

ML: Next Step

#1 Wstęp i podstawy

Plan zajęć



Forms

1

Wstęp i podstawy

2

Dane - kluczowy element ML

3

Klasyfikacja

4

Regresja

5

Grupowanie

Me: *uses machine learning*

Machine: *learns*

Me:

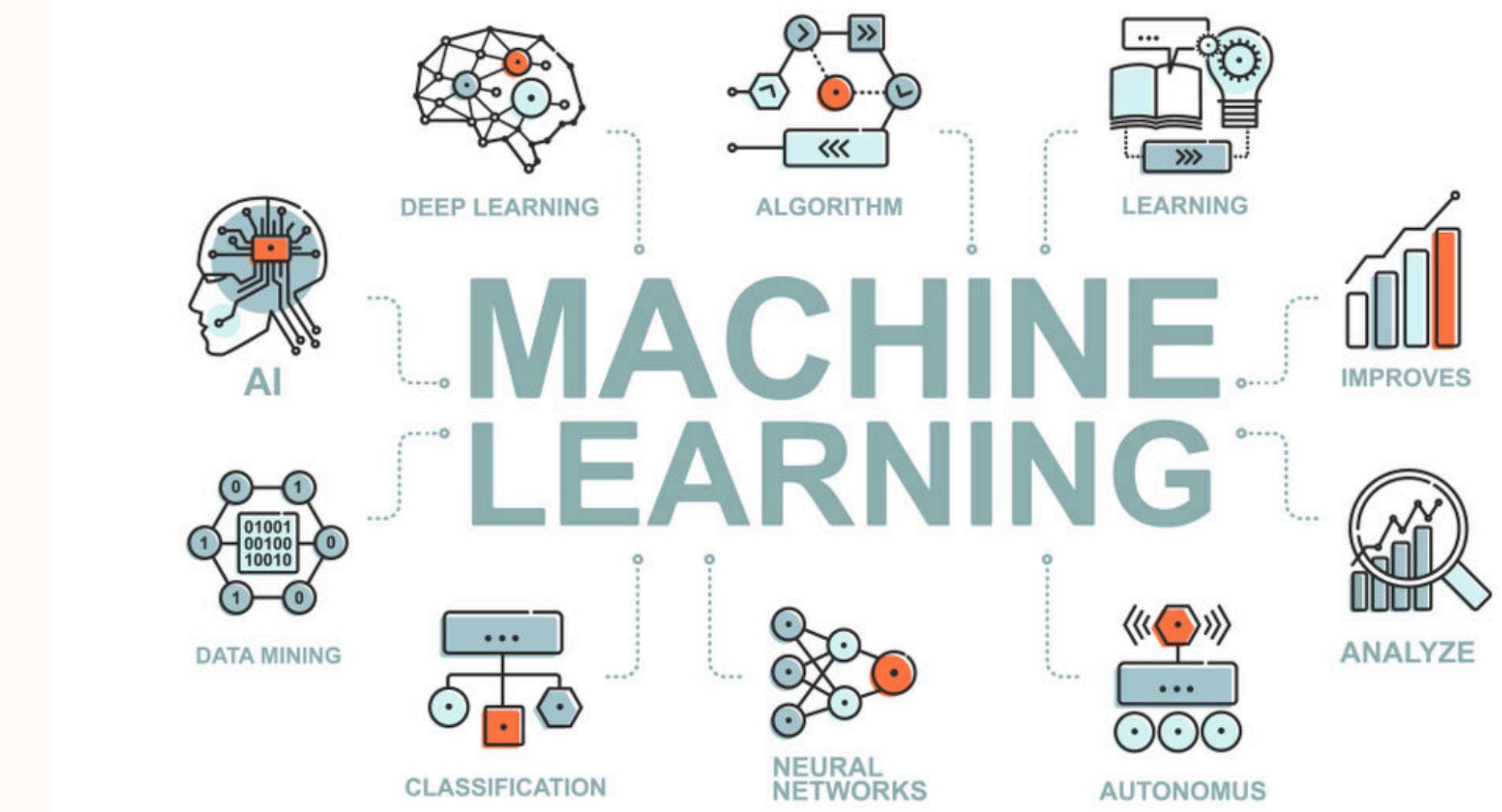


Wstęp i podstawy

Czym jest Uczenie Maszynowe?

Uczenie maszynowe (ML) to rodzaj sztucznej inteligencji (AI), która pozwala komputerom uczyć się bez sprecyzowanego (explicit) programowania . Polega na wprowadzaniu danych do algorytmów, które mogą następnie identyfikować wzory w datasetach i tworzyć prognozy na podstawie nowych danych.

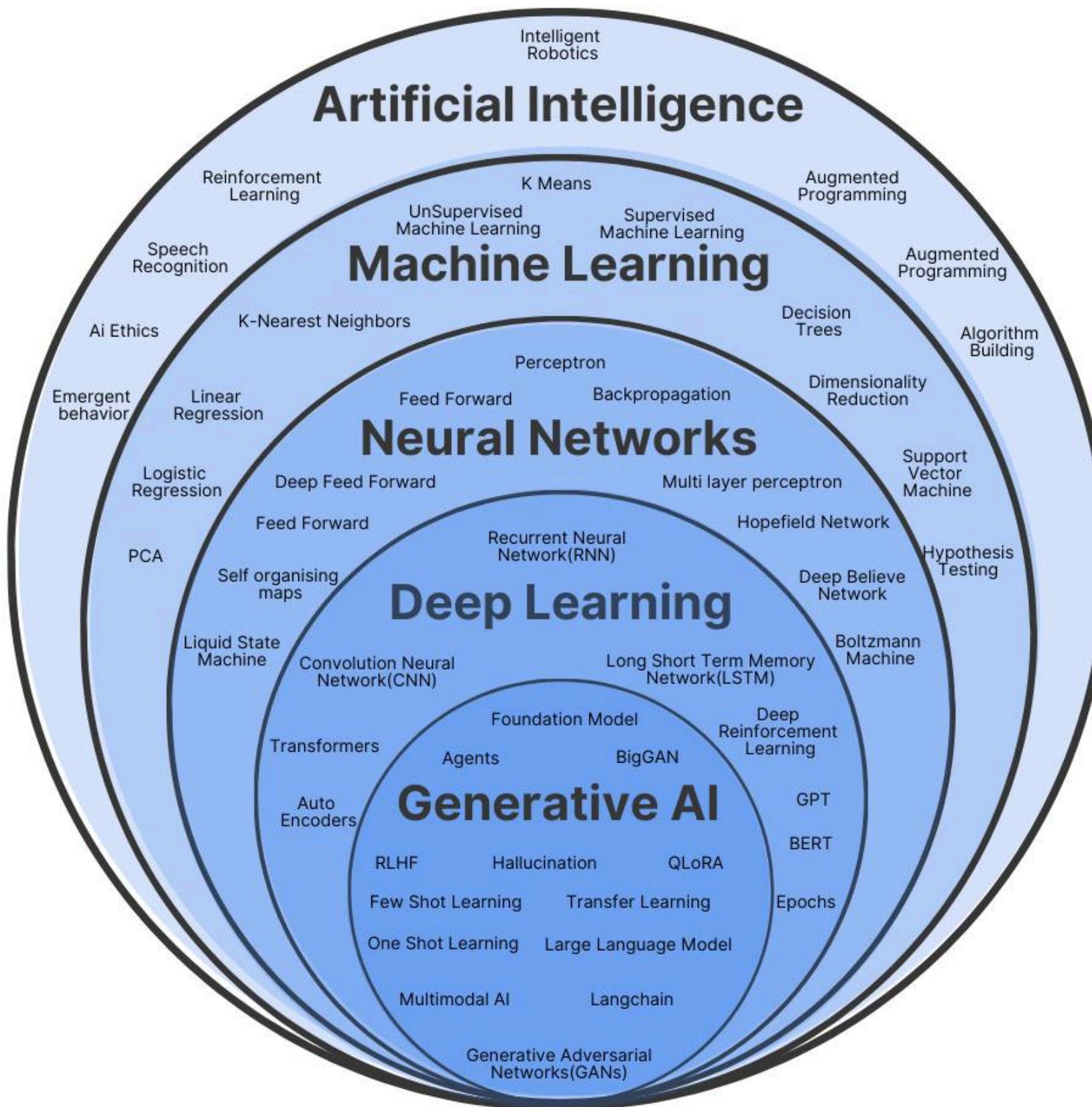
Główna idea: automatyczne znajdowanie wzorców w danych bez programowania "ręcznego"



Czym jest model?

Model ML to funkcja matematyczna, która na podstawie danych wejściowych przewiduje wynik. W procesie trenowania model uczy się odpowiednich wartości parametrów, np. współczynników w regresji.

The World of Artificial Intelligence



Artificial Intelligence
Engineering of making intelligent machines and programs.

Machine Learning
Ability to learn without being explicitly programmed.

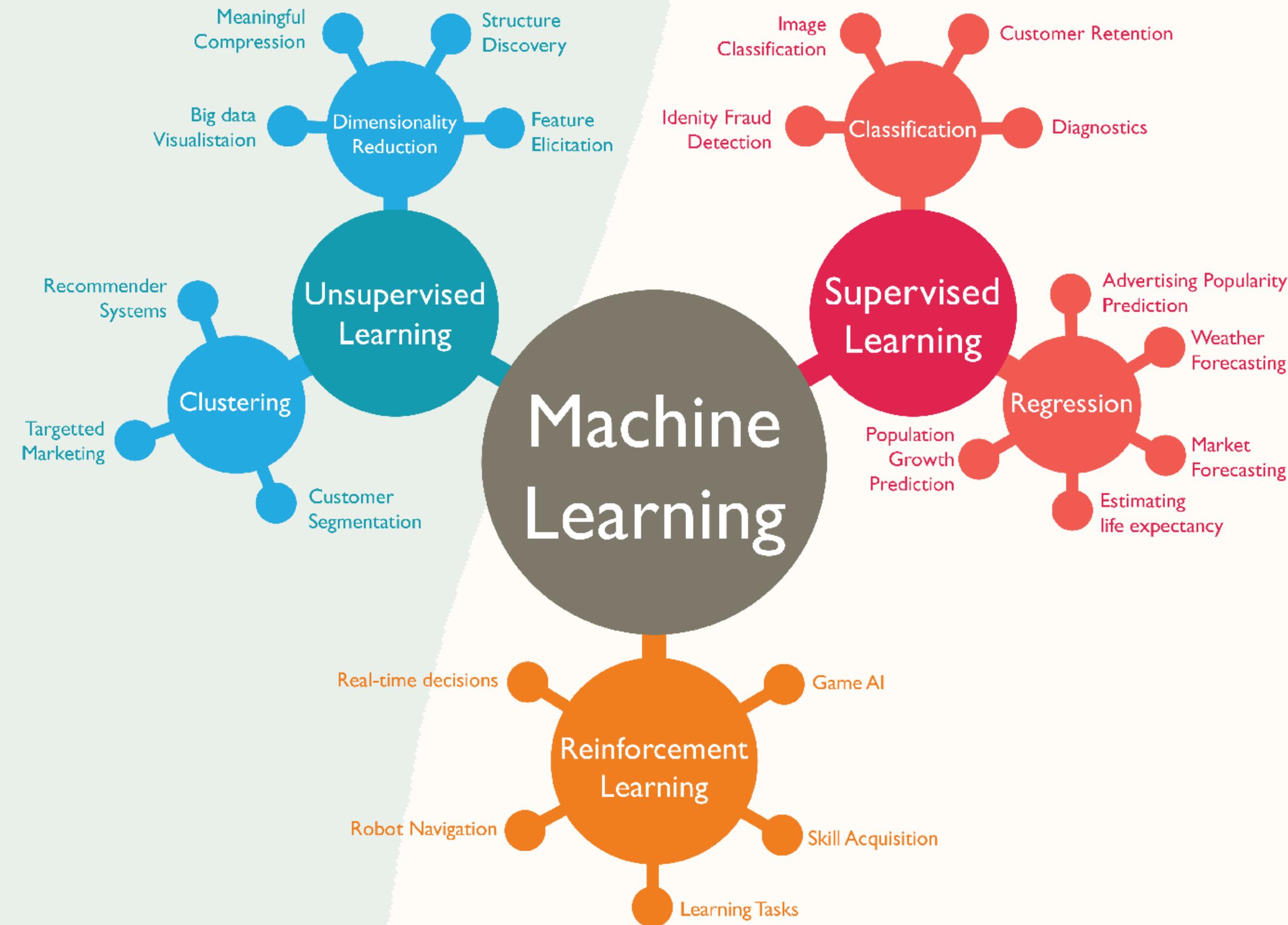
Deep Learning
Learning based on deep neural networks.

Przykłady zastosowań ML

1. Filtry antyspamowe w poczcie elektronicznej
2. Rozpoznawanie obrazów i twarzy
3. Systemy rekomendacyjne
4. Autonomiczne pojazdy
5. Personalizacja reklam
6. Analiza rynku i przewidywanie trendów w finansach
7. Analiza sentymentu w mediach społecznościowych

Rodzaje Uczenia Maszynowego

Machine Learning



Rodzaje Uczenia Maszynowego

- Uczenie nadzorowane (supervised): mamy dane + etykiety
- Uczenie nienadzorowane (unsupervised): tylko dane, bez etykiet
- Uczenie półnadzorowane (semisupervised): część danych ma etykiety, część ich nie ma
- Uczenie przez wzmacnianie (reinforcement): nagrody i kary

Proces Uczenia Maszynowego

- 1.Zbieranie danych
- 2.Przygotowanie danych
- 3.Budowanie i trenowanie modelu
- 4.Testowanie i optymalizacja modelu
- 5.Użycie modelu w rzeczywistych zastosowaniach

Wyzwania w Uczaniu Maszynowym

overfitting

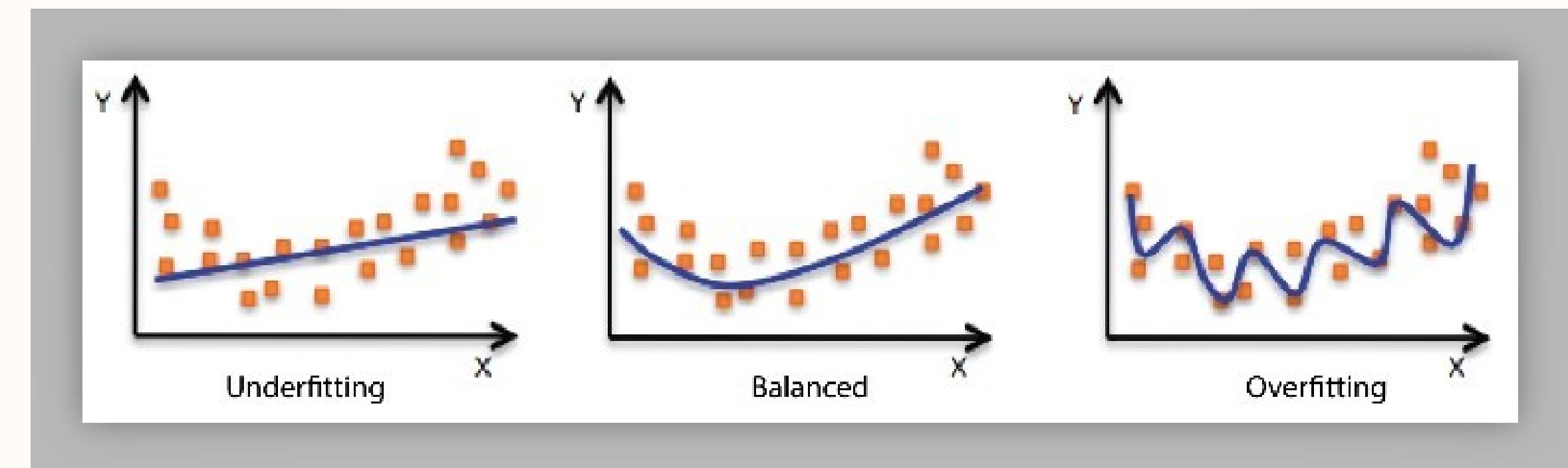
underfitting

Niska jakość lub brakujące dane

Problemy ze zróżnicowaniem
danych

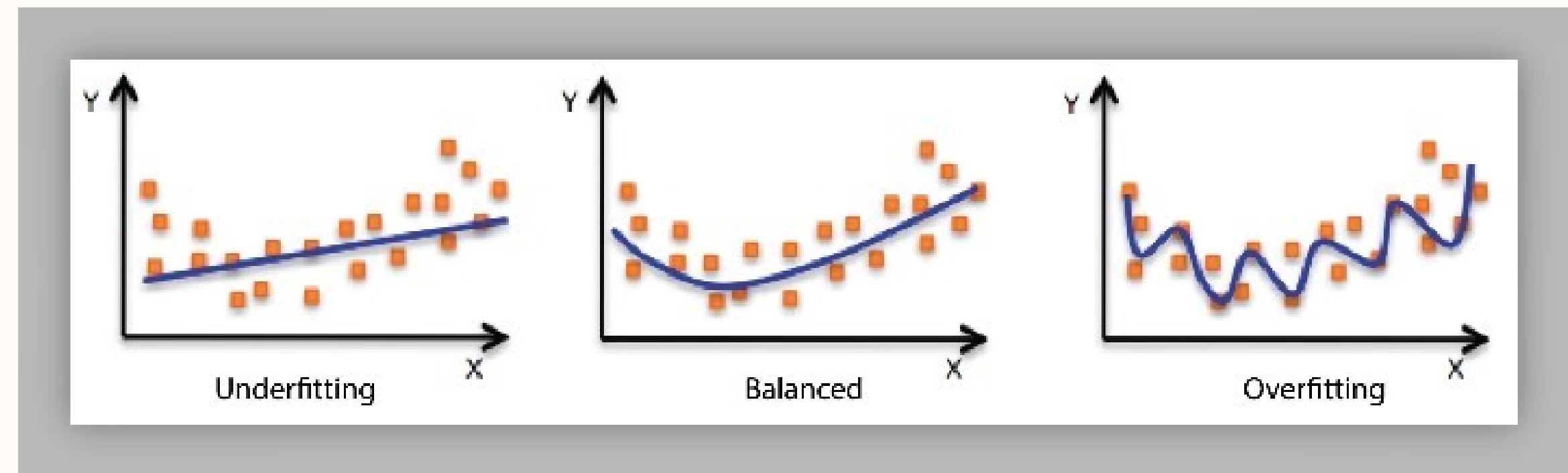
overfitting

Overfitting pojawia się, gdy model uczy się bardzo dobrze danych treningowych, ale traci zdolność generalizacji do nowych danych. Oznacza to, że model jest "zbyt dopasowany" do szczegółów zbioru treningowego, co prowadzi do słabych wyników na danych testowych. Problem ten często rozwiązuje się poprzez techniki regularizacji lub zwiększenie ilości danych treningowych.



underfitting

Underfitting pojawia się, gdy model nie jest w stanie dobrze dopasować się do danych treningowych, co skutkuje słabymi wynikami zarówno na danych treningowych, jak i testowych. Zwykle jest to efekt zbyt prostego modelu lub niewystarczającej liczby epok w procesie uczenia. Aby zmniejszyć ryzyko underfittingu, można zwiększyć złożoność modelu, dostarczyć więcej cech lub zmienić parametry modelu, tak aby lepiej uchwycił strukturę danych.



Niska jakość lub brakujące dane

Dane o niskiej jakości lub brakujące wartości utrudniają budowę wiarygodnych modeli, gdyż niepełne lub nieprecyzyjne informacje mogą prowadzić do błędnych prognoz. Często konieczne są metody wypełniania brakujących wartości lub techniki czyszczenia danych. Praca z takimi danymi wymaga także dobrego zrozumienia źródła danych, aby zminimalizować potencjalne błędy

Problem ze zróżnicowaniem danych

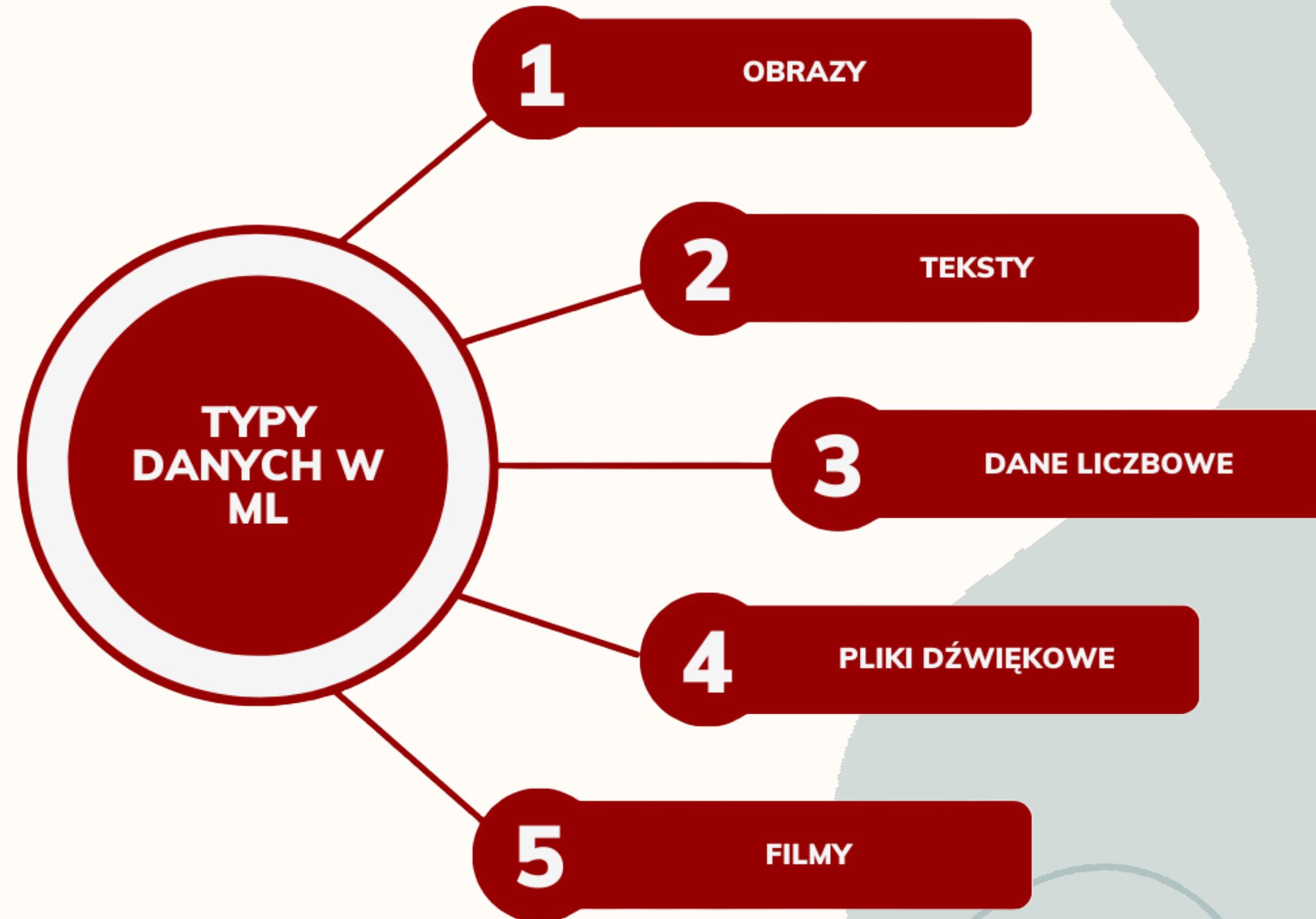
Niezrównoważone lub jednorodne dane mogą prowadzić do modeli, które nie są w stanie dobrze uogólnić. Jeśli np. pewna klasa jest nadreprezentowana, model może faworyzować tę klasę kosztem innych. W takim przypadku stosuje się metody balansowania klas lub wzbogacania zbioru danych, aby poprawić zdolność modelu do rozpoznawania rzadziej występujących wzorców.



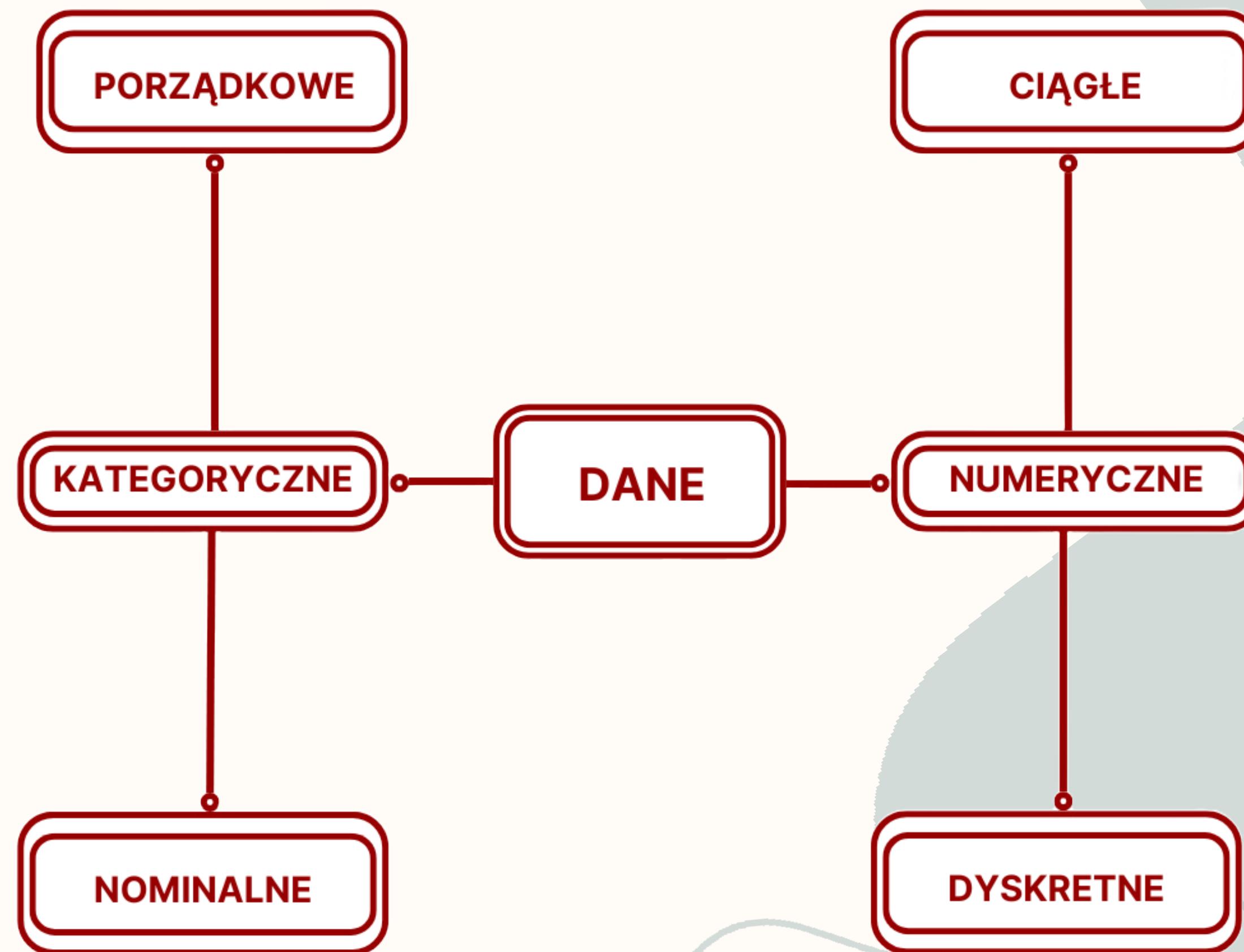
Dane

*kluczowy element w uczeniu
maszynowym*

Typy danych w uczeniu maszynowym

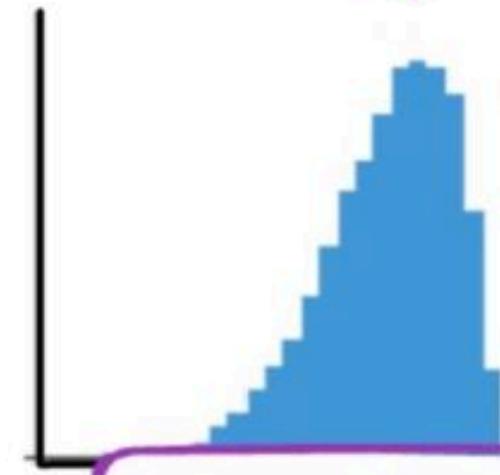


Typy danych w uczeniu maszynowym



EDA

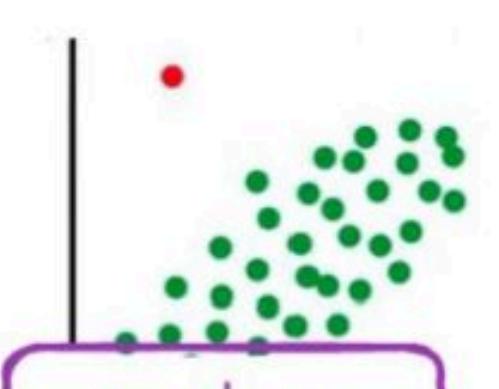
Exploratory Data Analysis



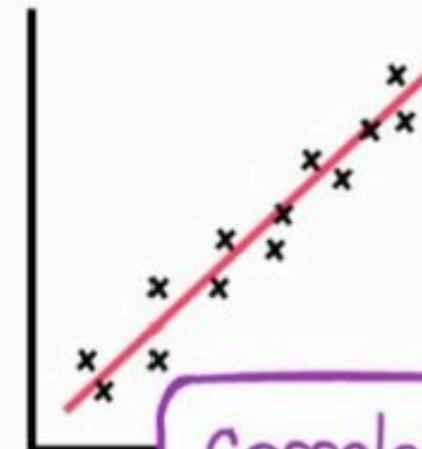
Data distribution

	F	G	H	I	J
A	0.620576	0.140053	1.352728	NaN	0.808078
B	NaN	0.526829	NaN	NaN	0.170902
C	NaN	0.458827	1.406713	0.071119	NaN
D	NaN	2.307197	NaN	NaN	NaN
E	0.203402	0.259913	NaN	0.505811	1.516755

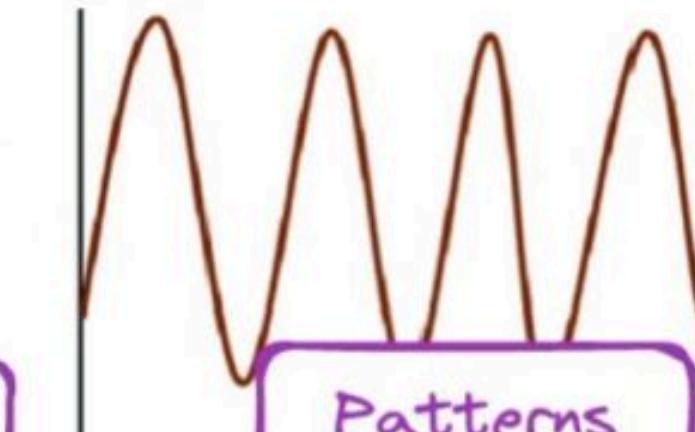
Missing data



Outliers



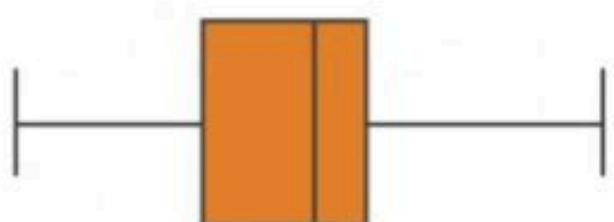
Correlation



Patterns

Cust_No
Cust_Name
Product_id
Product_cost
Purchase_Date
dtype: object

int64
object
int64
float64
datetime64[ns]



Data types



Data visualization

Data quality

OCZYSZCZANIE DANYCH

**USUWANIE LUB
UZUPEŁNIANIE
BRAKUJĄCYCH
WARTOŚCI**

**USUWANIE
DUPLIKATÓW**

**POPRAWIANIE
LITERÓWEK,
NIEZGODNYCH NAZW
KATEGORII**

**USUWANIE DANYCH
ODSTAJĄCYCH**

PRZETWARZANIE DANYCH

KODOWANIE
DANYCH
KATEGORYCZNYCH
(ONE-HOT ENCODING)

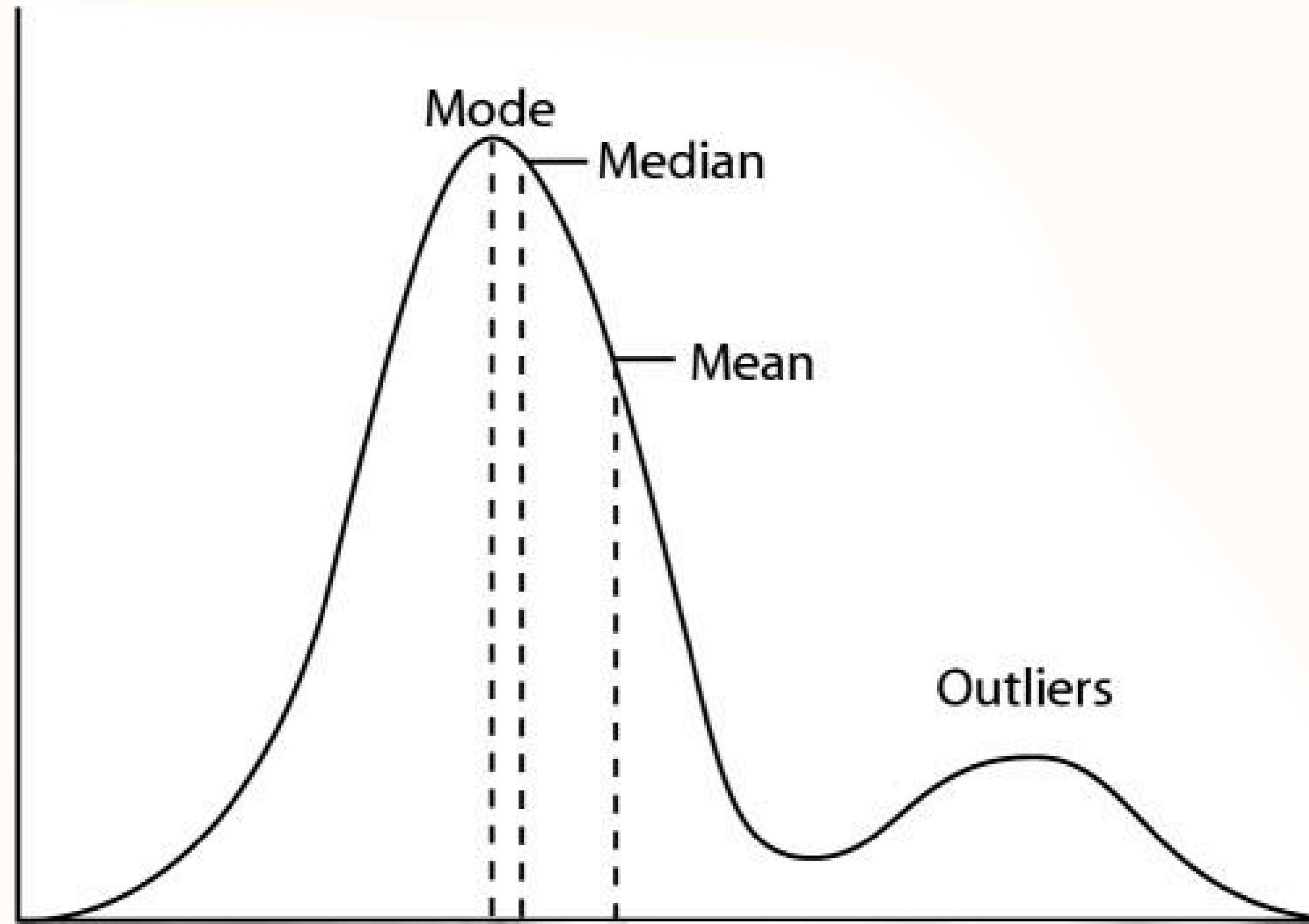
NORMALIZACJA LUB
STANDARDYZACJA
WARTOŚCI
NUMERYCZNYCH

TWORZENIE NOWYCH
CECH
(FEATURE ENGINEERING)

KONWERSJA TYPÓW
DANYCH

SKALOWANIE DANYCH

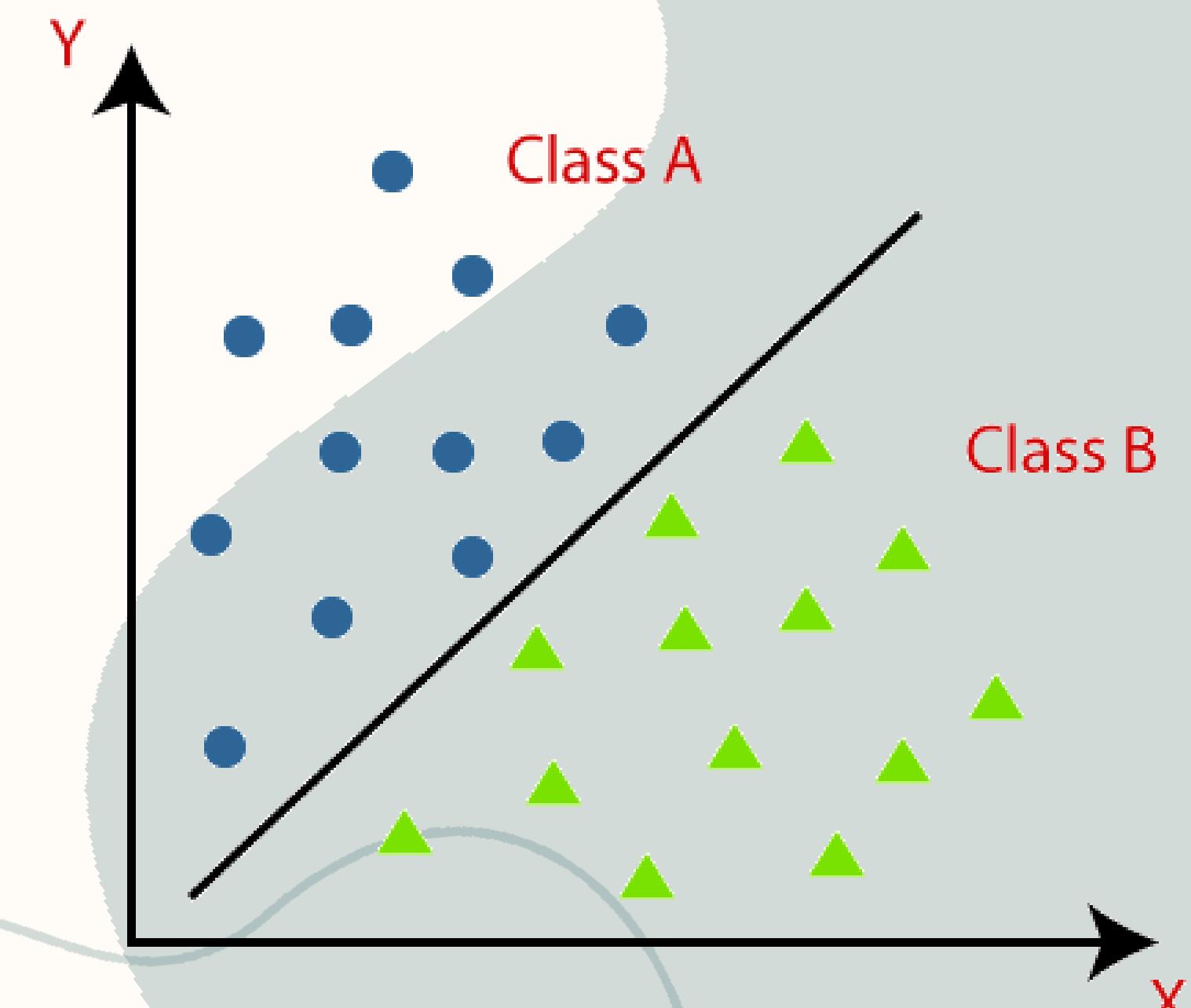
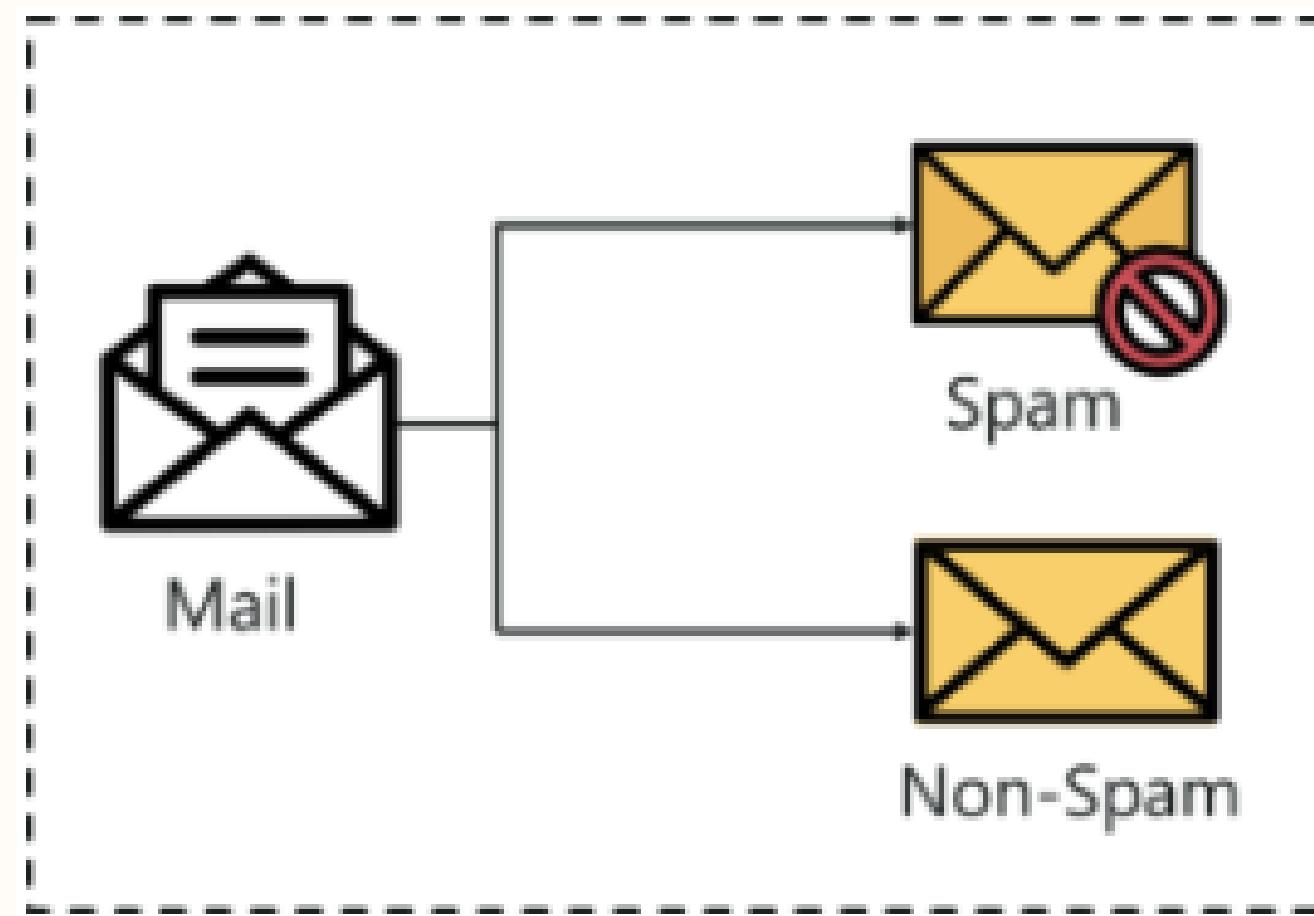
Wartości odstające (outliers)



Klasyfikacja

Czym jest klasyfikacja?

Klasyfikacja to przypisywanie obiektów do określonych kategorii na podstawie ich cech.



Przykłady zastosowań klasyfikacji

1. Bezpieczeństwo i filtrowanie
 - a. Filtry antyspamowe
 - b. Wykrywanie oszustw bankowych
2. Medycyna
 - a. Rozpoznawanie chorób
 - b. Rozpoznawanie chorób

Przykłady zastosowań klasyfikacji

3. Marketing i e-commerce

a. Segmentacja klientów

4. Technologie mobilne i autonomiczne

a. Rozpoznawanie obiektów na drodze

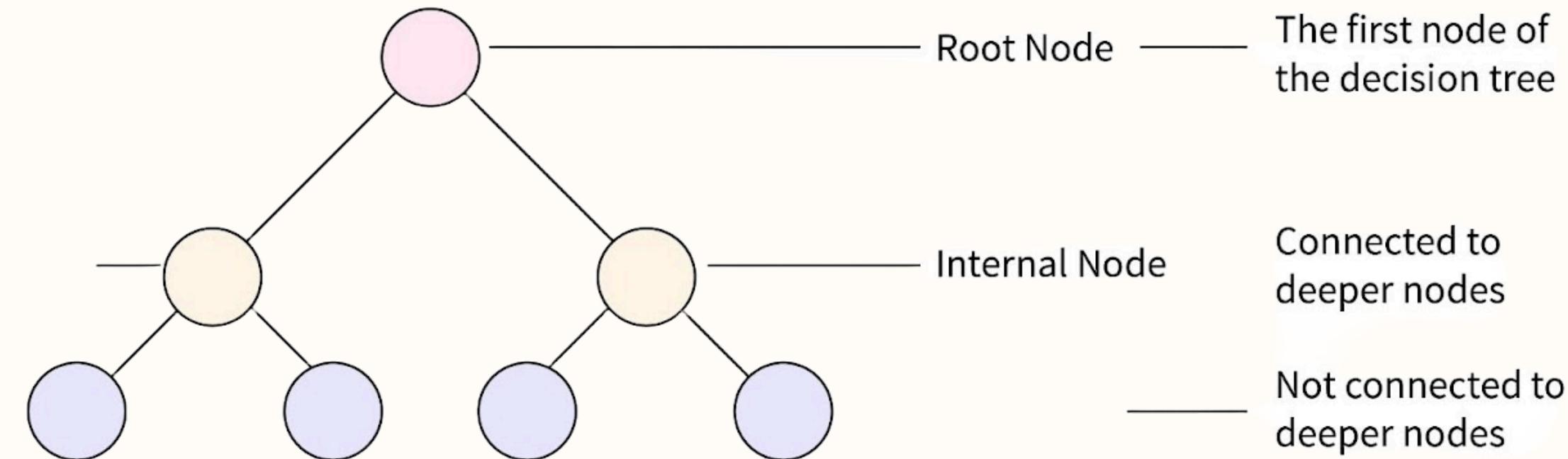
b. Autoryzacja użytkownika głosem lub twarzą

Rodzaje klasyfikacji

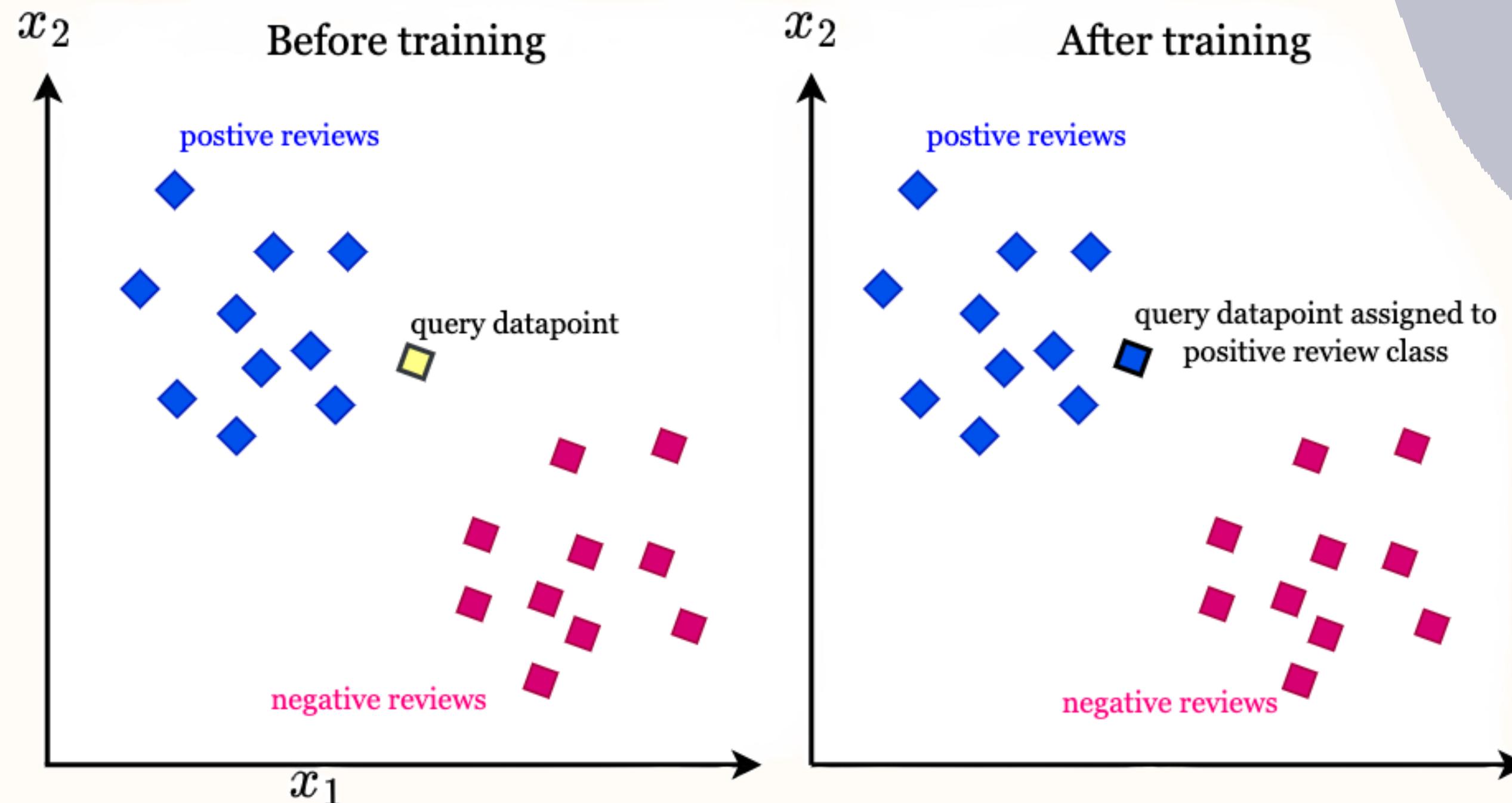
Wyróżniamy klasyfikację binarną, wieloklasową i multilabel.

W klasyfikacji binarnej mamy tylko dwie klasy, np. „tak” lub „nie”. Klasyfikacja wieloklasowa dotyczy więcej niż dwóch klas – np. rozpoznawanie gatunku kwiatu (setosa, versicolor, virginica). Multilabel to przypadek, w którym jedna próbka może należeć do wielu klas jednocześnie – np. zdjęcie zawierające kota i psa. Rodzaj klasyfikacji wpływa na wybór modelu i metryk oceny.

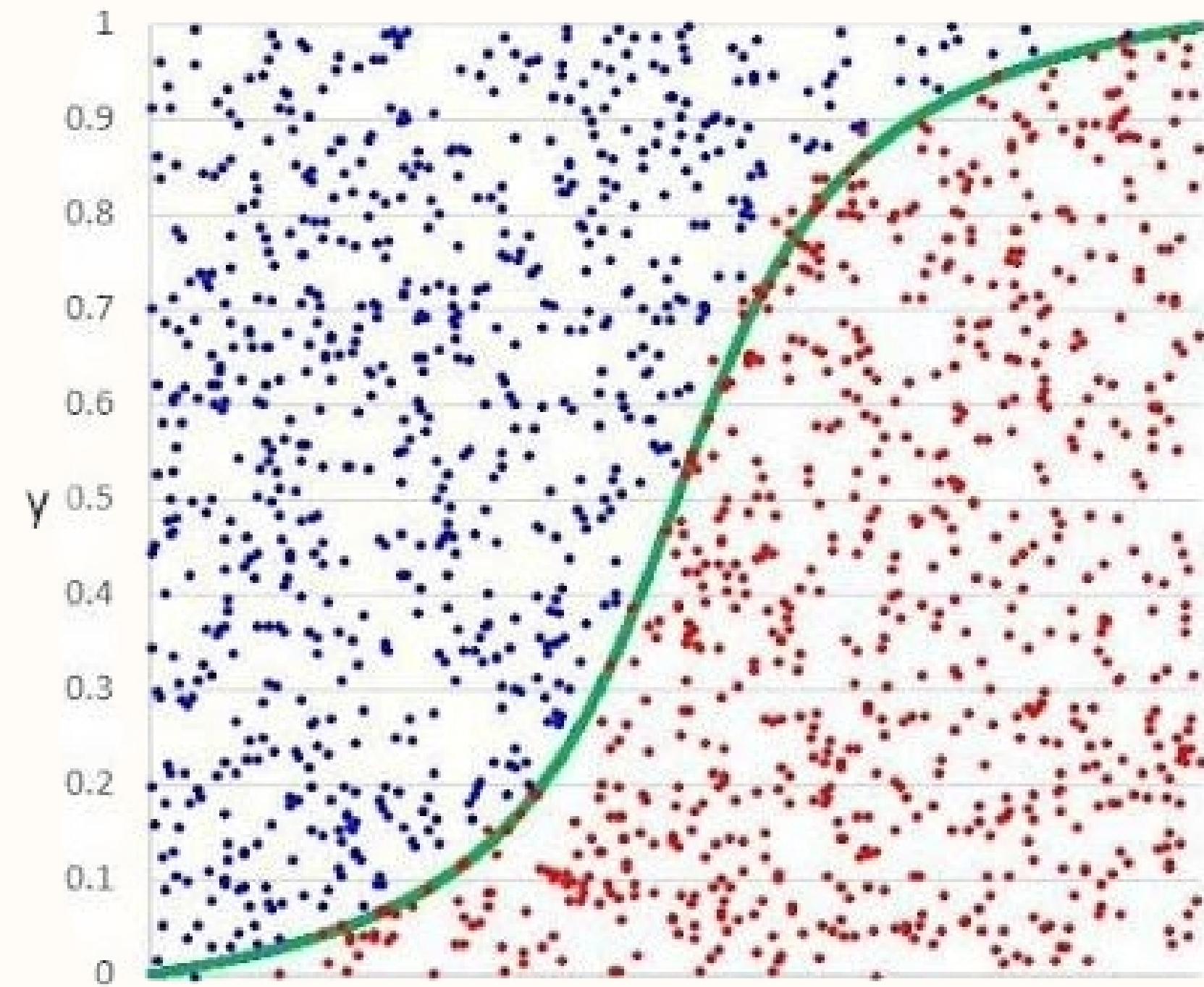
Klasyczne algorytmy klasyfikacji



Klasyczne algorytmy klasyfikacji



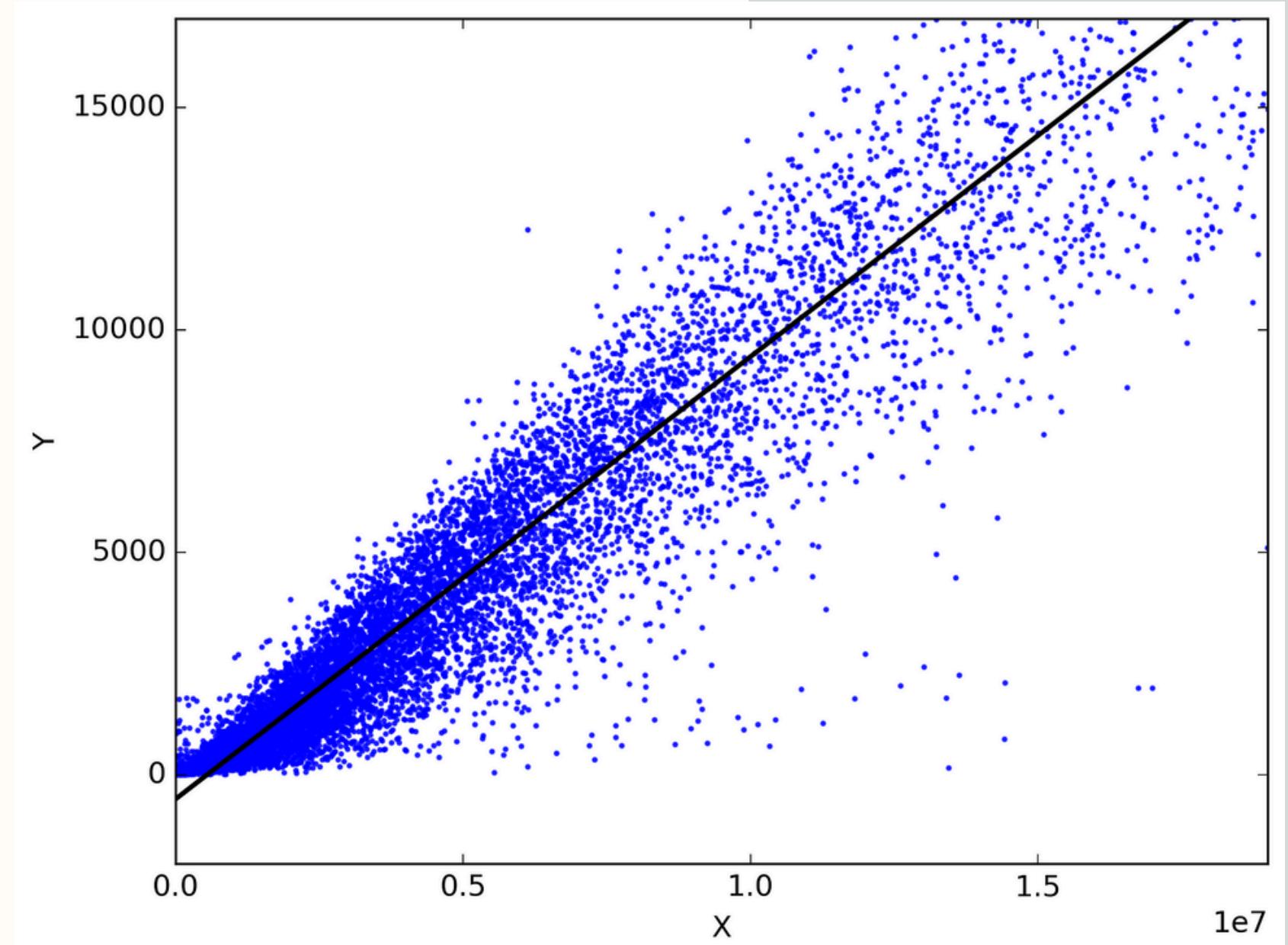
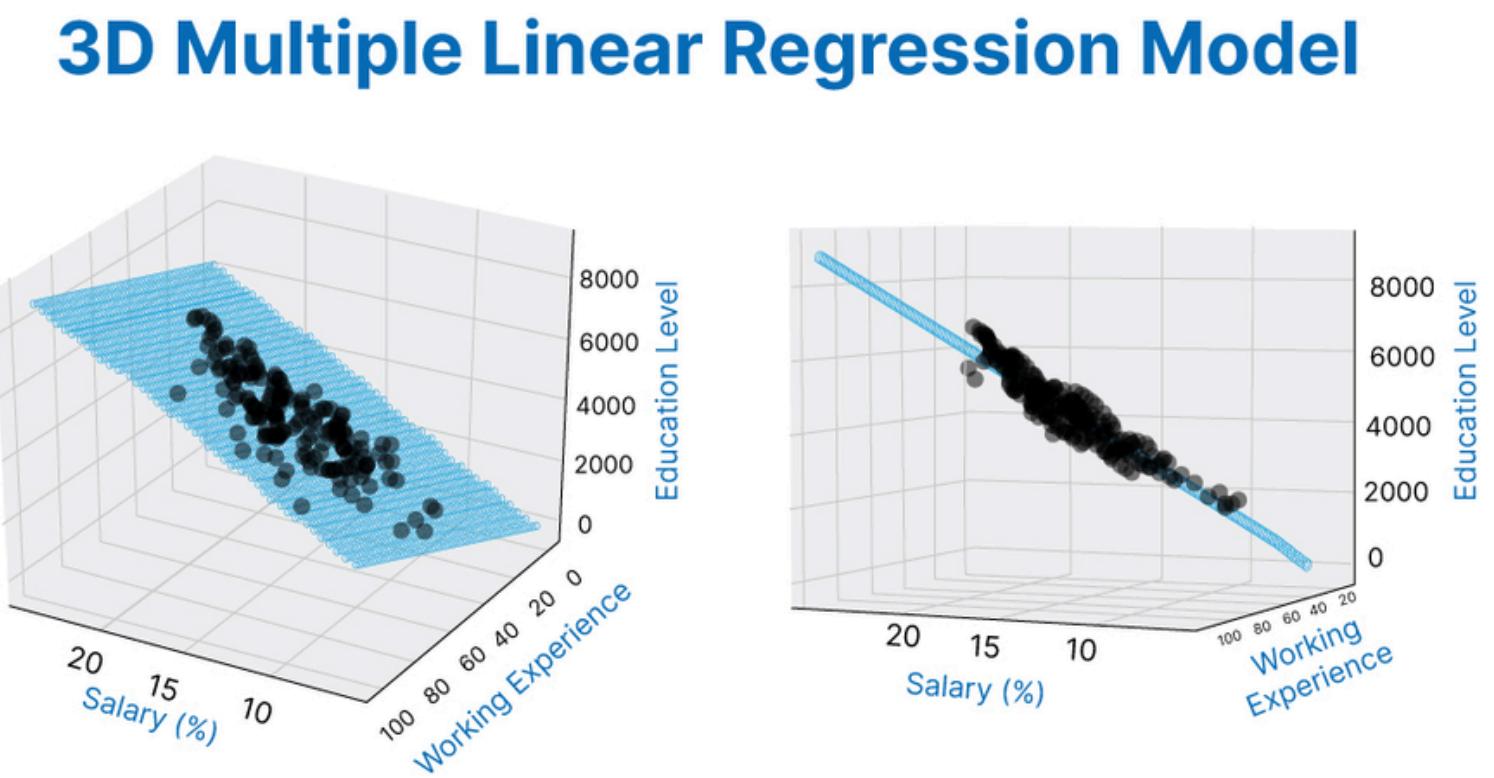
Klasyczne algorytmy klasyfikacji



Regresja

Czym jest regresja?

Regresja to przewidywanie wartości liczbowej na podstawie danych wejściowych.



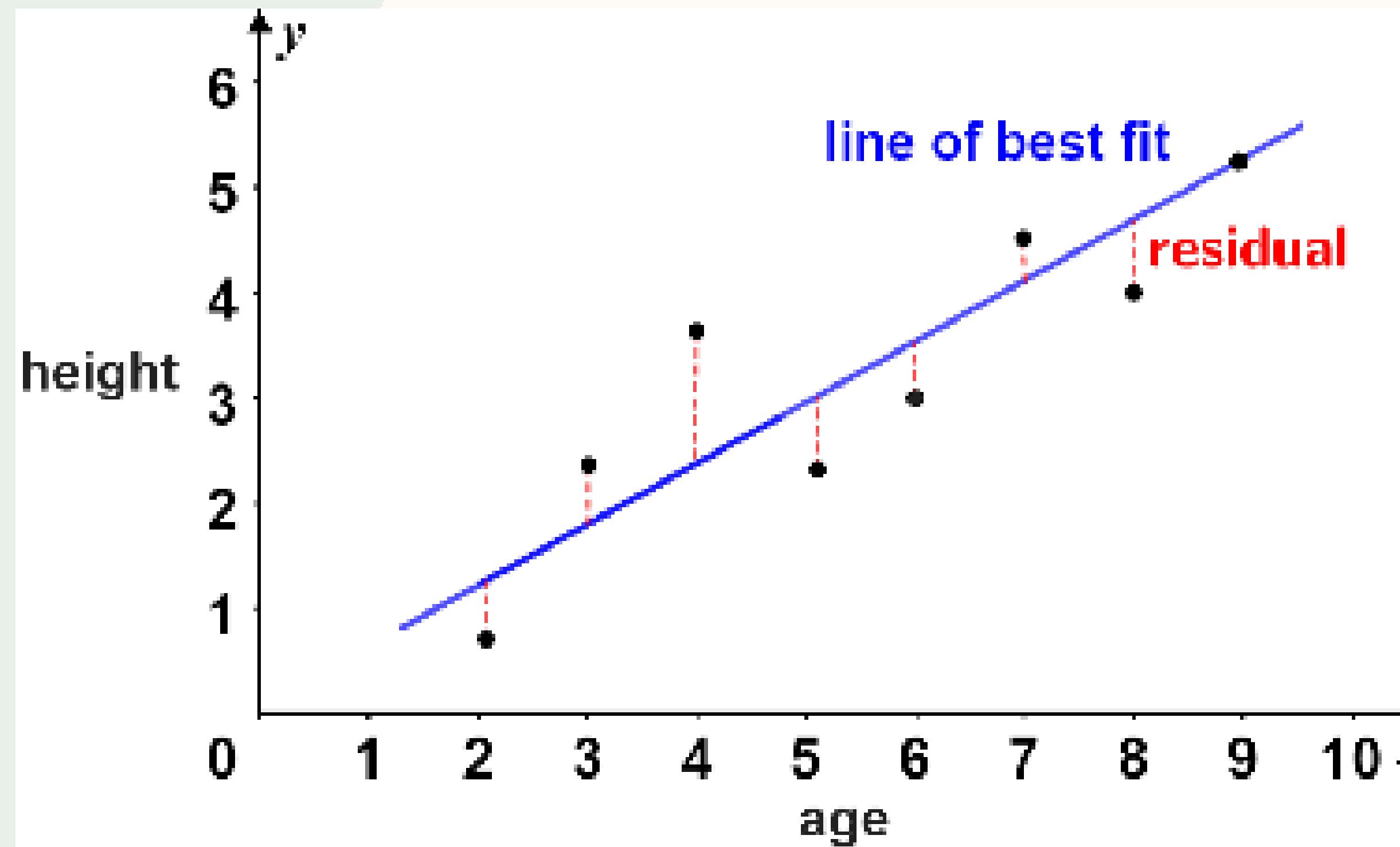
Przykłady zastosowań regresji

1. Prognozowanie
2. Analiza trendów
3. Modelowanie zjawisk

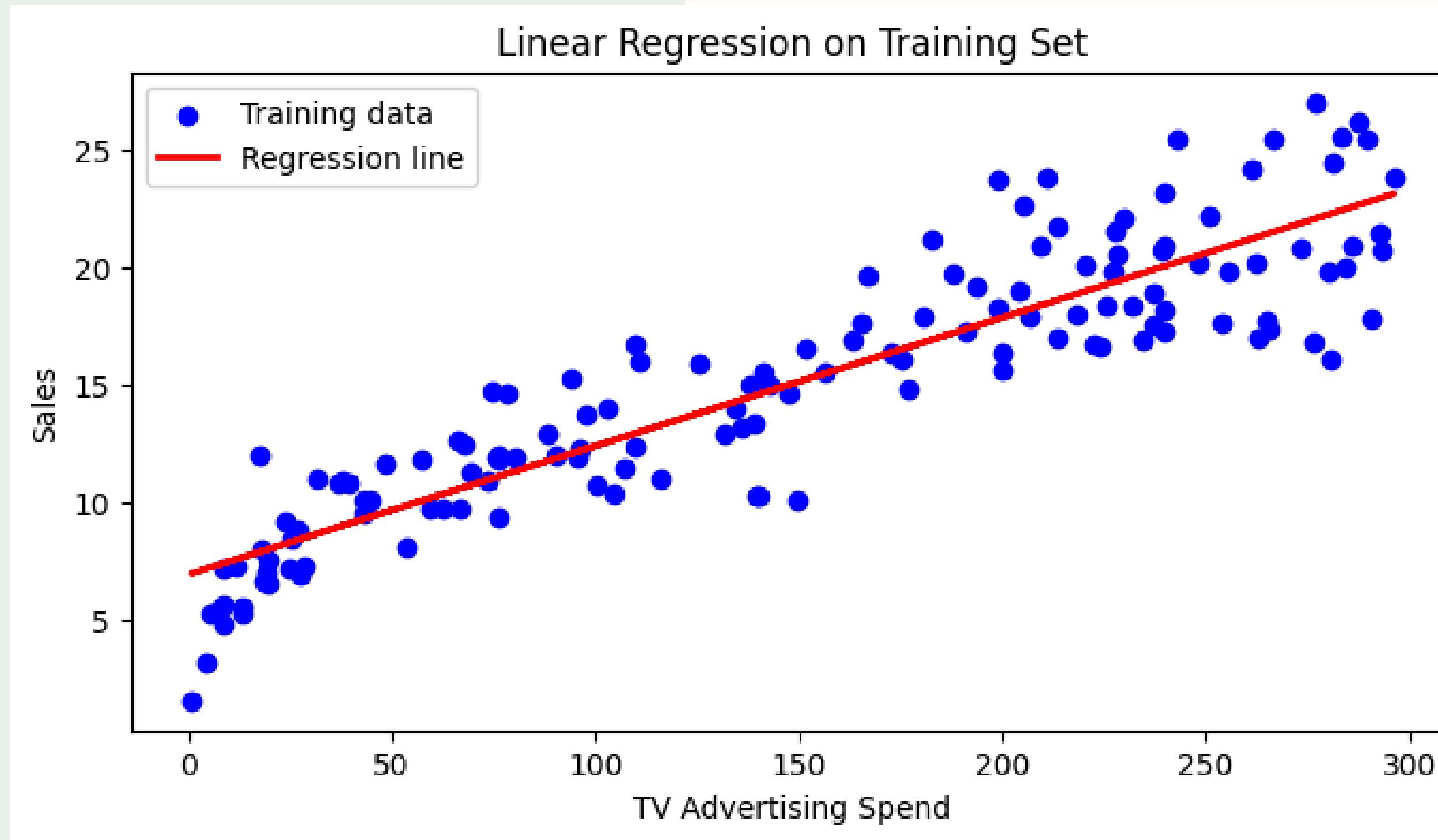
Rodzaje regresji

- 1. Regresja Liniowa:** Modelowanie liniowej zależności między zmiennymi.
- 2. Regresja Wielomianowa:** Modelowanie nieliniowych zależności poprzez wielomiany.
- 3. Regresja Logistyczna:** Modelowanie prawdopodobieństwa wystąpienia zdarzenia binarnego.
- 4. Regresja Ridge i Lasso:** Techniki regularizacji zapobiegające przeuczeniu modelu.

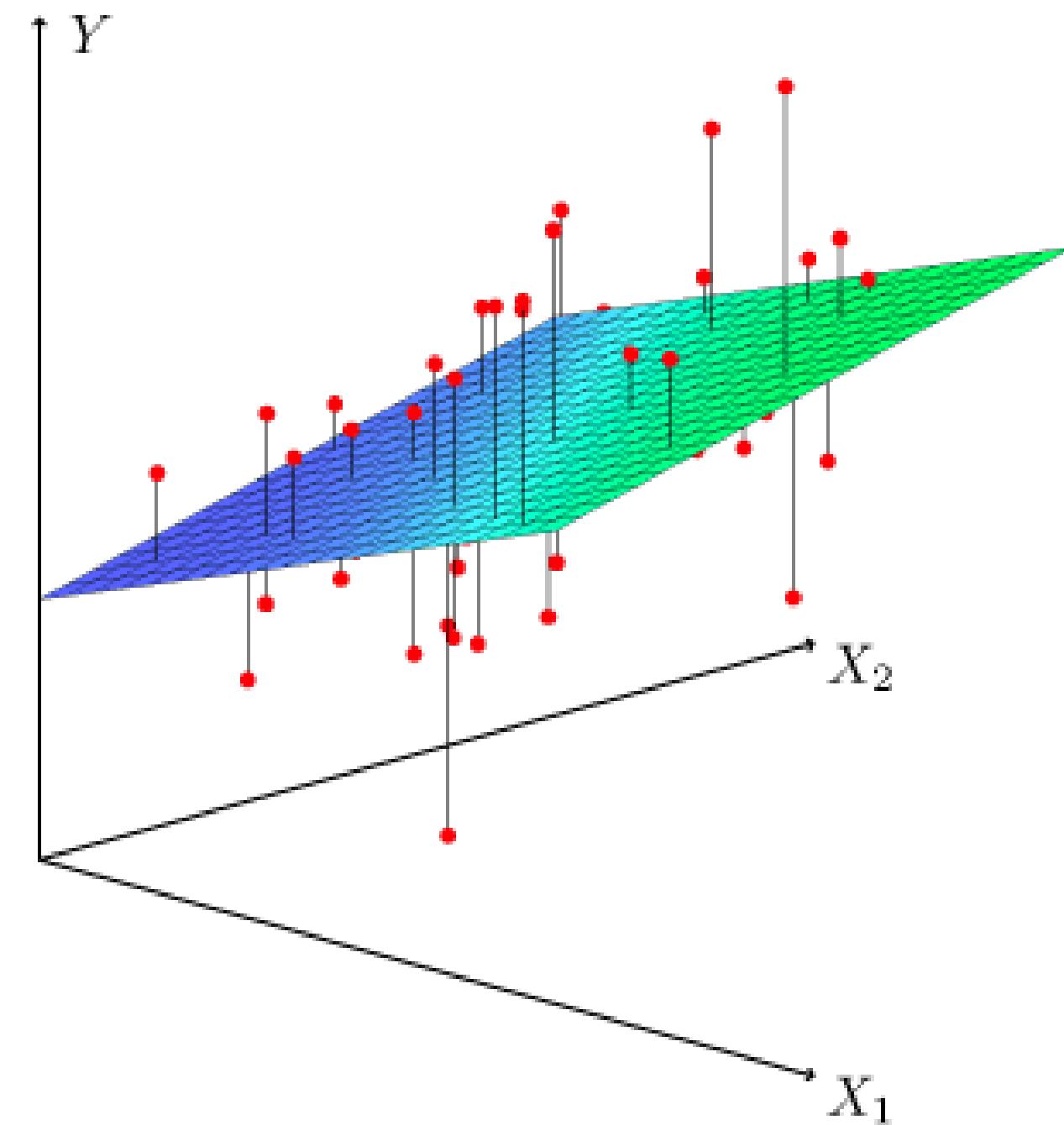
Błąd losowy - Reszty



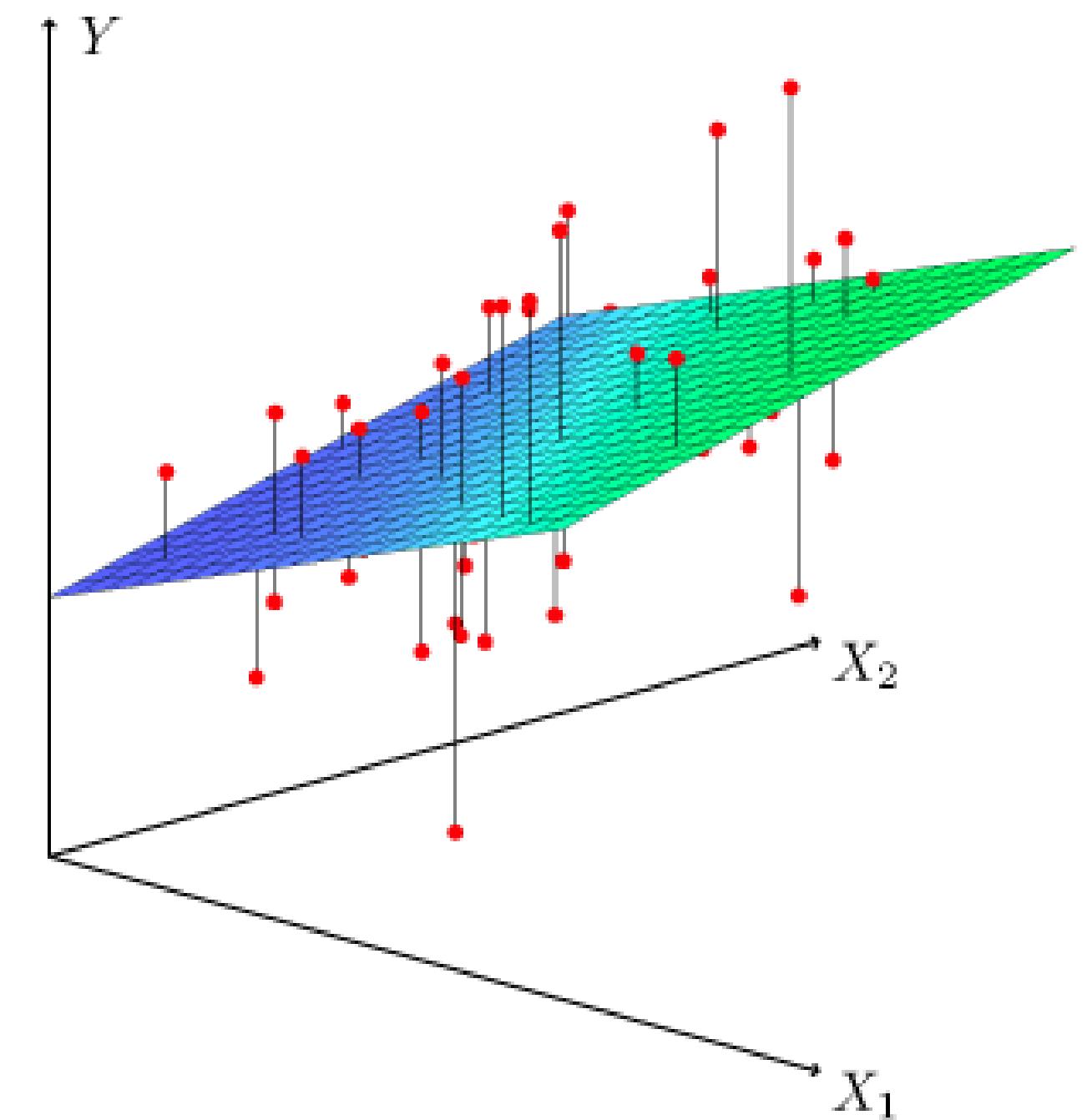
Prosta Regresja Liniowa



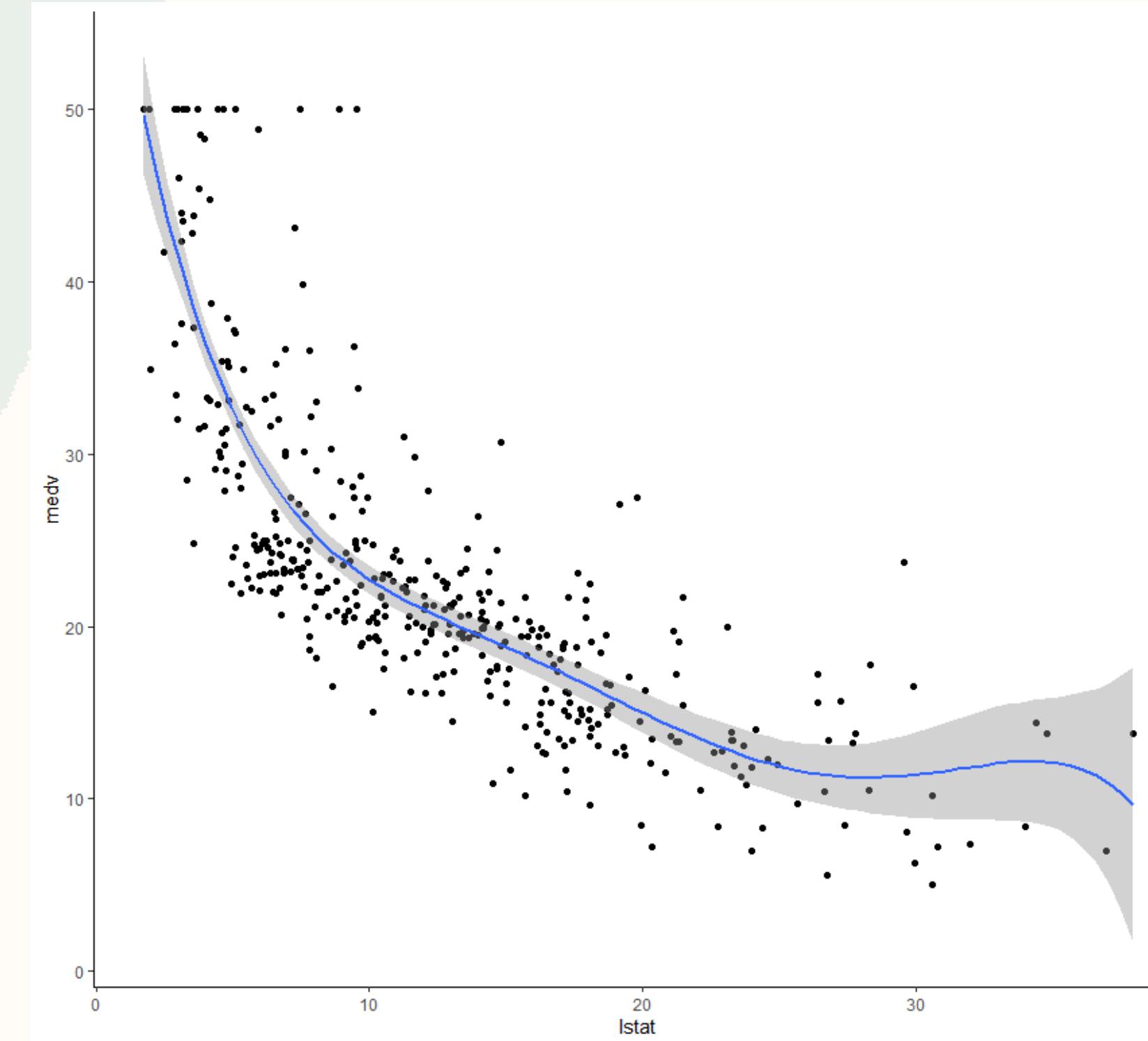
Multiple Linear Regression



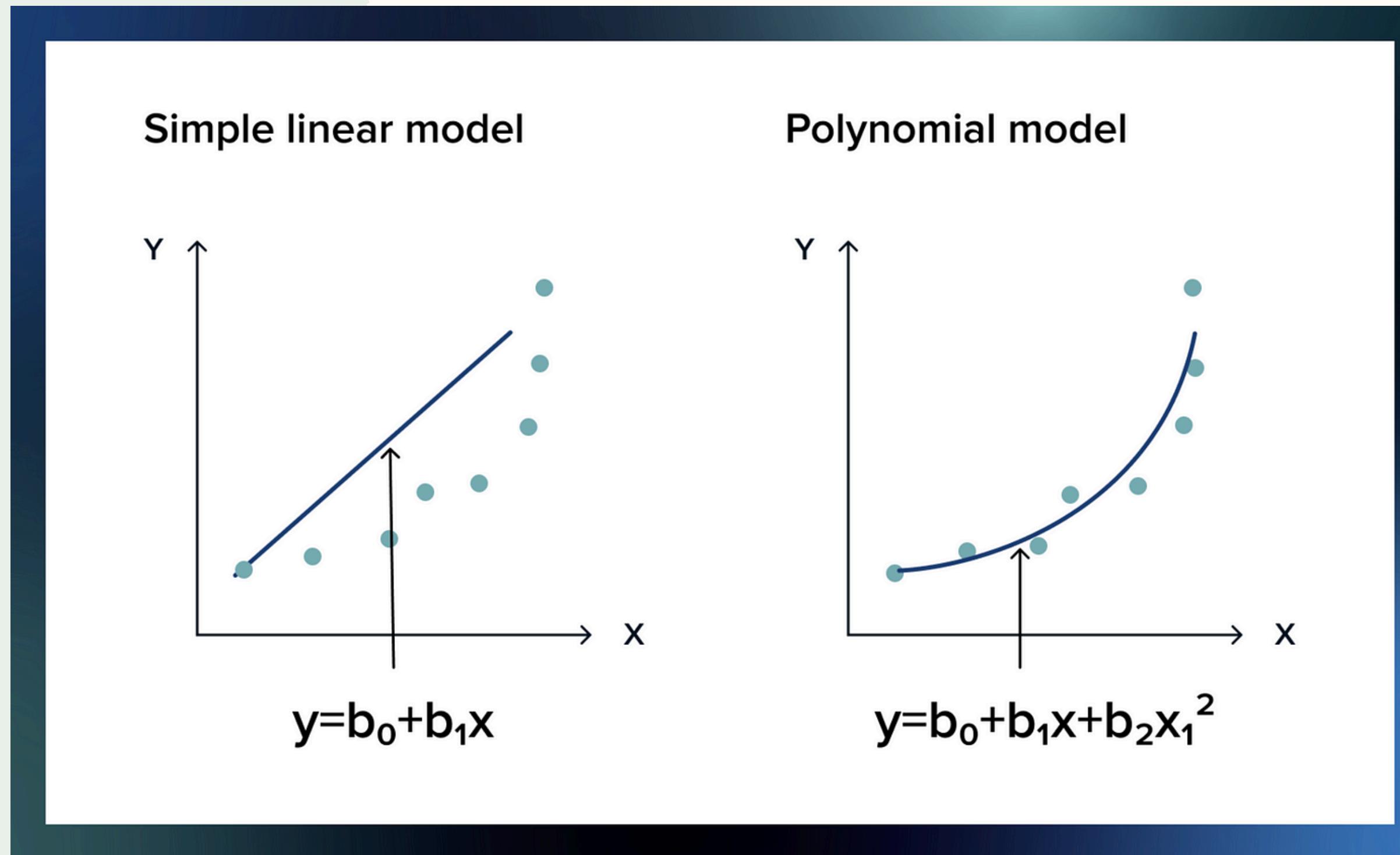
Wielozmiennowa Regresja Liniowa



Regresja wielomianowa (Polynomial Regression)



Regresja wielomianowa (Polynomial Regression)



Simple
Linear
Regression

$$y = b_0 + b_1 x_1$$

Multiple
Linear
Regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$$

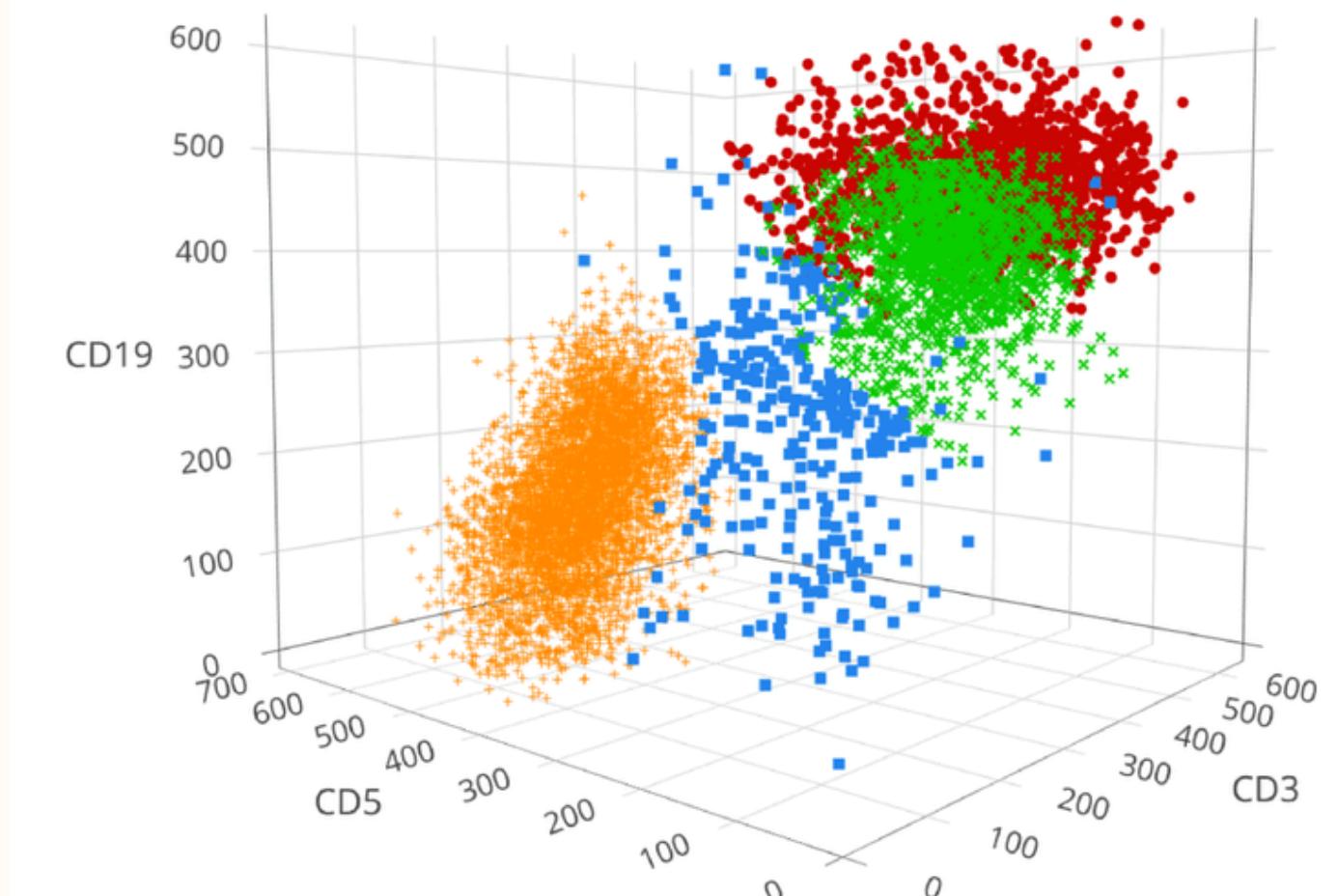
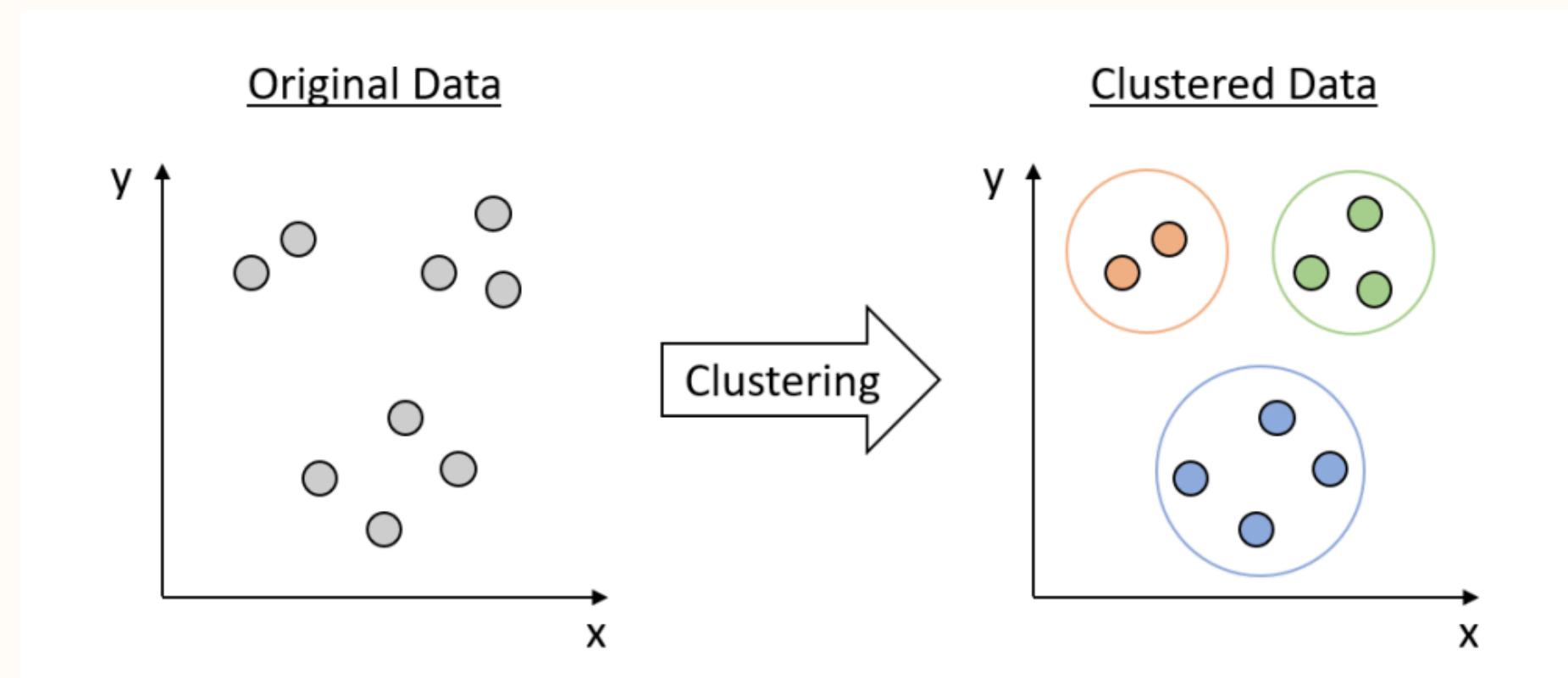
Polynomial
Linear
Regression

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \dots + b_n x_1^n$$

Klasteryzacja

Czym jest klasteryzacja?

Klasteryzacja polega na grupowaniu danych na podstawie ich podobieństw - bez znajomości etykiet.



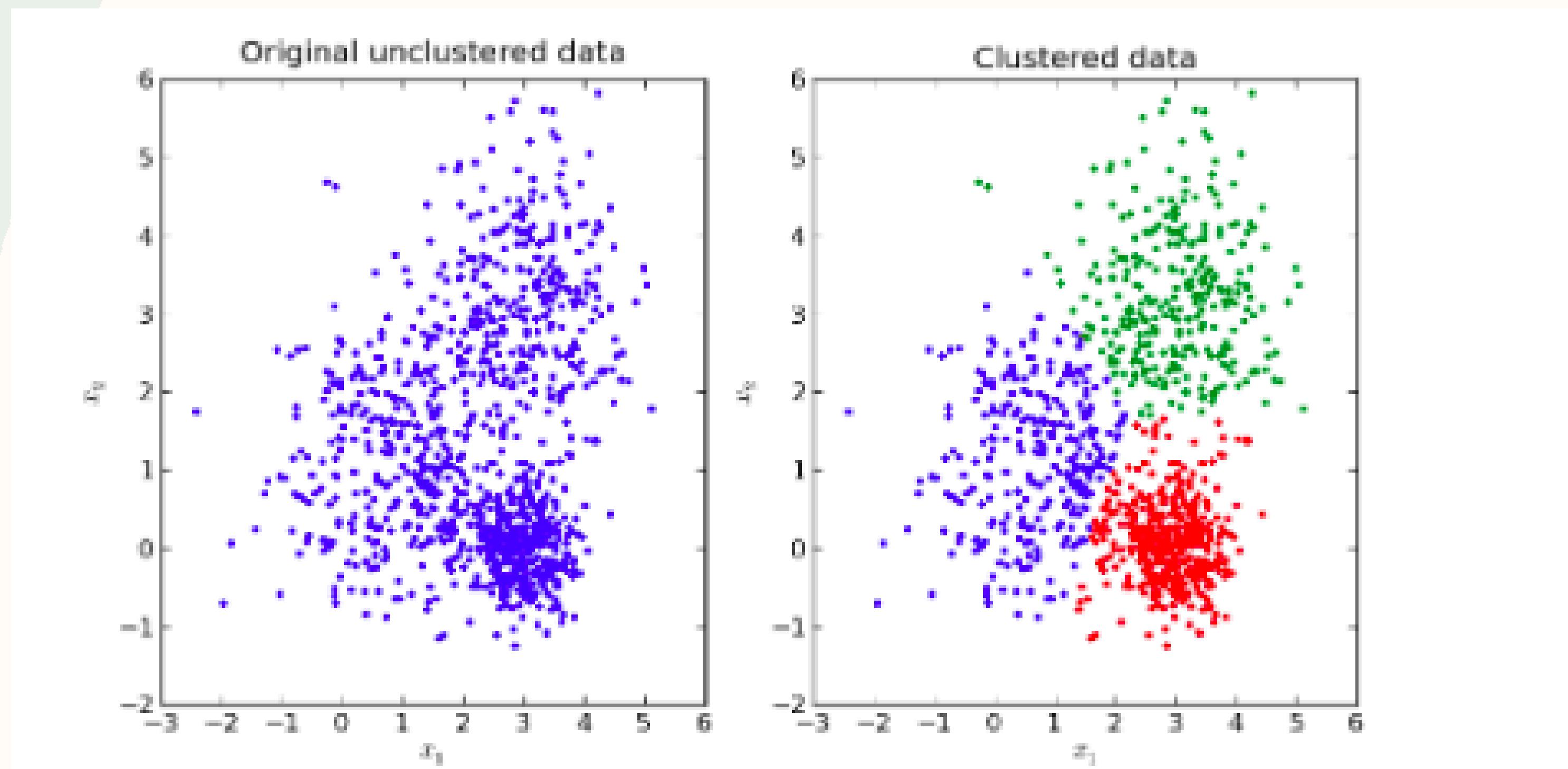
Przykłady zastosowań klasteryzacji

1. Segmentacja - segmentacja klientów
2. Analiza obrazów - rozpoznawanie wzorców
3. Biomedycyna - grupowanie genów
4. Analiza tekstu - grupowanie dokumentów

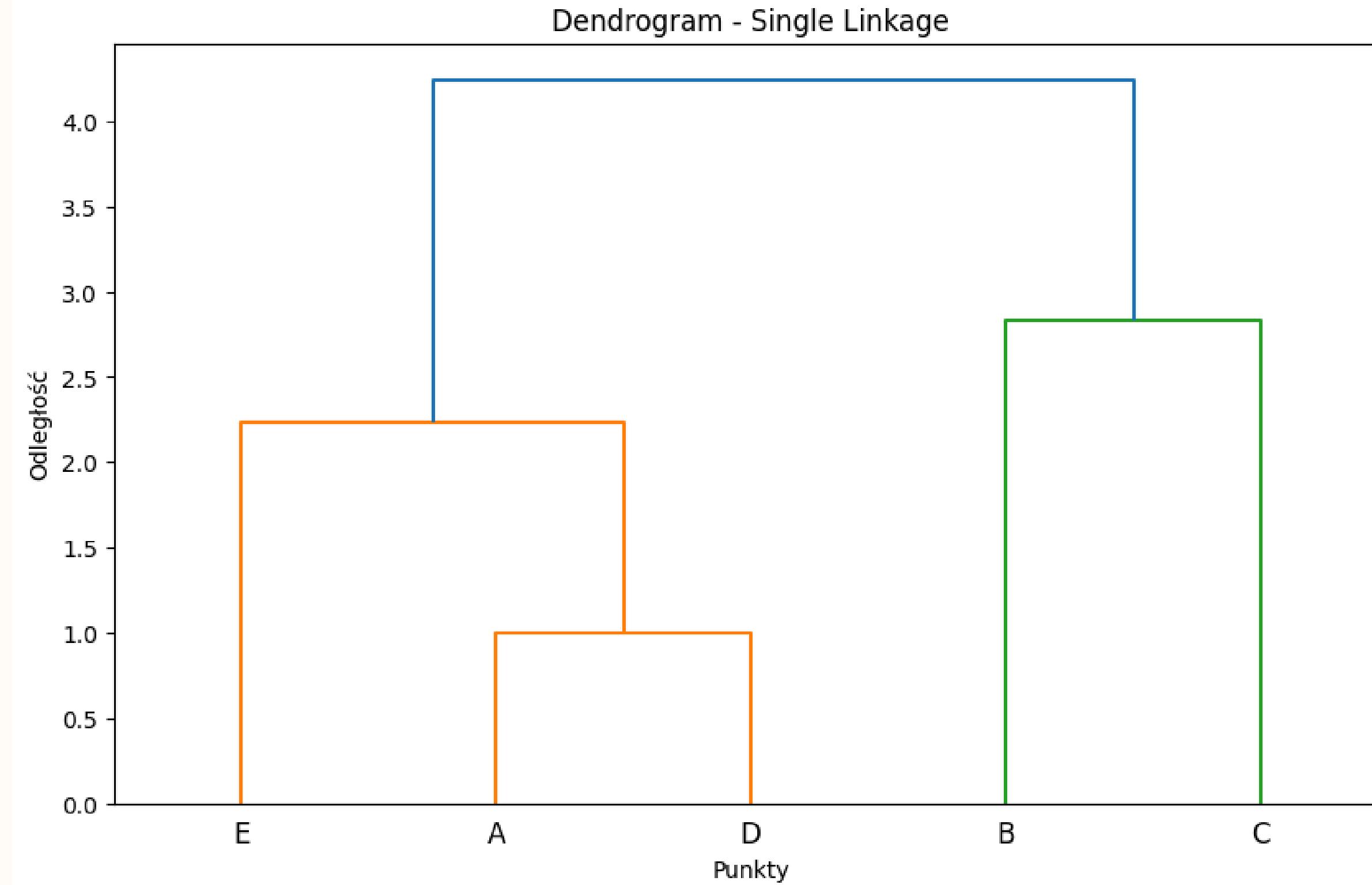


Różne typy
algorytmów
klasteryzacji

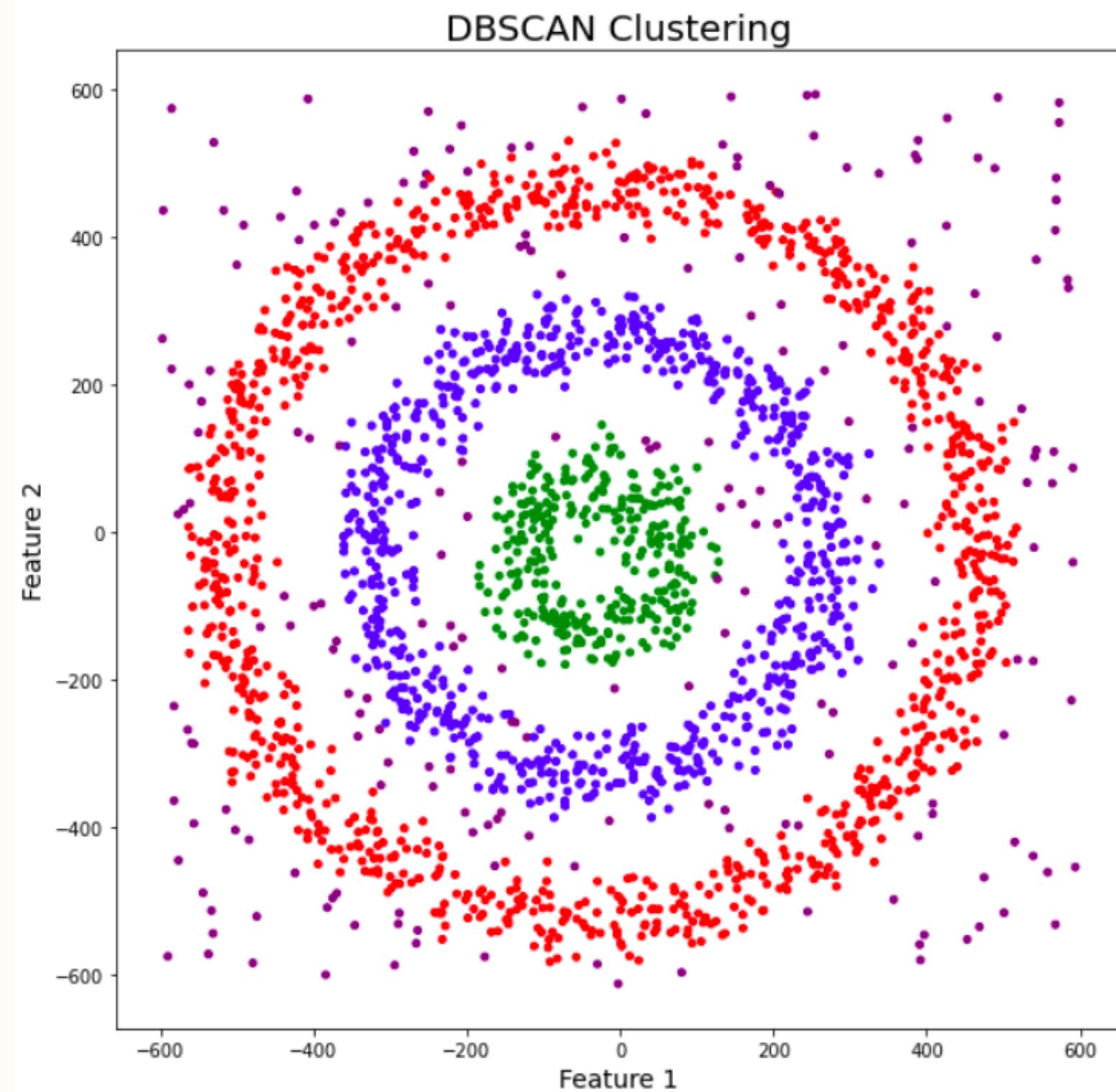
K-means



Klasteryzacja hierarchiczna



DBSCAN



DBSCAN



k-means



How to confuse machine learning:



Thank you!

Do zobaczenia za tydzień