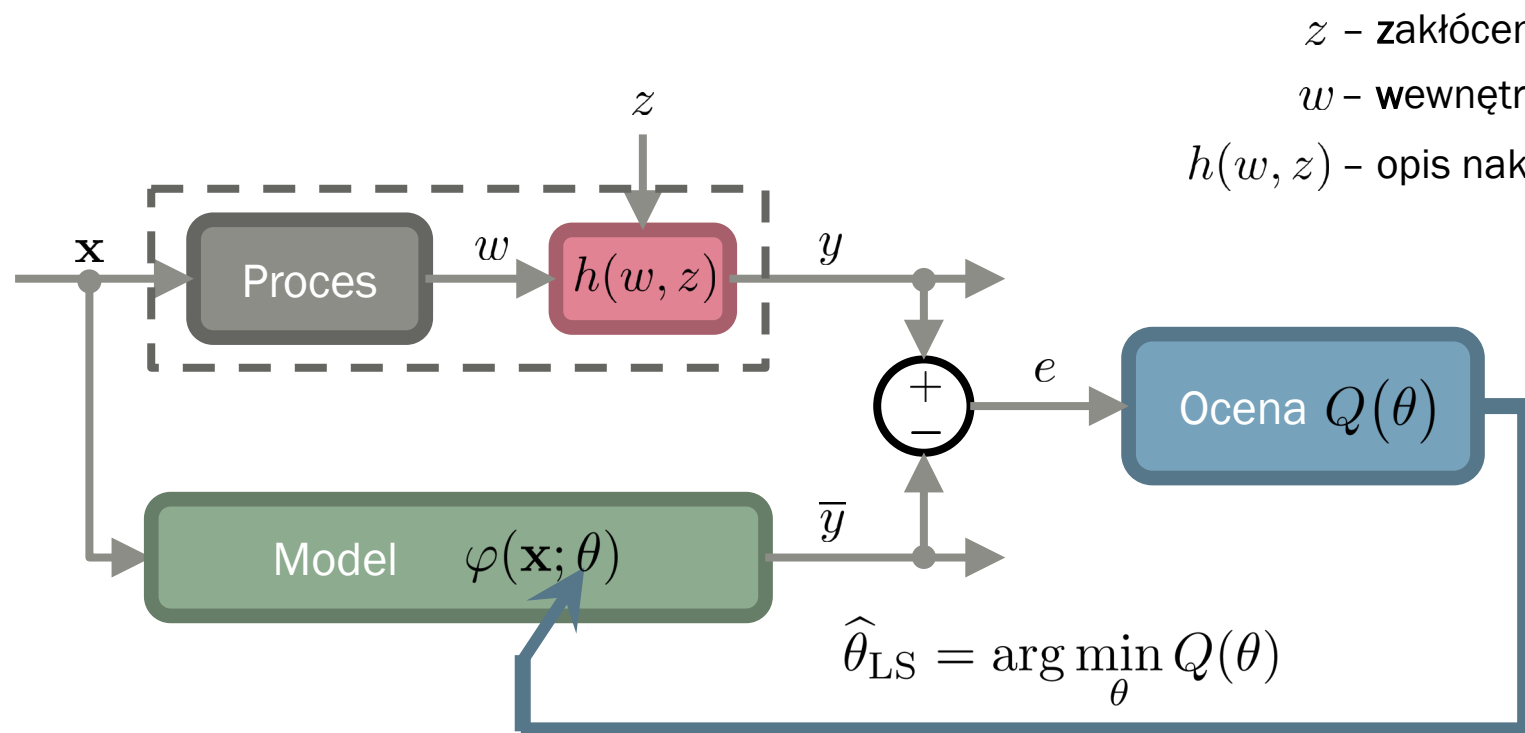
$$\frac{\partial}{\partial \theta} Q_2 = \mathbf{0}_K \quad \Rightarrow$$

$$\hat{\theta}_{\text{LS}} = (\mathbf{X}\mathbf{X}^T)^{-1} \mathbf{X}\mathbf{Y}^T$$



Czy to zawsze dobrze działa?

Założenia stojące u podstaw metody



Uogólnione modele liniowe

- $\bar{y} = \theta^T \mathbf{x}$
 $\bar{y} = ax + b$

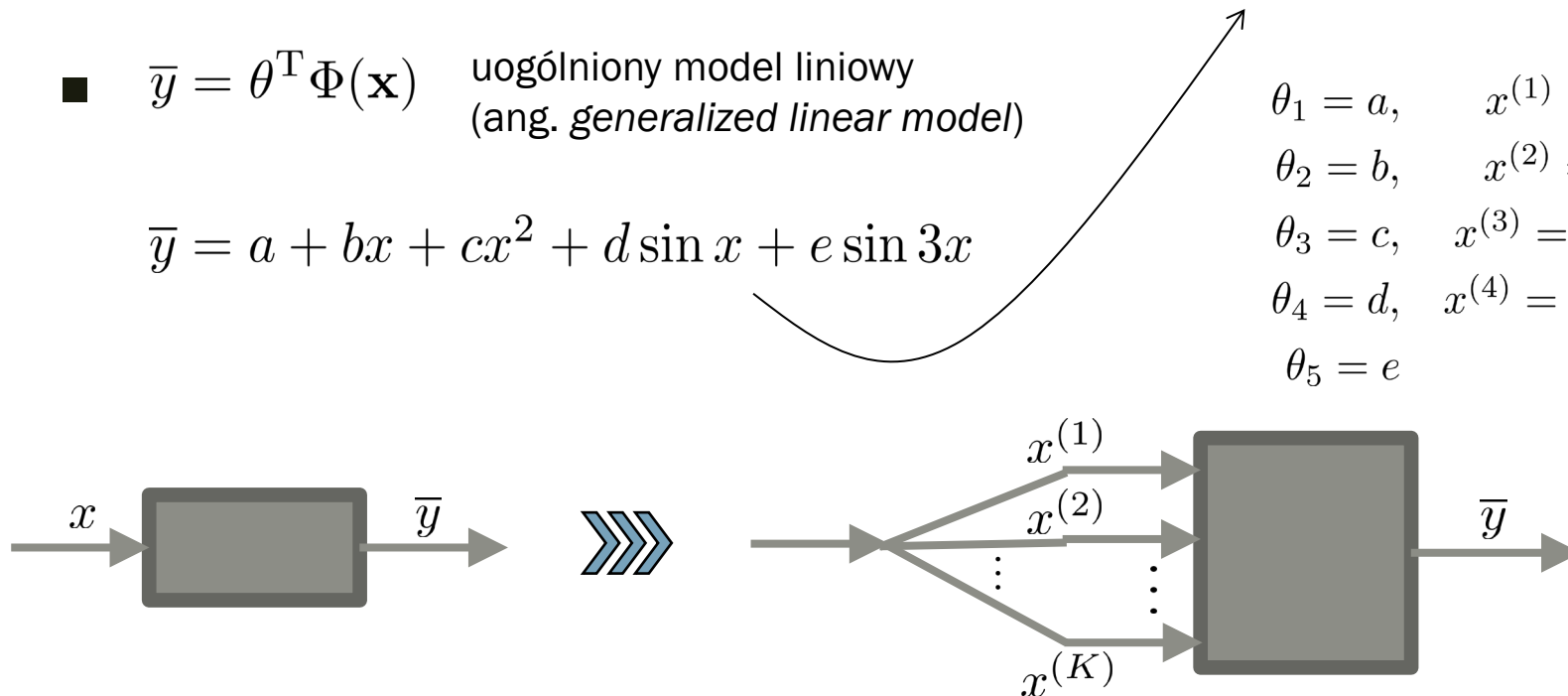
$$\bar{y} = \theta_1 x^{(1)} + \theta_2 x^{(2)} + \dots + \theta_K x^{(K)} = \sum_{k=1}^K \theta_k x^{(k)}$$

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_K \end{bmatrix}$$

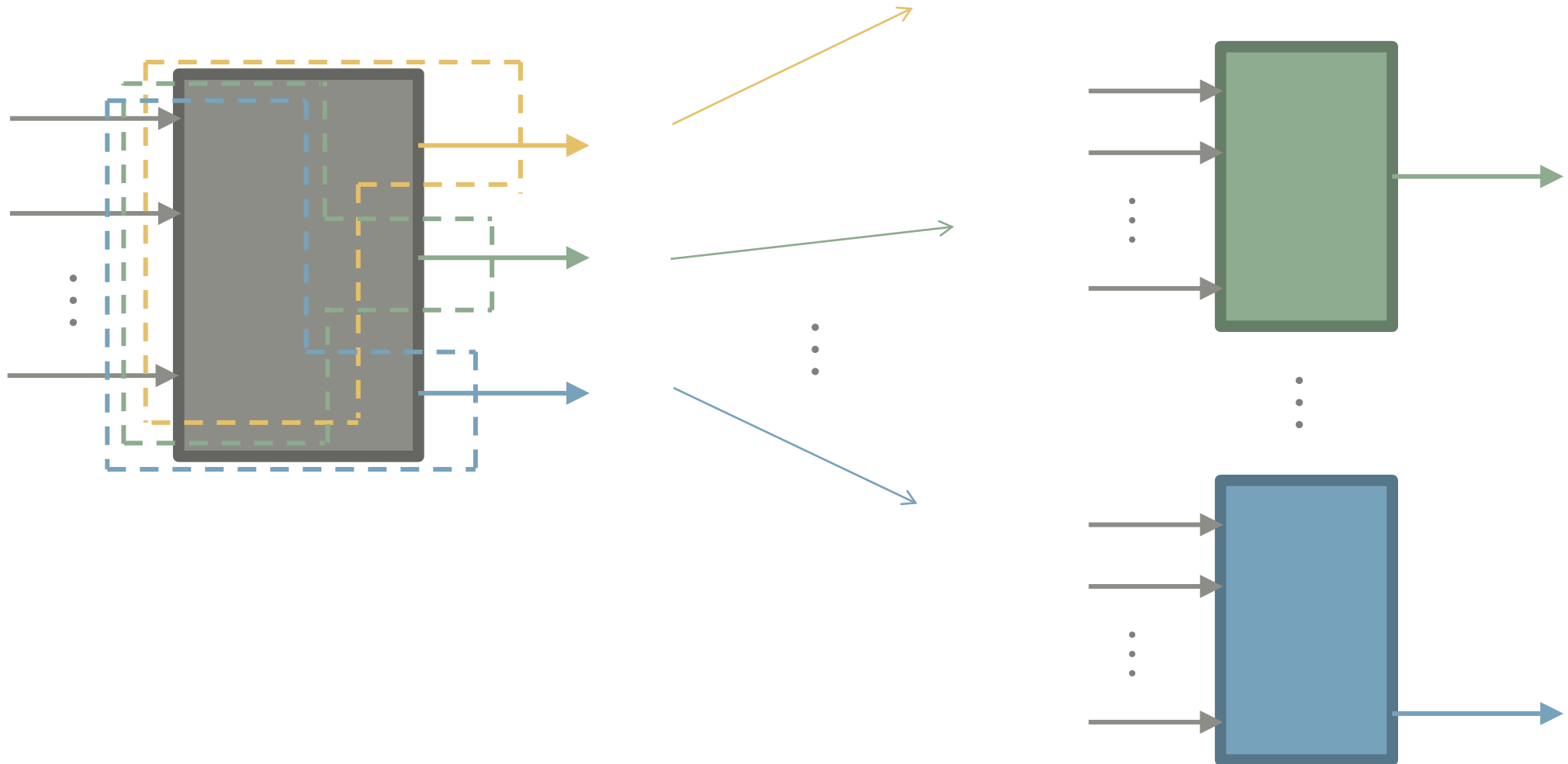
- $\bar{y} = \theta^T \Phi(\mathbf{x})$ uogólniony model liniowy
(ang. *generalized linear model*)

$$\bar{y} = a + bx + cx^2 + d \sin x + e \sin 3x$$

$$\begin{aligned} \theta_1 &= a, & x^{(1)} &= 1, \\ \theta_2 &= b, & x^{(2)} &= x, \\ \theta_3 &= c, & x^{(3)} &= \sin x, \\ \theta_4 &= d, & x^{(4)} &= \sin 3x, \\ \theta_5 &= e \end{aligned}$$

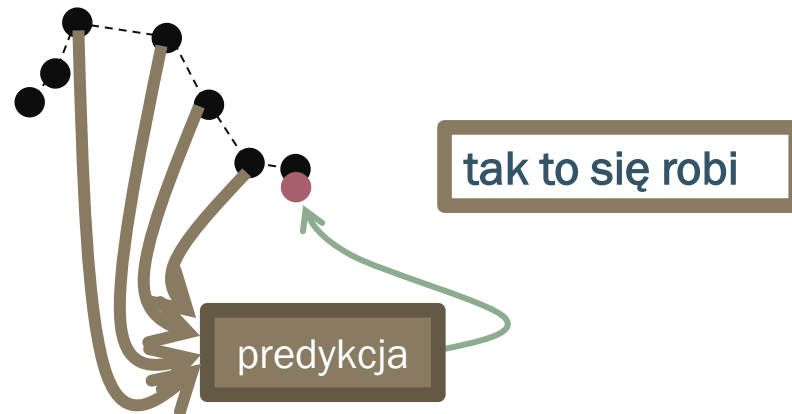
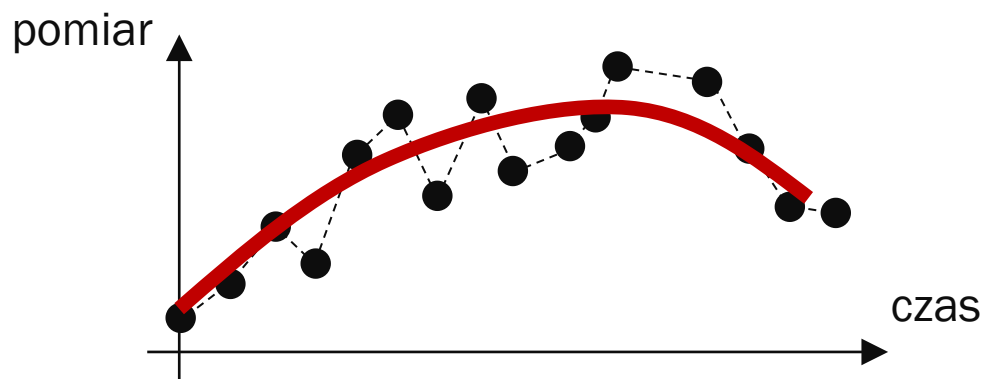


Wiele wyjść



Predykcja – częsty błąd

ten model nie ma sensu



Jak uczyć model
z wykorzystaniem aproksymatora?

Model odniesienia / model naiwny

■ Współczynnik determinacji

sum of squared errors

$$SSE = \sum_{i=1}^N e_i^2 = \sum_{i=1}^N (y_i - \bar{y}_i)^2$$

total sum of squares

$$TSS = \sum_{i=1}^N (y_i - \hat{y})^2$$

$$\bar{y}_i = \varphi(\mathbf{x}_i; \theta) \quad - \text{model docelowy}$$

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N y_i \quad - \text{model odniesienia}$$

coefficient of determination

$$R^2 = \frac{TSS - SSE}{TSS}$$



THE END