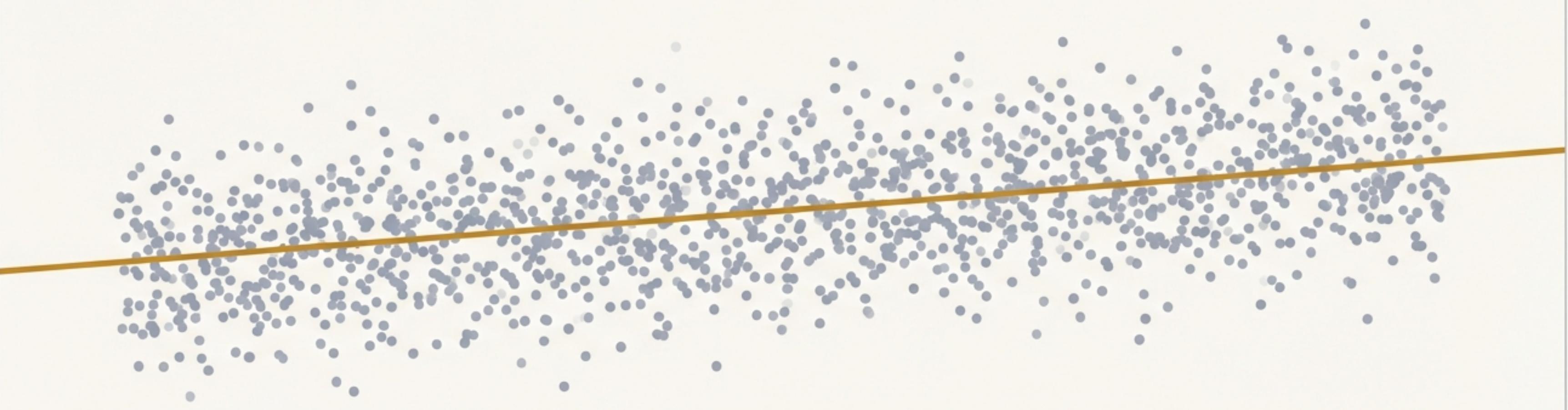


# Regresja Liniowa: Matka Wszystkich Algorytmów Uczzenia Maszynowego

Zrozumienie prostej, potężnej idei, która leży u podstaw nowoczesnej analityki i AI.

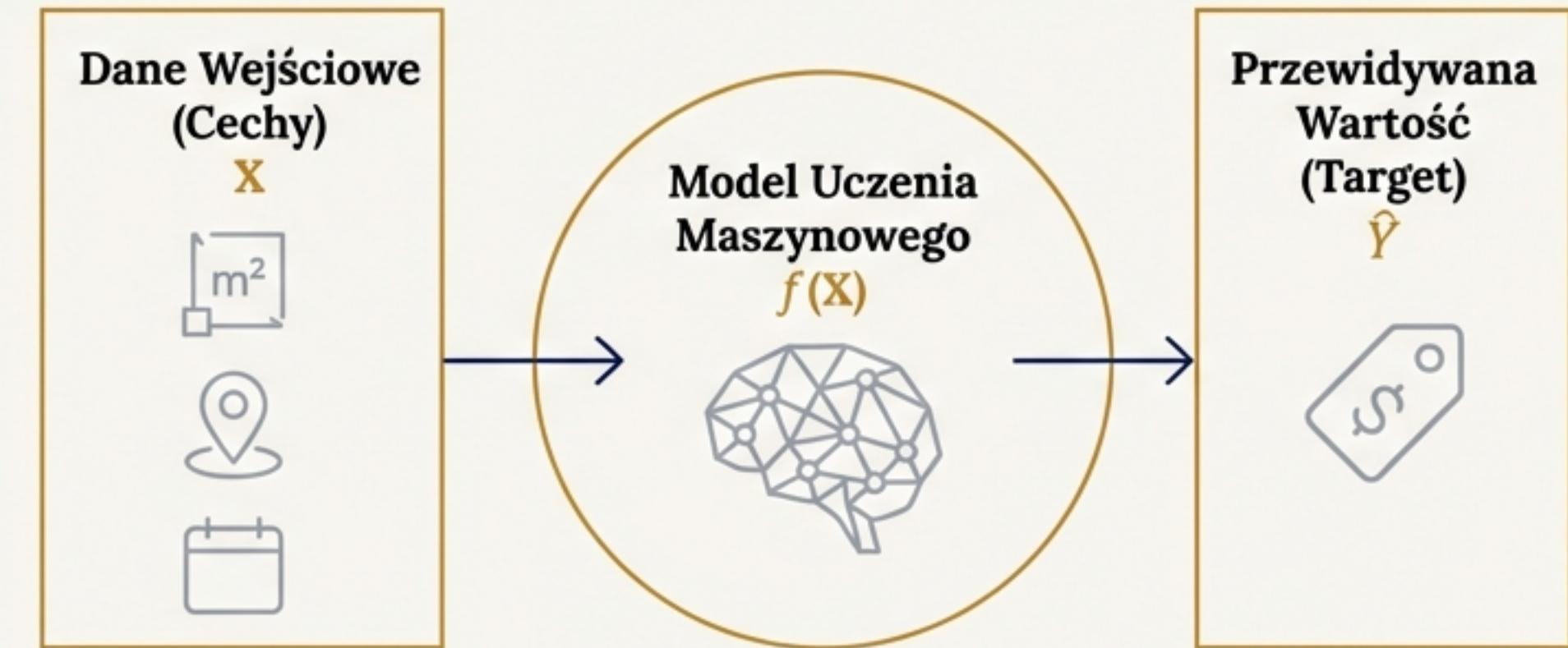


*„Zacznijmy od matki wszystkich algorytmów uczenia maszynowego: regresji liniowej.”*

Ta prezentacja wyjaśni, dlaczego ten prosty model jest nie tylko punktem wyjścia do nauki o danych, ale także potężnym i wysoce interpretowalnym narzędziem używanym do analizy przyczynowej i prognozowania.

# Główne Zadanie Uczzenia Nadzorowanego

W uczeniu nadzorowanym dysponujemy zbiorem danych z cechami wejściowymi ( $\mathbf{X}$ ) i zmienną docelową ( $\mathbf{Y}$ ), którą chcemy przewidzieć. Naszym celem jest znalezienie funkcji  $f$ , która najlepiej mapuje  $\mathbf{X}$  na  $\mathbf{Y}$ , tak aby  $\mathbf{Y} \approx f(\mathbf{X})$ . Ta funkcja, po „nauczeniu” na danych historycznych, pozwala nam przewidywać wartości  $\mathbf{Y}$  dla nowych, nieznanych danych.



**Regresja:** Przewidywanie ceny domu ( $\mathbf{Y}$ ) na podstawie jego powierzchni, lokalizacji, roku budowy ( $\mathbf{X}$ ).

**Klasyfikacja:** Określanie, czy e-mail ( $\mathbf{X}$ ) to spam, czy nie ( $\mathbf{Y}$ ).

# Regresja Liniowa: Potęga Prostej Linii

Regresja liniowa jest najbardziej fundamentalnym podejściem do modelowania relacji **między zmiennymi**. Zakłada, że związek ten można przybliżyć za pomocą prostej linii. Jest to narzędzie o podwójnym zastosowaniu:

- **Analiza przyczynowa:** Pomaga zrozumieć, które czynniki (X) mają statystycznie istotny wpływ na zmienną docelową (Y) i jak silny jest ten wpływ.
- **Prognozowanie:** Umożliwia przewidywanie przyszłych wartości Y na podstawie znanych wartości X.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

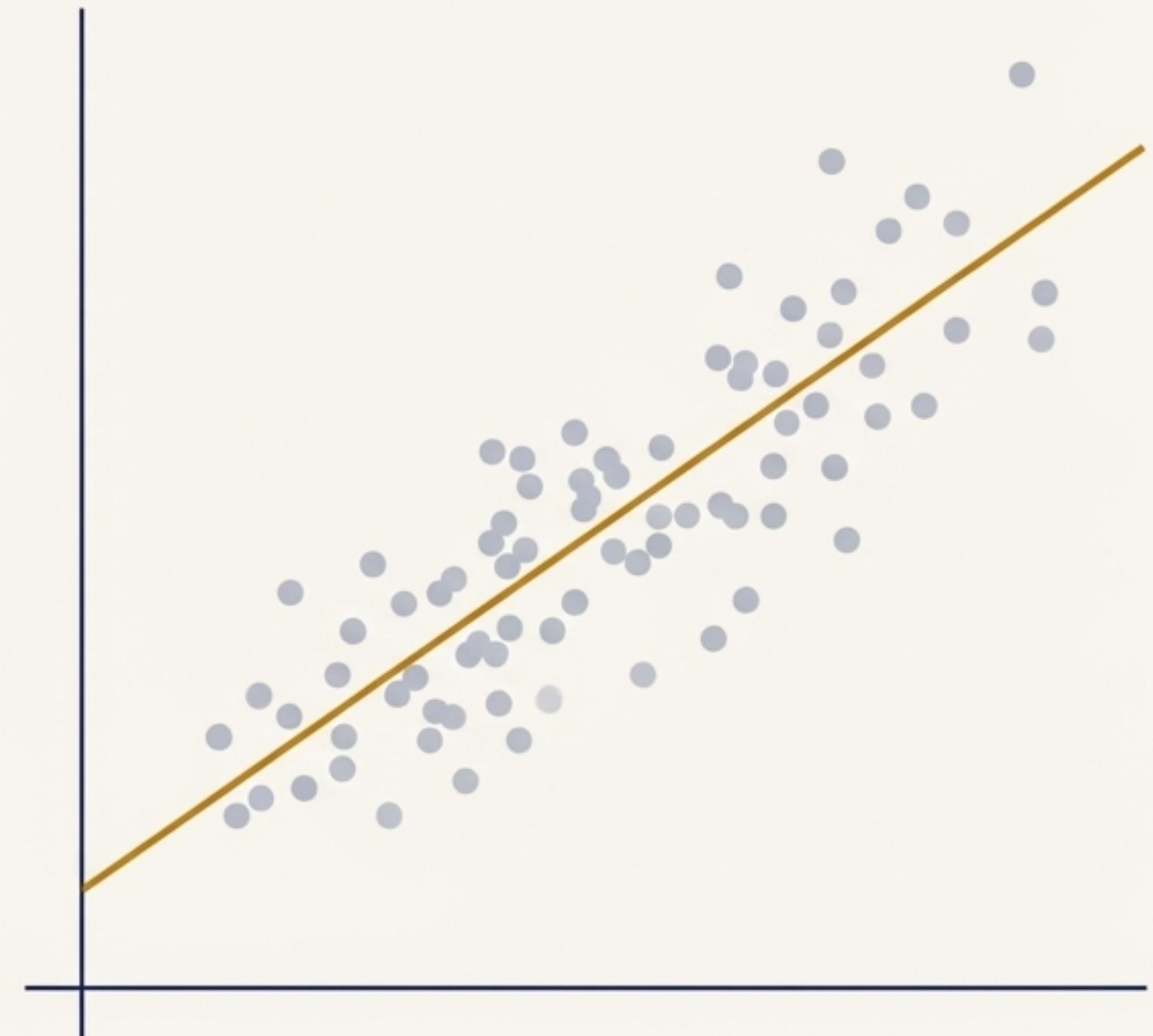
Y: Zmienna zależna (docelowa)

X: Zmienna niezależna (cecha)

$\beta_0$ : Wyraz wolny (intercept) – wartość Y, gdy X=0.

$\beta_1$ : Współczynnik nachylenia (slope) – zmiana w Y przy jednostkowej zmianie X.

$\epsilon$ : Błąd losowy (error term)



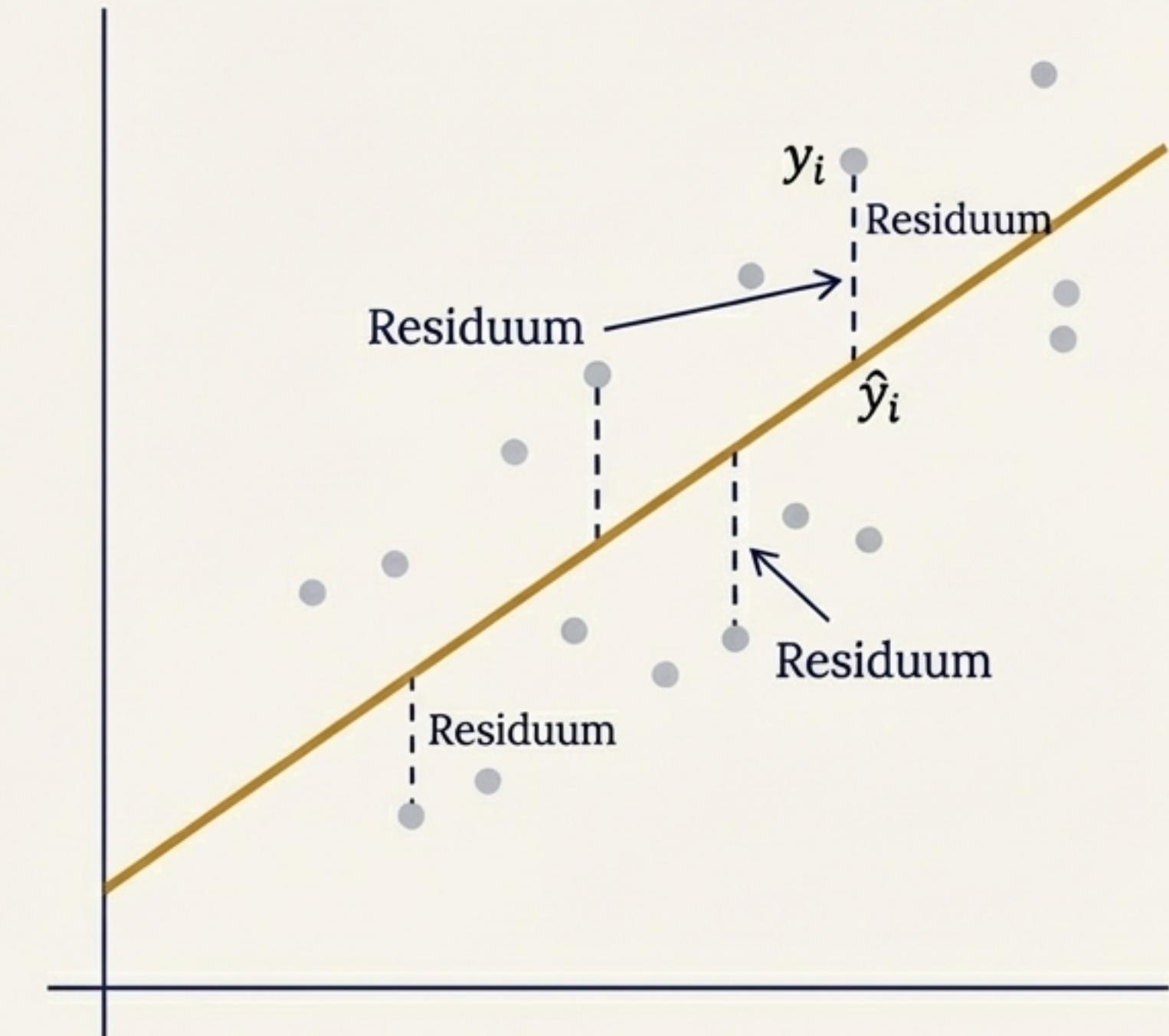
# Mechanika Modelu: Jak Znaleźć „Najlepszą” Linię?

Model liniowy może być reprezentowany przez nieskończoną liczbę linii. „Najlepsza” linia to ta, która znajduje się „najbliżej” wszystkich punktów danych jednocześnie. Aby to zmierzyć, używamy pojęcia **residuów**.

## Definicja Residuów (Błędów):

Residuum to różnica między obserwowaną, rzeczywistą wartością  $y_i$  a wartością przewidywaną przez model  $\hat{y}_i$  (czyli punktem na linii regresji). Jest to pionowa odległość od punktu do linii.

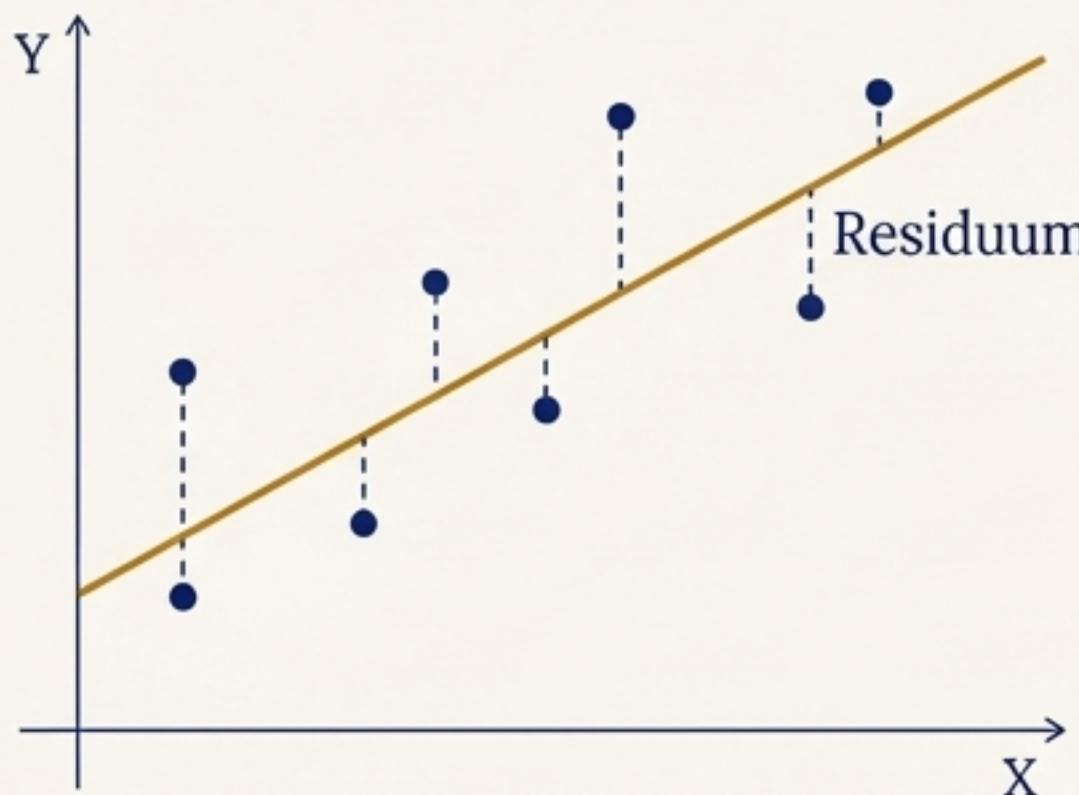
$$\text{Residuum}_i = y_i - \hat{y}_i$$



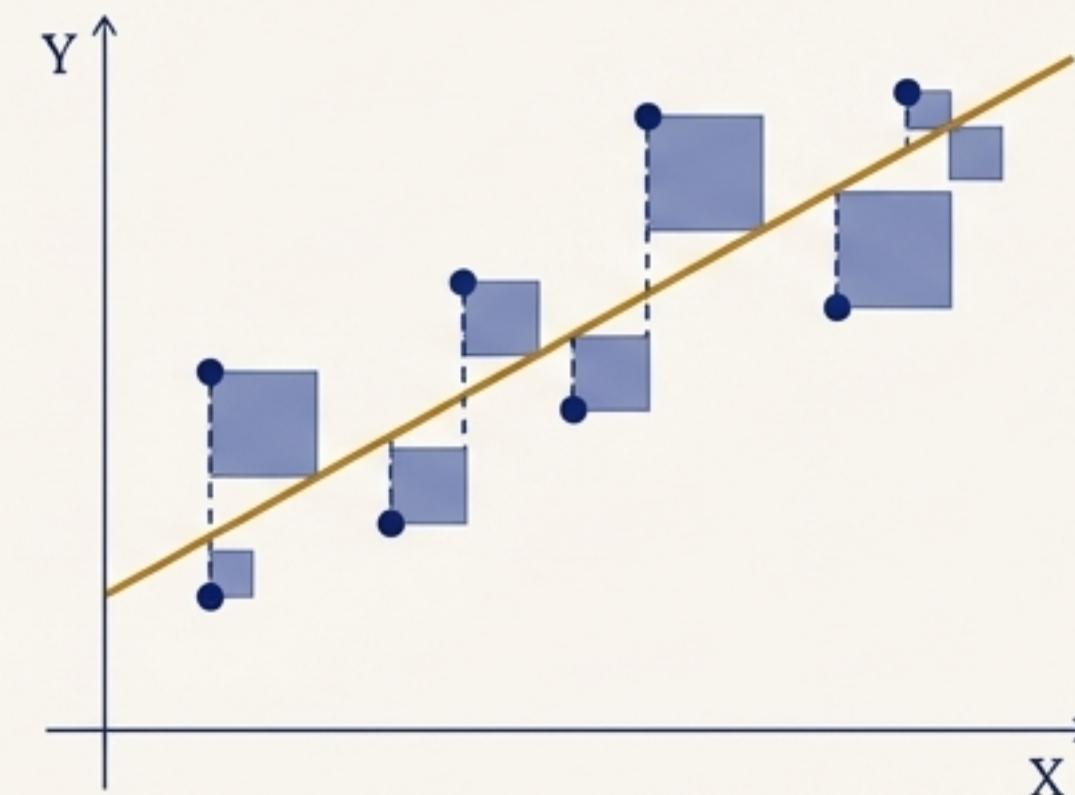
# Metoda Budowy: Zwykła Metoda Najmniejszych Kwadratów (OLS)

Zwykła Metoda Najmniejszych Kwadratów (Ordinary Least Squares, OLS) to najpopularniejsza technika estymacji parametrów ( $\beta_0$  i  $\beta_1$ ) w regresji liniowej. Celem OLS jest znalezienie takiej linii, dla której **suma kwadratów residuów ( błędów)** jest jak najmniejsza.

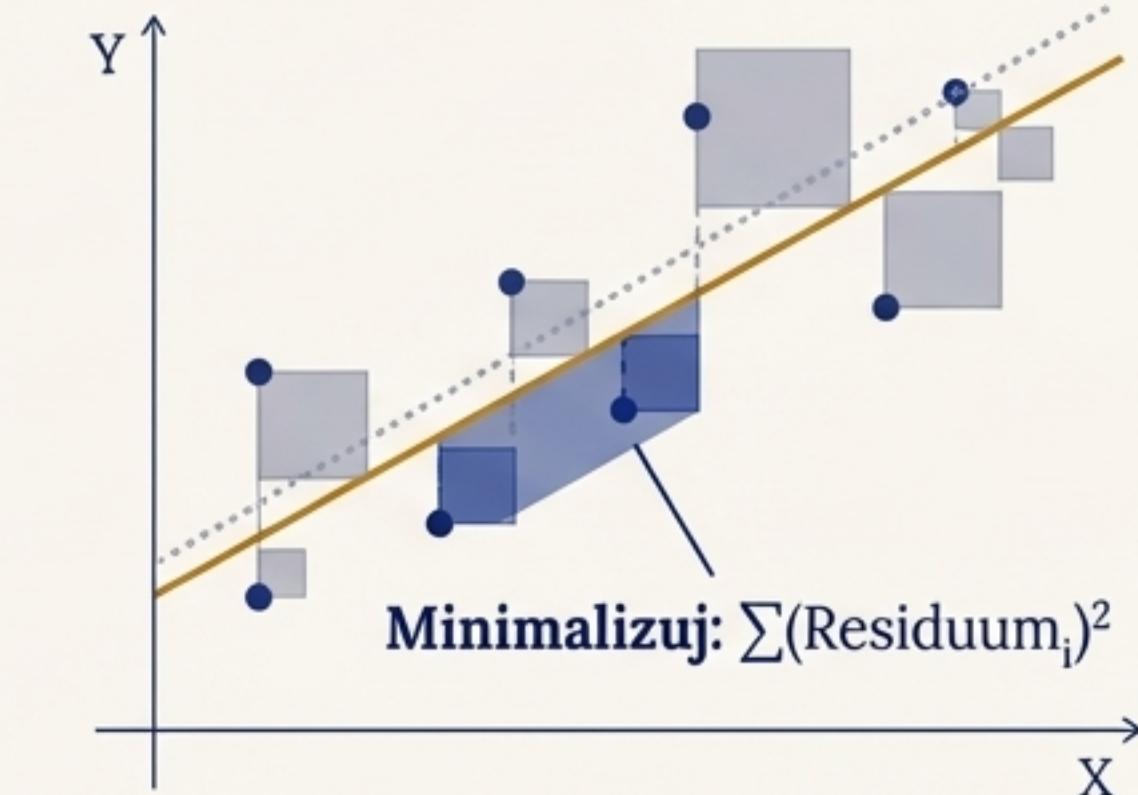
## 1. Oblicz residua ( błędy )



## 2. Podnieś błędy do kwadratu



## 3. Znajdź linię minimalizującą sumę kwadratów



## Dlaczego kwadraty? (1)

**Eliminacja znaku:** Podniesienie do kwadratu sprawia, że błędy dodatnie (nad estymacją) i ujemne (pod estymacją) nie znoszą się nawzajem.

## Dlaczego kwadraty? (2)

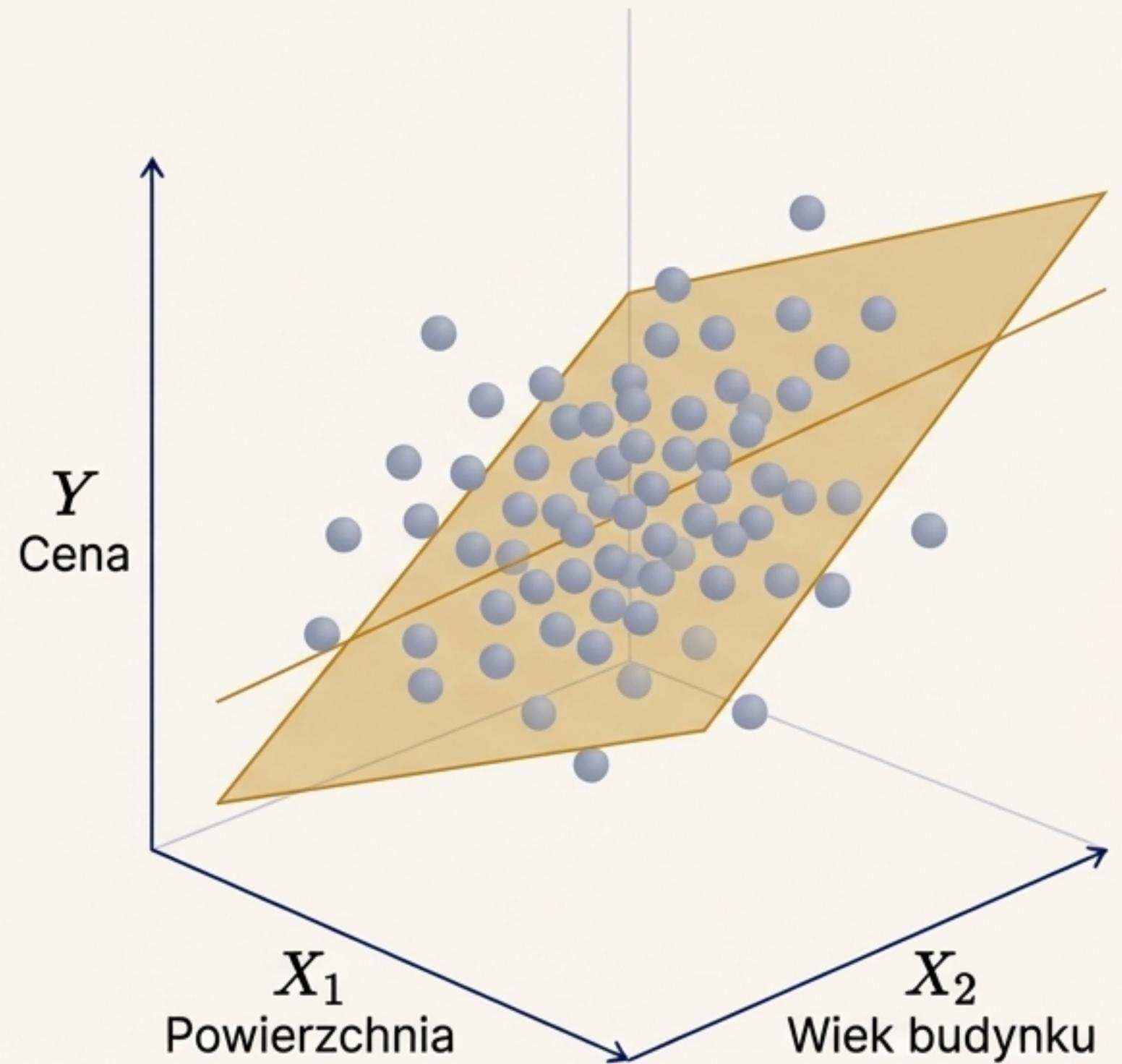
**Większa kara za duże błędy:** Kwadratowa funkcja straty „mocniej karze” duże odchylenia od linii niż małe, co sprawia, że model stara się unikać dużych pomyłek.

# Od Modelu Prostego do Wielorakiego

Rzadko kiedy jedna cecha ( $X$ ) w pełni wyjaśnia zmienną docelową ( $Y$ ). Regresja wieloraka (Multiple Linear Regression) rozszerza model prosty, pozwalając na uwzględnienie wielu zmiennych niezależnych jednocześnie. Zasada OLS pozostaje ta sama: minimalizujemy sumę kwadratów residiów, ale w wielowymiarowej przestrzeni.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

**Interpretacja współczynników  $\beta$ :** Każdy współczynnik  $\beta_j$  reprezentuje zmianę w  $Y$  związaną z jednostkową zmianą w  $X_j$ , przy założeniu, że **wszystkie pozostałe zmienne (predyktory) pozostają stałe**. To kluczowe dla analizy przyczynowej.

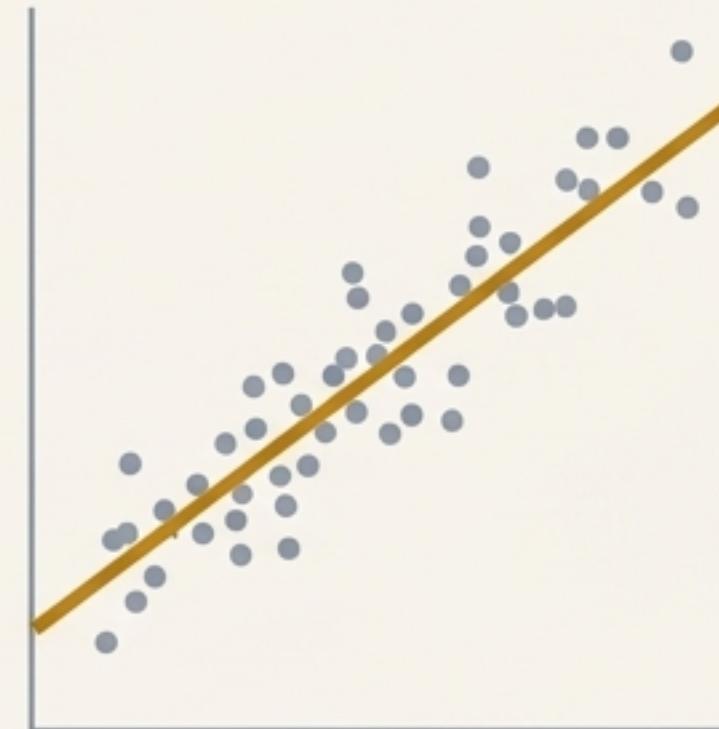


# Zasady Wiarygodności: Założenia Modelu OLS (Część 1)

Aby estymatory OLS były nieobciążone i efektywne, a wyniki modelu wiarygodne, dane powinny spełniać kilka kluczowych założeń. Traktuj je jako listę kontrolną dla rzetelnej analizy.

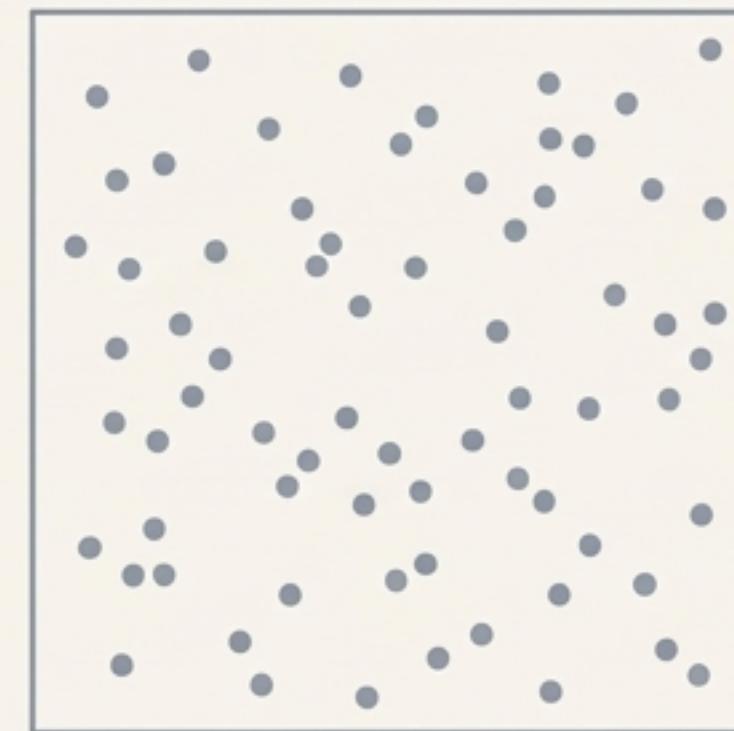
## 1. Liniowość

Relacja między zmiennymi niezależnymi a zmienną zależną jest liniowa. Model musi być „liniowy względem parametrów”.



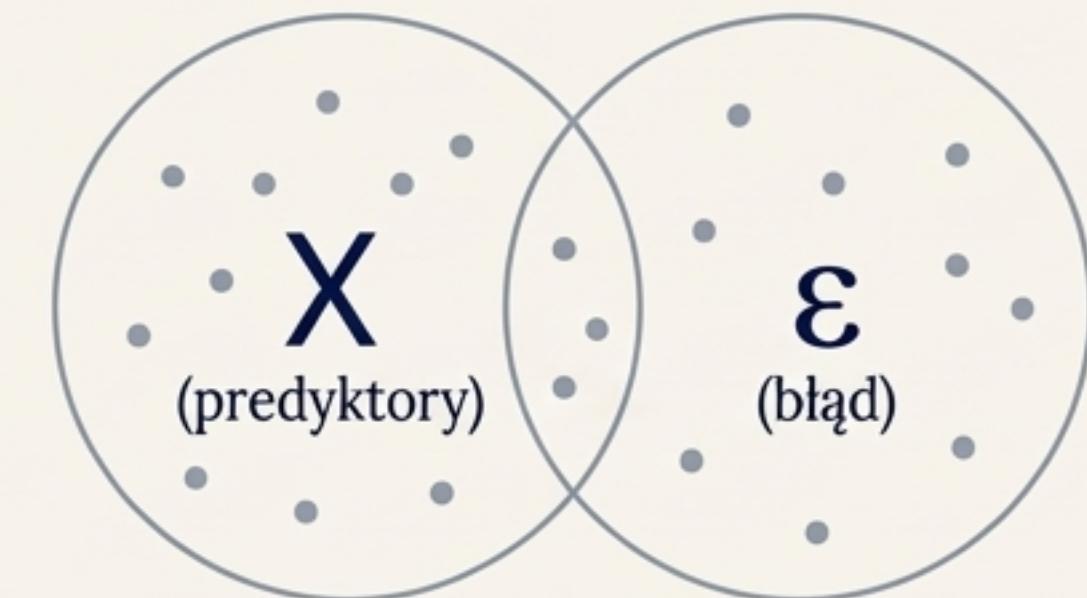
## 2. Losowość Próby / Niezależność Błędów

Obserwacje są losowo próbkiowane z populacji. Błędy (residua) poszczególnych obserwacji są od siebie niezależne.



## 3. Egzogeniczność

Zmienne niezależne (predyktory) nie są skorelowane ze składnikiem losowym (błędem). Oznacza to, że w błędzie nie ma ukrytych istotnych zmiennych, które są jednocześnie skorelowane z naszymi predyktorami (problem pominiętej zmiennej).

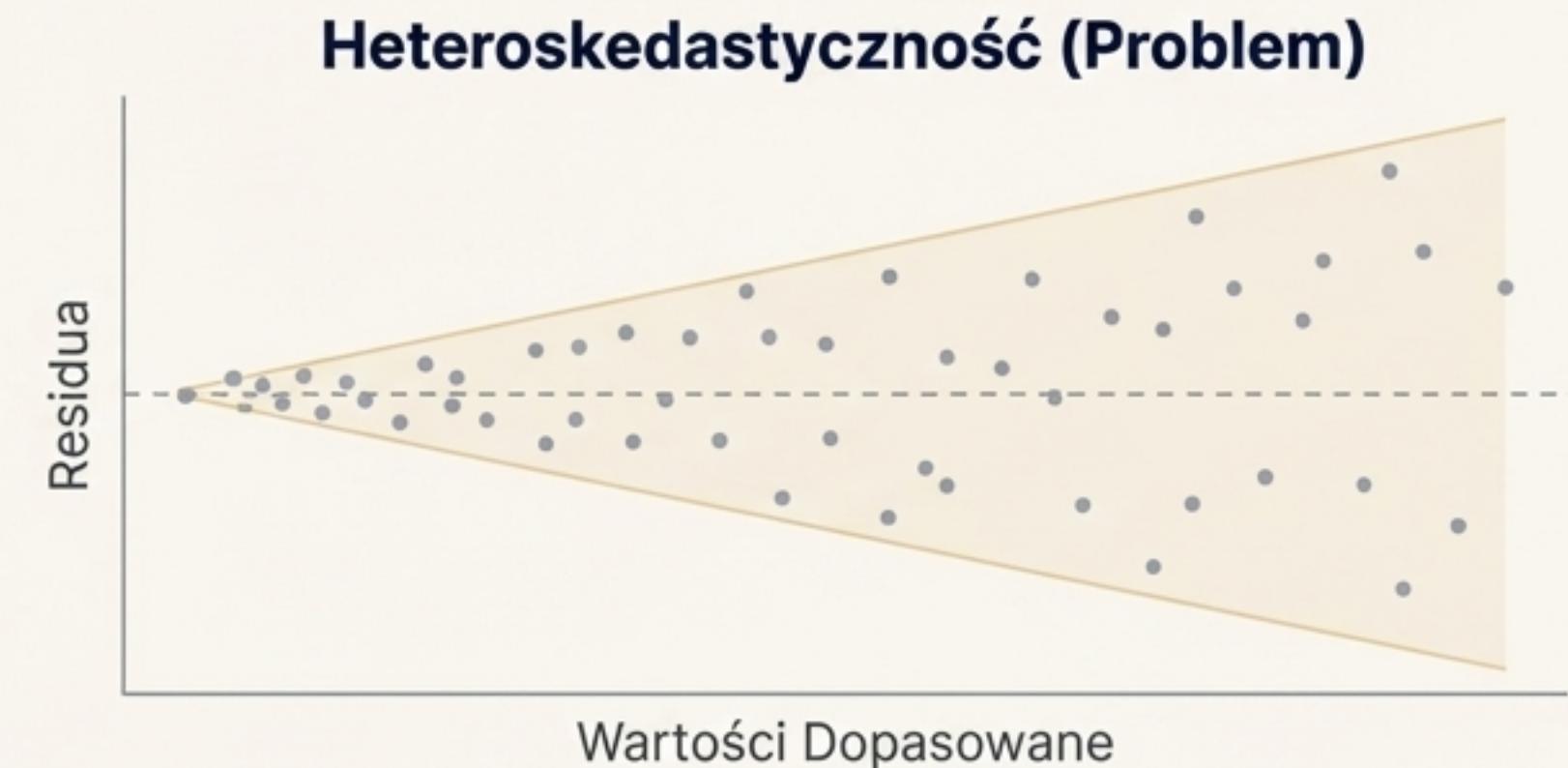


# Zasady Wiarygodności: Założenia Modelu OLS (Część 2)

## 4. Homoskedastyczność

Wariancja składnika losowego jest stała dla wszystkich wartości zmiennych niezależnych. Mówiąc prościej, rozrzut residuów powinien być taki sam na całej długości linii regresji.

**Problem:** **Heteroskedastyczność** (zmienna wariancja) prowadzi do nieefektywnych estymatorów i błędnych wnioskowań statystycznych (nieprawidłowe błędy standardowe i  $p$ -wartości).



## 5. Brak Doskonałej Współliniowości

Żadna ze zmiennych niezależnych nie jest idealną liniową kombinacją innych zmiennych niezależnych. Wysoka korelacja między predyktorami może powodować niestabilność oszacowań współczynników.

# Karta Oceny Modelu: Mierzenie Błędów Predykcji

Po zbudowaniu modelu musimy ocenić, jak dobrze działa. Podstawowe metryki oceniają średnią wielkość błędów predykcji.

## Porównanie Metryk Błędów

Metryka	Formuła	Interpretacja	Zalety/Wady
Błąd Średniokwadratowy (Mean Squared Error - MSE)	$MSE = \frac{1}{n} \sum (y_i - \hat{y}_i)^2$	Średnia kwadratów błędów. Jest to metryka, którą OLS bezpośrednio minimalizuje.	Jednostki są podniesione do kwadratu (np. dolary <sup>2</sup> ), co utrudnia interpretację. Wrażliwa na obserwacje odstające (outliers).
Pierwiastek Błędu Średniokwadratowego (Root Mean Squared Error - RMSE)	$RMSE = \sqrt{MSE}$	Wyrażony w tych samych jednostkach co zmienność docelowa (np. w dolarach), co czyni go intuicyjnym. Reprezentuje typową, średnią odległość między wartościami przewidywanymi a rzeczywistymi.	Bardziej zrozumiałym niż MSE. Nadal wrażliwy na obserwacje odstające, ale w mniejszym stopniu niż MSE.

# Karta Oceny Modelu: Współczynnik Determinacji ( $R^2$ )

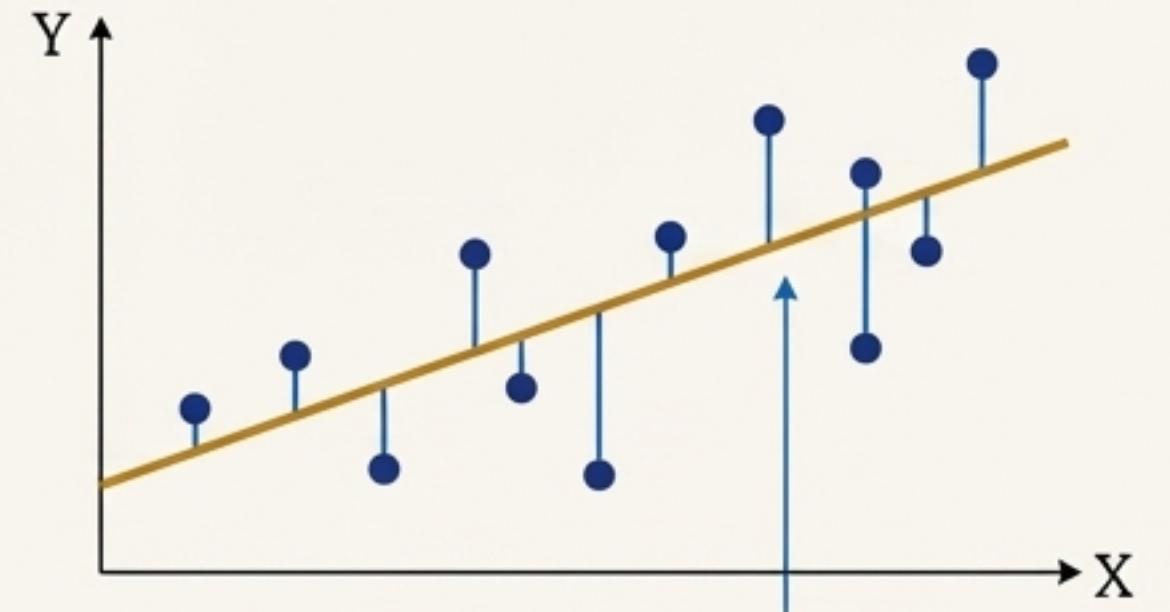
**R-kwadrat ( $R^2$ )** to metryka, która odpowiada na pytanie: „Jaki procent wariancji w zmiennej zależnej jest wyjaśniany przez nasz model?”. Porównuje błąd naszego modelu (**RSS**) do błędu modelu, który przewidywałby po prostu średnią wartość Y dla każdej obserwacji (**TSS** - Total Sum of Squares).

## Interpretacja:

- $R^2$  przyjmuje wartości od 0 do 1.
- $R^2$  przyjmuje wartości od 0 do 1.
- $R^2 = 0$ : Model nie wyjaśnia wariancji lepiej niż prosta średnia.
- $R^2 = 1$ : Model doskonale wyjaśnia całą wariancję w danych.
- Np.  $R^2 = 0.65$  oznacza, że 65% zmienności w Y można解释 za pomocą zmiennych X zawartych w modelu.

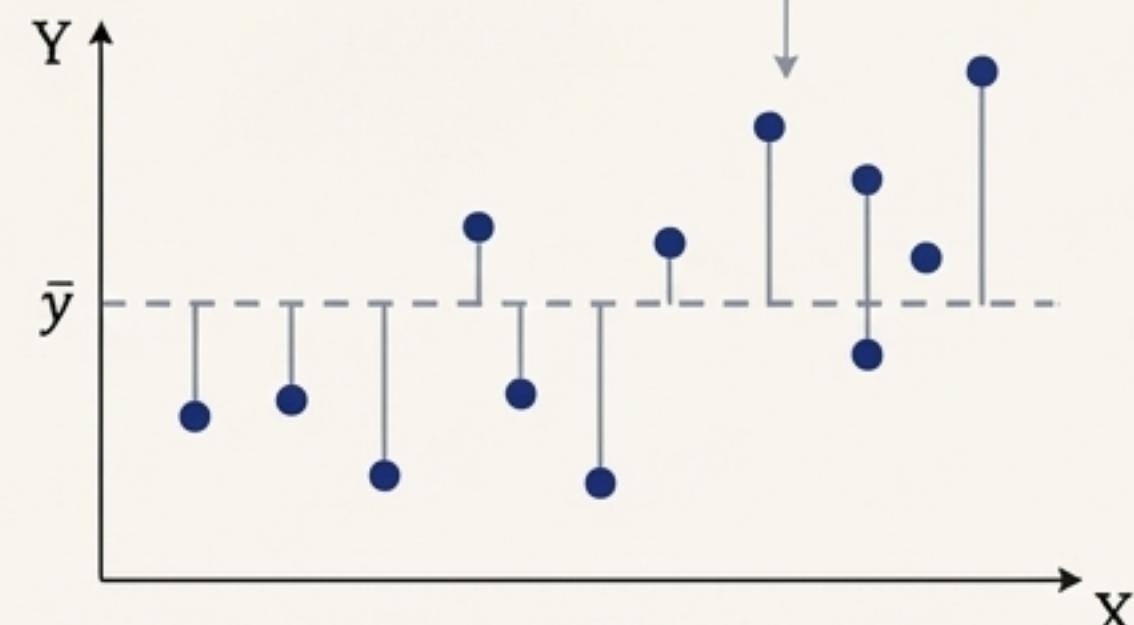
## Uwaga:

Wysoki  $R^2$  nie zawsze oznacza dobry model. Może on rosnąć wraz z dodawaniem kolejnych, nawet nieistotnych zmiennych. Dlatego często używa się **Skorygowanego  $R^2$  (Adjusted  $R^2$ )**, który uwzględnia liczbę predyktorów w modelu.



$$\text{Suma kwadratów residuów (RSS)} = \sum(y_i - \hat{y}_i)^2$$

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$



$$\text{Całkowita suma kwadratów (TSS)} = \sum(y_i - \bar{y})^2$$

# Zastosowanie w Praktyce: Analiza Cen Nieruchomości w Kalifornii

## Kontekst Problemu

Wykorzystamy publicznie dostępny zbiór danych „California Housing Prices” z 1990 roku. Każdy wiersz reprezentuje „grupę bloków” (najmniejszą jednostkę geograficzną spisu powszechnego).

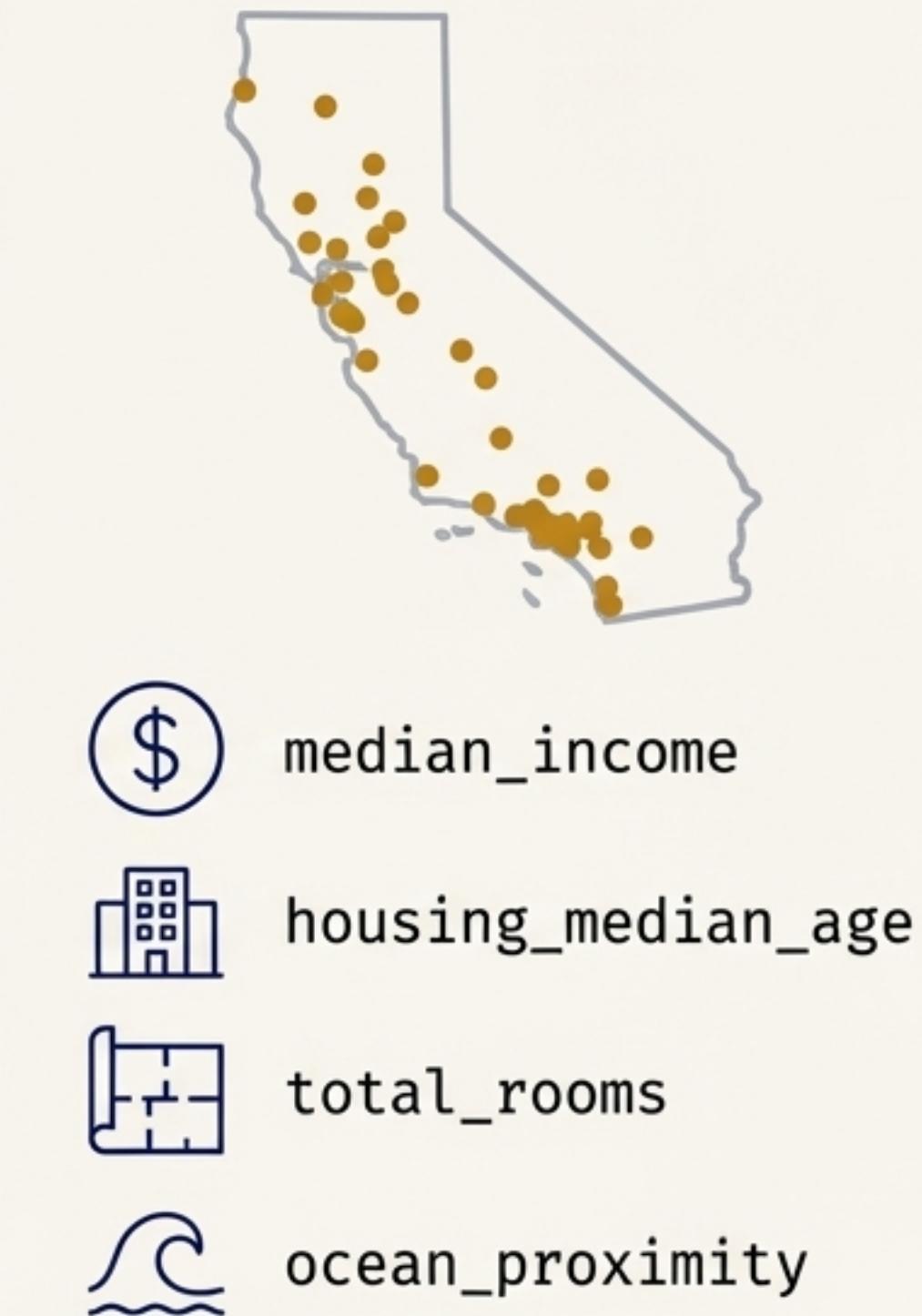
## Cele Analizy

- Cel Analityczny (Analiza Przyczynowa):** Zidentyfikować, które cechy bloków (np. średni dochód, wiek budynków, liczba pokoi) mają największy, statystycznie istotny wpływ na medianę wartości domów.
- Cel Predykcyjny:** Zbudować model, który będzie w stanie prognozować medianę wartości domów dla nowych bloków na podstawie ich cech.

## Zmienne

Zmienna Zależna (Y): `median\_house\_value`

Przykładowe Zmienne Niezależne (X): `median\_income`,  
`housing\_median\_age`, `total\_rooms`, `population`,  
`ocean\_proximity`, etc.



# Krok 1: Przygotowanie Danych do Analizy

Surowe dane rzadko kiedy są gotowe do modelowania. Dwa kluczowe kroki w naszym przypadku to obsługa brakujących wartości i analiza współliniowości.

## 1. Obsługa Brakujących Danych

- Problem:** Zmienna `total_bedrooms` zawiera 207 brakujących wartości (ok. 1% zbioru).
- Rozwiążanie:** Ze względu na niewielki odsetek, najprostszym i skutecznym rozwiązaniem jest usunięcie wierszy z brakującymi danymi.

## 2. Analiza Współliniowości

- Problem:** Założenie OLS o braku doskonałej współliniowości. Musimy sprawdzić, czy nasze predyktory nie są ze sobą silnie skorelowane.
- Narzędzie:** Macierz korelacji zwizualizowana jako mapa ciepła (heatmap).
- Wniosek:** Mapa ciepła wykazała bardzo wysoką korelację ( $>0.9$ ) między `total_rooms`, `total_bedrooms` i `households`.
- Rozwiążanie:** Aby uniknąć problemów z niestabilnością modelu, usunięto zmienną `total_bedrooms`.

## Macierz Korelacji (Fragment)

	total_rooms	total_bedrooms	population	households	median_income	
total_rooms		0.8	0.7	0.93		~0.1
total_bedrooms	0.8		0.5	0.88		~0.2
population	0.8	0.5		0.56		~0.1
households	0.93	0.93	0.61		0.93	
median_income	~0.1	~0.2	~0.1	0.93		

Low Correlation

High Positive Correlation

# Krok 2: Dopasowanie Modelu i Interpretacja Wyników

Po przygotowaniu danych, dopasowujemy model OLS przy użyciu biblioteki `statsmodels` w Pythonie. Poniżej znajduje się uproszczony i opatrzony adnotacjami wynik.

## Wrzygotowanie modelu

```
# Przygotowanie danych
X = data[['median_income', 'housing_median_age', ...]]
y = data['median_house_value']
X = sm.add_constant(X) # Dodanie wyrazu wolnego

# Podział na zbiór treningowy i testowy
X_train, X_test, y_train, y_test = train_test_split(...)

# Dopasowanie modelu
model = sm.OLS(y_train, X_train).fit()
print(model.summary())
```

## Wyniki z Adnotacjami

R-squared: **0.59**

Model wyjaśnia 59% wariancji cen domów.

Zmienna	Coeff.	Std.Err.	P> t
const	-2.1e+06	...	0.000
median_income	38449.2	...	<b>0.000</b>
housing_median_age	846.1	...	<b>0.000</b>
...	...	...	...

P<0.05, zmienna statystycznie istotna. Wzrost dochodu o 1 jednostkę (\$10k) zwiększa cenę domu o ~\$38k, ceteris paribus.

Starsze domy są droższe. Każdy dodatkowy rok wieku budynku to wzrost ceny o ~\$846.

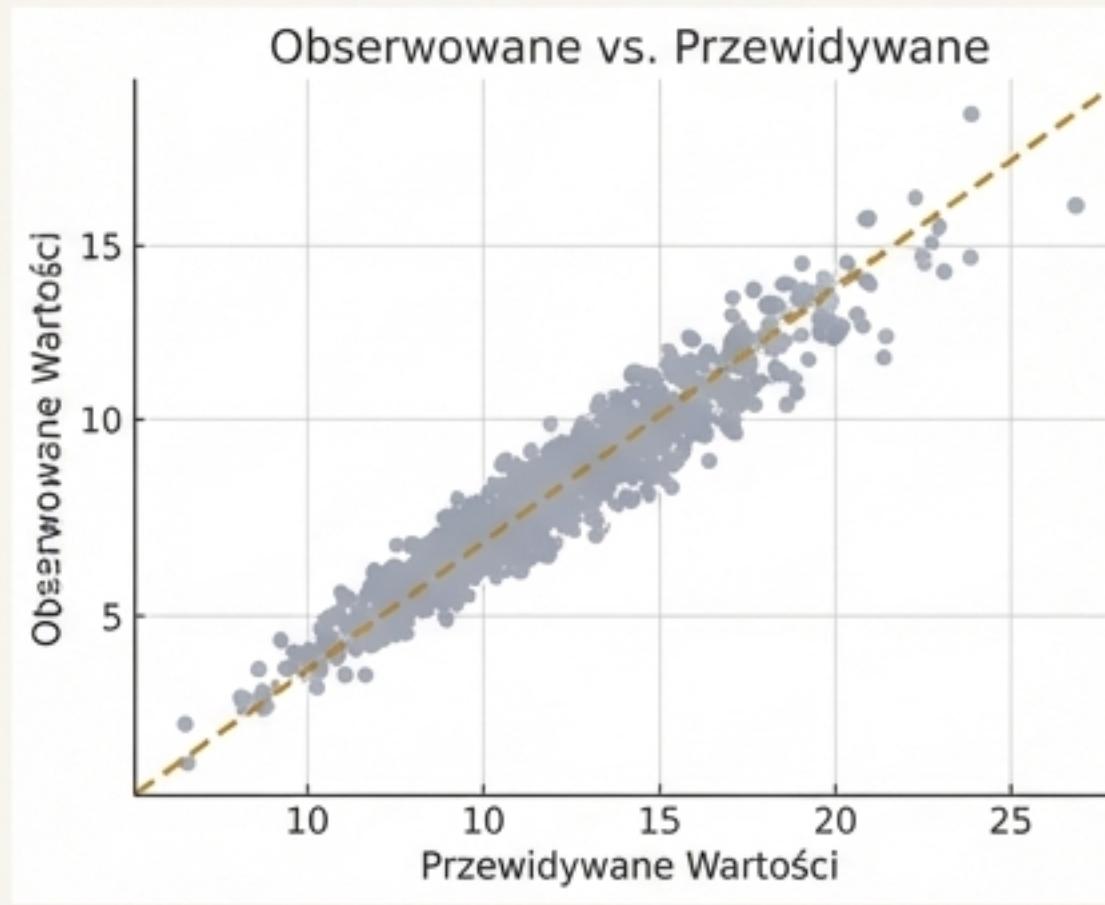
# Krok 3: Weryfikacja Modelu - Sprawdzanie Założeń

Dopasowanie modelu to nie koniec. Musimy sprawdzić, czy kluczowe założenia OLS zostały spełnione, aby móc ufać wynikom, zwłaszcza błędem standardowym i p-wartościom.

## 1. Weryfikacja Liniowości

**Metoda:** Wykres wartości obserwowanych ('y\_test') względem wartości przewidywanych.

**Wynik:** Punkty układają się wzduż linii prostej, co sugeruje, że założenie o liniowości jest w dużej mierze spełnione.

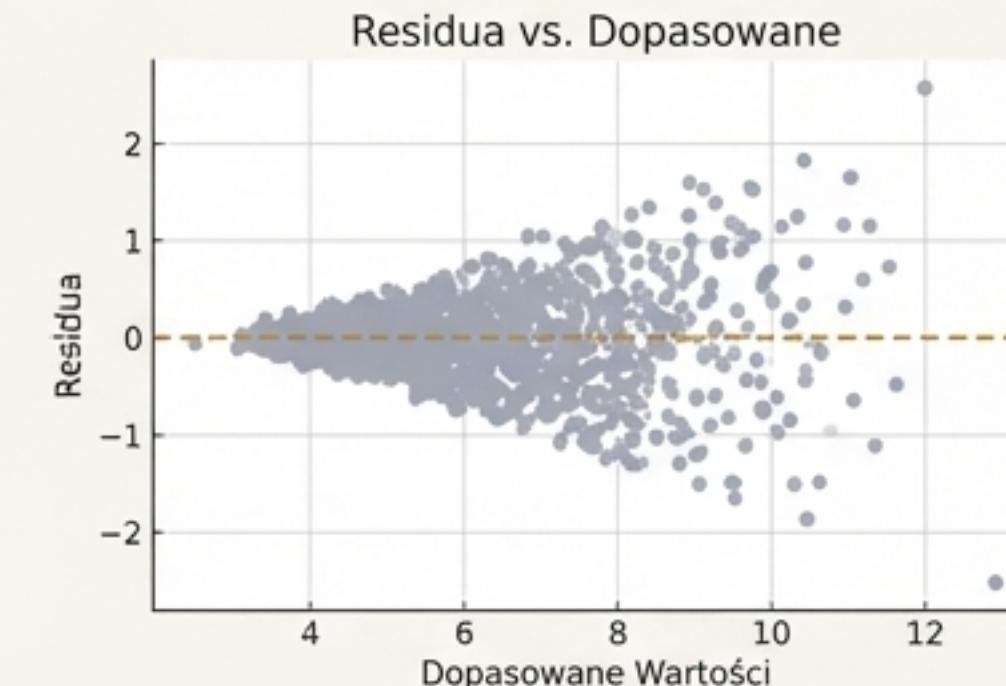


## 2. Weryfikacja Homoskedastyczności

**Metoda:** Wykres residuów względem wartości dopasowanych.

**Wynik:** Wykres pokazuje wyraźny kształt lejka (rozszerszający się rozrzut błędów dla wyższych przewidywanych cen). Jest to klasyczny objaw **heteroskedastyczności**.

**Implikacje:** Oznacza to, że nasze błędy standardowe i p-wartości z poprzedniego slajdu mogą być niewiarygodne. W praktyce wymagałyby to zastosowania bardziej zaawansowanych technik (np. odpornych błędów standardowych).

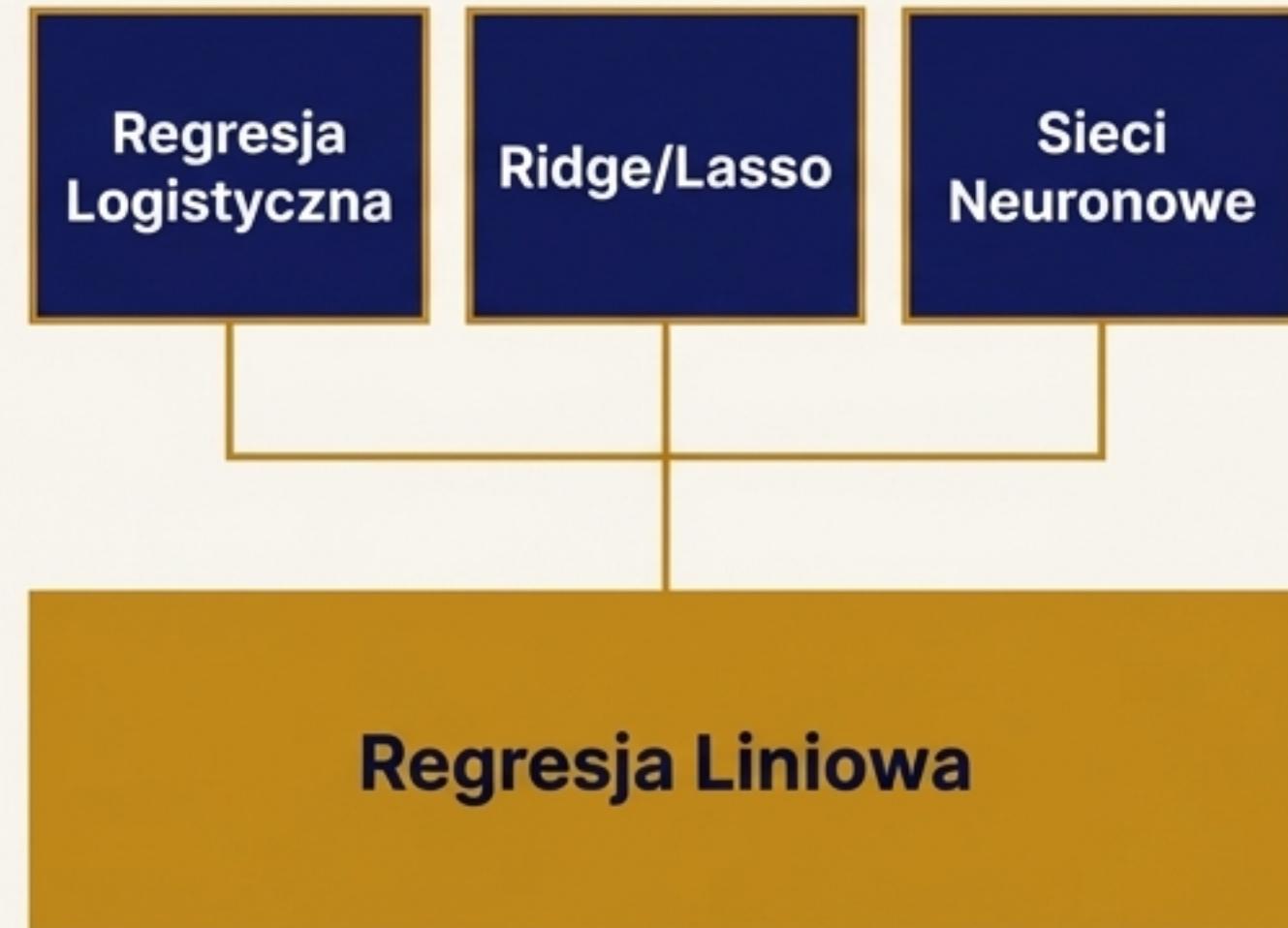


# Dziedzictwo Kamienia Węgielnego

Regresja liniowa, mimo swojej prostoty, pozostaje jednym z najważniejszych narzędzi w nauce o danych. Jej siła leży w równowadze między wydajnością a przejrzystością.

## Kluczowe Zalety

- **Interpretowalność:** Pozwala precyzyjnie zrozumieć i zmierzyć wpływ poszczególnych czynników. Jest to „biała skrzynka”, w przeciwieństwie do wielu złożonych modeli.
- **Prostota i Wydajność:** Jest szybka w obliczeniach i łatwa do wdrożenia, co czyni ją idealnym modelem bazowym (baseline).
- **Fundament:** Zrozumienie regresji liniowej jest kluczem do zrozumienia bardziej zaawansowanych modeli.



## Kluczowe Ograniczenia

- **Silne Założenia:** Wymaga spełnienia restrykcyjnych założeń, aby wyniki były w pełni wiarygodne.
- **Wrażliwość na Outliery:** Ekstremalne wartości mogą znacząco wpływać na linię regresji.
- **Modeluje tylko relacje liniowe:** Nie jest w stanie uchwycić złożonych, nieliniowych wzorców w danych.