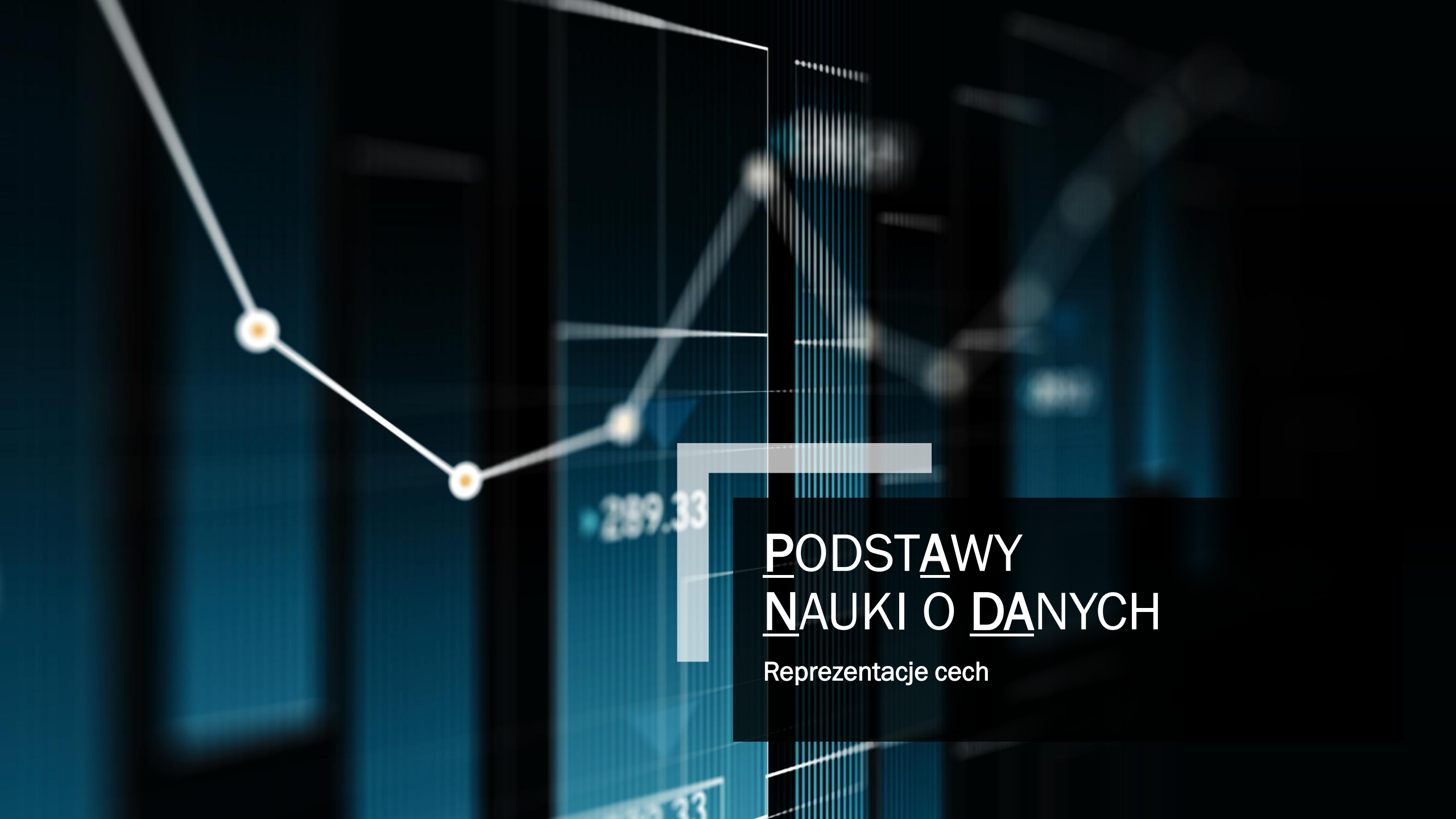




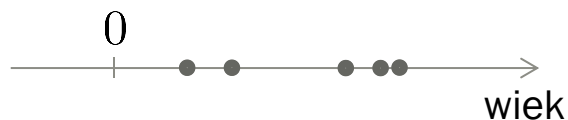
PODSTAWY NAUKI O DANYCH

Reprezentacje cech

PROSTE LICZBOWE REPREZENTACJE



Skale liczbowe



Absolutna: jest 0 i jednostka, np. ilość cukru, liczba subskrybentów
różnica i iloraz mają sens

Ilorazowa: sens ma stosunek wartości, np. temp. w Kelwinach, inflacja

Interwałowa / przedziałowa: 0 jest umowne, np. temp. w C, data

Porządkowa: kolejność ma znaczenie, np. miesiąc, wykształcenie

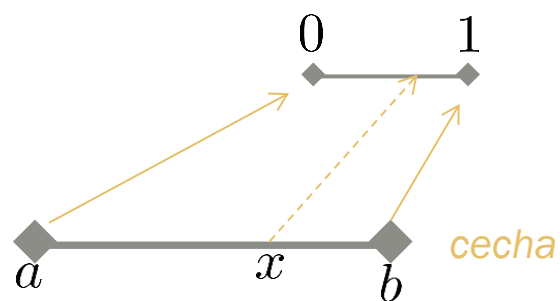
Nominalna: nie da się jednoznacznie uporządkować

• < • < • < • <

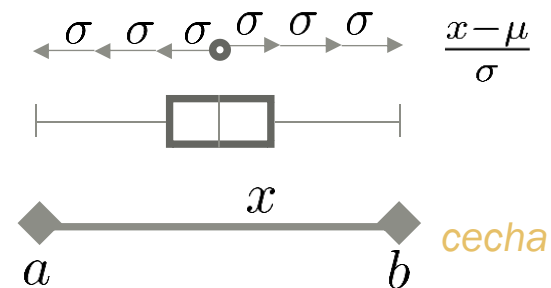
Wrocław
Kolbuszowa
Trzebnica
Rzeszów
Elbląg

Przekształcenia ciągłe $\mathbb{R} \rightarrow \mathbb{R}$

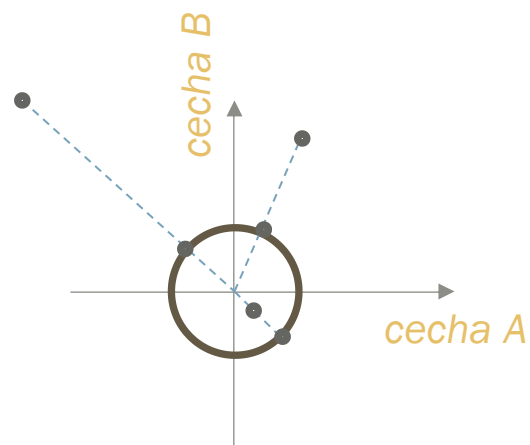
Skalowanie



Standaryzacja

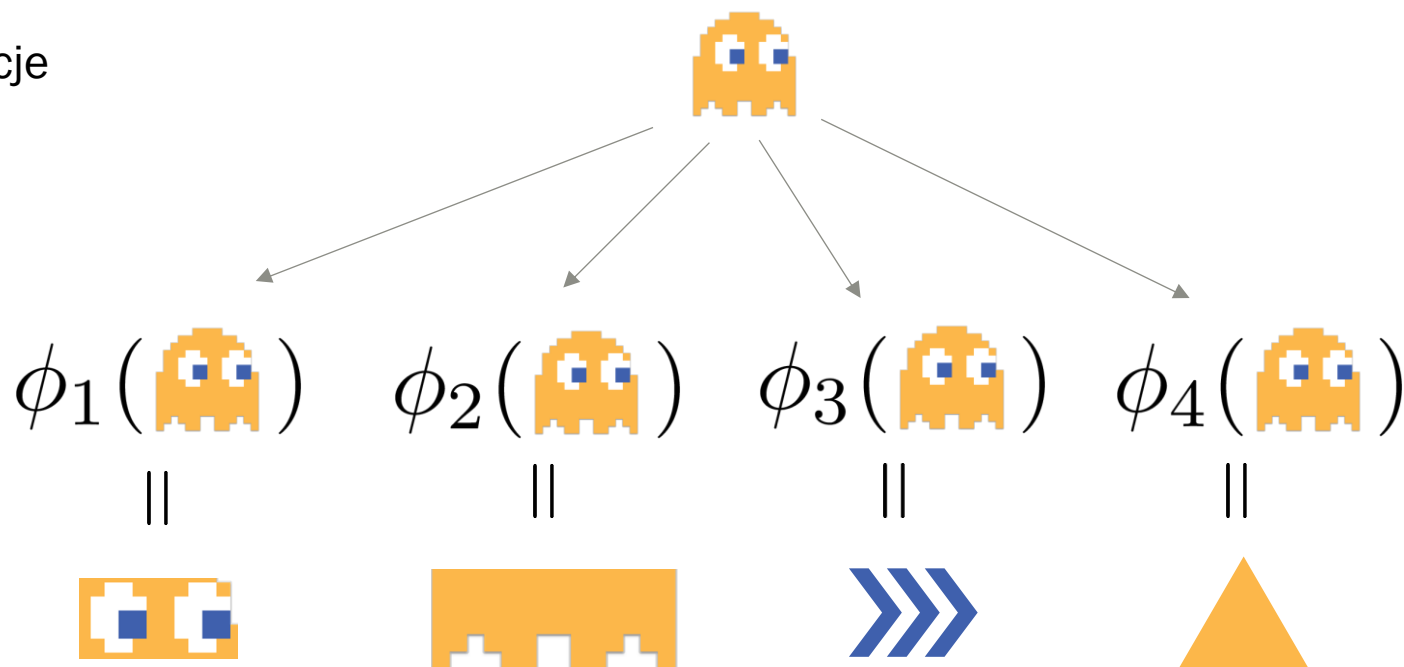


Normalizacja



Tworzenie cech pochodnych

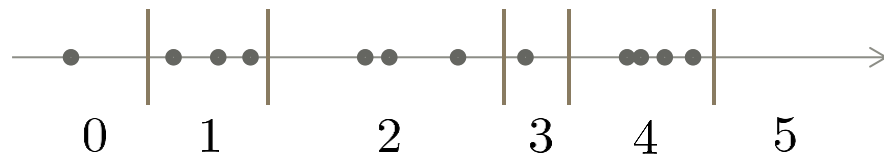
Transformacje



Przykład: *cechy wielomianowe, funkcje bazowe*

Tworzenie cech pochodnych

Dyskretyzacja



Tworzenie cech pochodnych

Kodowanie cech porządkowych

```
stan pacjenta = {  
  'zdrowy': 0,  
  'stabilny': 0.5,  
  'niestabilny': 1,  
  'chory': 4,  
  'poważnie chory': 4.5,  
  'zagrożenie życia': 8,  
}
```

```
glucose level = {  
  'Low': 0,  
  'Medium': 1,  
  'High': 2,  
}
```

Tworzenie cech pochodnych

Kodowanie cech kategorialnych / nadawanie etykiet

[lekarz, inżynier systemów, dyrektor, inżynier systemów]

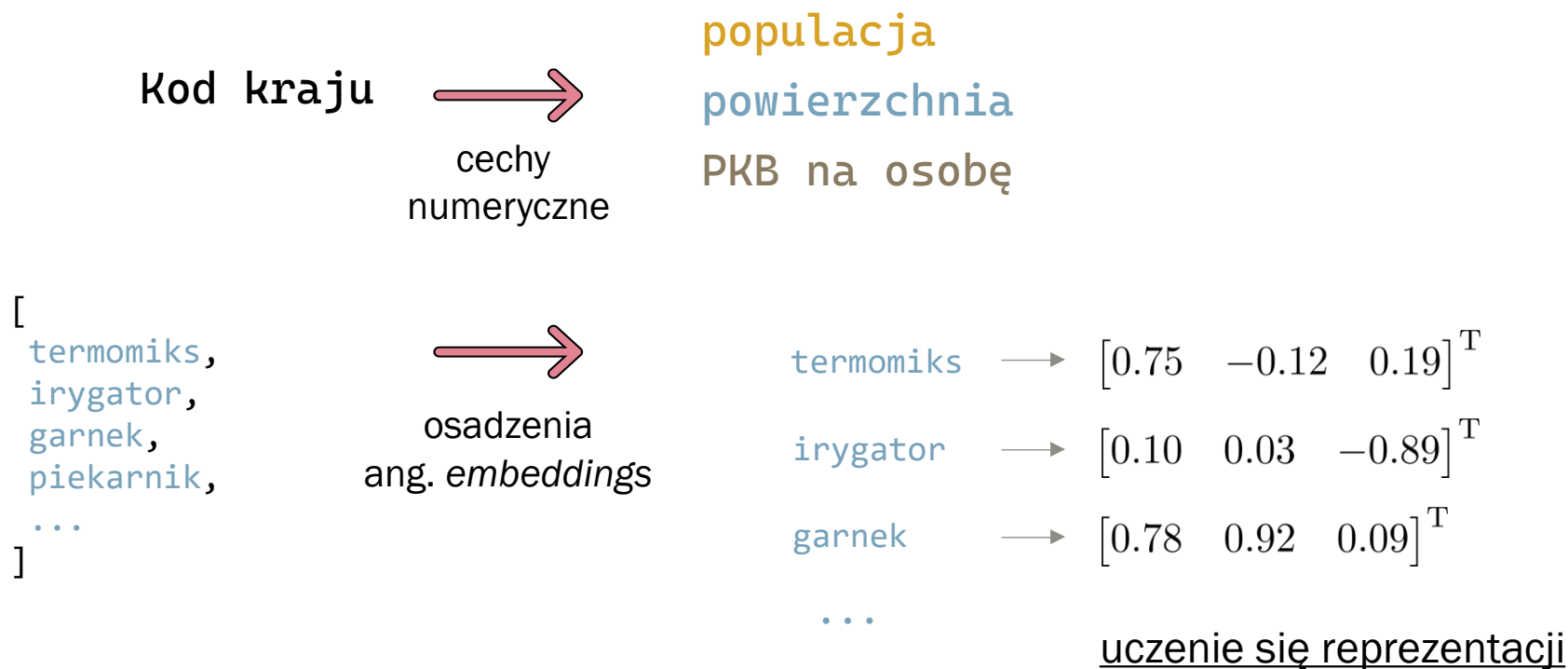
nr	lekarz	inżynier systemów	dyrektor
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

kodowanie gorącojedynekowe (ang. *one-hot encoding*)

Tworzenie cech pochodnych

Kodowanie cech kategoryalnych / nadawanie etykiet

Problem kodowania gorącojedynkowego **przy wielu kategoriach**



EKSTRAKCIJA CECH





Dane
po standaryzacji

$$\mathbf{X} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \dots \mathbf{x}_N] =$$

$x_1^{(1)}$	$x_2^{(1)}$	$\dots$	$x_N^{(1)}$	<i>cecha 1</i>
$x_1^{(2)}$	$x_2^{(2)}$	$\dots$	$x_N^{(2)}$	<i>cecha 2</i>
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$x_1^{(K)}$	$x_2^{(K)}$	$\dots$	$x_N^{(K)}$	<i>cecha K</i>
<i>obserwacja 1</i>	<i>obserwacja 2</i>	$\dots$	<i>obserwacja N</i>	

Macierz kowariancji

$$\text{Cov}[\mathbf{x}] = \frac{1}{N} \mathbf{X} \mathbf{X}^T = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1K} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2K} \\ \dots & \dots & \dots & \dots \\ \sigma_{K1} & \sigma_{K2} & \dots & \sigma_K^2 \end{bmatrix}$$

$$\Sigma_{\mathbf{x}} \equiv \frac{1}{N} \mathbf{X} \mathbf{X}^T$$

	<i>cecha 1</i>	<i>cecha 2</i>	<i>cecha 3</i>	<i>cecha 4</i>
<i>cecha 1</i>	1.00	-0.24	0.66	0.60
<i>cecha 2</i>	-0.24	1.00	-0.59	-0.47
<i>cecha 3</i>	0.66	-0.59	1.00	0.87
<i>cecha 4</i>	0.60	-0.47	0.87	1.00

$\text{Cov}[\mathbf{x}]$

Pytanie problemowe: Czy można tak obrócić osie układu współrzędnych,
aby **nowe cechy** wyglądały na zdekorelowane?

$$\mathbf{z}_i \leftarrow \text{zmiana bazy} \leftarrow \mathbf{x}_i \quad \text{dla } i = 1, 2, \dots, N$$

składowa główna = nowa cecha (w obróconej bazie)

$$\mathbf{Z} \leftarrow \text{zmiana bazy} \leftarrow \mathbf{X}$$

$$\mathbf{Z} = \mathbf{V}^T \mathbf{X}$$

$\mathbf{V}$ – macierz obrotu

$$\mathbf{V}^T = \mathbf{V}^{-1}$$

	składowa 1	składowa 2	składowa 3	składowa 4
składowa 1	2.76	0.00	0.00	-0.00
składowa 2	0.00	0.77	0.00	-0.00
składowa 3	0.00	0.00	0.37	0.00
składowa 4	-0.00	-0.00	0.00	0.11

$\text{Cov}[\mathbf{z}]$

	cecha 1	cecha 2	cecha 3	cecha 4
cecha 1	1.00	-0.24	0.66	0.60
cecha 2	-0.24	1.00	-0.59	-0.47
cecha 3	0.66	-0.59	1.00	0.87
cecha 4	0.60	-0.47	0.87	1.00

$\text{Cov}[\mathbf{x}]$

←
zmiana bazy

Rozwiązanie: 1) Wyznaczyć macierz kowariancji danych po obrocie w zależności od $\mathbf{V}$
2) Znaleźć taką macierz obrotu $\mathbf{V}$, dla której macierz kowariancji danych jest diagonalna

$$\text{Cov}[\mathbf{z}] = \frac{1}{N} \mathbf{Z} \mathbf{Z}^T = \frac{1}{N} \mathbf{V}^T \mathbf{X} (\mathbf{V}^T \mathbf{X})^T = \frac{1}{N} \mathbf{V}^T \mathbf{X} \mathbf{X}^T \mathbf{V} = \mathbf{V}^T \Sigma_{\mathbf{x}} \mathbf{V}$$

Nowe cechy (składowe główne) mają być zdekorelowane:

$$\mathbf{D} \equiv \text{Cov}[\mathbf{z}] = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_K \end{bmatrix}$$

$$\mathbf{V}^T \Sigma_{\mathbf{x}} \mathbf{V} = \mathbf{D}$$

$$\Sigma_{\mathbf{x}} \mathbf{V} = \mathbf{V} \mathbf{D}$$



$\mathbf{V}$ - wektory własne macierzy kowariancji $\Sigma_{\mathbf{x}}$

$\text{diag}(\mathbf{D})$ - wartości własne macierzy kowariancji $\Sigma_{\mathbf{x}}$

Analiza składowych głównych

Principal Component Analysis



$$\mathbf{V} = [\mathbf{v}_1 \quad \dots \quad \mathbf{v}_K]$$

$$\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_S = \lambda_{\min}$$



– ciach! –

wariancja
rzutowanych
danych

$$\lambda_1 + \dots + \lambda_D + \lambda_{D+1} + \dots + \lambda_K$$

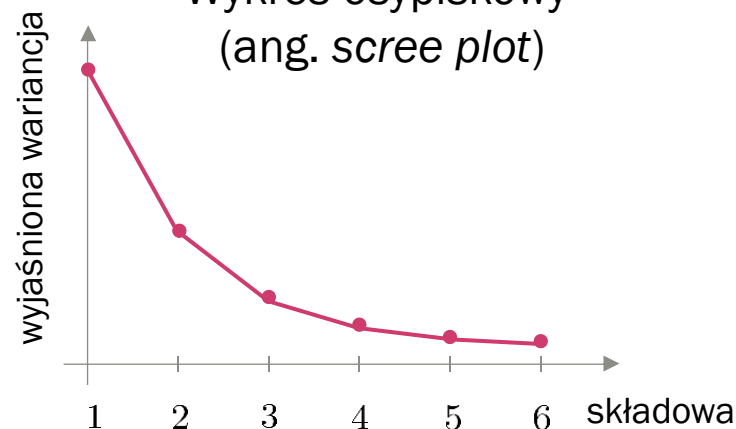
$$\lambda_1 + \lambda_2 + \dots + \lambda_K$$

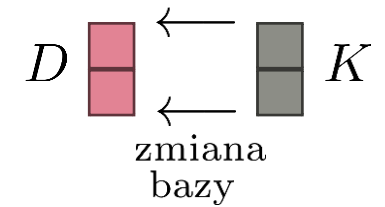
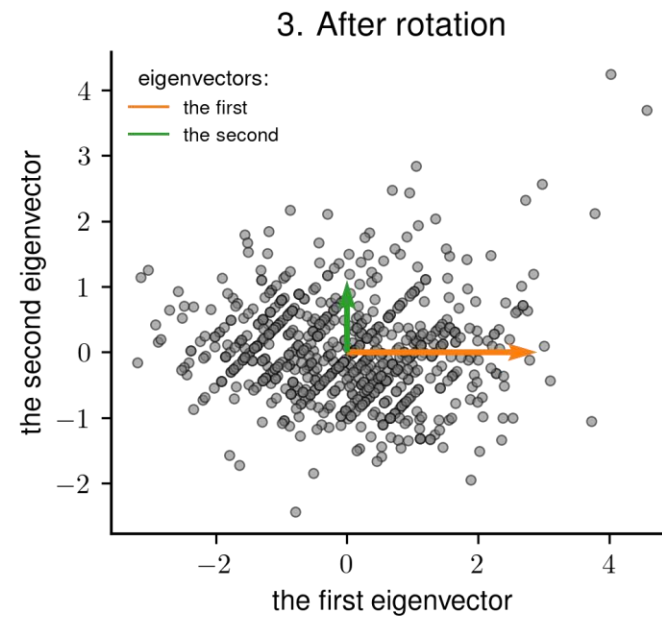
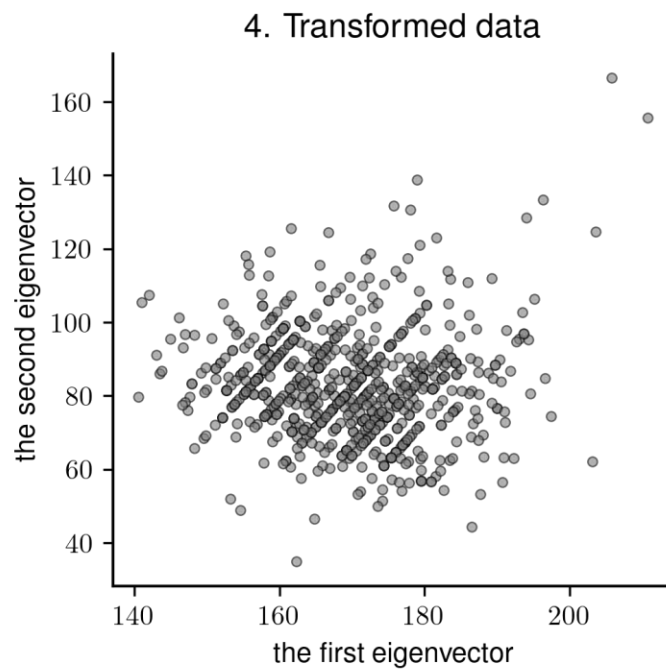
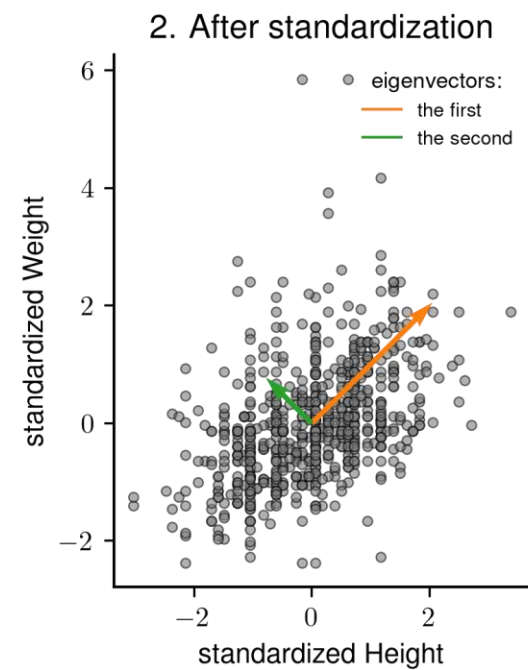
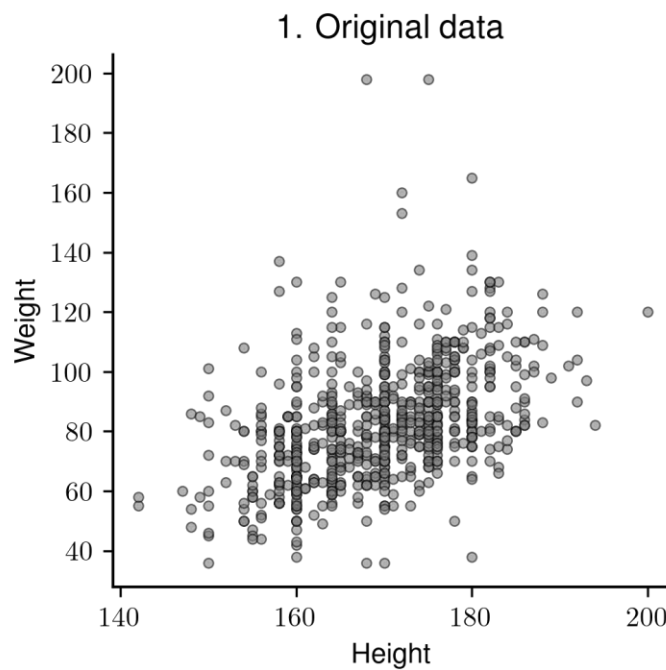
błąd
aproksymacji
danych

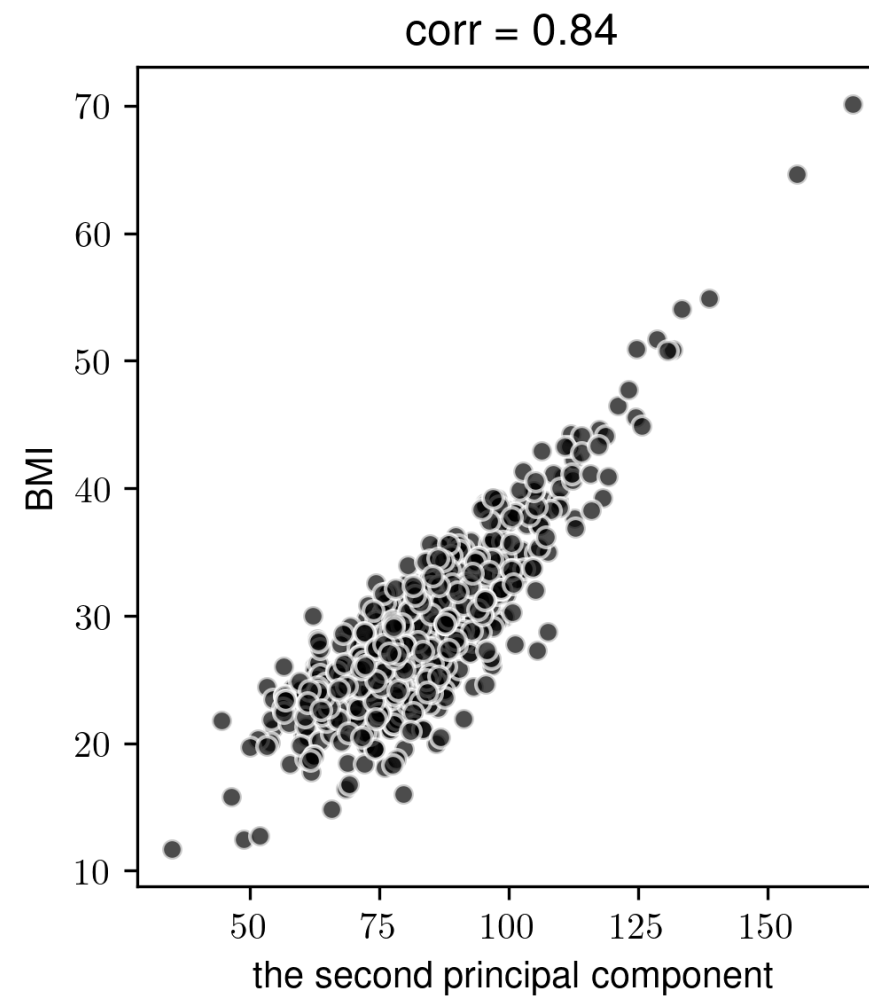
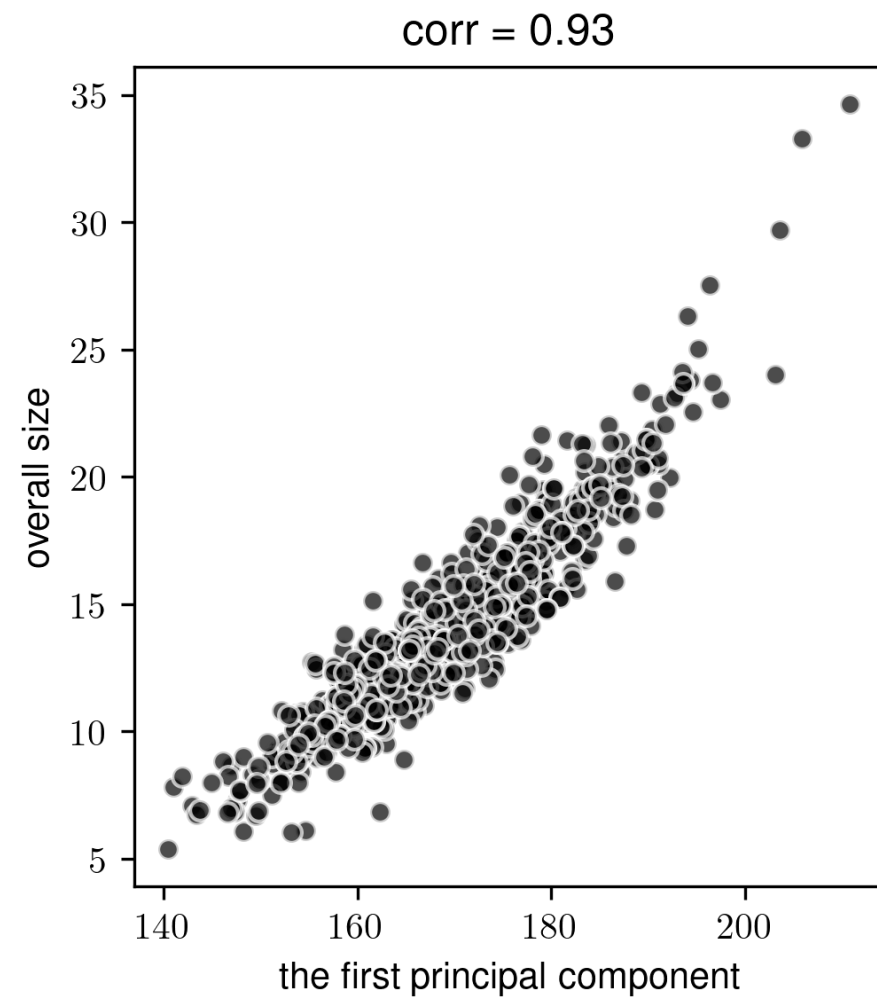
$$\frac{\lambda_1 + \dots + \lambda_D}{\lambda_1 + \lambda_2 + \dots + \lambda_K}$$

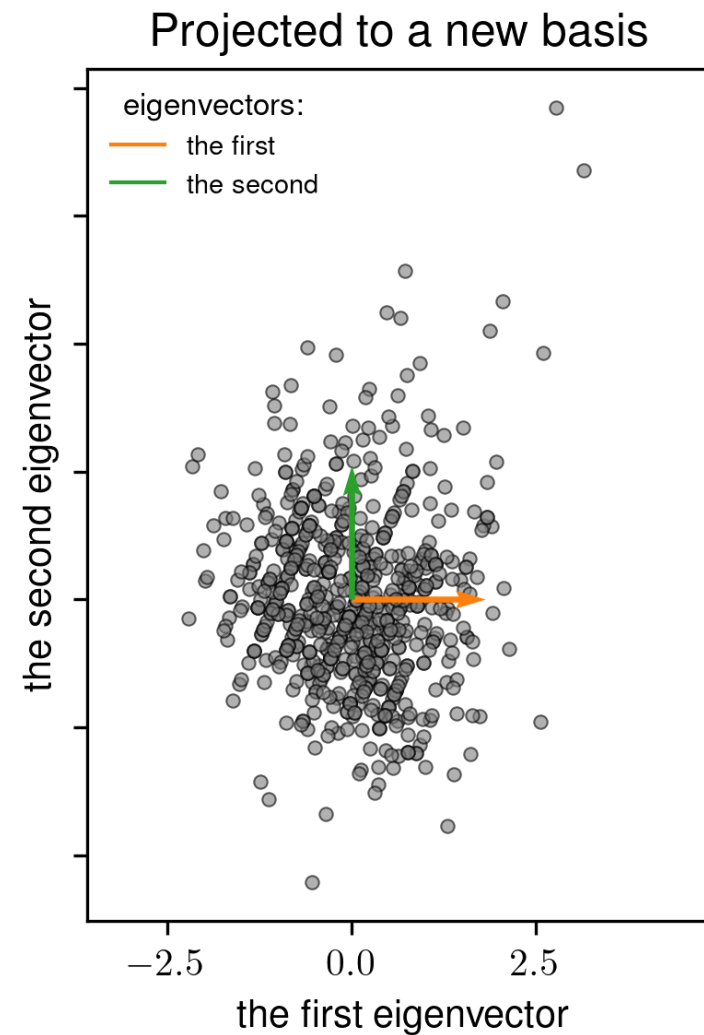
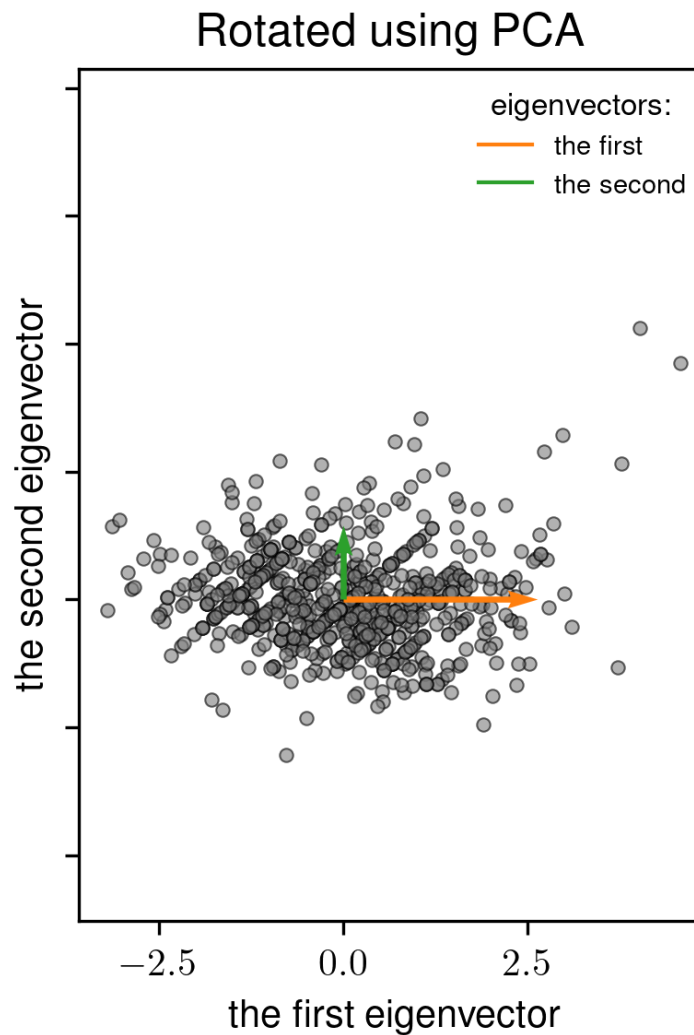
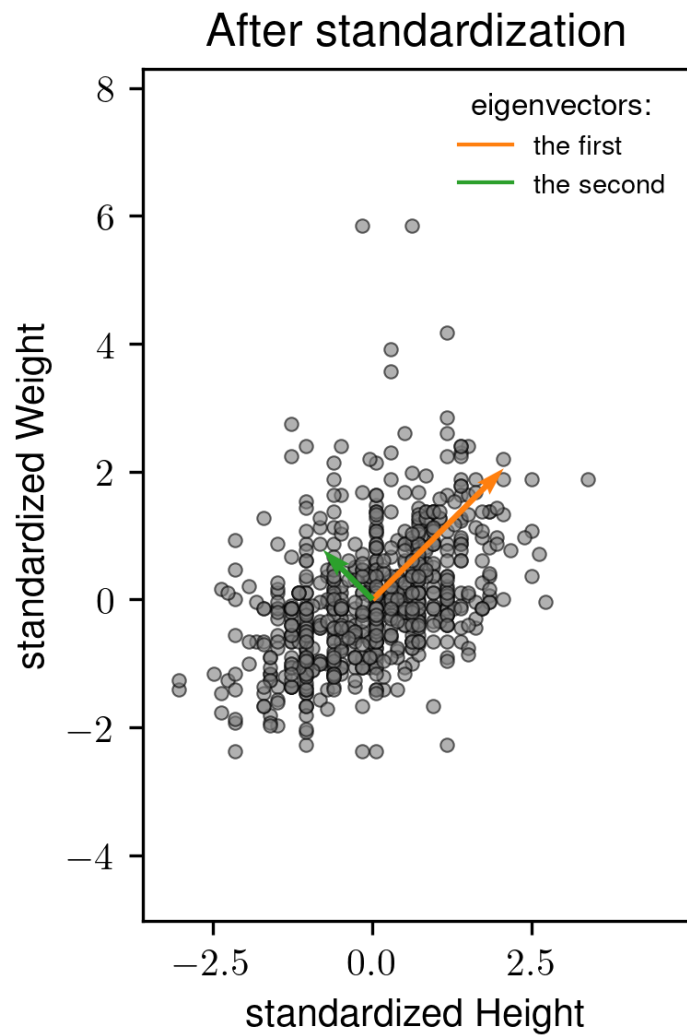
Wyjaśniona wariancja

Wykres osypiskowy
(ang. *scree plot*)





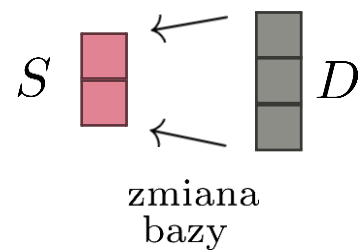
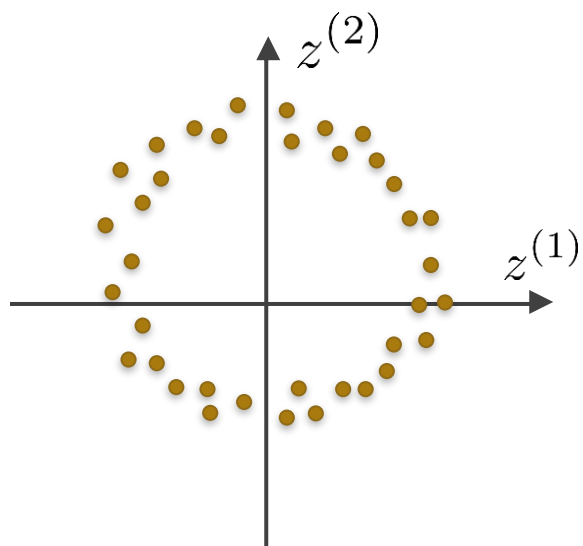




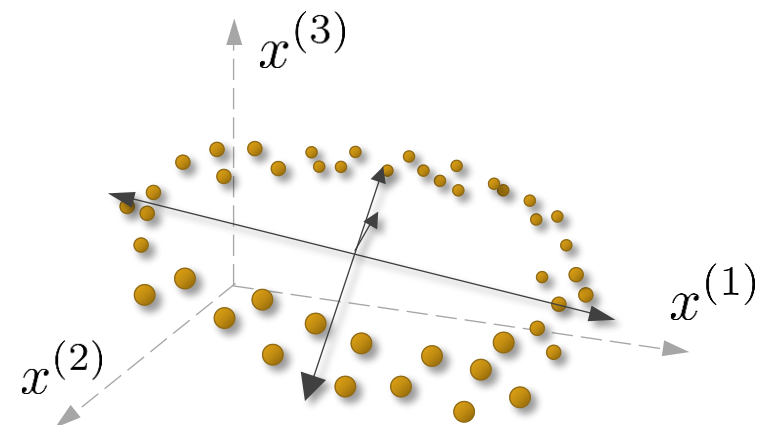
REDUKCJA WYMIAROWOŚCI



Redukcja cech dla modelu

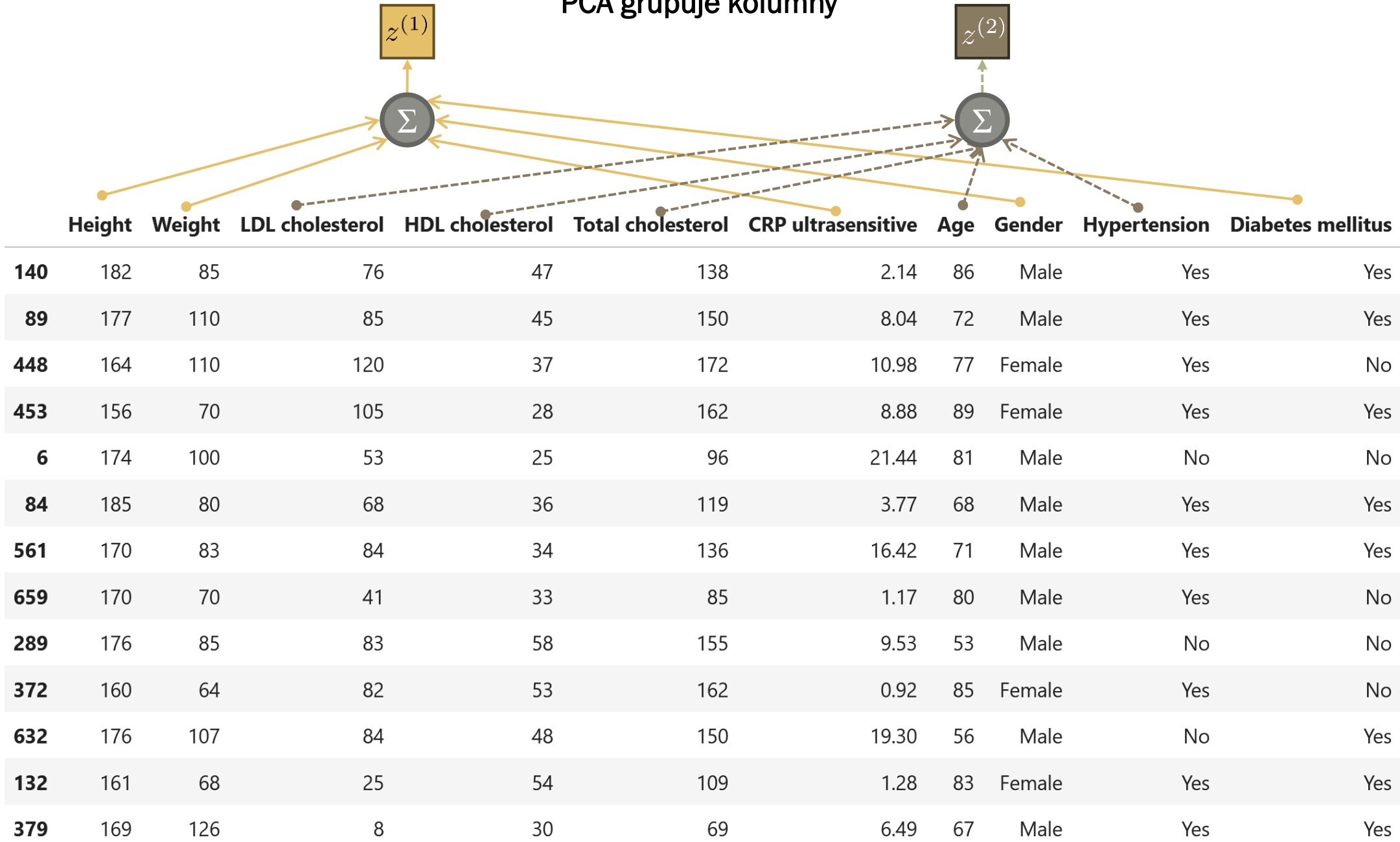


$$\mathbf{z} = \begin{bmatrix} z^{(0)} \\ z^{(1)} \\ \vdots \\ z^{(S)} \end{bmatrix}$$



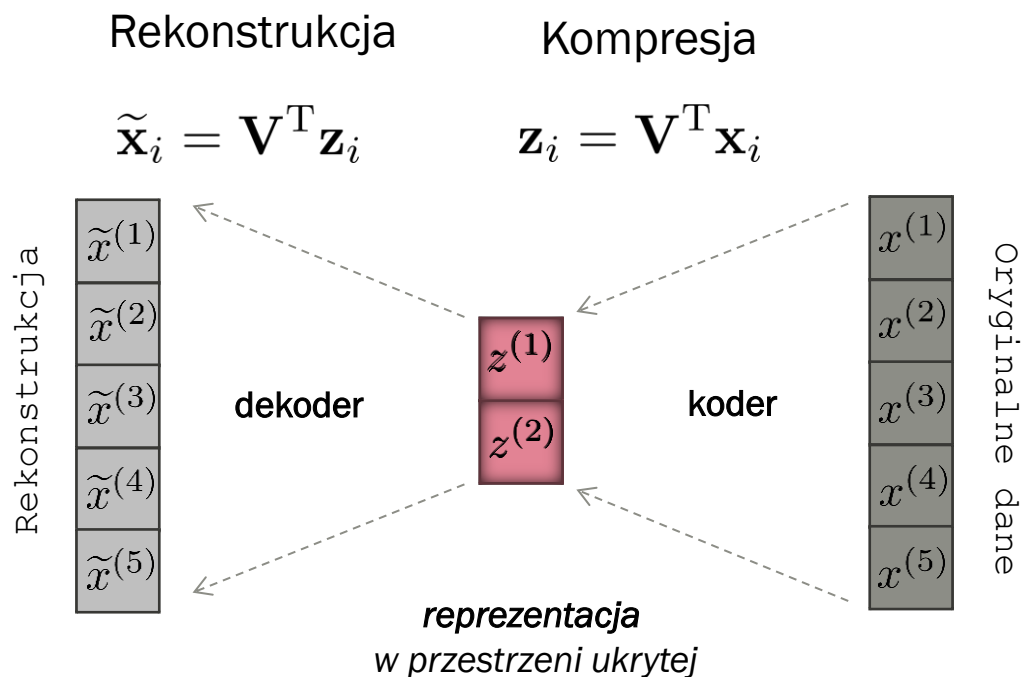
$$\mathbf{x} = \begin{bmatrix} x^{(0)} \\ x^{(1)} \\ \vdots \\ x^{(D)} \end{bmatrix}$$

PCA grupuje kolumny



Grupowanie grupuje dane

Schemat koder – dekodek



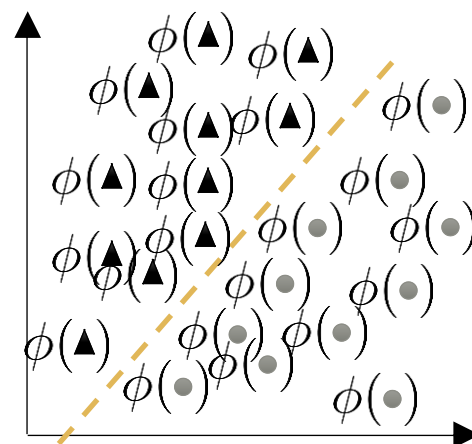
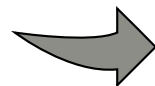
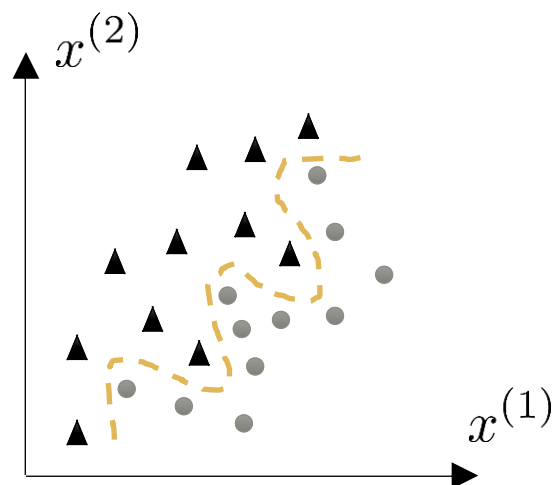
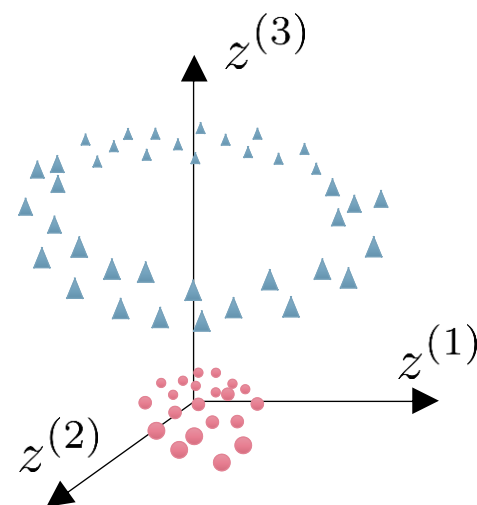
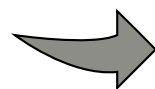
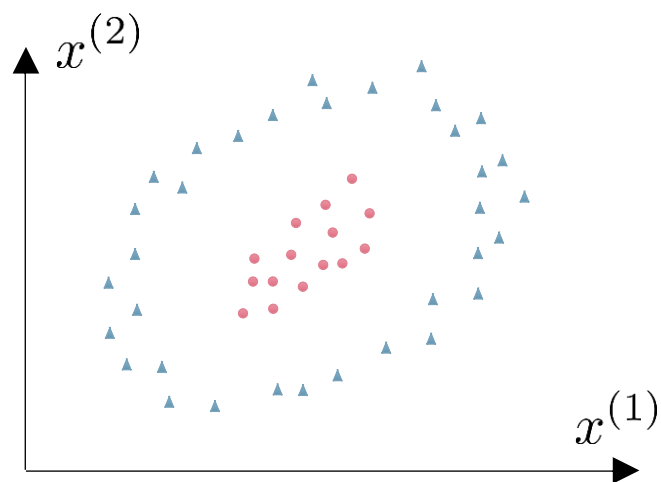
PCA minimalizuje błąd rekonstrukcji

$$\sum_i \|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|_2^2$$

Karhunen-Loève Expansion

- gdy są dostępne etykiety C
- równoważne PCA z podziałem na $\Sigma(C)$ dla każdej klasy C oddzielnie

Jądrowa PCA (ang. *kernel PCA*)



WEKTORY WŁASNE



Przykład 1

- <https://psycnet.apa.org/record/1981-32589-001>
- 1634 studentów z Los Angeles
- Każda substancja oceniona w skali 1-5

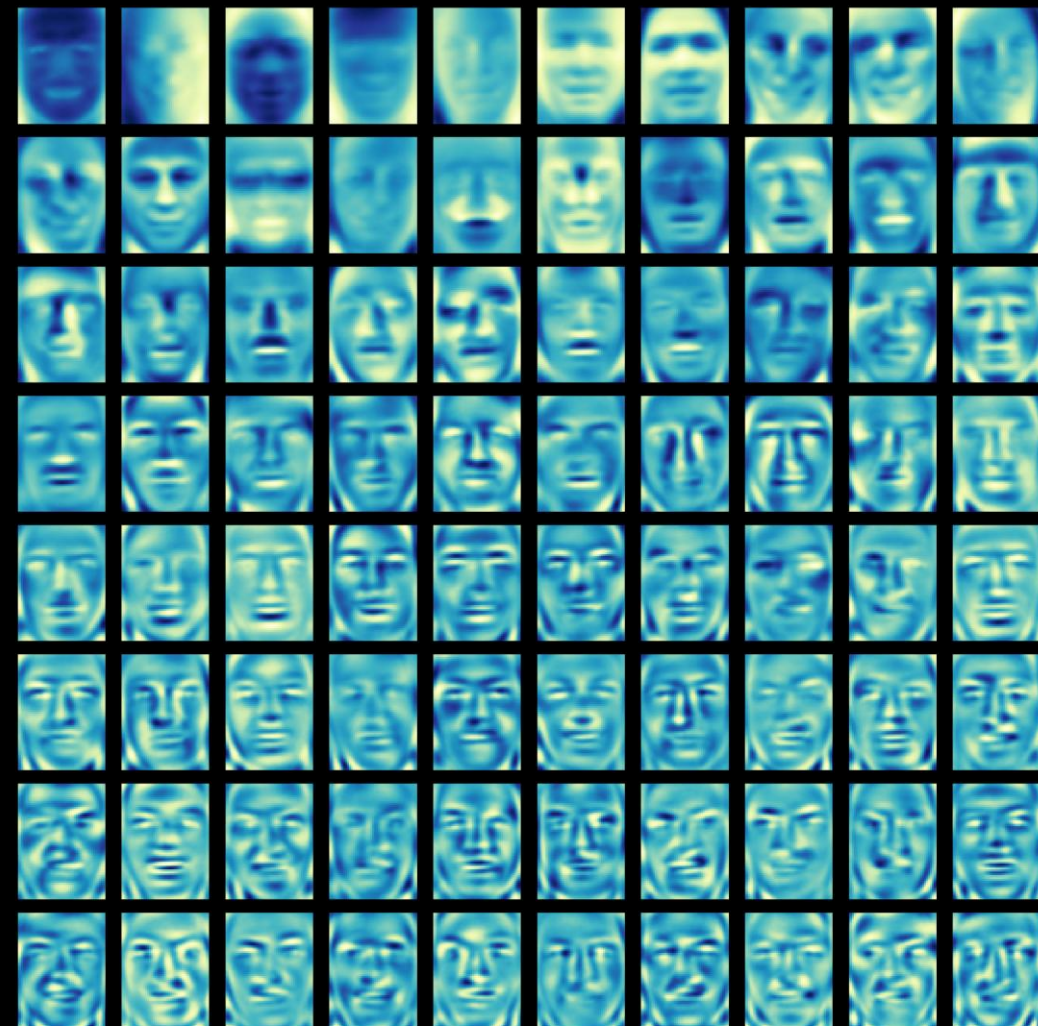
zażywanie	punktacja
nigdy	1
tylko raz	2
kilka razy	3
wiele razy	4
regularnie	5

substancja	v_1	v_2
Papierosy	0.28	0.28
Piwo	0.29	0.40
Wino	0.26	0.39
Wódka	0.32	0.32
Kokaina	0.20	-0.29
Środki uspokajające	0.29	-0.26
Leki odurzające	0.18	-0.19
Heroina i opiaty	0.20	-0.32
Marihuana	0.34	-0.16
Haszysz	0.33	-0.05
Kleje itp.	0.28	-0.17
Halucynogeny	0.25	-0.33
Amfetamina	0.33	-0.23

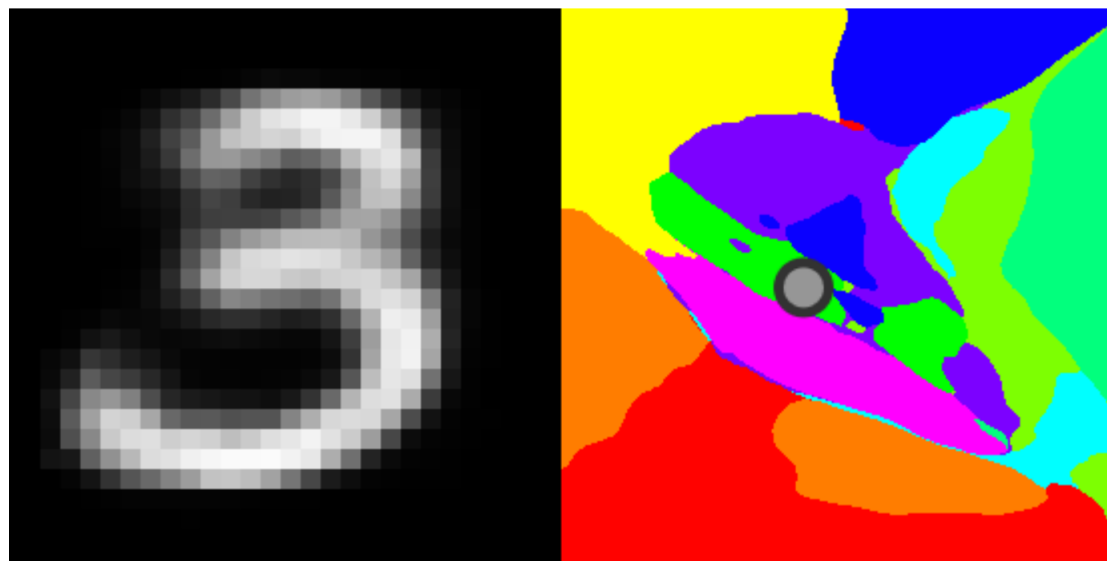
$$62 \times 47 = 2917$$



$$62 \times 47 = 2917$$



„eigen-twarze”



Liniowa analiza dyskryminacyjna

Linear Discriminant Analysis

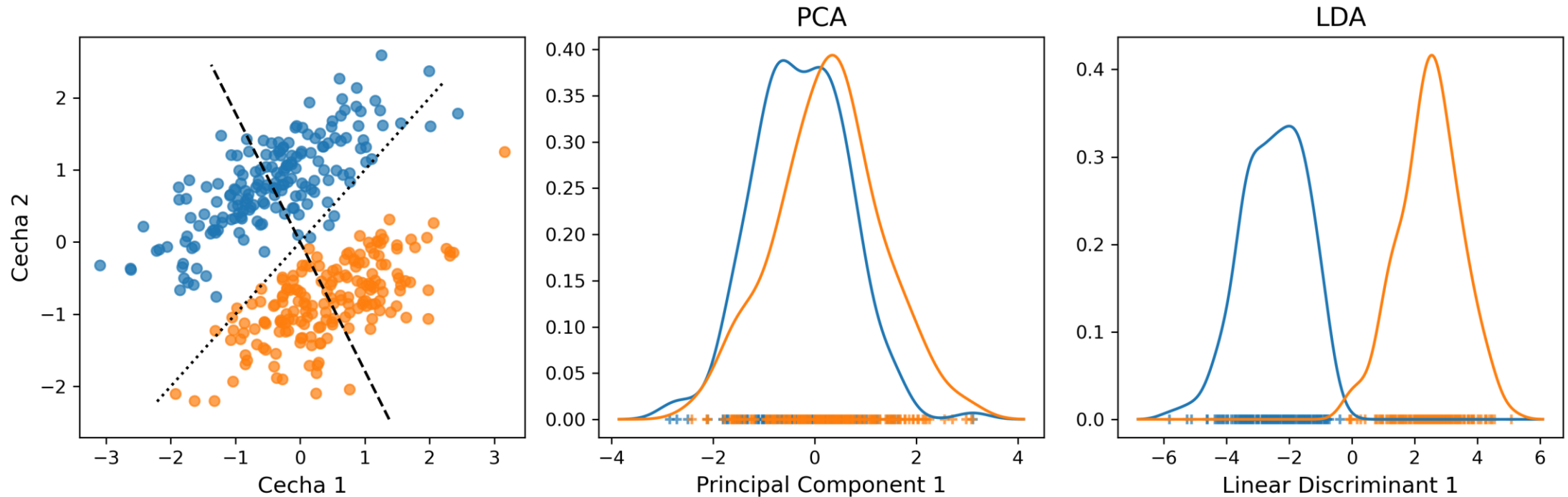
PCA

- nienadzorowana

LDA

- nadzorowana

Szukamy takiej podprzestrzeni cech, która **najlepiej separuje klasy**



The End