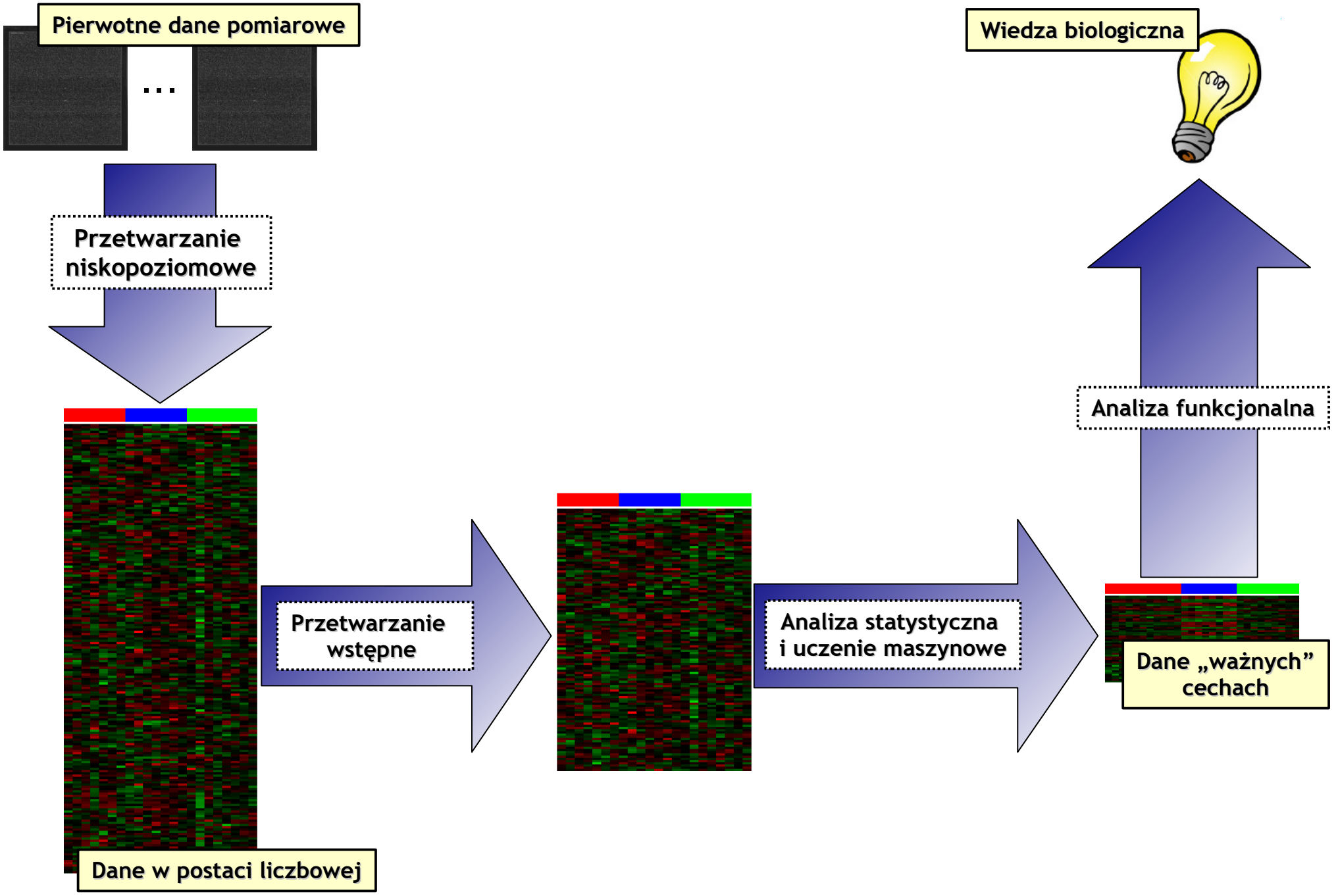
Uczenie maszynowe w bioinformatyce

Wykład 5: przetwarzanie wstępne danych

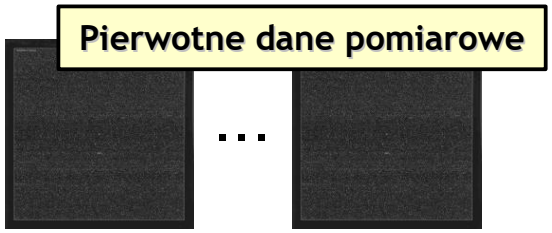
Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

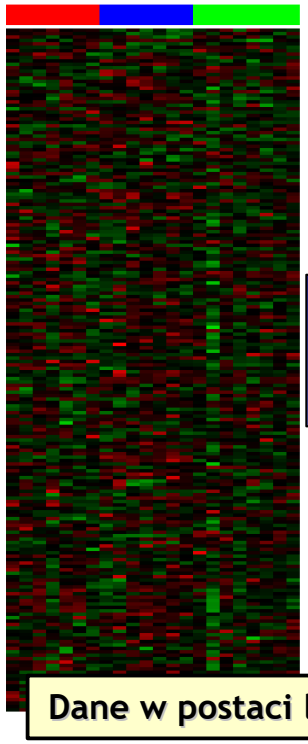
Etapy przetwarzania i analizy danych



Etapy przetwarzania i analizy danych



Przetwarzanie niskopoziomowe



Dane w postaci liczbowej

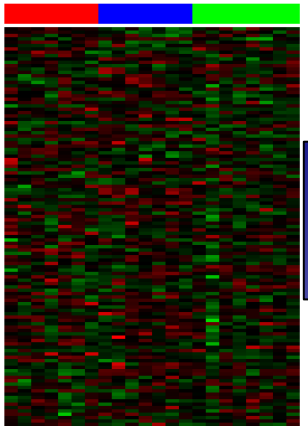
Przetwarzaniem wstępnym określa się ogół czynności mających na celu poprawę jakości danych (eliminację artefaktów pomiarowych oraz wpływu czynników o niebiologicznym pochodzeniu) i doprowadzenie ich do postaci ułatwiającej dalszą analizę.

Kroki przetwarzania wstępnego i stosowane metody są zazwyczaj niezależne od celu dalszej analizy danych. Zwykle też na tym etapie nie wykorzystuje się wiedzy o przynależności próbek do grup badanych.

Typowymi krokami przetwarzania danych z technik wielkoskalowych są:

- eliminacja brakujących i odstających wartości;
- usunięcie cech nieistotnych z punktu widzenia analizy;
- ograniczenie zakresu wartości i zmiana ich skali;
- normalizacja.

Przetwarzanie wstępne



Analiza statystyczna i uczenie maszynowe

Dane „ważnych” cechach

Uczenie maszynowe w bioinformatyce

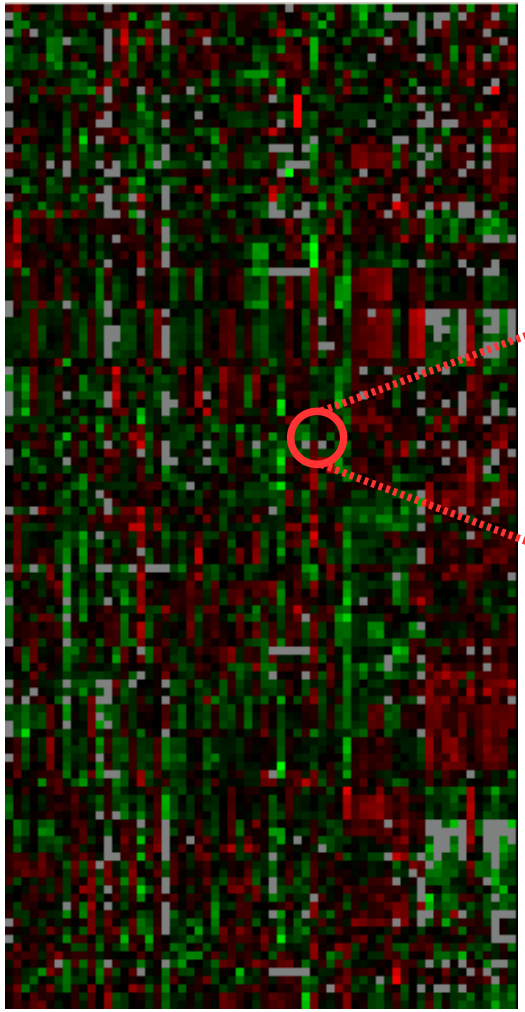
Wykład 5: przetwarzanie wstępne danych (imputacja brakujących wartości)

Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Przetwarzanie wstępne: imputacja brakujących wartości

Dane mogą zawierać brakujące wartości, wynikające np. z niedoskonałości procesu pomiarowego lub błędów na wcześniejszych etapach przetwarzania. W zależności od typu danych udział brakujących wartości może sięgać nawet do kilkunastu procent.



	14.14	
	10.52	
.82	8.96	10.9
7.98	10.02	10.75
8.99	8.73	7.24
10.36	11.28	12.66
11.11	11.33	13.41
NaN	10.37	NaN
10.66	12.20	13.46
11.07	10.39	12.12
9.01	8.89	10.54
7.43	8.05	9.81
.99	8.52	9.8
	11.17	
	8.51	

Przetwarzanie wstępne: imputacja brakujących wartości

Większość metod analizy statystycznej i uczenia się maszyn wymaga danych bez brakujących wartości. Aby uniknąć konieczności usuwania niekompletnych wierszy można użyć metod imputacji, czyli odtworzenia brakujących wartości na podstawie pozostałych, prawidłowo określonych elementów macierzy danych.

Trywialnymi sposoby imputacji brakujących wartości są np.:

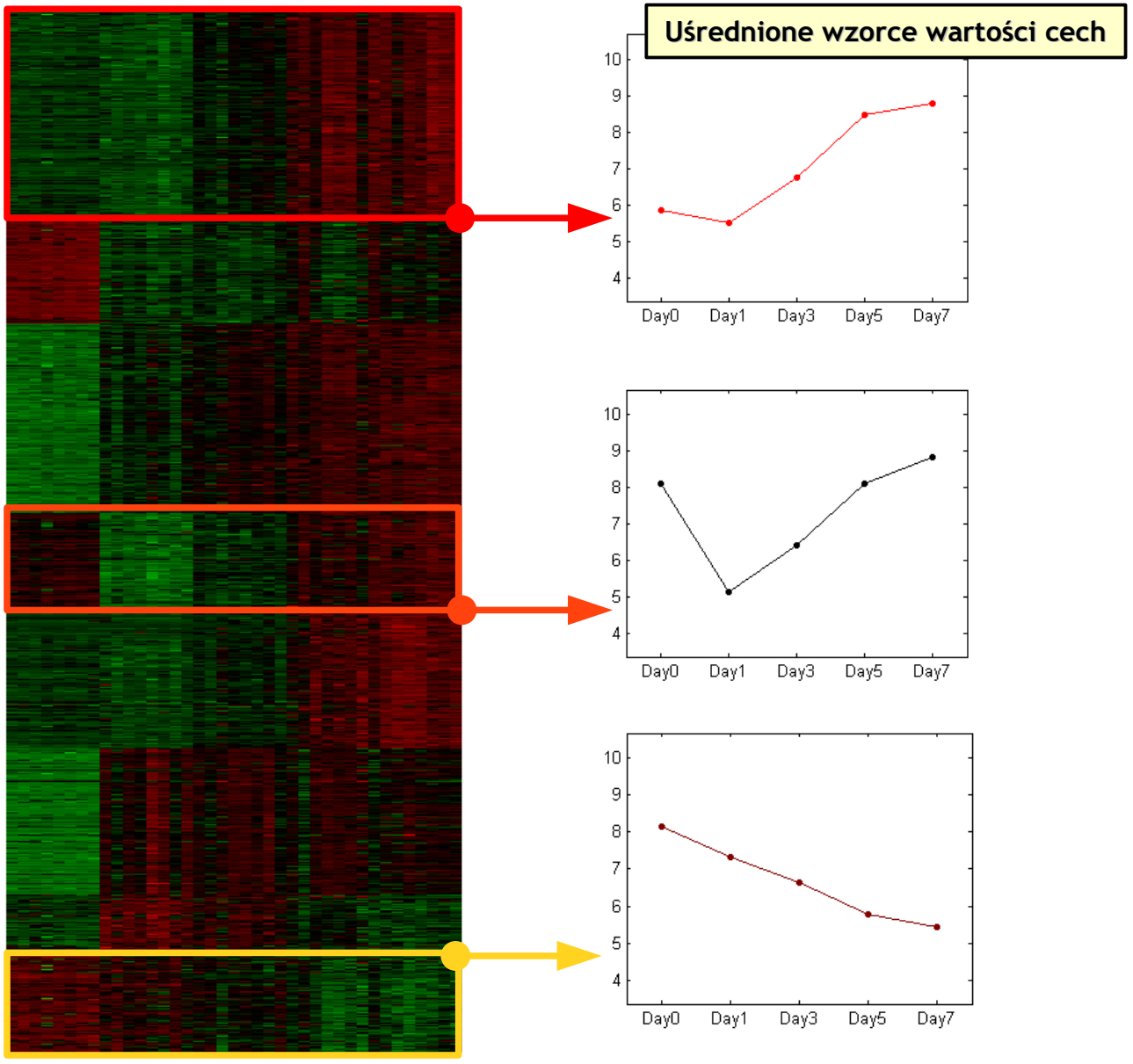
- **zastąpienie ustaloną wartością (np. zerem);**
- **zastąpienie średnią arytmetyczną lub medianą prawidłowo określonych wartości z tego samego wiersza.**

Przykładami skuteczniejszych metod imputacji brakujących wartości są:

- **algorytm *K*-najbliższych sąsiadów;**
- **algorytm E-M (*Expectation-Maximization*);**
- **użycie rozkładu na wartości szczególne (SVD – *Singular Value Decomposition*).**

Przetwarzanie wstępne: imputacja brakujących wartości (algorytm *K*-NN)

Algorytm *K*-najbliższych sąsiadów (*K-Nearst Neighbours, K-NN*), działa w oparciu o założenie, że istnieją duże grupy cech (np. genów), których wartości są skorelowane.



Przetwarzanie wstępne: imputacja brakujących wartości (algorytm *K*-NN)

Aby oszacować brakującą wartość cechy i w próbce j, szukamy K cech o najbardziej podobnych wzorcach, a następnie poszukiwaną wartość $\tilde{x}_{ij}$ wyznaczamy jako średnią ważoną wartości w próbce j dla cech należących do sąsiedztwa:

$$\tilde{x}_{ij} = \frac{\sum_{k=1}^{K} w_k x_{kj}}{\sum_{k=1}^{K} w_k}$$

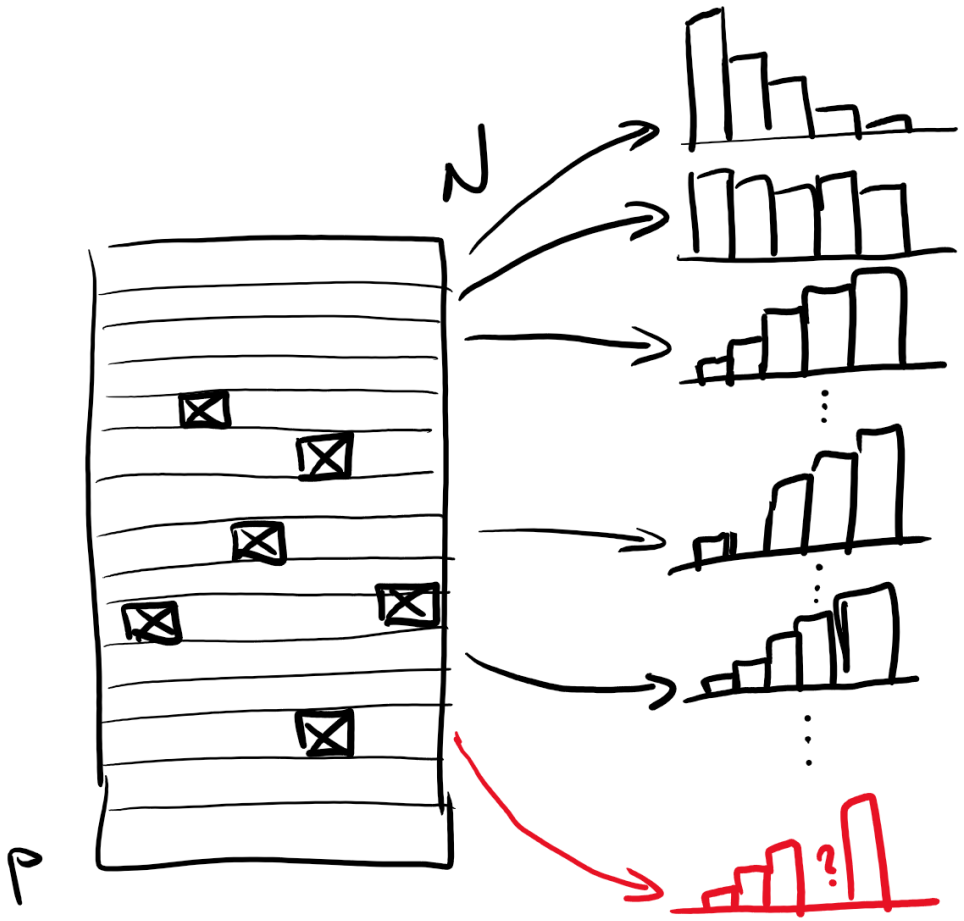
Wagi określane są jako odwrotności odległości Euklidesa $d_E(i,k)$ między daną cechą, a cechami z sąsiedztwa (pomiędzy i-tym a k-tymi wierszami macierzy danych X):

$$w_k = \frac{1}{d_E(i,k)} = \frac{1}{\sqrt{\left(\sum_{j=1}^{N} (x_{ij} - x_{kj})^2\right)}}$$

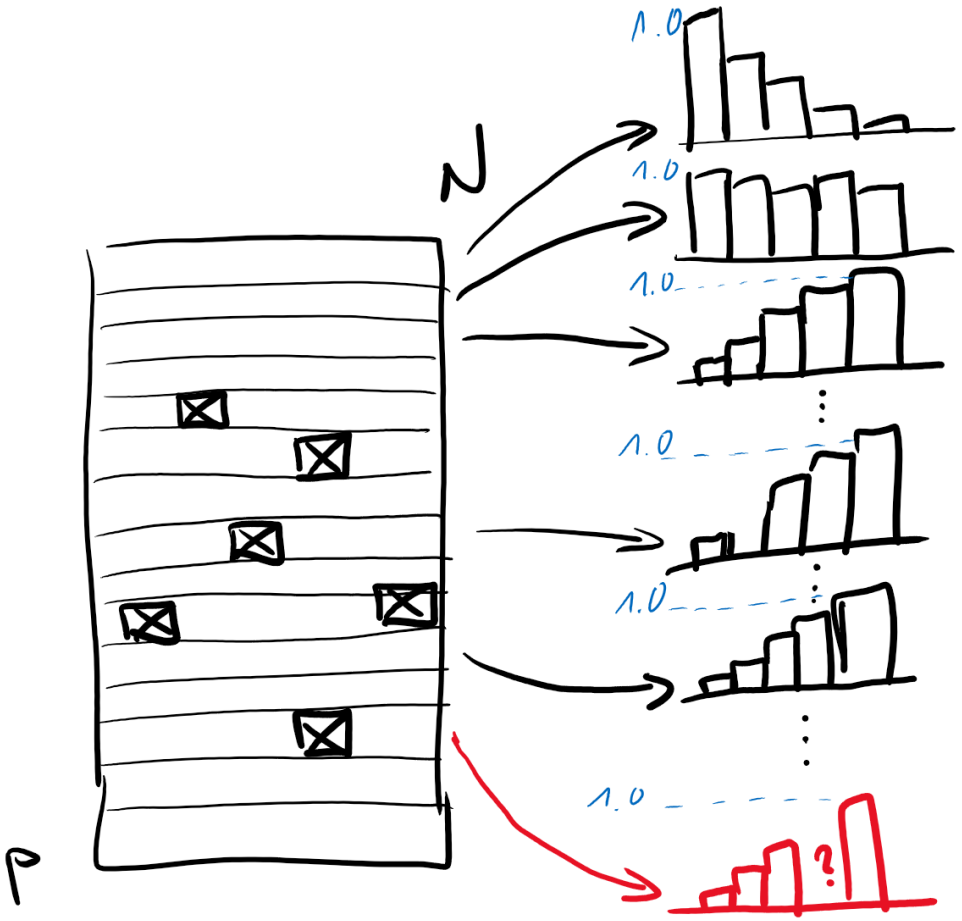
Poprawę skuteczności można uzyskać stosując dynamiczny rozmiar sąsiedztwa (wtedy K jest jego maksymalną wielkością) w połączeniu z miarą niepodobieństwa równą $1 - r_{ik}$, gdzie r_{ik} jest współczynnikiem korelacji i-tego i k-tego wiersza macierzy.

	MATLAB	R
KDE:	`knnimpute`	`knnImputation`

Przetwarzanie wstępne: imputacja brakujących wartości (algorytm *K*-NN)

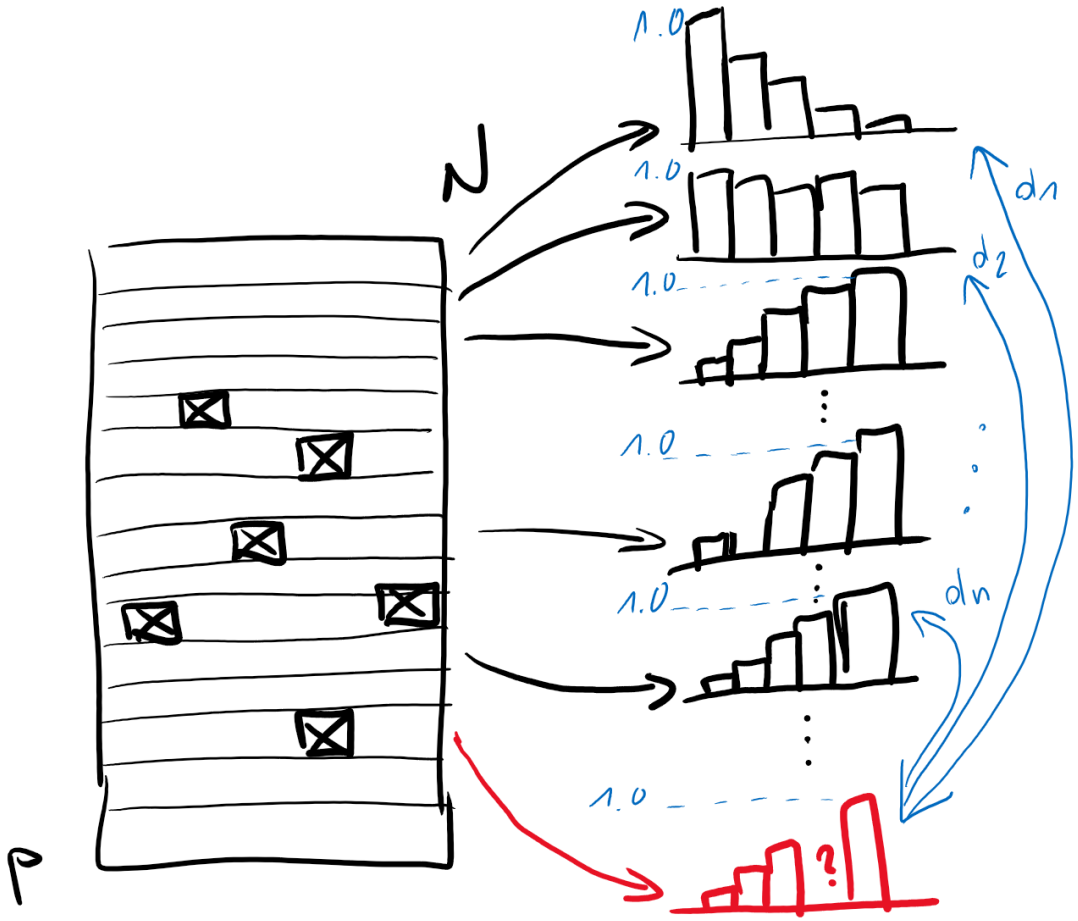


Przetwarzanie wstępne: imputacja brakujących wartości (algorytm *K*-NN)



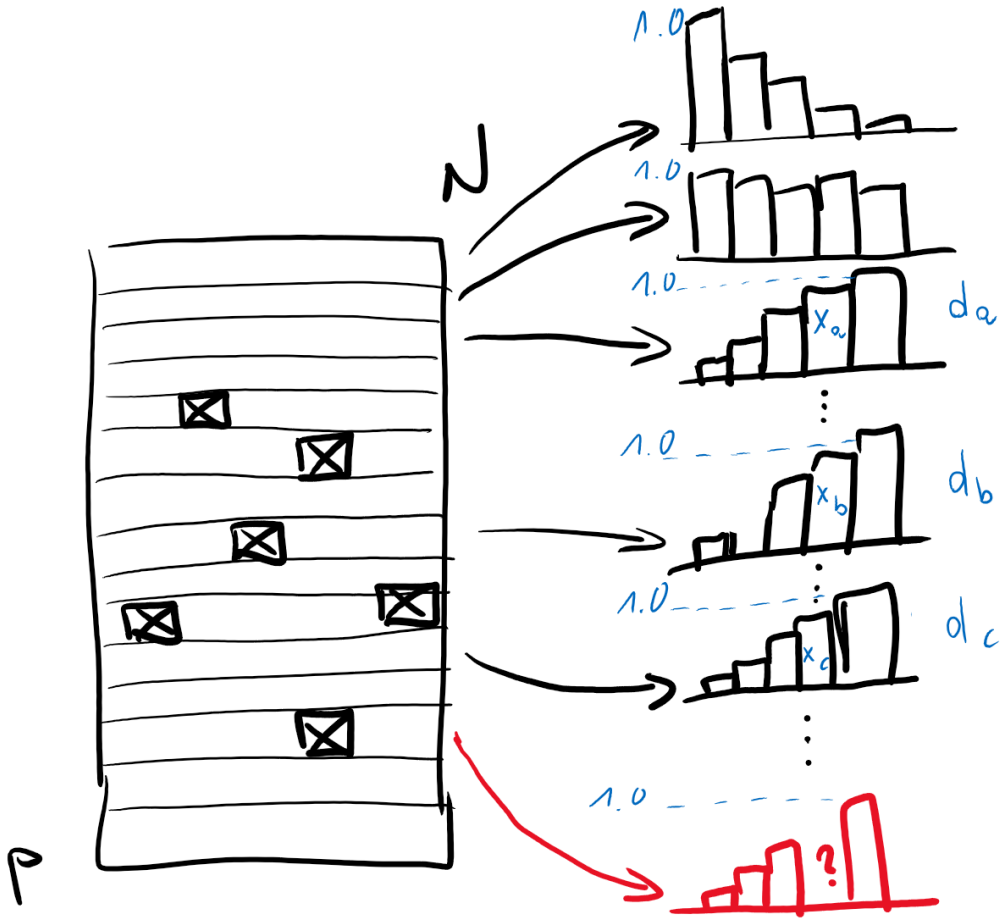
1. PODZIELENIE PRZEZ MAX

Przetwarzanie wstępne: imputacja brakujących wartości (algorytm *K*-NN)



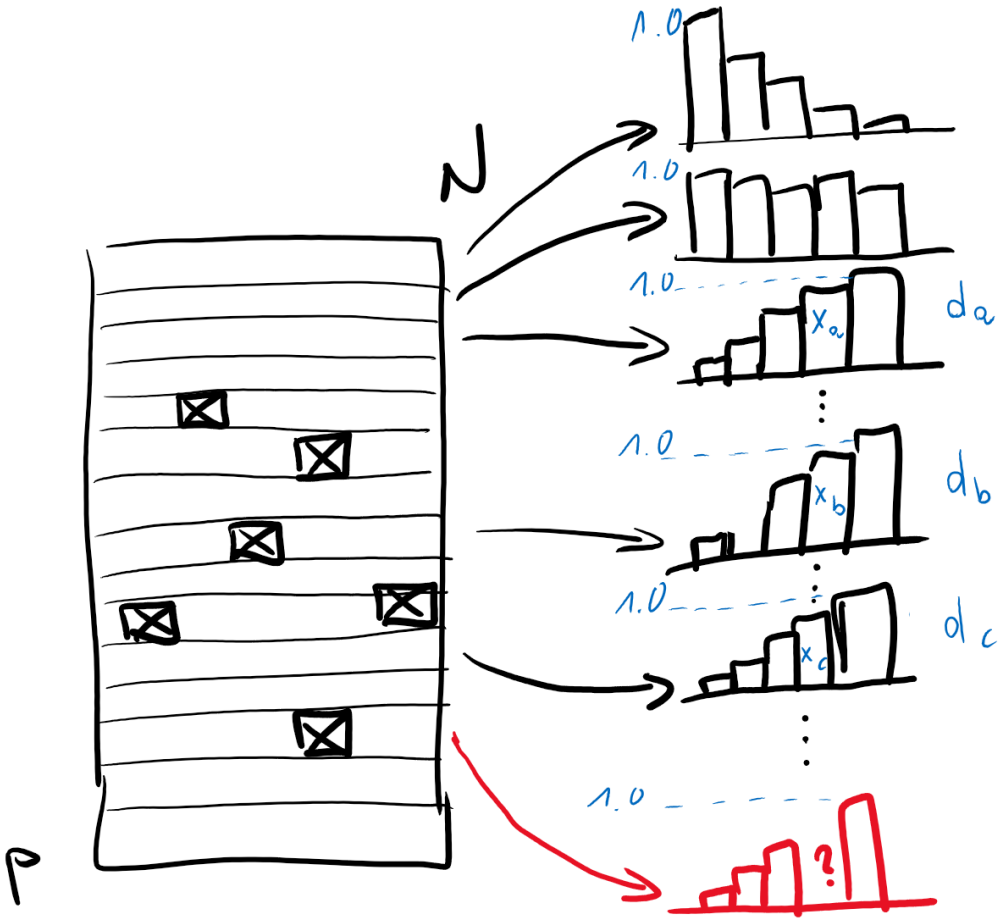
1. PODZIELENIE PRZEZ MAX
2. WYZNACZENIE ODLEGŁOŚCI

Przetwarzanie wstępne: imputacja brakujących wartości (algorytm *K*-NN)



1. PODZIELENIE PRZEZ MAX
2. WYZNACZENIE ODLEGŁOŚCI
3. WYBÓR K NAJBLIŻSZYCH (O MIN. ODLEGŁOŚCIACH)

Przetwarzanie wstępne: imputacja brakujących wartości (algorytm *K*-NN)

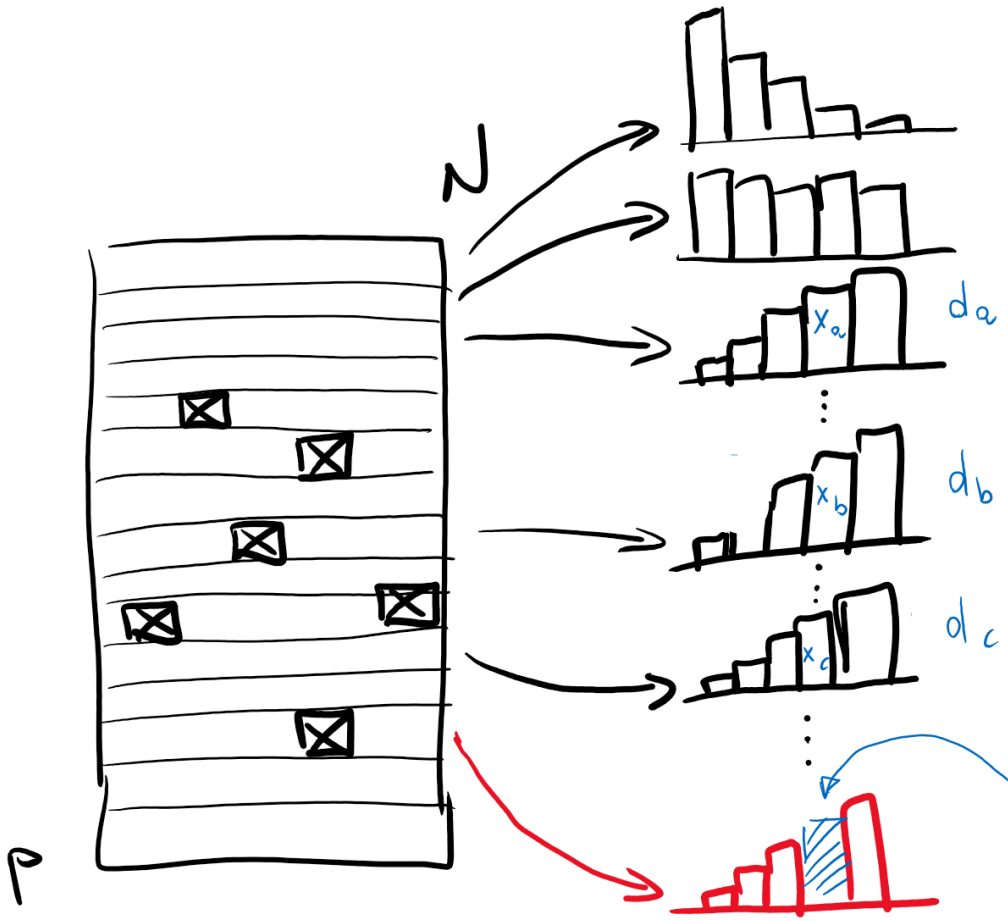


1. PODZIELENIE PRZEZ MAX
2. WYZNACZENIE ODLEGŁOŚCI
3. WYBÓR K NAJBLIŻSZYCH (O MIN. ODLEGŁOŚCIACH)
4. POLICZENIE ŚREDNIEJ WAŻONEJ

$$x_? = \frac{w_a \cdot x_a + w_b \cdot x_b + w_c \cdot x_c}{w_a + w_b + w_c}$$

$$w_i = \frac{1}{d_i}$$

Przetwarzanie wstępne: imputacja brakujących wartości (algorytm *K*-NN)



1. PODZIELENIE PRZEZ MAX
2. WYZNACZENIE ODLEGŁOŚCI
3. WYBÓR K NAJBLIŻSZYCH (O MIN. ODLEGŁOŚCIACH)
4. POLICZENIE ŚREDNIEJ WAŻONEJ

$$x_? = \frac{w_a \cdot x_a + w_b \cdot x_b + w_c \cdot x_c}{w_a + w_b + w_c}$$

$$w_i = \frac{1}{d_i}$$

5. PRZYWRÓCENIE PIERWOTNEJ SKALI

$$x_? = x_? \cdot max$$

Uczenie maszynowe w bioinformatyce

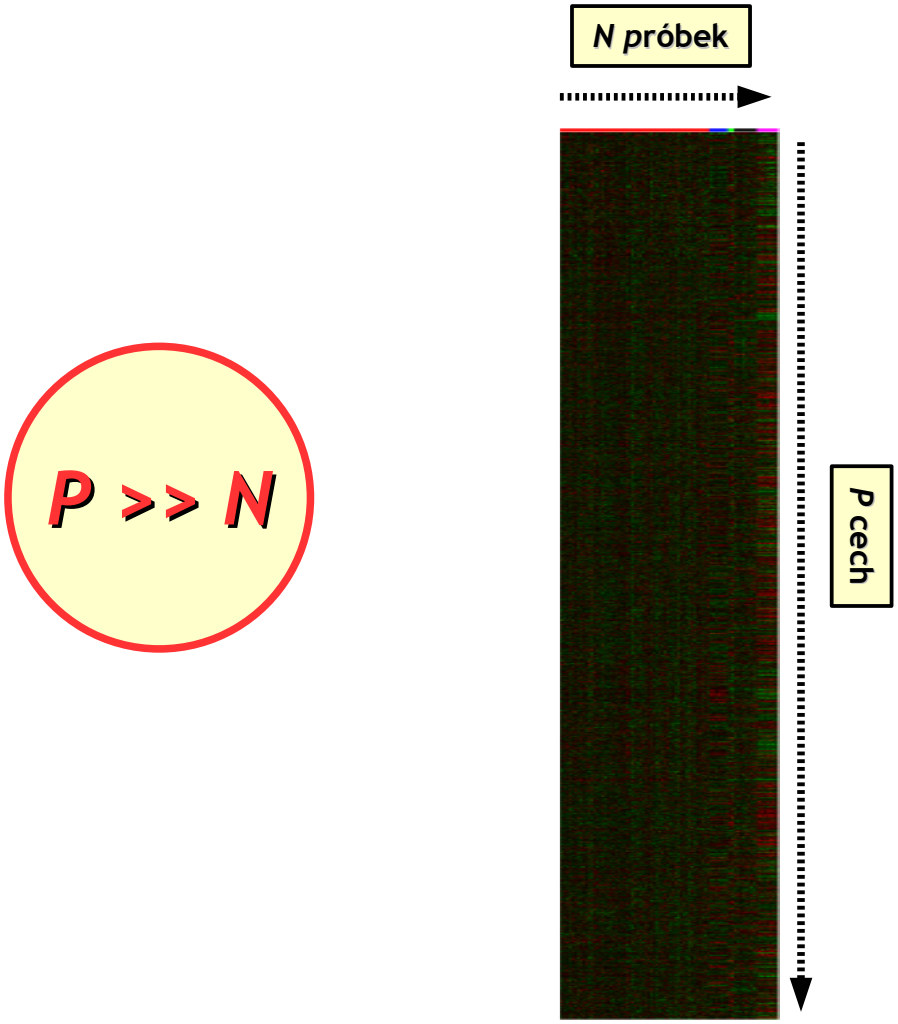
Wykład 5: przetwarzanie wstępne danych (eliminacje cech nieistotnych)

Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Przetwarzanie wstępne: eliminacja cech nieistotnych

Liczba cech opisujących badane obiekty zależy od rodzaju i pochodzenia danych. W skrajnych przypadkach, takich jak dane z wielkoskalowych technik pomiarowych, może ona dochodzić do setek tysięcy i znacząco przekraczać liczbę próbek. Jednak nie każda cecha musi być równie istotna z punktu widzenia dalszych kroków analizy.



Przetwarzanie wstępne: eliminacja cech nieistotnych

W macierzy danych mogą występować wiersze reprezentujące cechy nieistotne dla celu eksperymentu. Przykładami mogą być:

- cechy kontrolne, których wartości służą jedynie do oceny jakości pomiarów;
- **cechy o „płaskich" wzorcach**, czyli mające w przybliżeniu stałe wartości we wszystkich próbkach. Będą one charakteryzować się małym **współczynnikiem zmienności (*Coefficient of Variation*, CV)**, wyznaczanym jako stosunek odchylenia standardowego do wartości średniej. Dla i-tego wiersza szacowany jest jako:

$$CV_i = \frac{\hat{\sigma}_{x_i}}{\hat{\mu}_{x_i}}$$

UWAGA: usunięcie takich cech jest zwykle korzystne z punktu widzenia dalszych kroków analizy (m. in. ogranicza związany z nimi koszt obliczeniowy), ale wymaga ustalenia „sensownego" progu CV, co nie zawsze jest możliwe bez dodatkowej wiedzy a priori o badanych próbkach.

Uczenie maszynowe w bioinformatyce

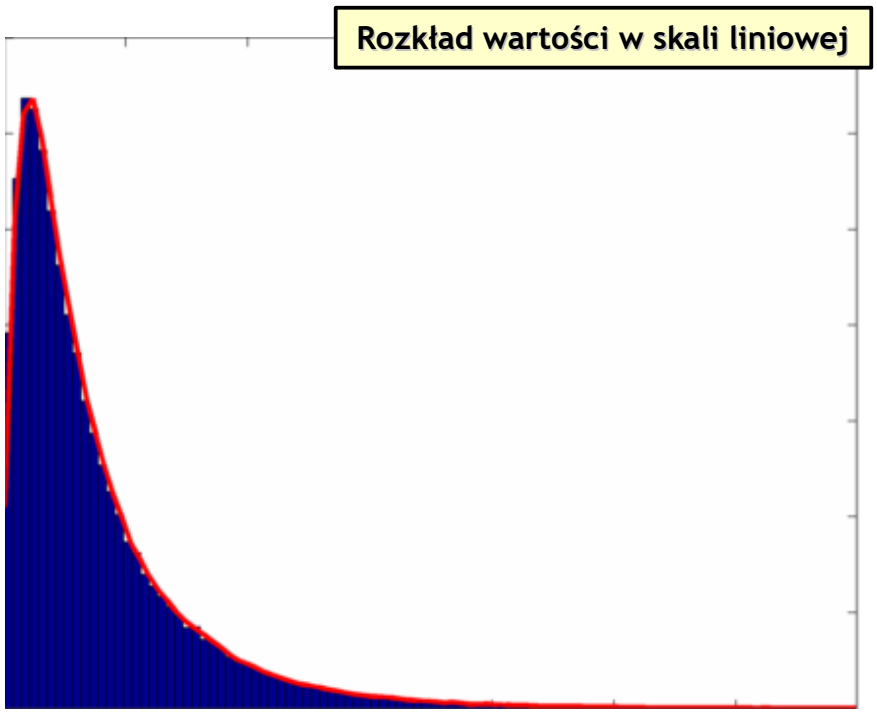
Wykład 5: przetwarzanie wstępne danych (zmiana skali danych)

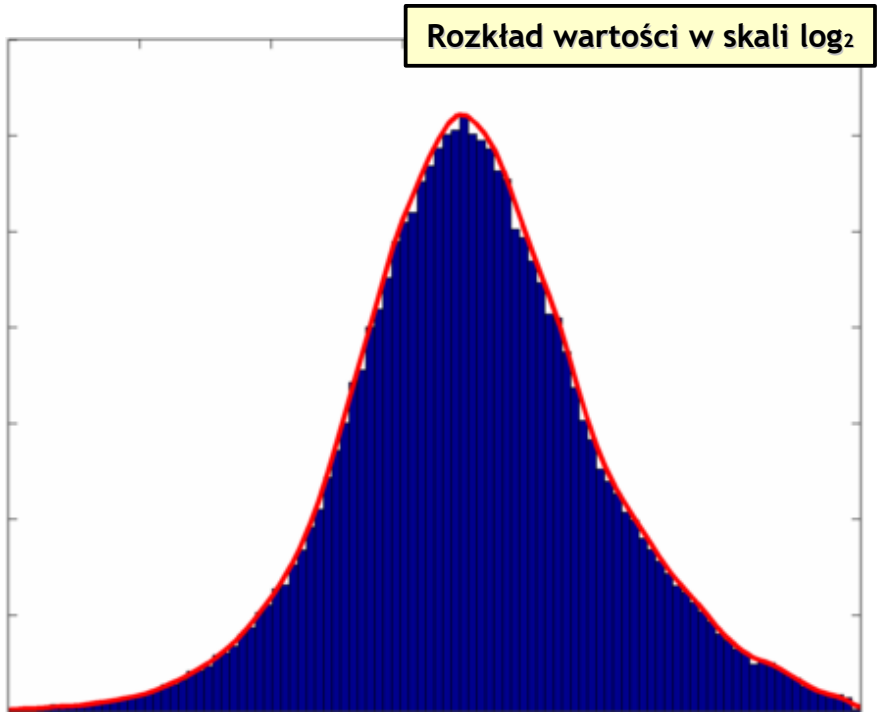
Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Przetwarzanie wstępne: zamiana skali na logarytmiczną

Dokładna postać rozkładu wartości cech w zbiorze danych zależy od użytej metody pomiarowej i przetwarzania niskopoziomowego. Jednak często rozkład ten wykazuje lewostronną asymetrię, czyli przesunięcie maksimum w kierunku małych wartości (odpowiadającym np. genom o niskim poziomie ekspresji). W takim przypadku dane można poddać logarytmowaniu, aby uzyskać bardziej symetryczny rozkład wartości i ograniczyć ich zakres dynamiczny.





Przetwarzanie wstępne: zamiana skali na logarytmiczną

Zalety użycia skali logarytmicznej:

- **zmiana rozkładu wartości na w przybliżeniu symetryczny,** o charakterze podobnym do rozkładu normalnego, co nadaje bardziej intuicyjny sens średniej i wariancji;

- **ograniczenie zakresu wartości**, co zmniejsza ryzyko „zdominowania" zbioru danych przez cechy o dużych wartościach, które niekoniecznie muszą być interesujące;

- **osłabienie zależności pomiędzy wariancjami a wartościami średnimi cech** i **zamiana efektów multiplikatywnych na addytywne** (obie właściwości są korzystne z punktu widzenia modeli statystycznych używanych do opisu danych).

Uczenie maszynowe w bioinformatyce

Wykład 5: przetwarzanie wstępne danych (normalizacja cech)

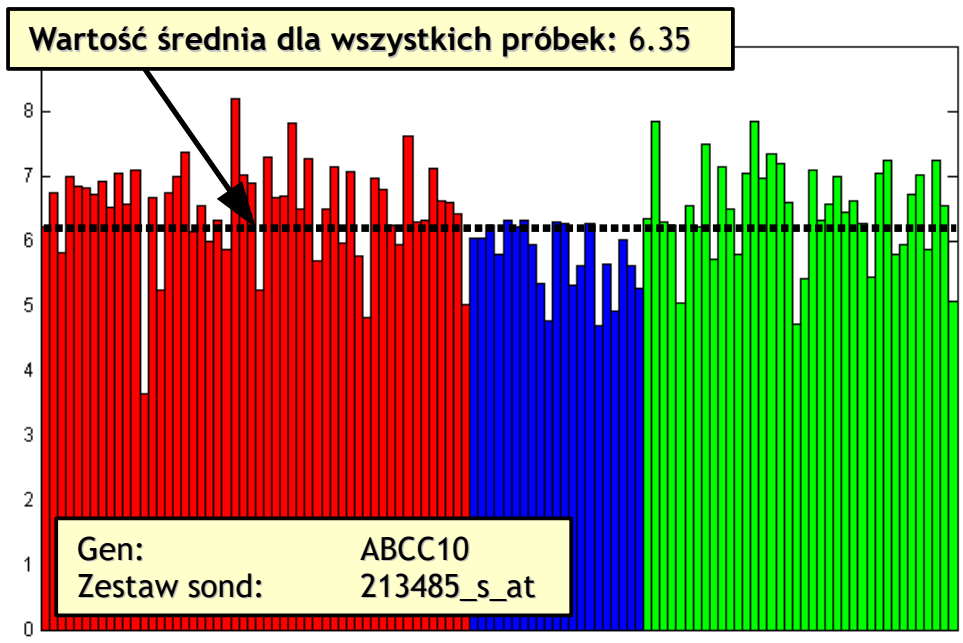
Tymon Rubel

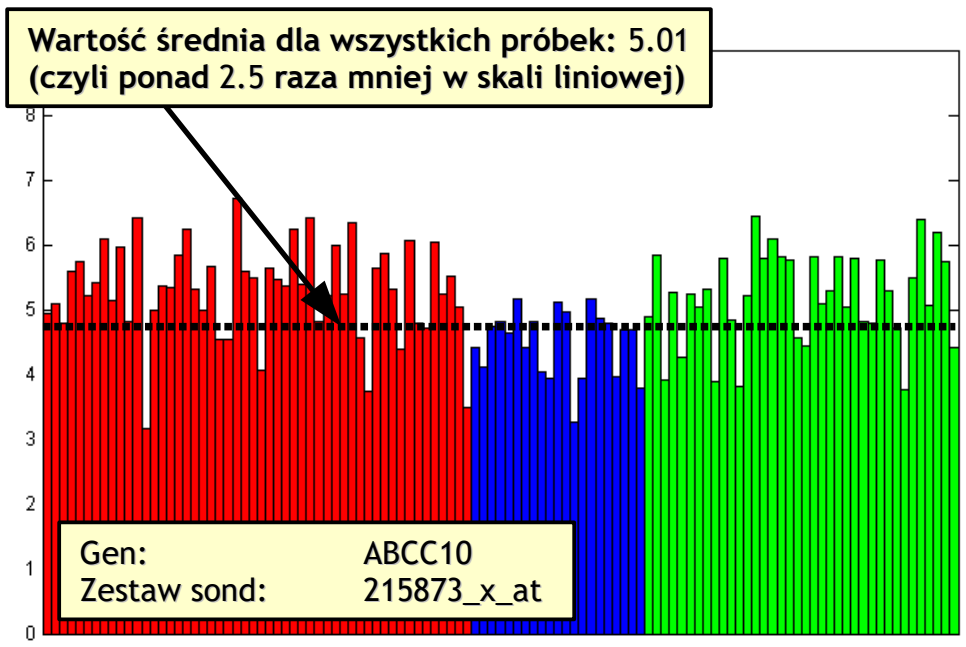
Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Przetwarzanie wstępne: normalizacja cech

Wartości cech mogą mieć charakter względny, co oznacza, że możliwe jest jedynie porównywanie wartości tej samej cechy w różnych próbkach (tzn. w ramach jednego wiersza macierzy danych), a nie różnych cech w obrębie jednej próbki.

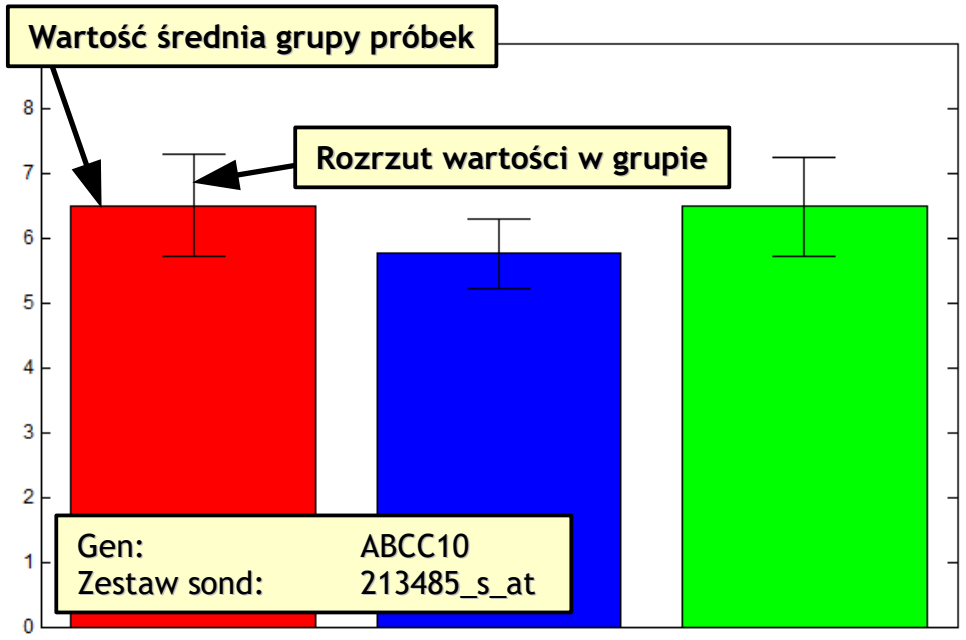
Przykładem mogą być wartości miary ekspresji zestawów sond reprezentujących ten sam gen na mikromacierzy DNA:

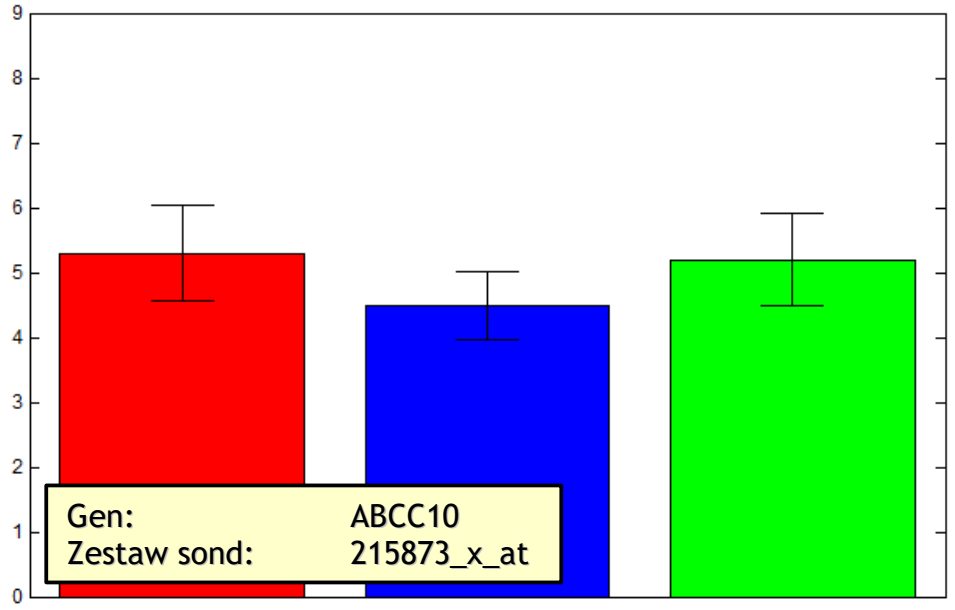




Przetwarzanie wstępne: normalizacja cech

W takim przypadku, pierwotna skala wartości danej cech (wyrażana np. średnią lub zakresem od minimum do maksimum) nie jest „ciekawym” parametrem, a podstawą wnioskowania są różnice pomiędzy średnimi w porównywanych grupach próbek, odniesione do rozrzutu wartości (wariancji) w obrębie tych grup.





Przetwarzanie wstępne: normalizacja cech

Normalizacja cech oznacza przeskalowanie ich wartości (w ramach wierszy macierzy X) w celu ujednolicenia zakresów zmienności i wyrównania wpływu poszczególnych cech na cały zbiór danych.

Często stosowaną techniką normalizacji jest **standaryzacja**, polegająca na odjęciu od każdej z cech wartości średniej i podzieleniu przez odchylenie standardowe (jedno i drugie estymowane z próby). Dla i-tej cechy:

$$x_i^{STD} = \frac{x_i - \hat{\mu}_{x_i}}{\hat{\sigma}_{x_i}}$$

Po wykonaniu standaryzacji każda cecha ma zerową wartość średnią i jednostkową wariancję (czyli wnosi tyle samo do całkowitej wariancji zbioru danych).

Inną bardzo popularną metodą jest **skalowanie min-max**:

$$x_i^{MINMAX} = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$$

W efekcie skalowania min-max każda z cech ma wartości z zakresu $[0; 1]$.

	MATLAB	R
KDE:	`normalize`	`scale/preProcess`

Przetwarzanie wstępne: normalizacja cech

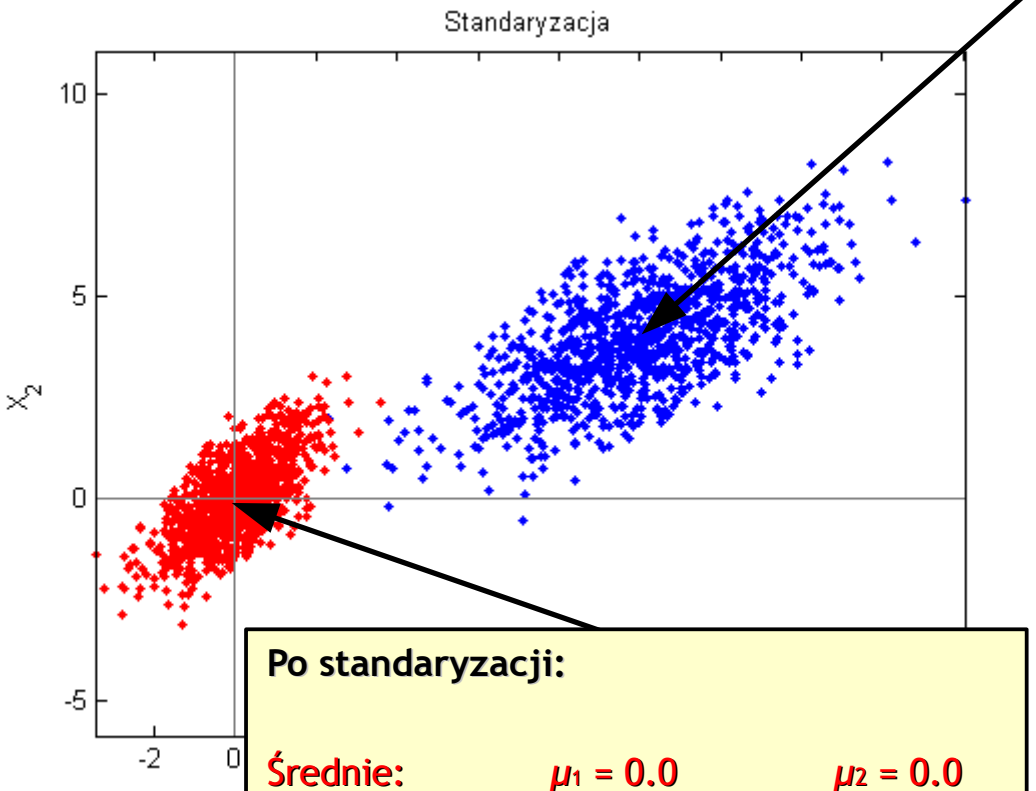
Zbiór danych przed normalizacją cech

Średnie: $\mu_1 = 10.0$ $\mu_2 = 4.0$

Odchylenia std.: $\sigma_1 = 5.0$ $\sigma_2 = 2.0$

Wsp. korelacji: $r_{1,2} = 0.7$

Zakres wartości: [2.4; 17.3] [-0.9; 8.3]



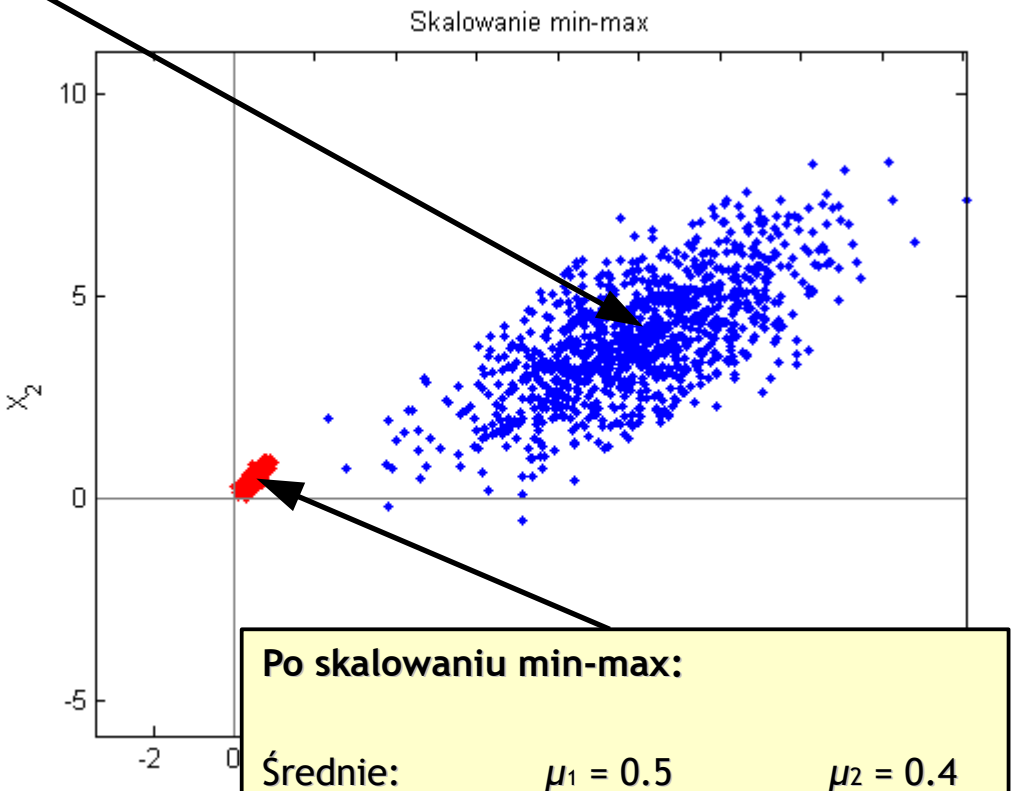
Po standaryzacji:

Średnie: $\mu_1 = 0.0$ $\mu_2 = 0.0$

Odchylenia std.: $\sigma_1 = 1.0$ $\sigma_2 = 1.0$

Wsp. korelacji: $r_{1,2} = 0.7$

Zakres wartości: [-2.7; 2.9] [-3.0; 3.5]



Po skalowaniu min-max:

Średnie: $\mu_1 = 0.5$ $\mu_2 = 0.4$

Odchylenia std.: $\sigma_1 = 0.2$ $\sigma_2 = 0.1$

Wsp. korelacji: $r_{1,2} = 0.7$

Zakres wartości: [0.0; 1.0] [0.0; 1.0]