POLITECHNIKA WROCŁAWSKA
WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI

METODY ANALIZY I EKSPLORACJI DANYCH

Wykład 5 - Redukcja wymiaru. Metody liniowe

DR INŻ. AGATA MIGALSKA

01

REDUKCJA WYMIARÓW

Motywacja. Metody liniowe i nieliniowe.

02

PCA

Analiza głównych składowych (Principal Component Analysis)

03

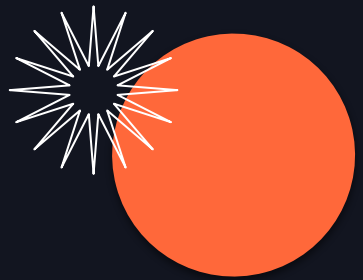
LDA

Liniowa analiza dyskryminacyjna (Linear Discriminant Analysis)

04

ZASTOSOWANIA I PRZYKŁADY

Eigenfaces vs Fisher faces.
PCA i LDA na Iris dataset.



01

REDUKCJA WYMIARÓW

KLĄTWA WYMIAROWOŚCI

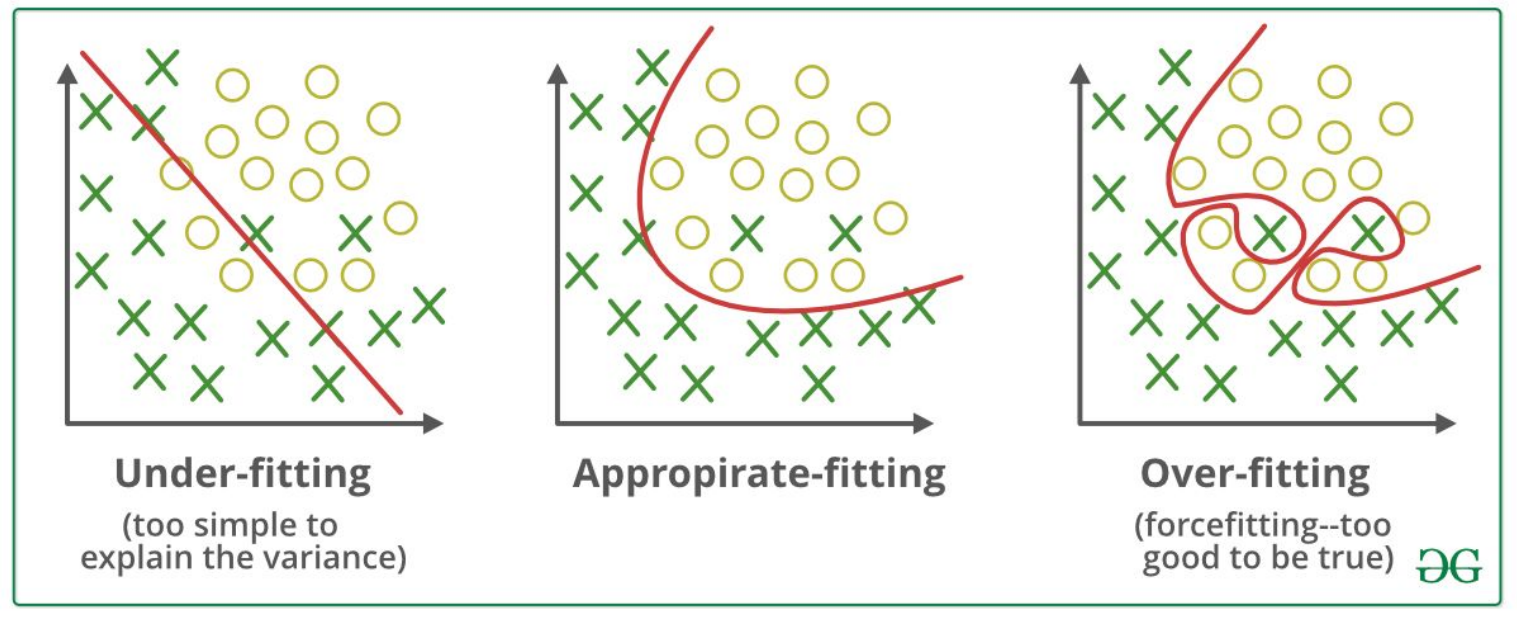
W uczeniu maszynowym, aby wyłapać przydatne wskaźniki i uzyskać dokładniejszy wynik, na początku dodajemy jak najwięcej cech do modelu.

Klątwa wymiarowości (Curse of Dimensionality)

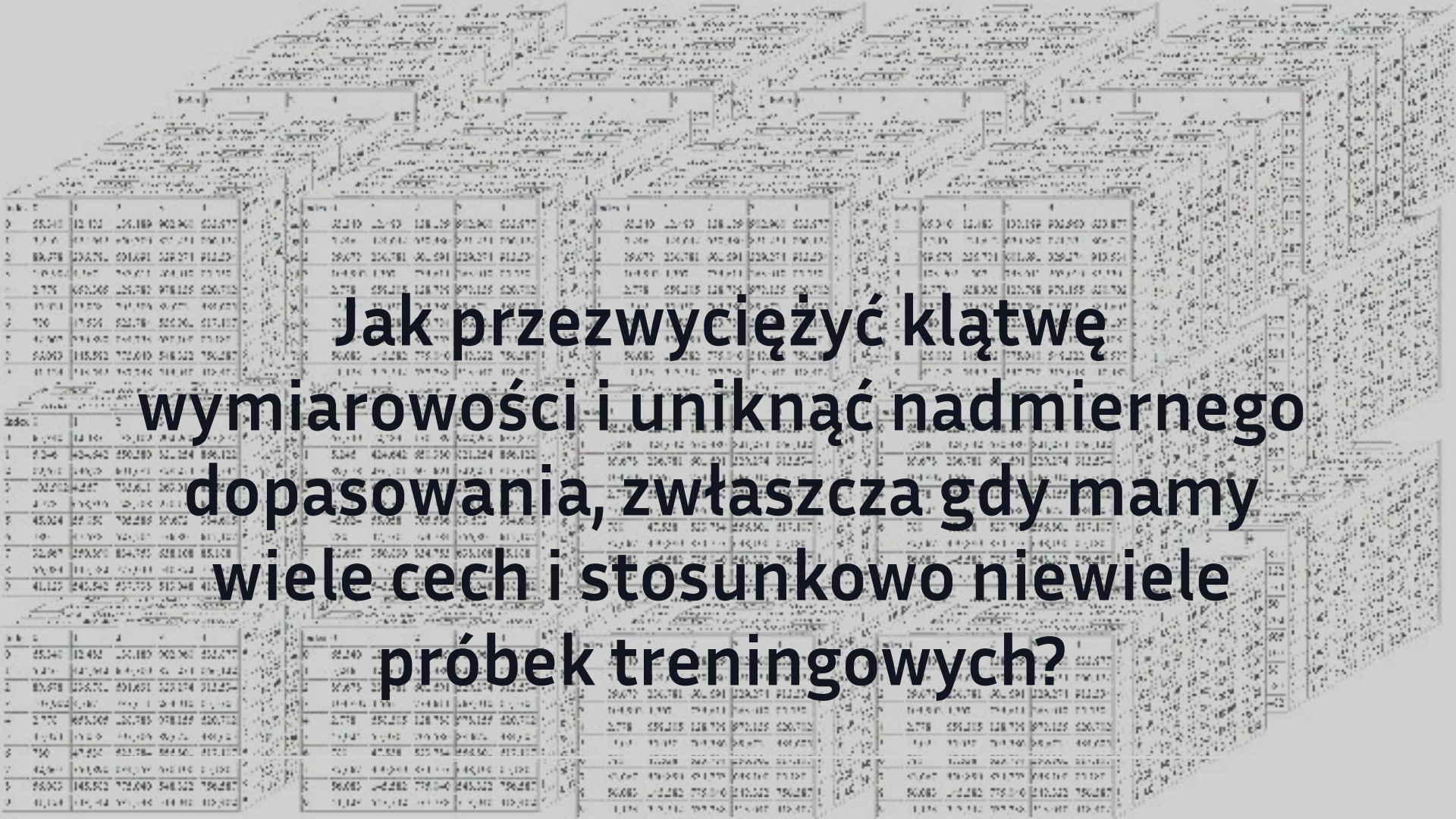
Gdy zwiększa się wymiarowość:

- objętość przestrzeni rośnie tak szybko, że dostępne dane stają się rzadkie. Aby uzyskać wiarygodny wynik, ilość potrzebnych danych często rośnie wykładniczo wraz z wymiarowością.
    - Dane rzadko są losowo rozłożone w dużych wymiarach i często są silnie skorelowane.
    - Odległości między najbliższym i najdalszym punktem danych mogą stać się równe w dużych wymiarach, co może utrudnić dokładność niektórych narzędzi do analizy opartej na odległości.
- od pewnego momentu wydajność modeli spada wraz ze wzrostem liczby zmiennych.
    - W rzadkich danych znacznie łatwiej jest znaleźć „idealne” rozwiązanie, co często prowadzi do nadmiernego dopasowania modelu do danych (overfittingu).

OVER- I UNDERFITTING



Overfitting ma miejsce, gdy model zbyt ściśle odpowiada określonemu zestawowi danych i nie uogólnia dobrze. Przesadnie dopasowany model działałby zbyt dobrze na uczącym zestawie danych, przez co nie sprawdzałby się na przyszłych danych i sprawiał, że prognoza była niewiarygodna.


Jak przezwyciężyć klątwę wymiarowości i uniknąć nadmiernego dopasowania, zwłaszcza gdy mamy wiele cech i stosunkowo niewiele próbek treningowych?

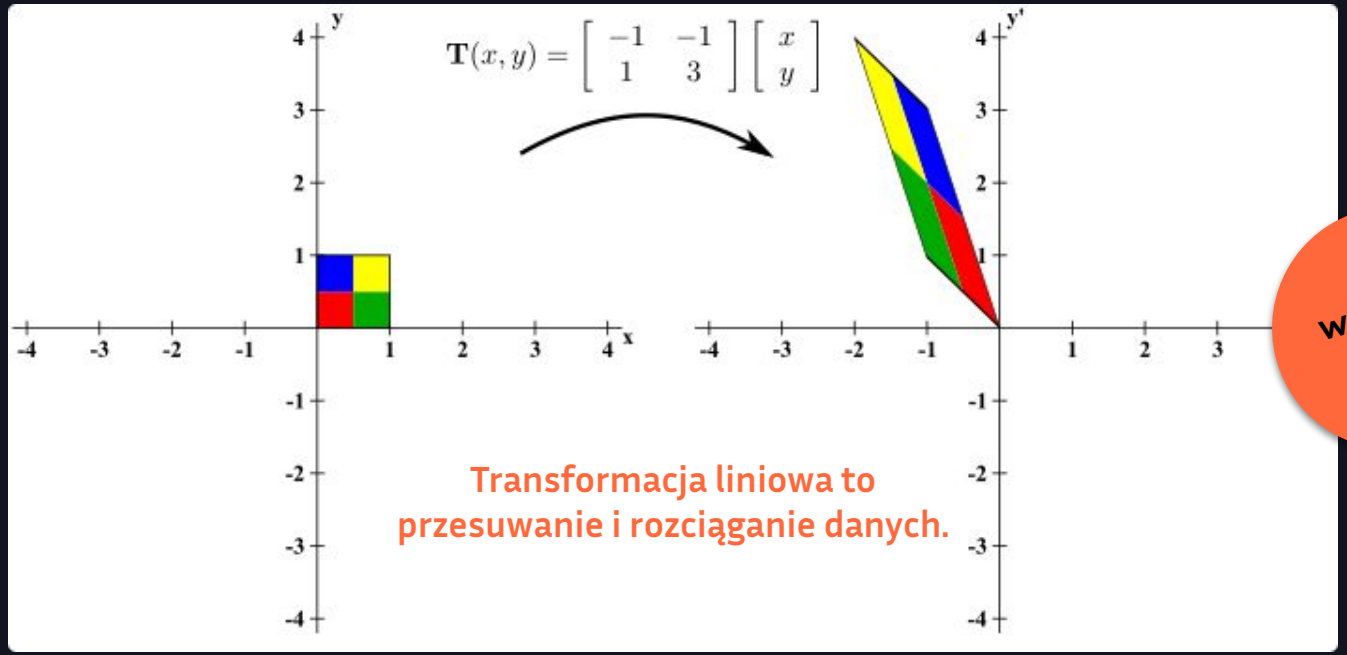
REDUKCJA WYMIAROWOŚCI

proces zmniejszania wymiarowości przestrzeni cech z uwzględnieniem uzyskania zbioru cech głównych

Dodatkowe korzyści:

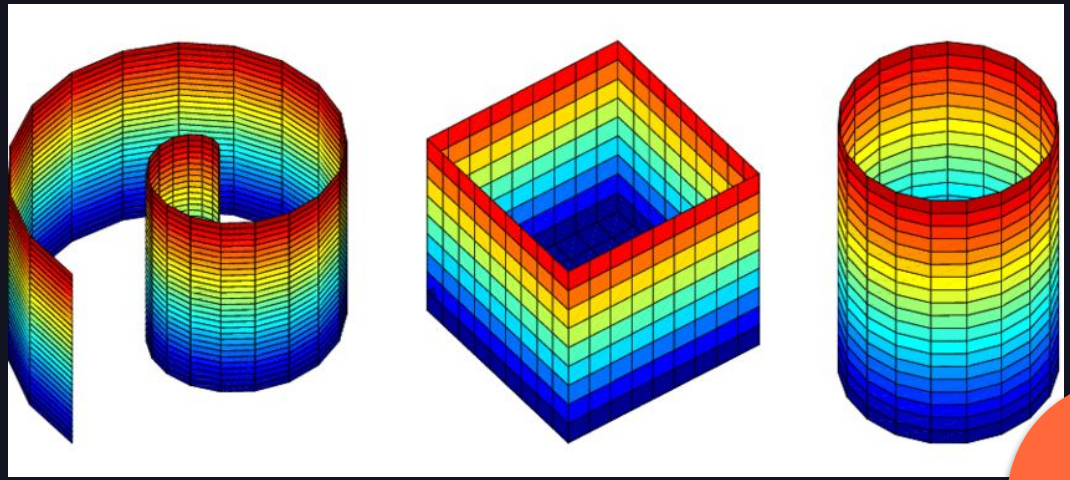
- mniejsze zapotrzebowanie na zasoby
  - redukcja kosztów
  - czasami jedyny sposób, żeby w ogóle wytrenować model
- łatwiejsza wizualizacja danych

METODY LINIOWE I NIELINIOWE REDUKCJI



Wykład 5

METODY LINIOWE I NIELINIOWE REDUKCJI



Transformacja nieliniowa to bardziej "radykalne" zmiany kształtu.

Wykład 6

Źródło grafiki: https://www.researchgate.net/figure/Three-typical-manifolds-the-Swiss-roll-left-the-open-box-middle-and-the-cylinder_fig1_2551965

TYPOWE PRZYPADKI UŻYCIA

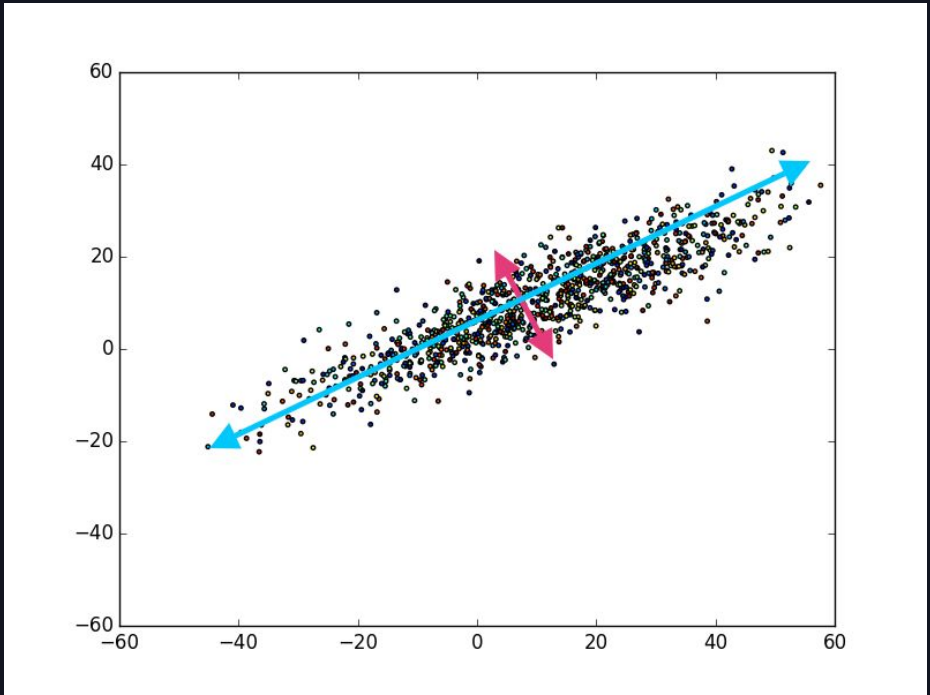
- **Redukcja wymiarowości**: znajdź niskowymiarowe przybliżenie X rozmiaru n × k (gdzie k jest znacznie mniejsze niż p) przy zachowaniu większości wariancji, jako etap wstępnego przetwarzania do klasyfikacji lub wizualizacji.
- **Inżynieria cech** (feature engineering): utwórz nową reprezentację X z liniowo nieskorelowanymi elementami.
- **Uczenie nienadzorowane**: wyodrębnij k głównych cech (gdzie k jest często znacznie mniejsze niż p). Zrozum zestaw danych, analizując, w jaki sposób oryginalne cechy wpływają na wartość wyodrębnionych cech.

02

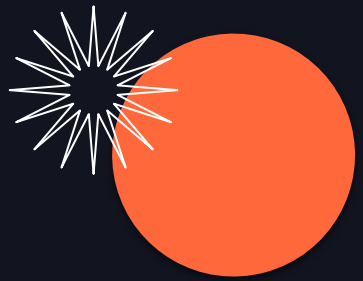
PCA
ANALIZA GŁÓWNYCH SKŁADOWYCH

PCA

PCA jest procedurą transformacji, która przekształca macierz cech pierwotnych, potencjalnie skorelowanych ze sobą, w zestaw liniowo nieskorelowanych zmiennych zwanych głównymi składowymi.



DEMO

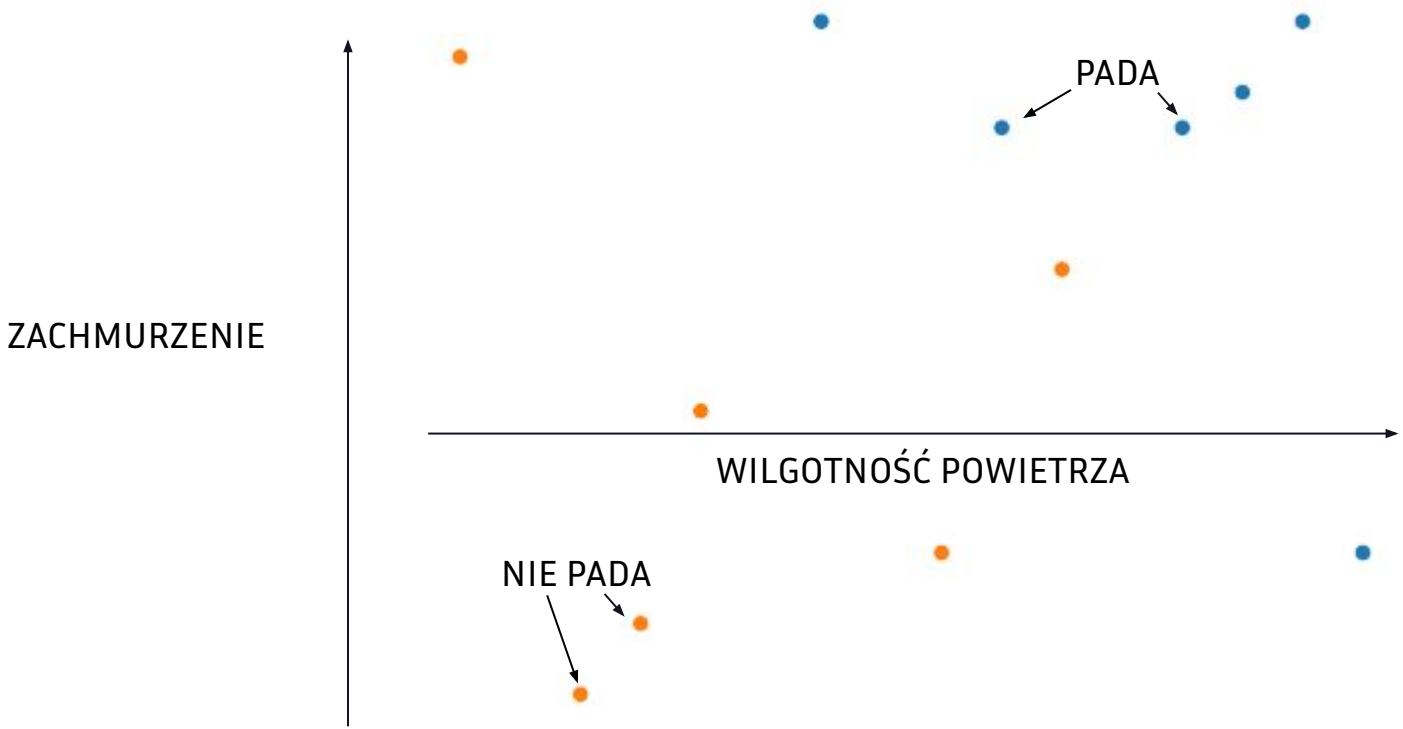


03

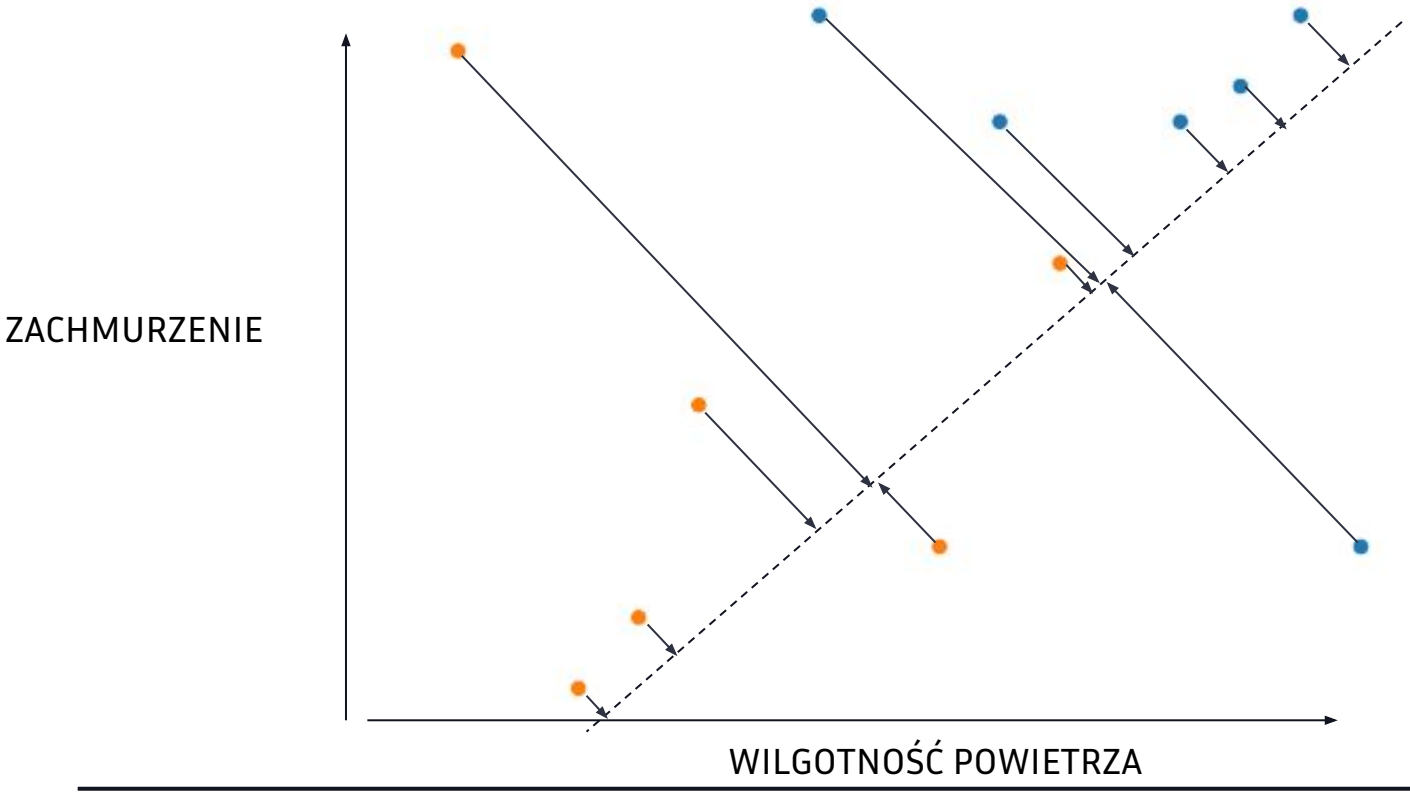
LDA (LINEAR DISCRIMINANT ANALYSIS)

LDA Z JEDNĄ ZMIENNĄ

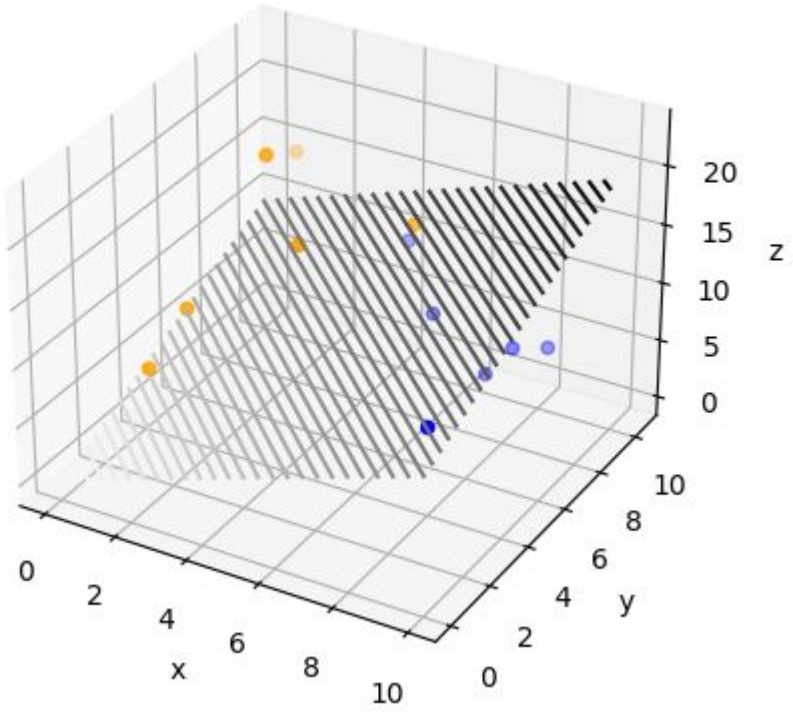
LDA Z DWOMA ZMIENNYMI



LDA Z DWOMA ZMIENNYMI



LDA Z TRZEMA ZMIENNYMI...

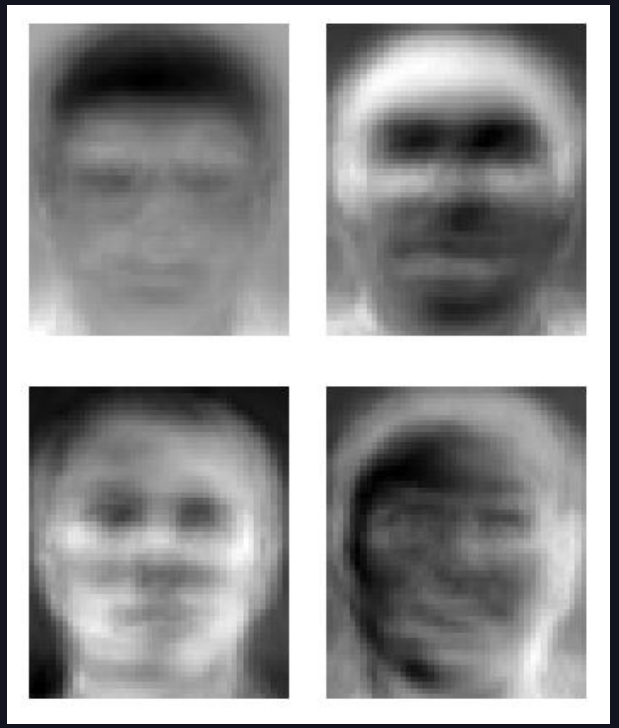


DEMO

04

ZASTOSOWANIA I PRZYKŁADY

EIGENFACES vs FISHERFACES



THANKS!

DZIĘKUJĘ ZA UWAGĘ