KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

dr inż. Artur Klepaczko

Eksploracja danych
Redukcja wymiaru danych

Zadanie nr 13 – Studia podyplomowe „Przetwarzanie i analiza obrazów biomedycznych”

Prezentacja multimedialna
współfinansowana przez Unię Europejską
w ramach Europejskiego Funduszu Społecznego
w projekcie

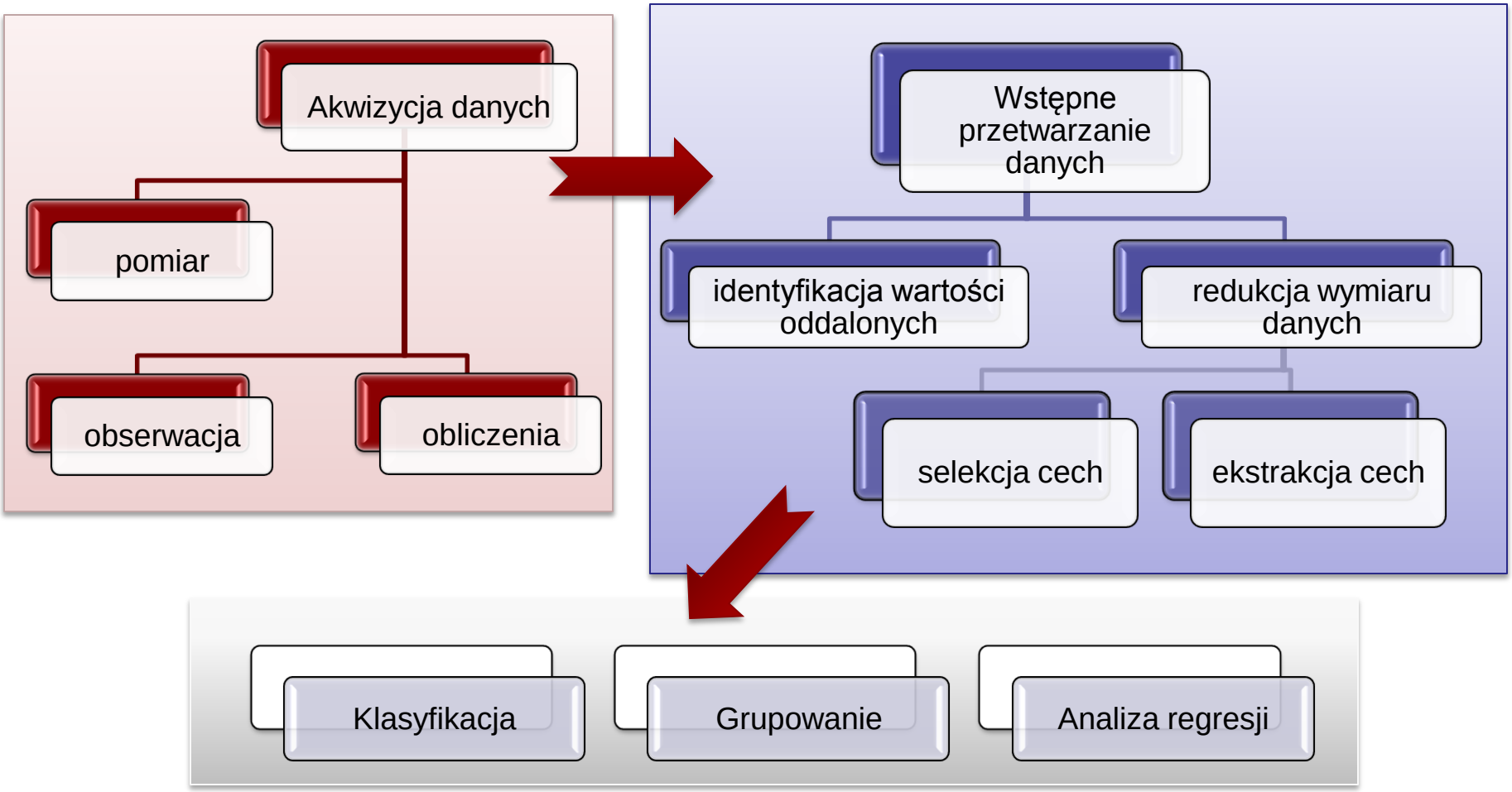
„Innowacyjna dydaktyka bez ograniczeń – zintegrowany rozwój Politechniki Łódzkiej – zarządzanie Uczelnią, nowoczesna oferta edukacyjna i wzmacniania zdolności do zatrudniania osób niepełnosprawnych”

Politechnika Łódzka
Instytut Elektroniki

90-924 Łódź, ul. Żeromskiego 116,
tel. 042 631 28 83
www.kapitalludzki.p.lodz.pl

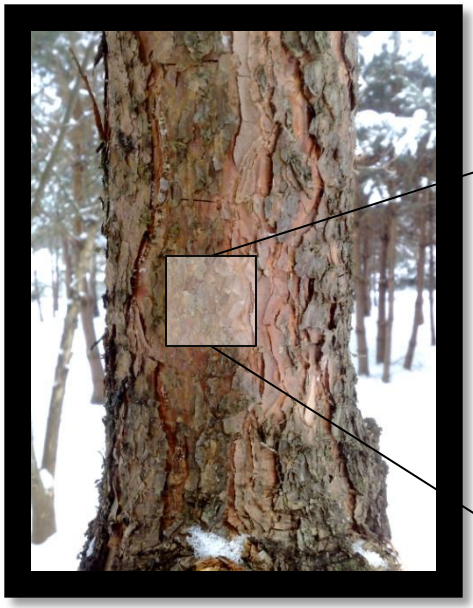
Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Redukcja wymiaru — wstępny etap eksploracji danych

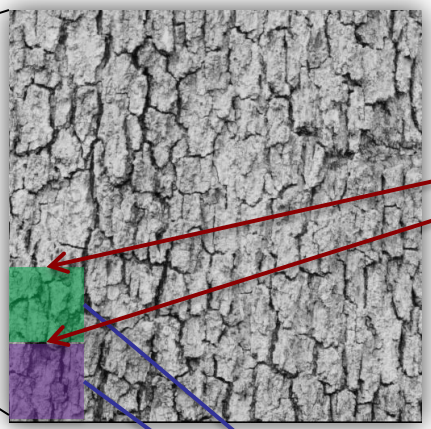


Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Obiekty i cechy



Obiekt



Obraz obiektu

Obszary zainteresowania

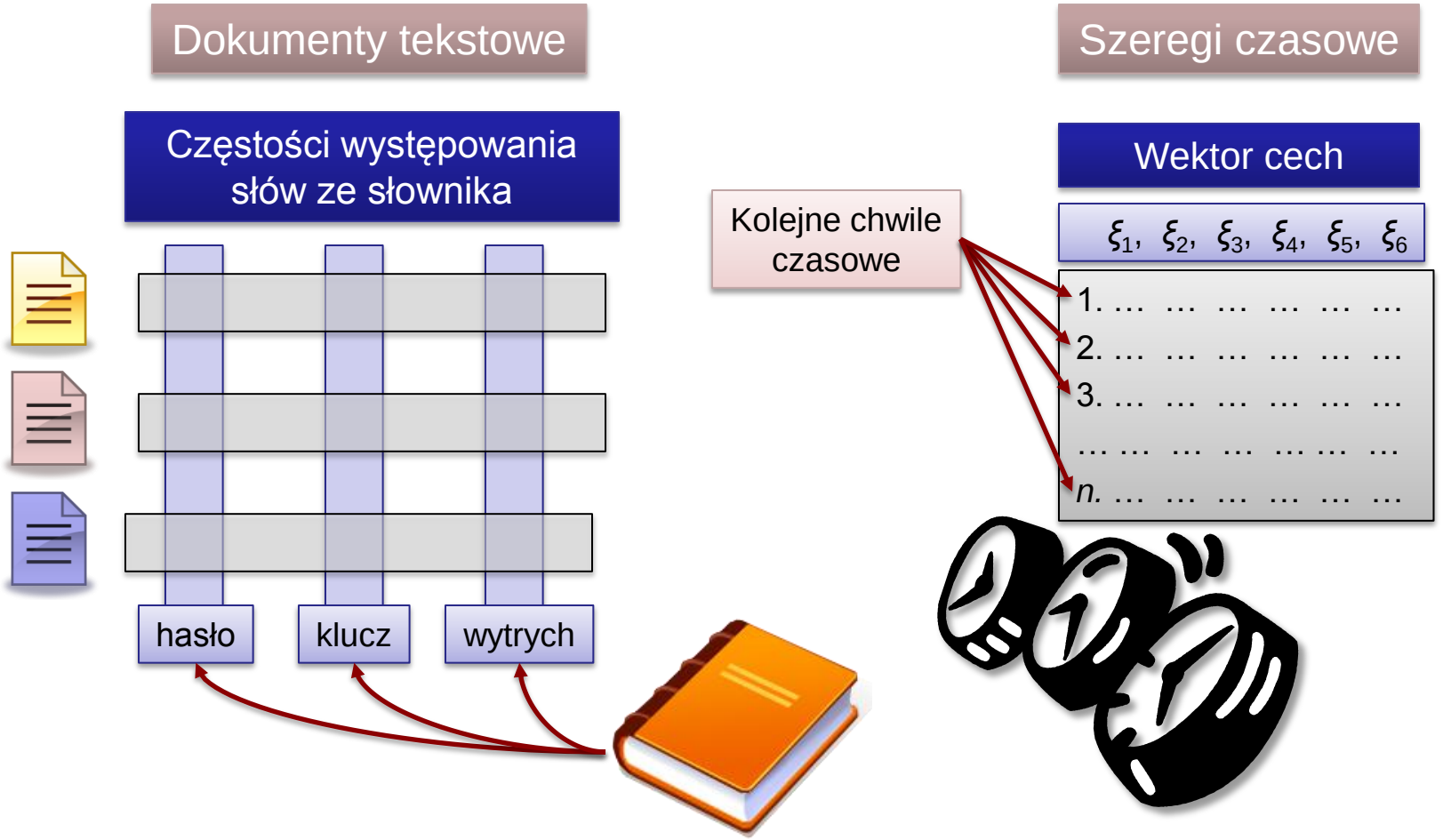
Cechy tekstury

Wektory cech

-0.339775, 0.591024, 0.998889, 1.009308, 42.27055,…
-0.473273, 0.640661, 0.996667, 1.395732, 65.171616,…
…

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

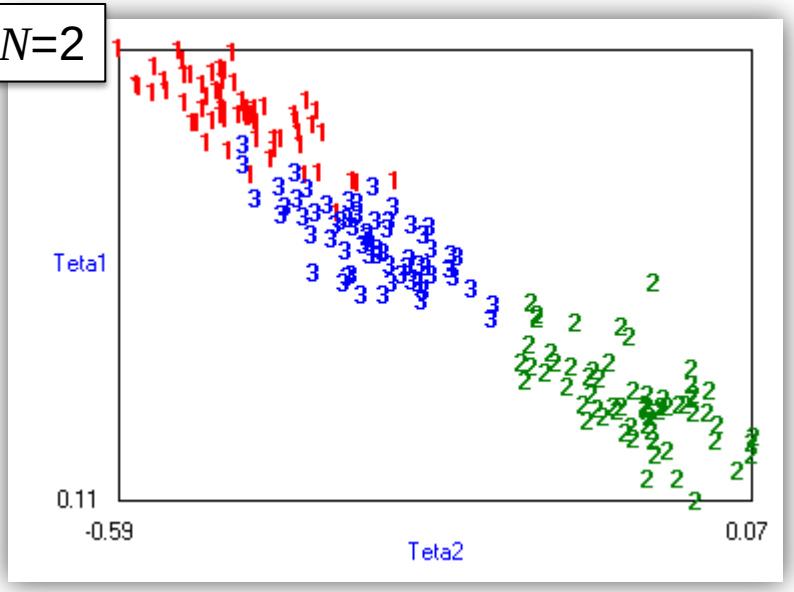
Inne przykłady obiektów i cech



Wymiar danych

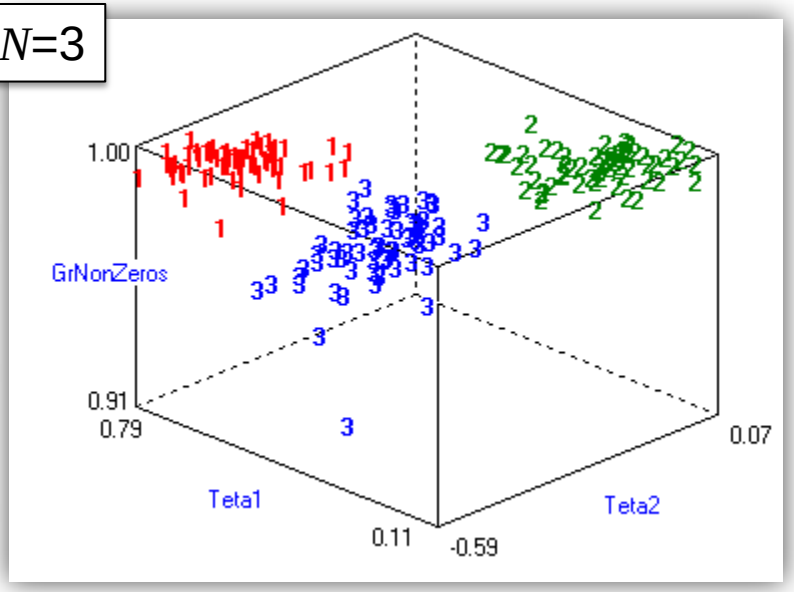
N — długość wektora cech, tzn. liczba parametrów opisujących obiekty

N=2



Przestrzeń dwuwymiarowa

N=3

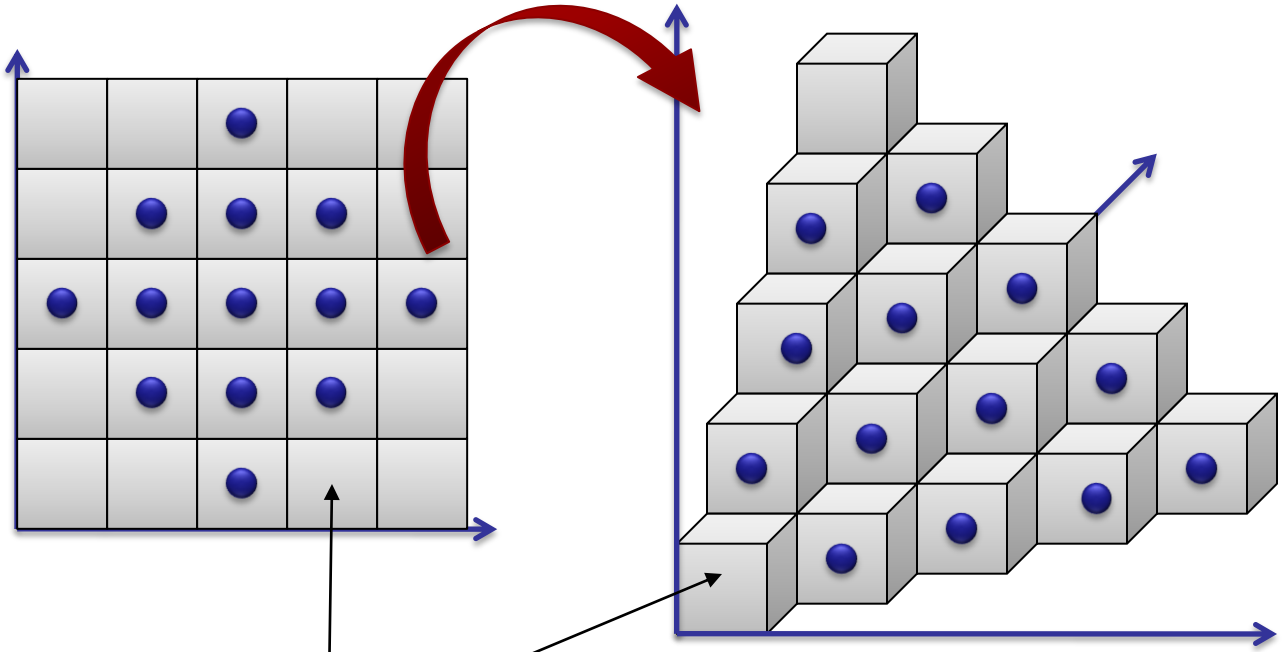


Przestrzeń trójwymiarowa

Ogólnie mówi się, że wektory danych umieszczone są w N–wymiarowej przestrzeni cech

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Przekleństwo wymiarowości (ang. *curse of dimensionality*)



Jednostkowy hipersześcian

Liczba wektorów na jednostkę objętości

N=2
g=13/25

N=3
g=13/125

W przestrzeni o większej wymiarowości potrzeba więcej wektorów danych, aby utrzymać ich stałe zagęszczenie g.

Co to oznacza?

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Duży wymiar danych — skutki

Każdy dodatkowy wymiar sprawia, że wektory danych stają się coraz bardziej od siebie oddalone.

Rośnie liczba wektorów danych niezbędna do prawidłowego oszacowania ich rozkładu prawdopodobieństwa.

Większa złożoność obliczeniowa (zarówno ze względu na liczbę wektorów jak i ich wymiar).

Wnioskowanie na podstawie zbyt małej próby losowej wektorów danych może być niewiarygodne.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Skąd się bierze duży wymiar?

Skoro duży wymiar oznacza…

1. Przekleństwo wymiarowości
2. Złożoność obliczeniową (może rosnąć wykładniczo wraz ze wzrostem N)

…to dlaczego dopuszczamy do zaistnienia problemu na etapie akwizycji danych?

1. Nie wiadomo z góry, które cechy obiektów będą odpowiednie do konkretnego zadania eksploracji danych (cechy mogą być znaczące lub nie).
2. Niektóre atrybuty nadają się do dyskryminacji, inne do reprezentacji danych.
3. Czasami, specyfika badanych obiektów (np. dokumenty tekstowe, strony internetowe) narzuca dużą liczbę cech.

Jaka na to rada?

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Metody redukcji wymiaru danych

Nr	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	Klasa
1	9	1.3	0.5	102	60	A
2	10	1.3	0.4	116	101	A
3	11	1.0	0.7	59	58	B
…	…	…	…	…	…	…
Q	10	1.1	0.8	58	91	B

Ekstrakcja cech

Transformacja oryginalnej przestrzeni cech do nowej przestrzeni złożonej z tzw. agregatów cech.

Nr	K_1	K_2	Klasa
1	10.5	10	A
2	10.4	11	A
3	6.7	1	B
…	…	…	…
Q	6.8	2	B

Selekcja cech

Wybór podzbioru cech znaczących spośród całego zbioru atrybutów.

Nr	ξ_3	ξ_4	Klasa
1	0.5	102	A
2	0.4	116	A
3	0.7	59	B
…	…	…	…
Q	0.8	58	B

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Cechy znaczące

Istotność cech zależy od celu zadania eksploracji danych.

Zadanie eksploracji danych	Interpretacja istotności cech	Funkcja cech znaczących
Klasyfikacja, grupowanie	Cechy dające najmniejszy błąd klasyfikacji	Cechy dyskryminacyjne
Szacowanie funkcji rozkładu prawdopodobieństwa	Cechy dające największy stopień wiarygodności modelu	Cechy reprezentatywne
Analiza regresji	Cechy dające najlepsze dopasowanie danych do aproksymowanej funkcji	Cechy reprezentatywne

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Selekcja cech w trybie uczenia nadzorowanego

Zaetykietowany zbiór danych treningowych

Nr	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	Klasa
1	9	1.3	0.5	102	60	A
2	10	1.3	0.4	116	101	A
3	11	1.0	0.7	59	58	B
…	…	…	…	…	…	…
Q	10	1.1	0.8	58	91	B

Zbiór treningowy pozwala powiązać wartości cech z etykietami kategorii.

Uczenie nadzorowane — zbiór wektorów danych zawiera jednoznaczną informację o tym, które cechy najlepiej nadają się do dyskryminacji obiektywnie różnych klas. Stosowane miary istotności cech:

1. Błąd klasyfikacji
2. Średnia odległość między klasami
3. Szerokość marginesu rozdzielającego klasy
4. …

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Selekcja cech w trybie uczenia nienadzorowanego

Niezaetykietowany zbiór danych treningowych

Nr	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5
1	9	1.3	0.5	102	60
2	10	1.3	0.4	116	101
3	11	1.0	0.7	59	58
…	…	…	…	…	…
Q	10	1.1	0.8	58	91

Brak bezpośredniej przesłanki pozwalającej powiązać wartości cech z prawdziwym podziałem wektorów danych na klastery.

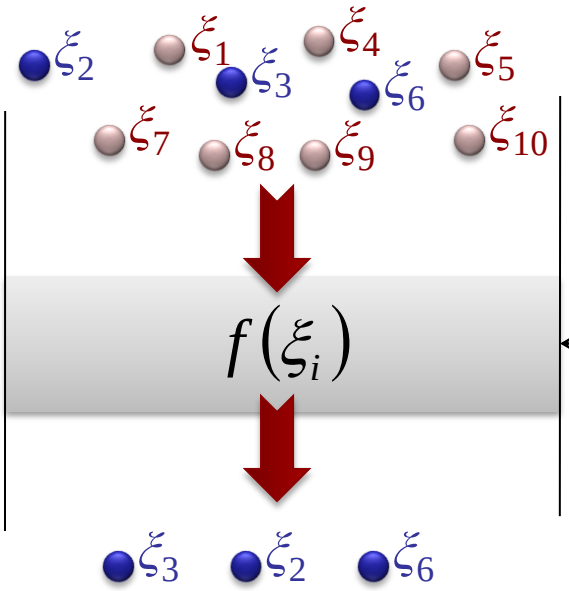
Uczenie nienadzorowane — selekcja cech odbywa się na podstawie:

1. subiektywnych miar oceny wyników grupowania (np. jakość klasteryzacji),
2. miar wymagających przyjęcia wstępnych założeń dotyczących modelu danych (np. kryterium MDL), albo
3. miar, które niekoniecznie prowadzą do uzyskania cech najbardziej dyskryminacyjnych (np. entropia zbioru danych).

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Selekcja cech — metody typu *filtr*

Metody typu *filtr* — niezależne od schematu



1. Każda cecha badana jest pojedynczo.
2. Mała złożoność obliczeniowa.
3. Cechy pozostają znaczące niezależnie od stosowanego algorytmu klasyfikacji.
4. Ranking cech — pozwala wyróżnić cechy najlepsze, dobre i zupełnie nieistotne.

$$f(\xi_3) > f(\xi_2) > f(\xi_6)$$

Funkcja kryterialna

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

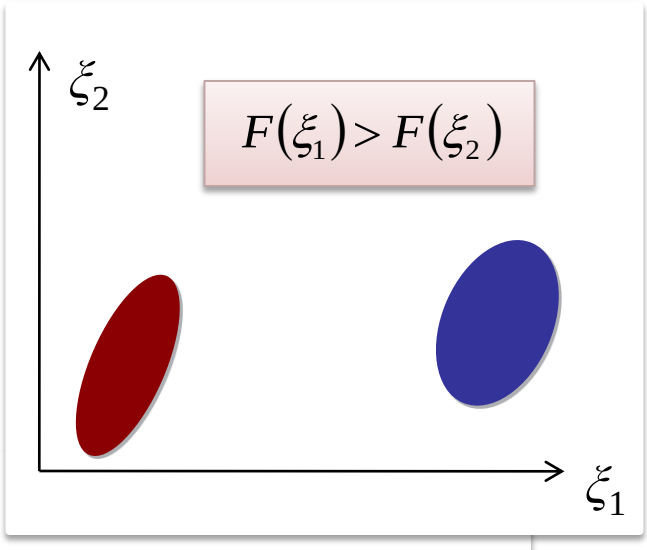
Współczynnik Fishera

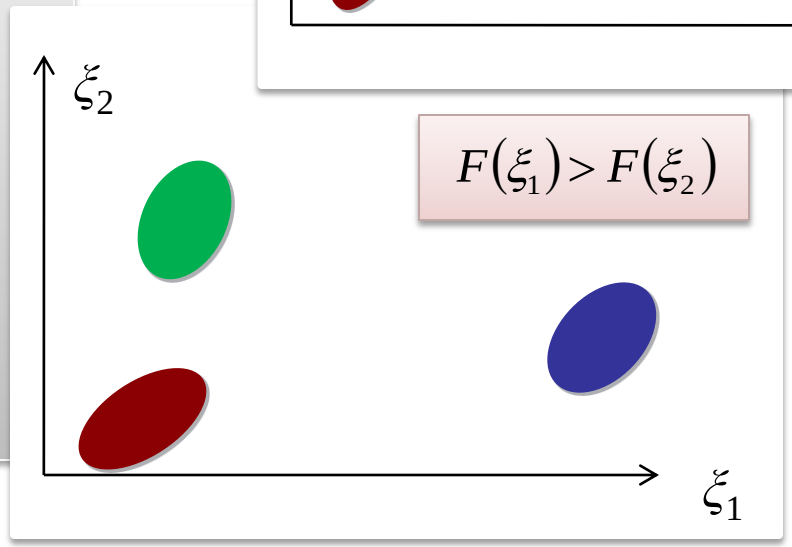
$$F = \frac{D}{V}$$

Rozproszenie międzyklasowe

Wariancja wewnątrzklasowa

Dokładna definicja znajduje się w podręczniku użytkownika programu MaZda.

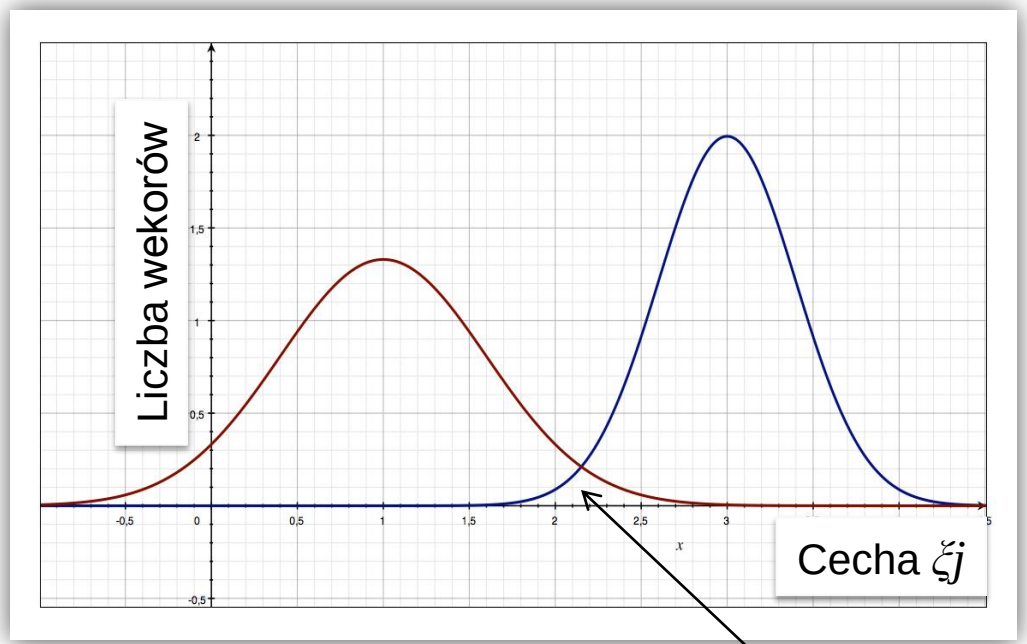




Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Prawdopodobieństwo błędu

Współczynnik *POE* (ang. probability of error)

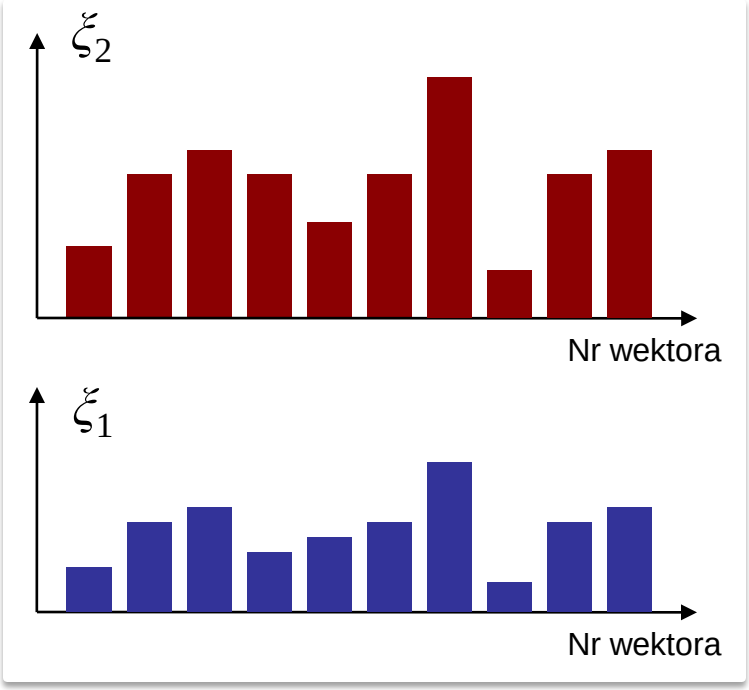


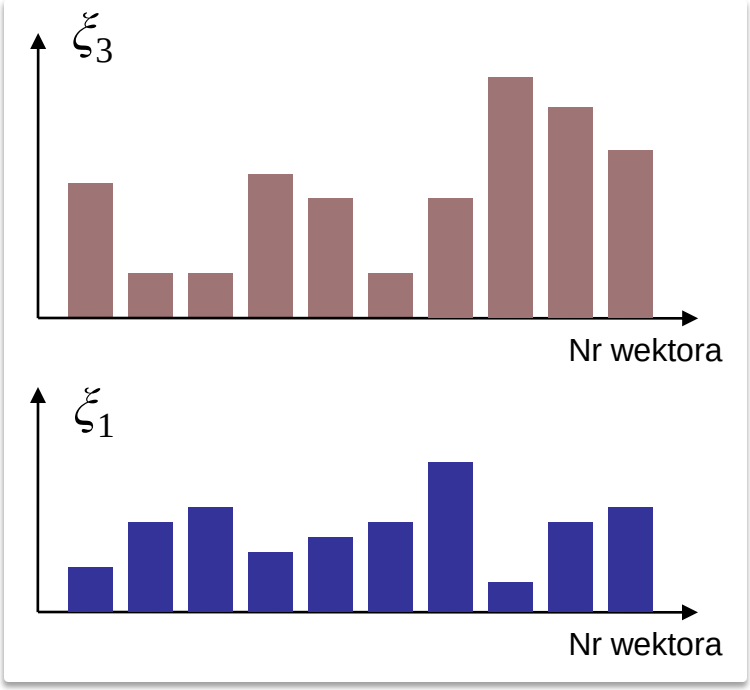
$$POE(\xi_j) = \frac{Q_e^{\xi_j}}{Q}$$

Q_e – liczba błędnie zaklasyfikowanych wektorów

Q – całkowita liczba wektorów danych

Współczynnik korelacji





$$CC(\xi_1, \xi_2) > CC(\xi_1, \xi_3)$$

Zmienność wartości cech ξ_1 i ξ_2 jest zbliżona.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Współczynnik POE + ACC

POE

Znajdujemy cechę ξ^1 zapewniającą minimalną wartość współczynnika *POE*

POE + CC

Ze zbioru cech wykluczamy ξ^1 i następnie znajdujemy cechę ξ^2, która jednocześnie zapewnia minimum prawdopodobieństwa błędu i współczynnika korelacji $CC(\xi^2, \xi^1)$.

POE + ACC

Każda kolejna cecha minimalizuje sumę współczynników *POE* (obliczanego bez uwzględniania już wybranych atrybutów) oraz współczynnika *ACC* (czyli uśrednionej korelacji między daną cechą a już wybranym podzbiorem cech).

Współczynnik ACC (ang. *accumulated correlation coefficient*) — średnia wartość współczynnika korelacji jednej cechy z podzbiorem *d* innych cech.

$$ACC\left(\xi_j, \left\{\xi^i : i = 1..d\right\}\right) = \frac{1}{d}\sum_{i=1}^{d} CC\left(\xi_j, \xi^i\right)$$

Informacja wzajemna

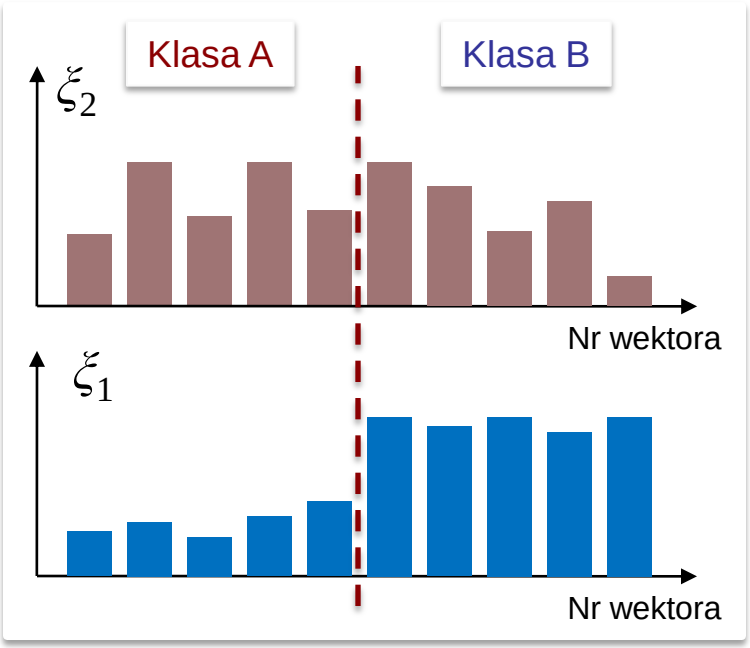
Entropia — ilość informacji w zmiennej losowej (stopień nieuporządkowania próby losowej)

$$MI(\lambda,\xi_j)=H(\lambda)+H(\xi_j)-H(\lambda,\xi_j)$$

Kategorie wektorów

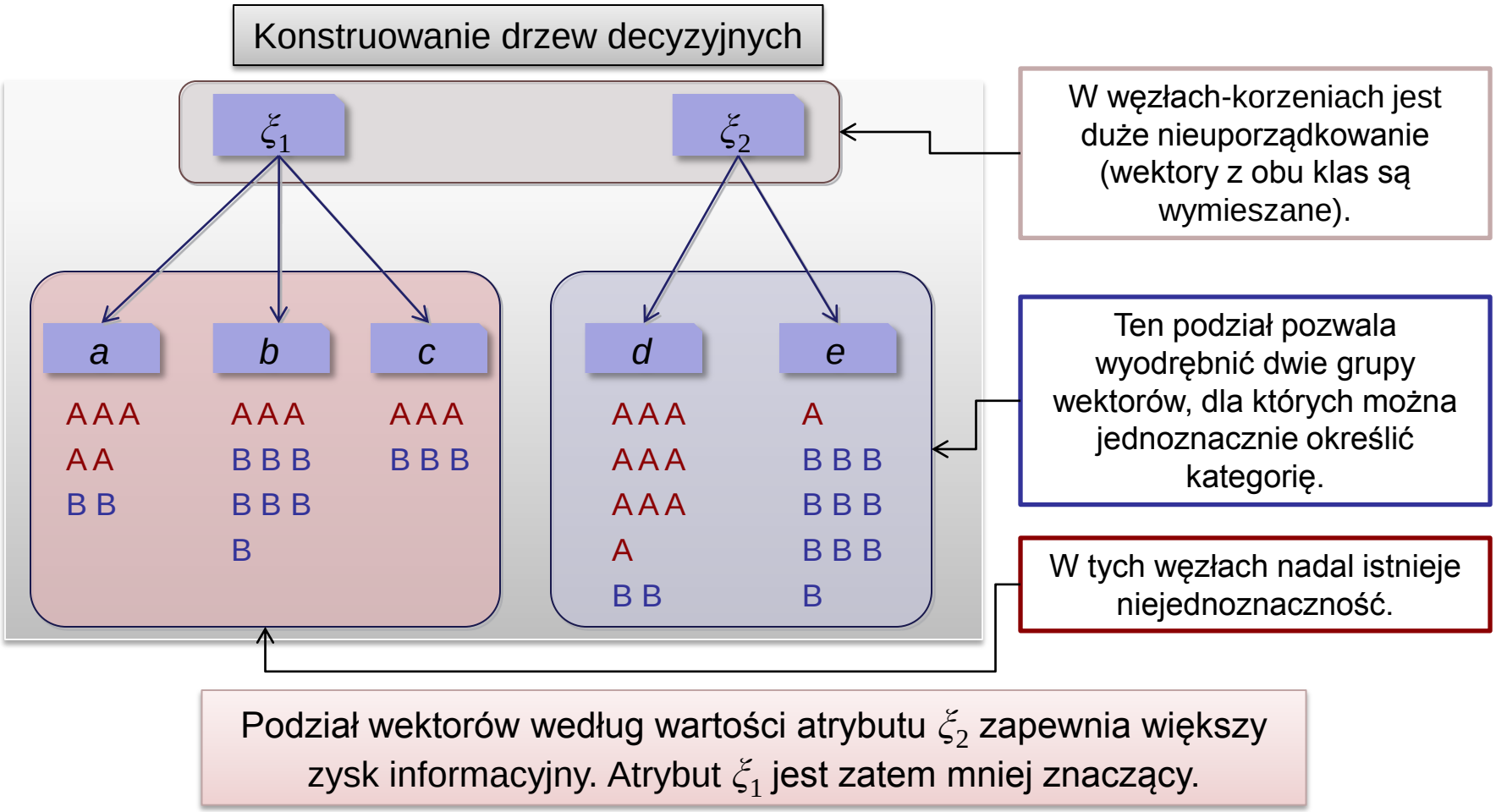
Wartości cechy ξ_j

Jeśli ξ_j w znikomym stopniu zależy od λ to ich informacja wzajemna jest mała, a cecha ξ_j jest nieznacząca.



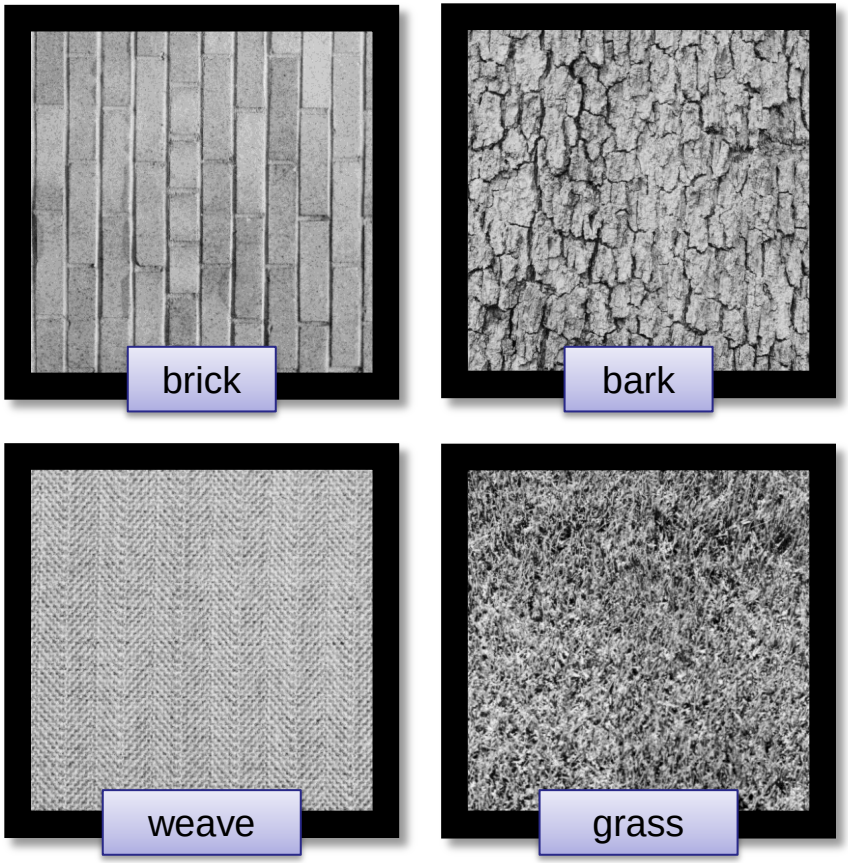
$$MI(\lambda,\xi_1)>MI(\lambda,\xi_2)$$

Zysk informacyjny



Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Metody typu *filtr* w zastosowaniu do zbioru tekstur



*Brodatz P., *A photographic album for artists and designers*, Dover, New York 1966.

Współczynnik Fishera

Cecha	F
Teta1	28.93
GrVariance	25.51
GrMean	19.85
Teta2	14.38
GrNonZeros	13.70

Błąd klasyfikatora 1-NN: 13.67%

Współczynnik informacji wzajemnej

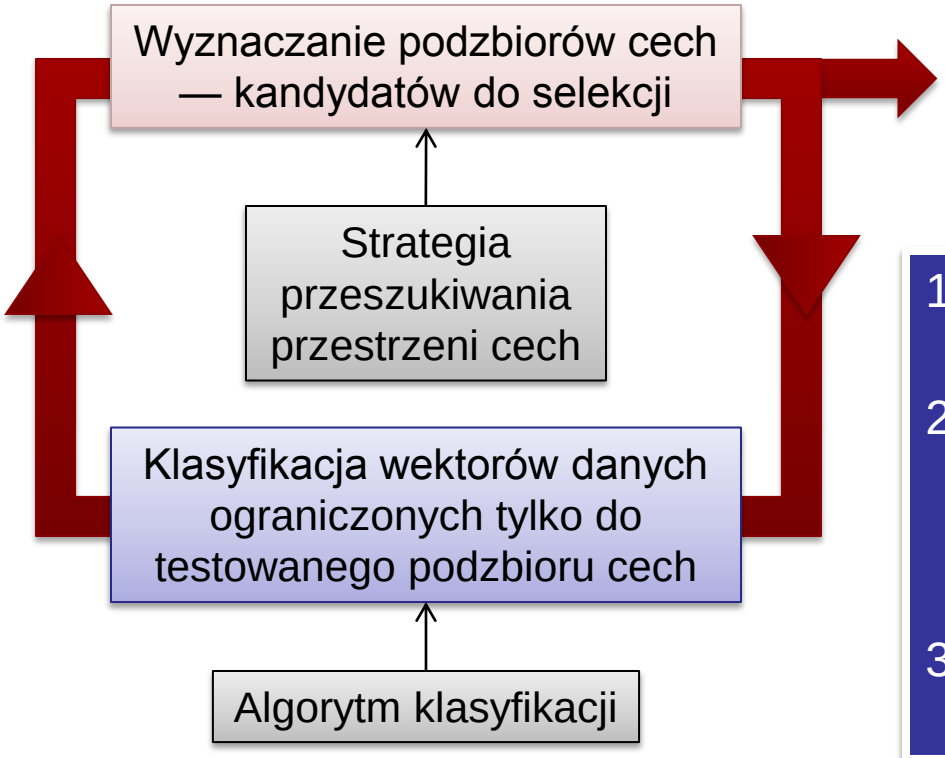
Cecha	MI
S(0,5)Contrast	1.30
S(0,4)Contrast	1.26
S(0,5)Correlat	1.25
S(0,4)Correlat	1.23
S(0,5)DifEntrp	1.22

Błąd klasyfikatora 1-NN: 21.88%

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Selekcja cech — metody typu *powłoka*

Metody typu *powłoka* (ang. *wrapper*) — zależne od schematu



Wybór podzbioru zapewniającego największą jakość klasyfikacji.

1. Złożoność obliczeniowa może być bardzo duża.
2. Wybrane cechy mogą nie nadawać się do klasyfikacji wektorów przy użyciu innych algorytmów klasyfikacji.
3. Cechy jednak badane są w podzbiorach a nie pojedynczo.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Przeszukiwanie przestrzeni cech

Zadanie optymalizacji

$$\hat{\Xi} = \arg\max_{\Xi \in \Omega} J(\Xi)$$

Kryteria istotności cech:
- błąd klasyfikacji
- szerokość marginesu rozdzielającego klasy
- entropia zbioru danych
- wiarygodność modelu

Ω — zbiór wszystkich cech

Ξ — podzbiór Ω

Przeszukiwanie przestrzeni cech → wyznaczanie kolejnych podzbiorów Ξ

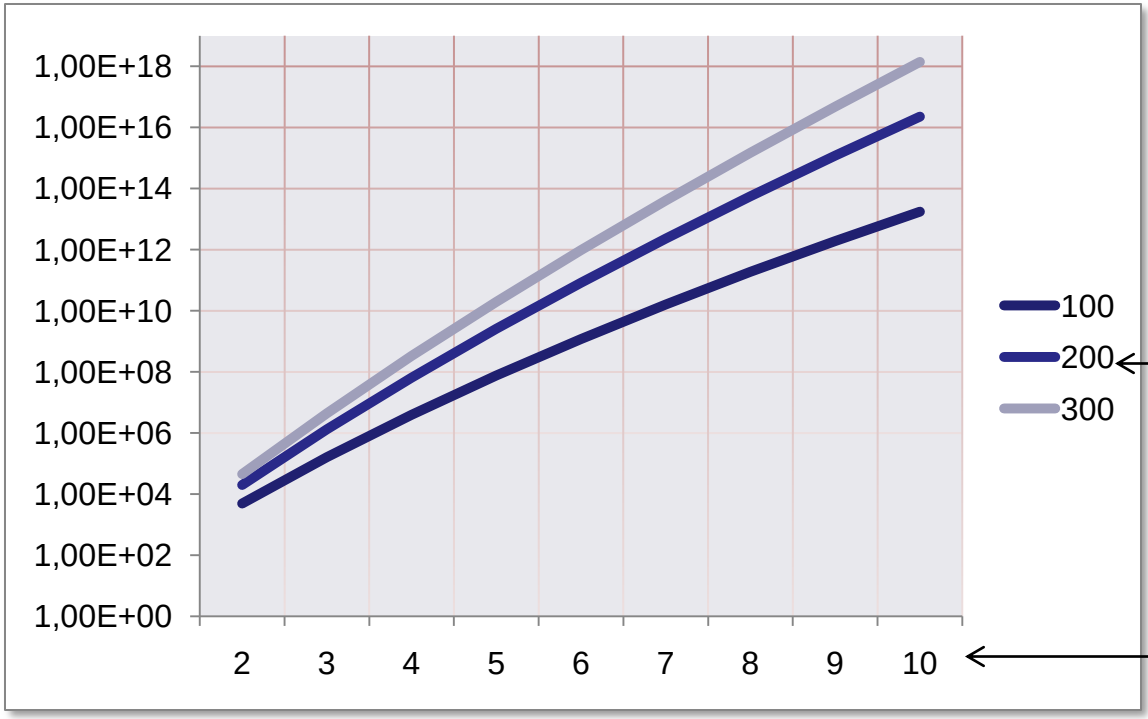
Strategie przeszukiwania

1. Kompletne — pełnego przeglądu lub metodą podziałów i ograniczeń
2. Sekwencyjne — wprzód lub wstecz
3. Losowe — algorytmy genetyczne

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Przeszukiwanie metodą pełnego przeglądu

Liczba wszystkich możliwych podzbiorów cech o określonej długości



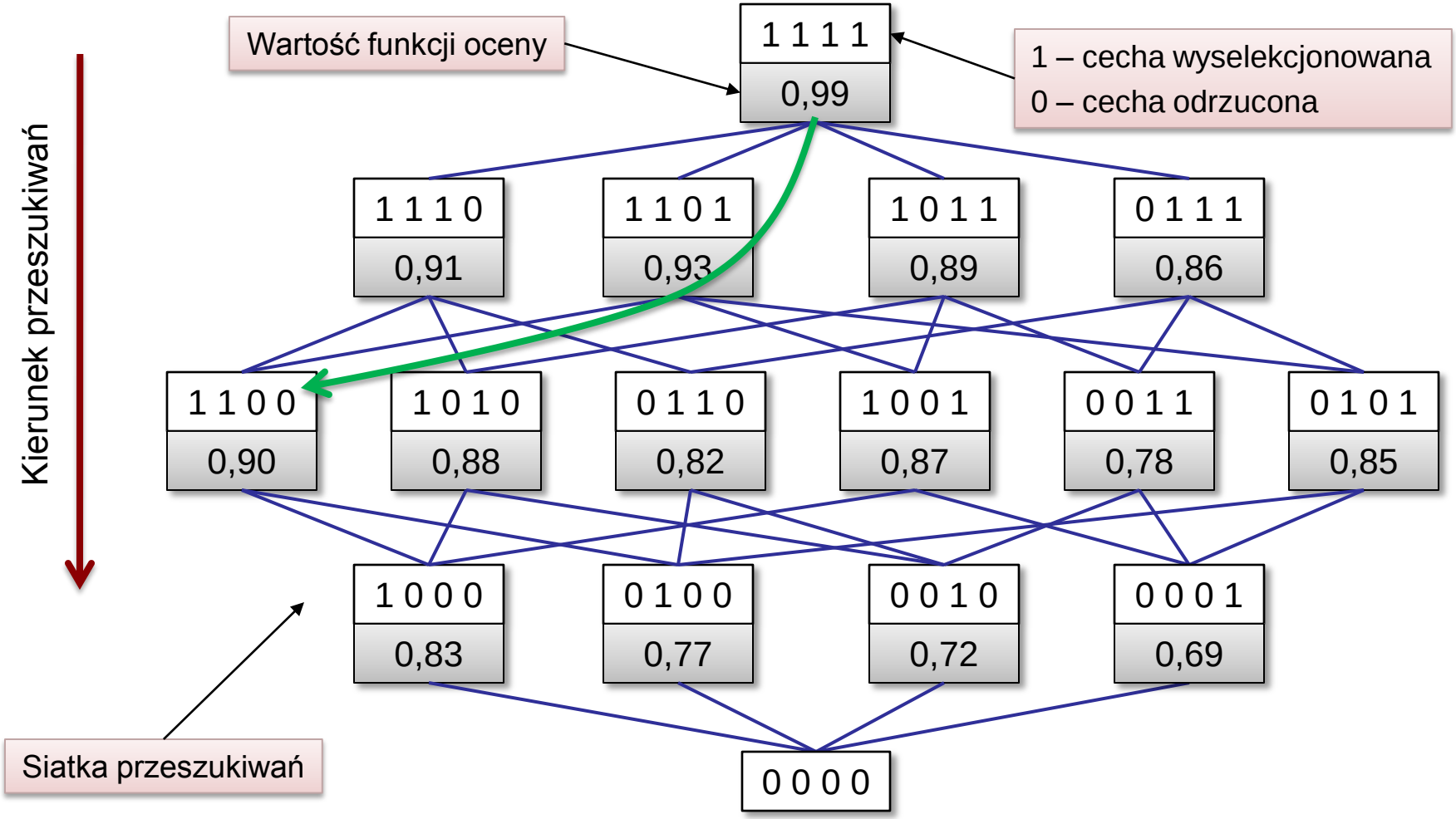
Liczba wszystkich cech

Liczba cech w podzbiorach

Metoda pełnego przeglądu nadaje się tylko do przestrzeni cech o stosunkowo małej wymiarowości.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

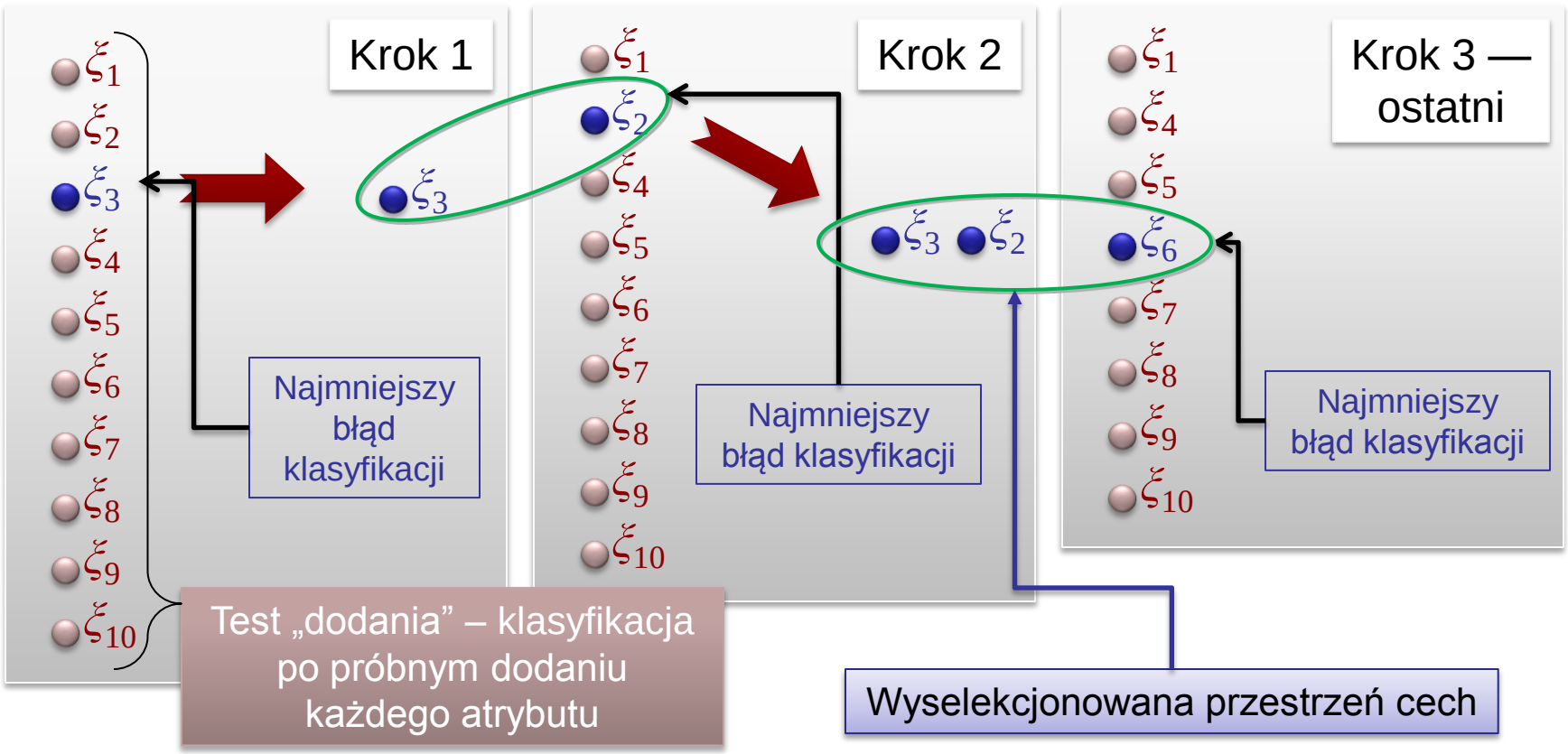
Przeszukiwania metodą podziałów i ograniczeń



Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

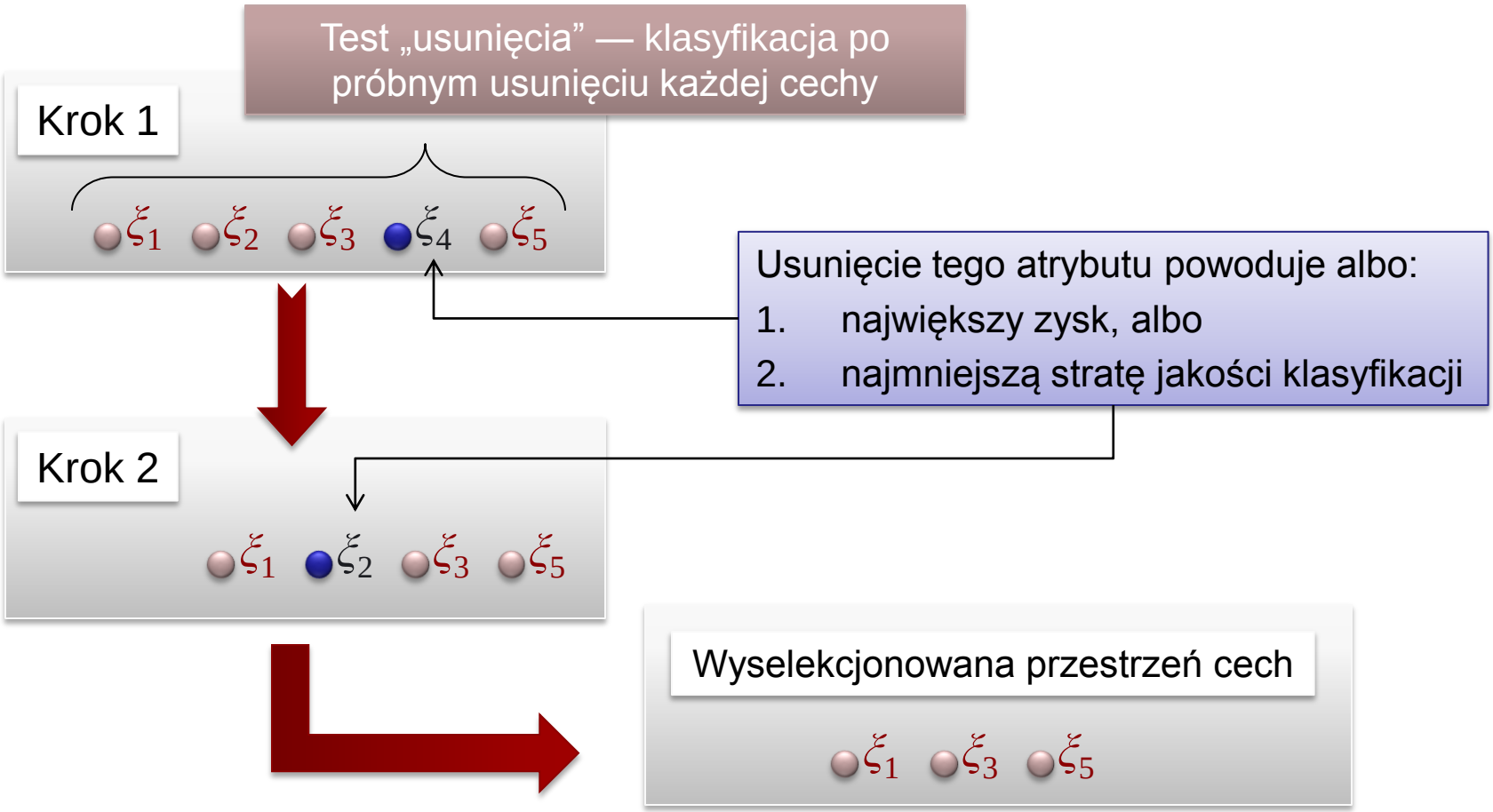
Przeszukiwanie wprzód (*sequential forward selection*)

Krok 0: Założenie wymiaru przestrzeń docelowej (np. 3 cechy)



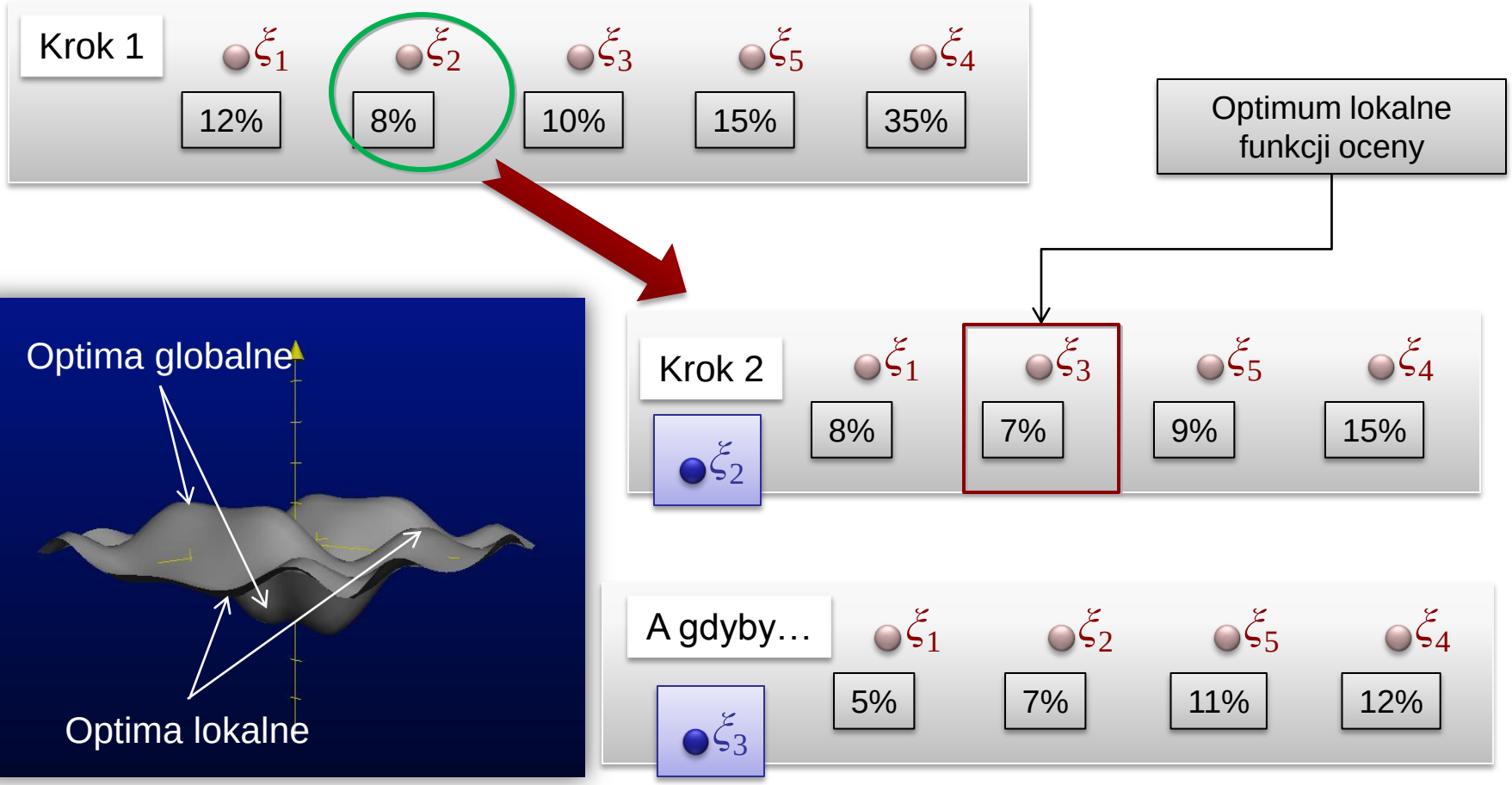
Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Przeszukiwanie wstecz (*sequential backward selection*)



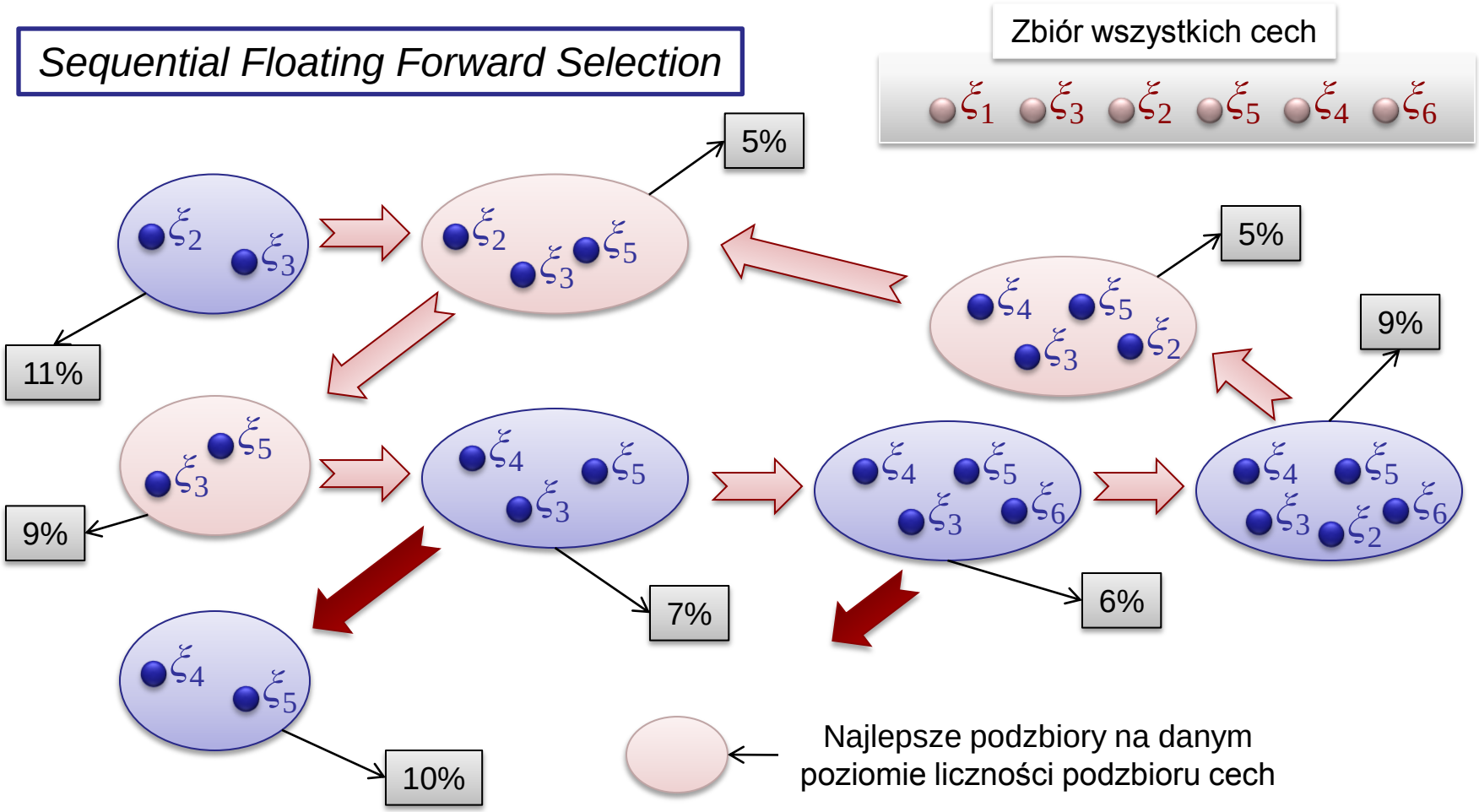
Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Proste przeszukiwanie zawodzi



Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Przeszukiwanie sekwencyjne z korektą



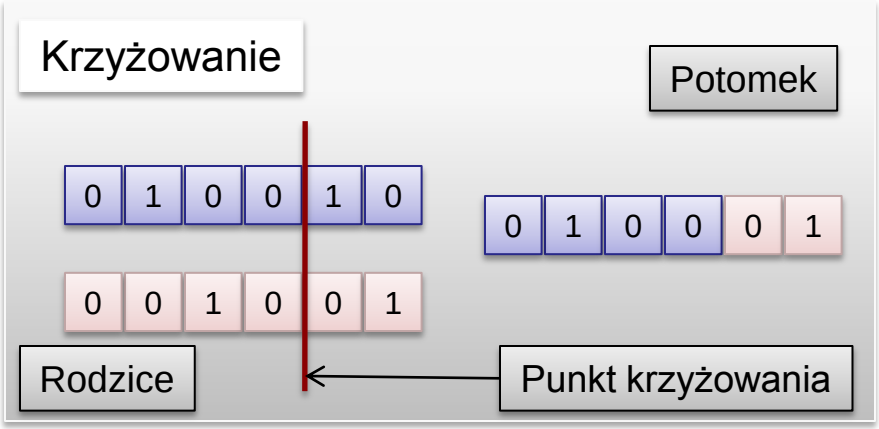
Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Algorytmy genetyczne — podstawowe definicje

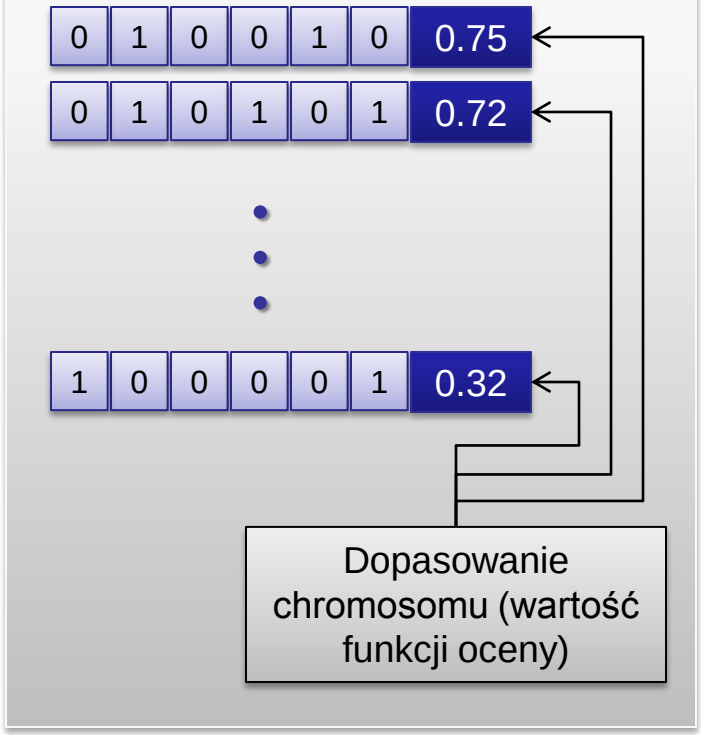
Chromosom

$g_i=1 \Leftrightarrow$ *i*-ta cecha jest znacząca
$g_i=0 \Leftrightarrow$ *i*-ta cecha jest nieznacząca

Krzyżowanie

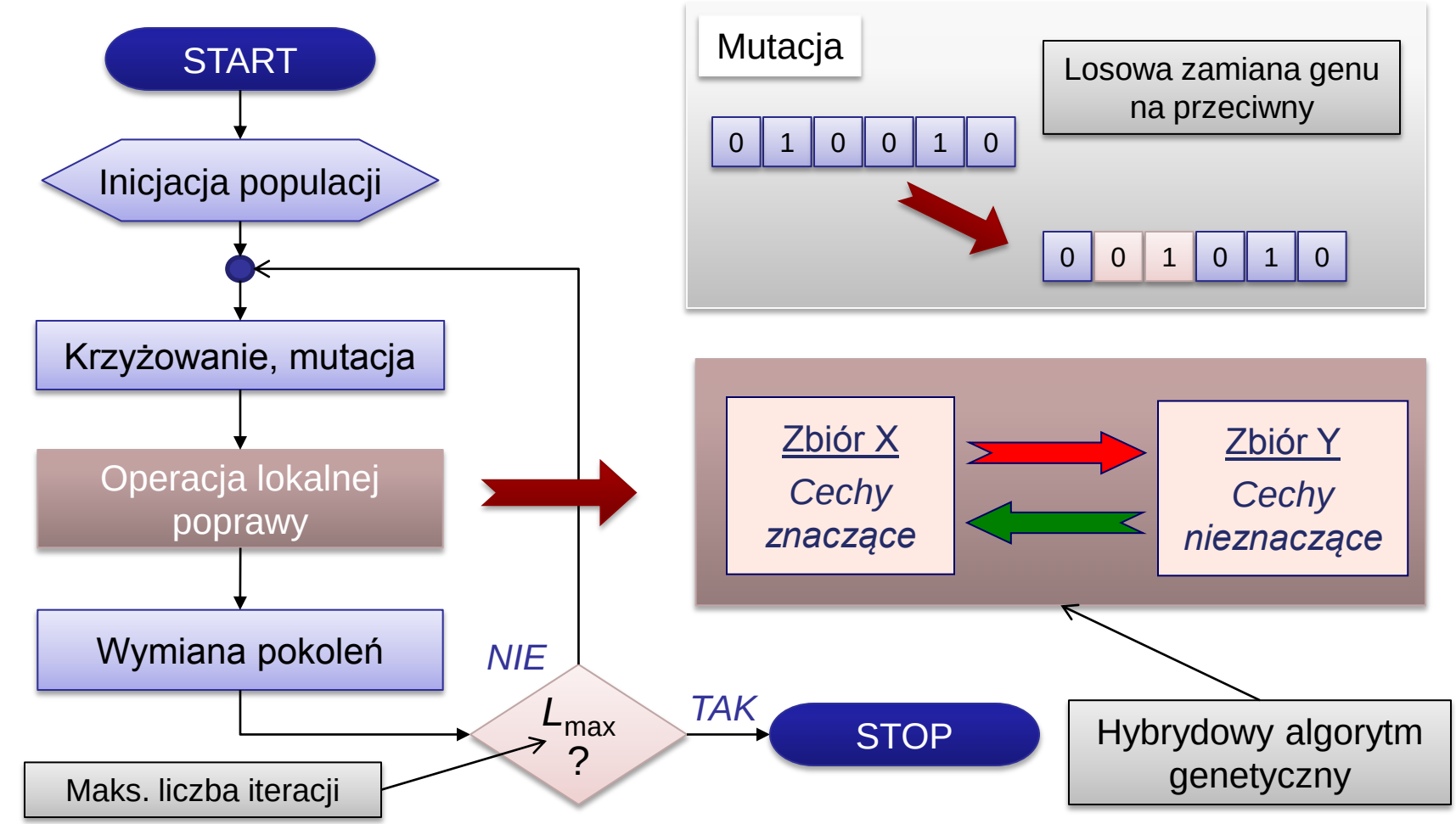


Populacja



Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Schemat algorytmu genetycznego



Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Porównanie metod przeszukiwania

Dokładność klasyfikacji za pomocą algorytmu 1-NN [%]

Zbiór danych	d_{red}	Optimum	SFS	SFFS	GA	HGA
Segmentation (*D*=19)	4	92,81	**92,81**	**92,81**	**92,81**	**92,81**
	8	92,95	**92,95**	**92,95**	**92,95**	**92,95**
	11	92,95	**92,95**	**92,95**	**92,95**	**92,95**
	15	92,57	**92,57**	**92,57**	**92,57**	**92,57**
Satellite (*D*=36)	7	–	86,85	**88,55**	87,89	88,44
	14	–	89,45	90,10	90,61	**90,87**
	22	–	90,45	**91,45**	91,36	91,44
	29	–	90,40	90,95	91,10	**91,24**
Sonar (*D*=60)	12	–	87,02	92,31	92,40	**94,71**
	24	–	89,90	93,75	95,49	**95,96**
	36	–	88,46	93,27	95,09	**95,82**
	48	–	91,82	91,35	92,02	**93,17**

Oh I.S., Lee J.S., Moon B.R., *Hybrid Genetic Algorithms for Feature Selection*, IEEE Trans. PAMI 26(11), 1424–1437, 2004