

---

# Analiza Skupień - Grupowanie



JERZY STEFANOWSKI  
Inst. Informatyki PP  
Wersja 2016 /akt 2020

# Organizacja wykładu

- Wprowadzenie i możliwe zastosowania
- Dobór parametrów algorytmów:
  - Hierarchiczne (AHC)
  - k - średnich
- Studium przypadku użycia



# Elementy terminologiczne

---

Trochę uwag:

- Cluster Analysis → Analiza skupień, grupowanie.
- Numerical taxonomy → Metody taksonomiczne (ekonomia)
  - Uwaga: znaczenie taksonomii w biologii może mieć inny kontekst (podział systematyczny oparty o taksony)
- Cluster → Skupienie, skupisko, grupa / klasa / pojęcie
- **Nigdy nie mów:** klaster, klastering, klastrowanie!

...

# Referencje do literatury (przykładowe)

---

- Koronacki J. Statystyczne systemy uczące się, WNT 2005.
- Pocięcha J., Podolec B., Sokołowski A., Zając K. „Metody taksonomiczne w badaniach społeczno-ekonomicznych”. PWN, Warszawa 1988,
- Stąpor K. „Automatyczna klasyfikacja obiektów” Akademska Oficyna Wydawnicza EXIT, Warszawa 2005.
- Hand, Mannila, Smyth, „Eksploracja danych”, WNT 2005.
- Larose D: „Odkrywanie wiedzy z danych”, PWN 2006.
- Kucharczyk J. „Algorytmy analizy skupień w języku ALGOL 60” PWN Warszawa, 1982,
- Materiały szkoleniowe firmy Statsoft.

# Przykłady zastosowań analizy skupień

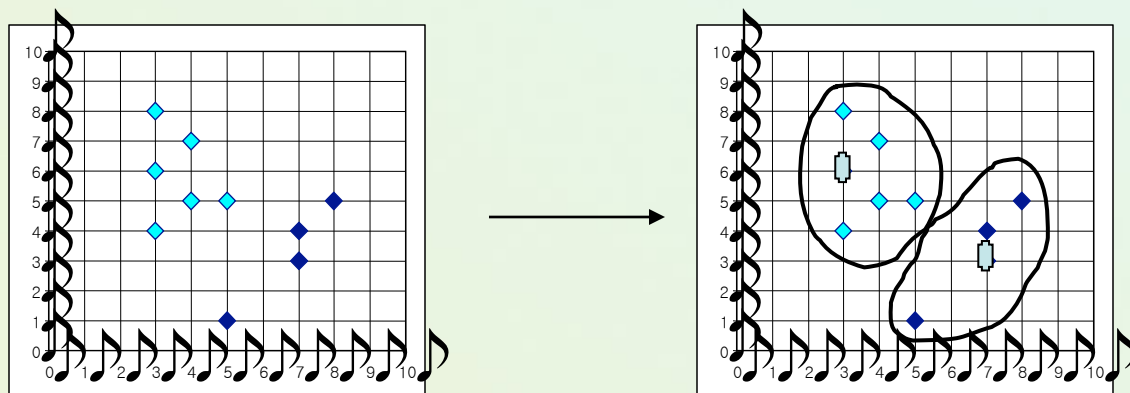
---

- Zastosowania ekonomiczne:
  - Identyfikacja grup klientów bankowych (np. właścicieli kart kredytowych wg. sposobu wykorzystania kart oraz stylu życia, danych osobowych, demograficznych) → cele marketingowe.
  - Systemy rekomendacji produktów i usług.
  - Rynek usług ubezpieczeniowych (podobne grupy klientów).
  - Analiza sieci sprzedaży (np. czy punkty sprzedaży podobne pod względem społecznego sąsiedztwa liczby personelu, itp., przynoszą podobne obroty).
  - Poszukiwanie wspólnych rynków dla produktów.
  - Planowanie, np. nieruchomości
- Badania naukowe (biologia, medycyna, nauki społeczne)
- Analiza zachowań użytkowników serwisów WWW
- Rozpoznawanie obrazów, dźwięku
- Wiele innych

# Wielowymiarowe statystyczne spojrzenie

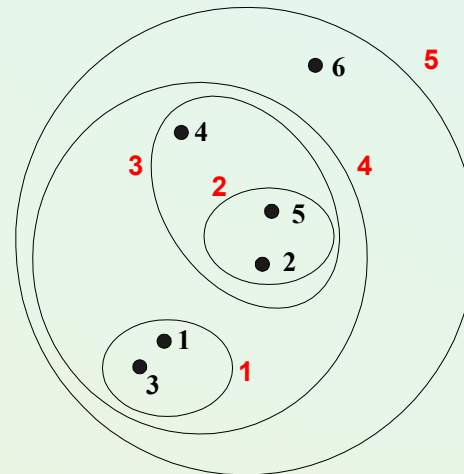
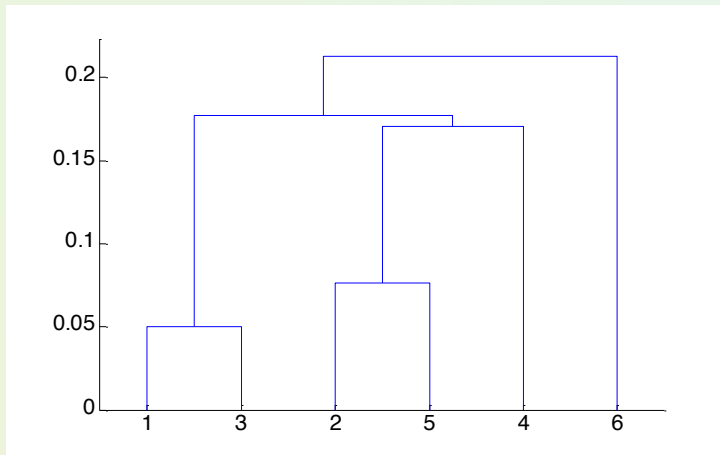
- obiekt opisany za pomocą  $n$  zmiennych  $X_1, X_2, \dots, X_n$  jest punktem  $x=(x_1, \dots, x_n)$  w  $n$ -wymiarowej przestrzeni  $\Omega$
- Cel podziału na grupy (S)  $\rightarrow$  obiekty podobne (reprezentowane przez punkty znajdujące się blisko siebie w przestrzeni) przydzielone do tej samej grupy, a obiekty niepodobne (reprezentowane przez punkty leżące w dużej odległości w przestrzeni) znajdują się w różnych grupach

$$S_i \cap S_j = \emptyset \quad (i \neq j; \quad i, j = 1, \dots, p) \quad \bigcup_{i=1}^p S_i = \Omega$$



# Grupowanie hierarchiczne

- Tworzy się stopniowo hierarchię zawierających się skupisk
  - Połączenie lub podział podzbiorów obiektów
- Wizualizacja – struktura drzewa nazwana **dendrogramem**



# Podział znanych metod

---

- Podziałowo- optymalizacyjne: Znajdź podział na zadaną liczbę skupień wg. zadanego kryterium.
- Metody hierarchiczne: Zbuduj drzewiastą strukturę skupień.
- Gęstościowo (Density-based): Poszukuj obszarów o większej gęstości występowania obserwacji
- Grid-based: wykorzystujące wielowymiarowy podział przestrzeni siatką ograniczeń
- Model-based: hipoteza co do własności modelu pewnego skupienia i procedura jego estymacji.



# Problemy do rozstrzygnięcia

---

- Jak odwzorować obiekty w przestrzeni?
  - Wybór zmiennych
  - Normalizacja zmiennych
- Jak mierzyć odległości między obiektami?
- Jaką metodę grupowania zastosować?

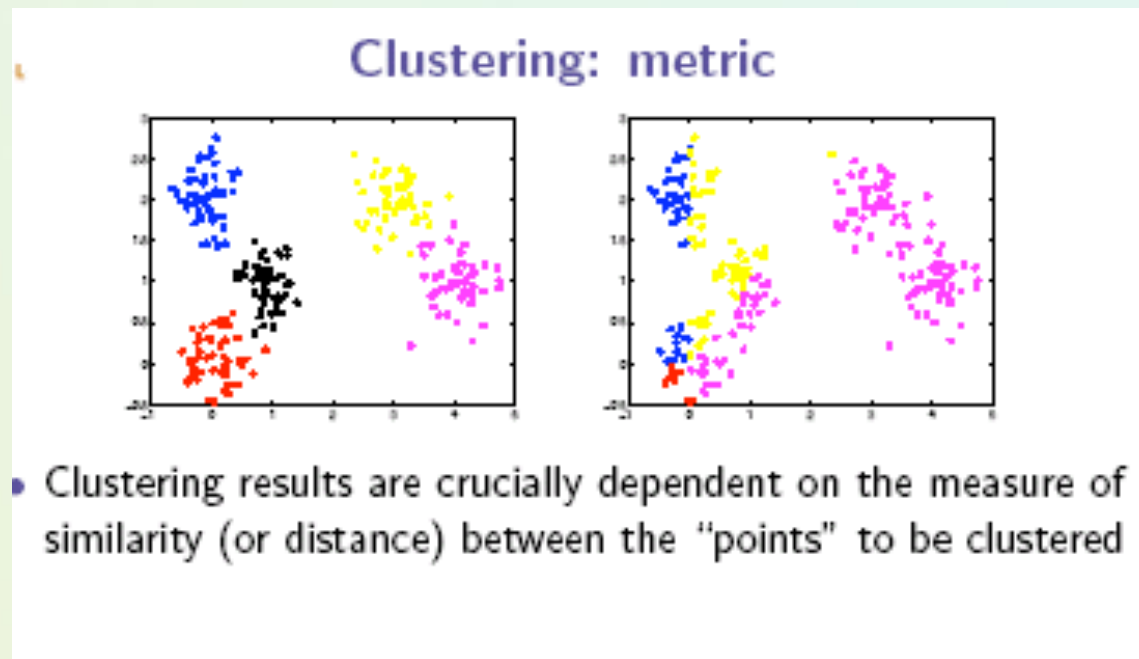
## Najczęściej stosowane miary odległości

Nazwa miary odległości	Definicja miary	Uwagi
Odległość euklidesowa	$d_{(rs)} = \left[ \sum_{k=1}^p (x_{(r)k} - x_{(s)k})^2 \right]^{\frac{1}{2}}$	Jest to odległość geometryczna w przestrzeni wielowymiarowej.
Kwadrat odległość euklidesowej	$d_{(rs)} = \left[ \sum_{k=1}^p (x_{(r)k} - x_{(s)k})^2 \right]^{\frac{1}{4}}$	Odległość euklidesową podnosi się do kwadratu, aby przypisać większą wagę obiektom, które są bardziej oddalone
Odległość miejska (Manhattan, City block)	$d_{(rs)} = \sum_{k=1}^p  x_{(r)k} - x_{(s)k} $	Jest to przeciętna różnica mierzona wzdłuż wymiarów. W większości przypadków ta miara odległości daje podobne wyniki, jak zwykła odległość euklidesowa. W przypadku tej miary, wpływ pojedynczych dużych różnic (przypadków odstających) jest stłumiony (ponieważ nie podnosi się ich do kwadratu).

# Problem doboru miary odległości / podobieństwa

---

- Nietrywialny i silnie wpływa na wynik
- Różne miary odległości



# Algorytmy podziałowo - optymalizacyjne

---

- Zadanie: Podzielenie zbioru obserwacji na  $K$  zbiorów elementów (skupień  $C$ ), które są jak najbardziej jednorodne
- Jednorodność – funkcja oceny
- Intuicja → zmienność wewnątrzskupieniowa  $wc(C)$  i zmienność międzyskupieniowa  $bc(C)$ 
  - Możliwe są różne sposoby zdefiniowania
  - np. wybierzmy środki skupień  $\mathbf{r}_k$  (centroidy)  $\mathbf{r}_k = \frac{1}{n_k} \sum_{\mathbf{x} \in C_k} \mathbf{x}$
  - Wtedy

$$wc(C) = \sum_{k=1}^K \sum_{\mathbf{x} \in C_k} d(\mathbf{x}, \mathbf{r}_k)^2$$

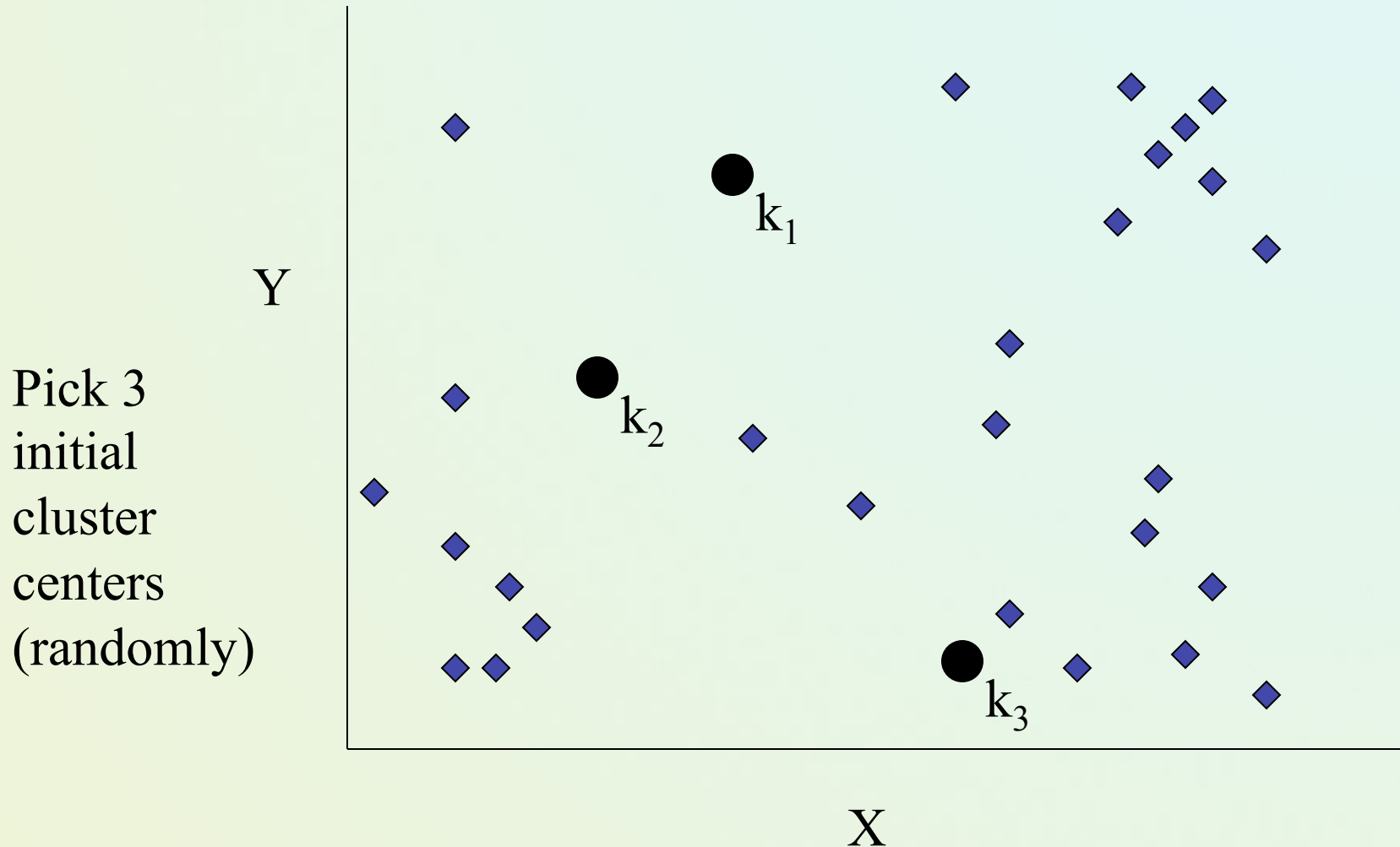
$$bc(C) = \sum_{1 \leq j < k \leq K} d(\mathbf{r}_j, \mathbf{r}_k)^2$$

# Podstawowe algorytmy podziałowe

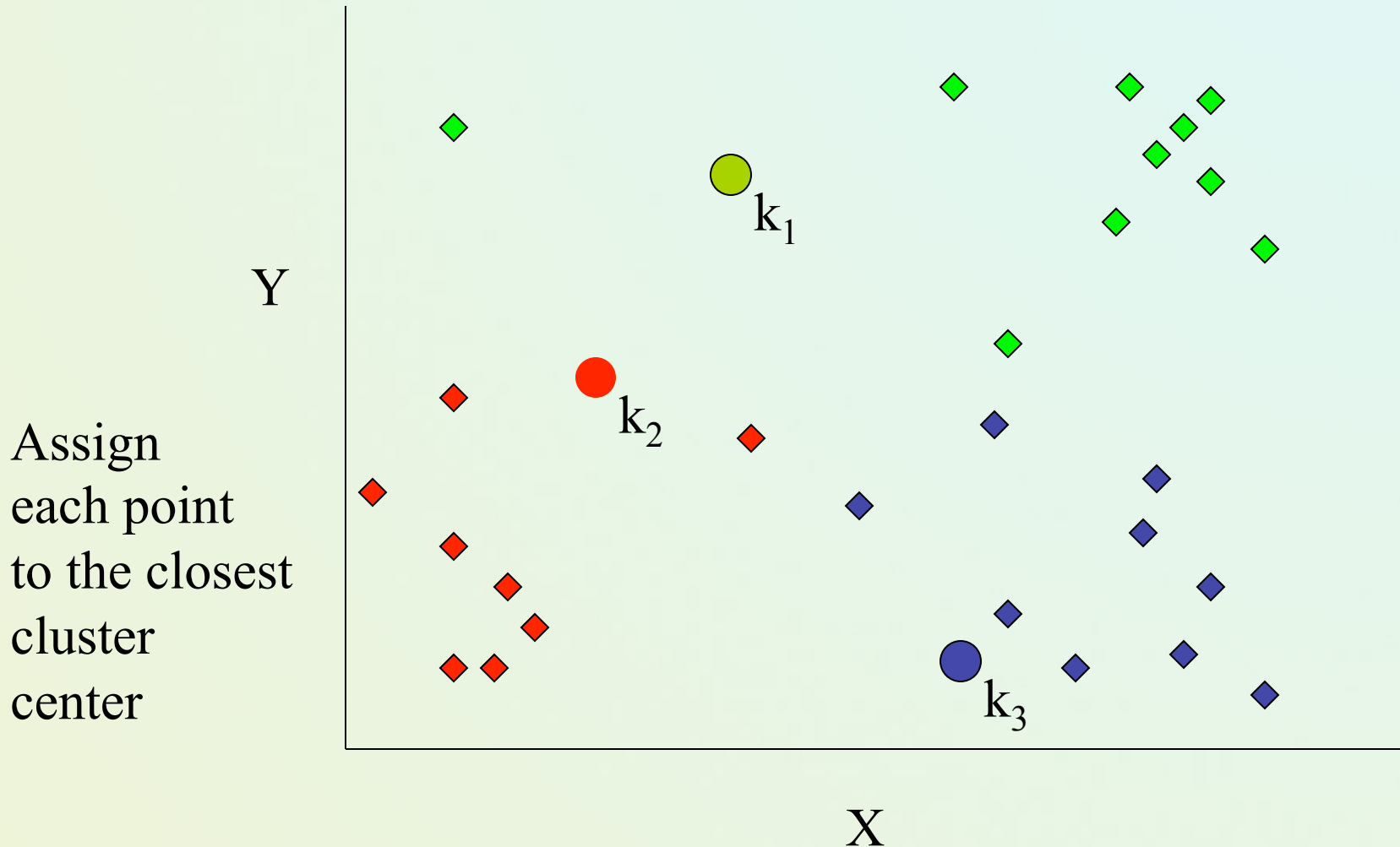
---

- Metoda  $K$  - średnich  $\rightarrow$  minimalizacja  $wc(C)$
- Przeszukiwanie przestrzeni możliwych przypisań  $\rightarrow$  bardzo kosztowne (oszacowanie w ks. Koronackiego)
- Problem optymalizacji kombinatorycznej  $\rightarrow$  systematyczne przeszukiwanie metodą iteracyjnego udoskonalania:
  - Rozpocznij od rozwiązania początkowego (losowego).
  - Ponownie przypisz punkty do skupień tak, aby otrzymać największą zmianę w funkcji oceny.
  - Przelicz zaktualizowane środki skupień, ...
  - Postępuj aż do momentu, w którym nie ma już żadnych zmian w funkcji oceny lub w składzie grup.
- Zachłanne przeszukiwanie  $\rightarrow$  proste i prowadzi do co najmniej lokalnego minimum. Różne modyfikacje, np. rozpoczynania od kilku rozwiązań startowych
- Złożoność algorytmy  $K$  - średnich  $\rightarrow O(KnI)$

# K-means example, step 1

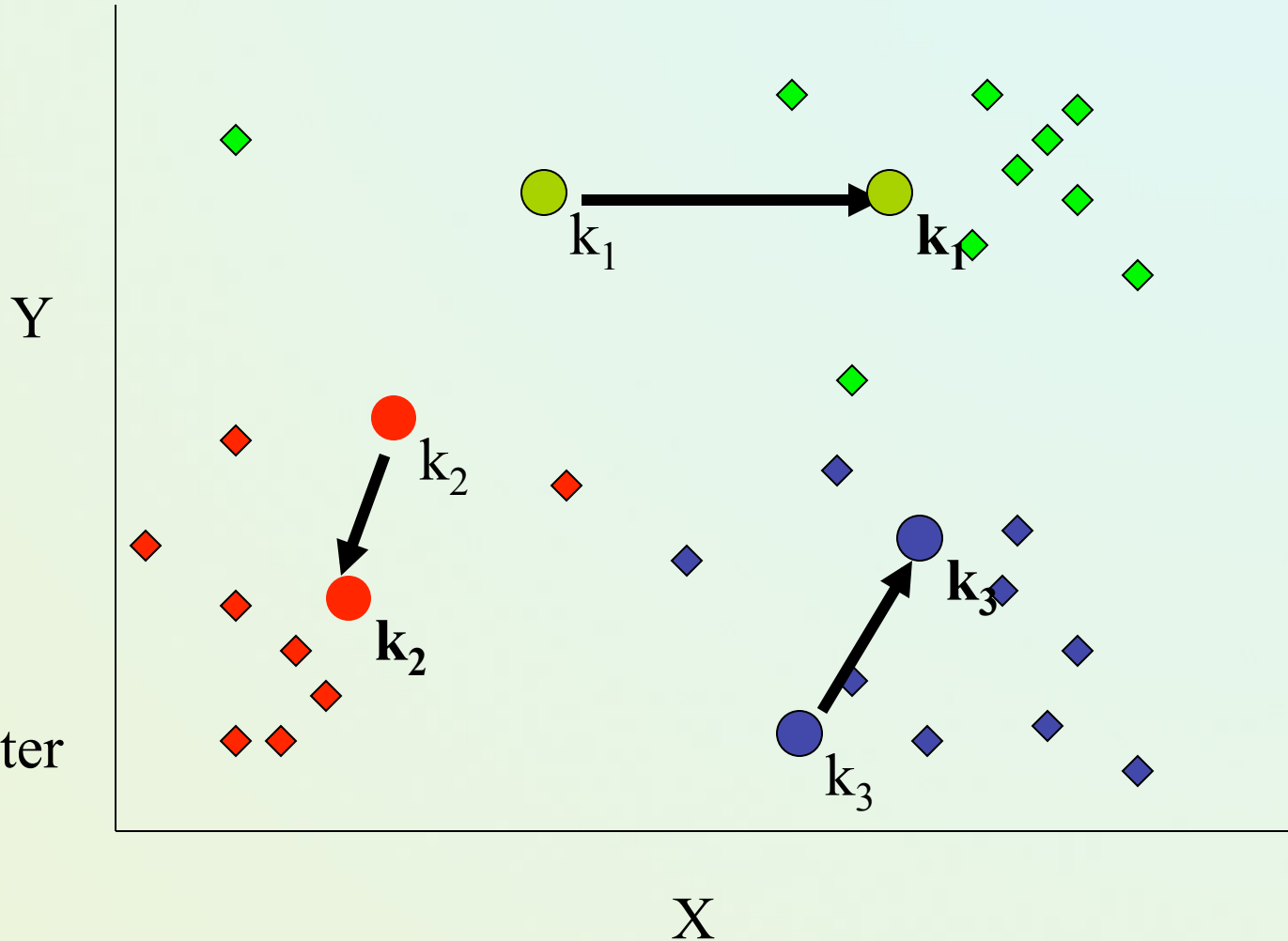


# K-means example, step 2



# K-means example, step 3

Move  
each cluster  
center  
to the mean  
of each cluster

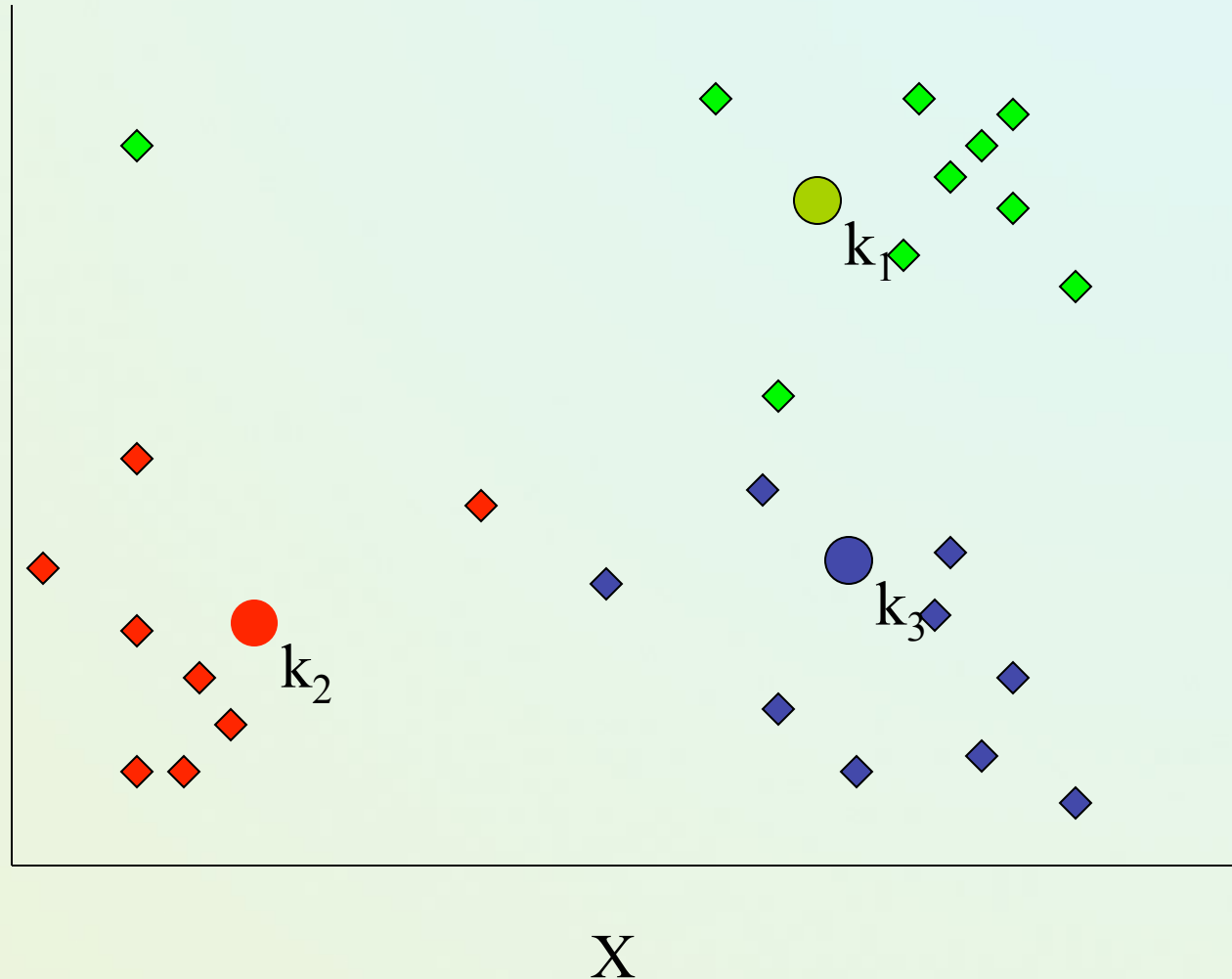




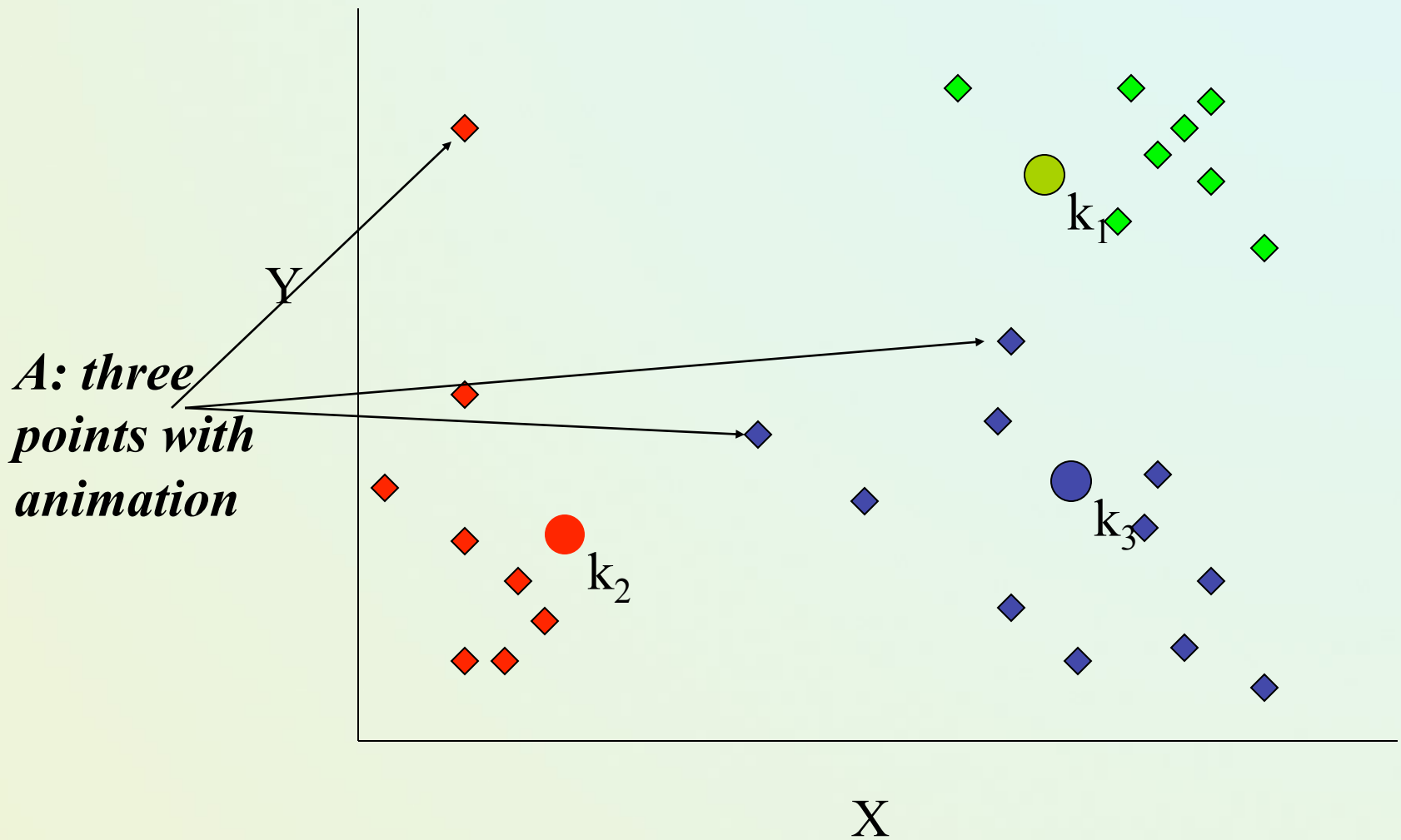
# K-means example, step 4

Reassign  
points  
closest to a  
different new  
cluster center

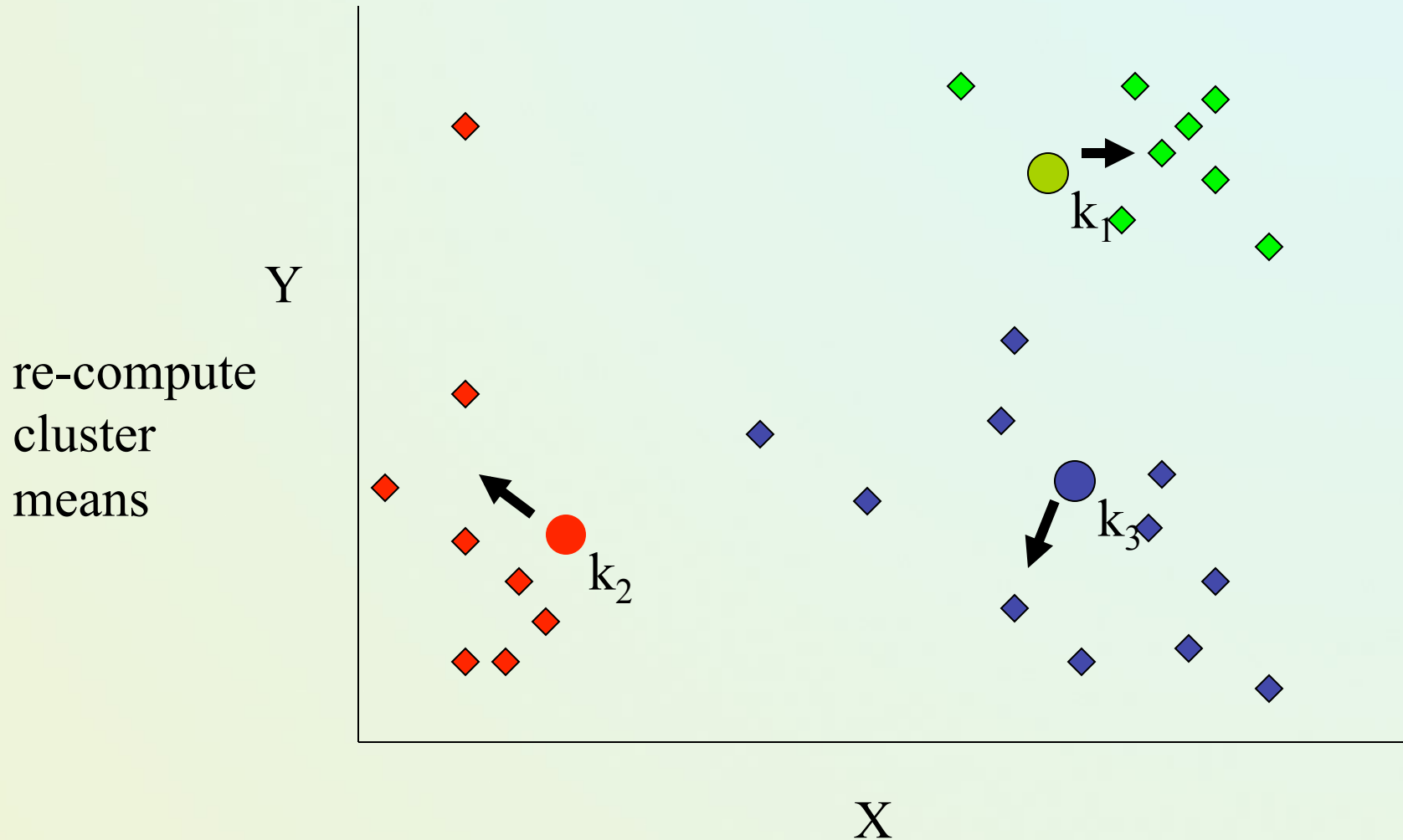
*Q: Which  
points are  
reassigned?*



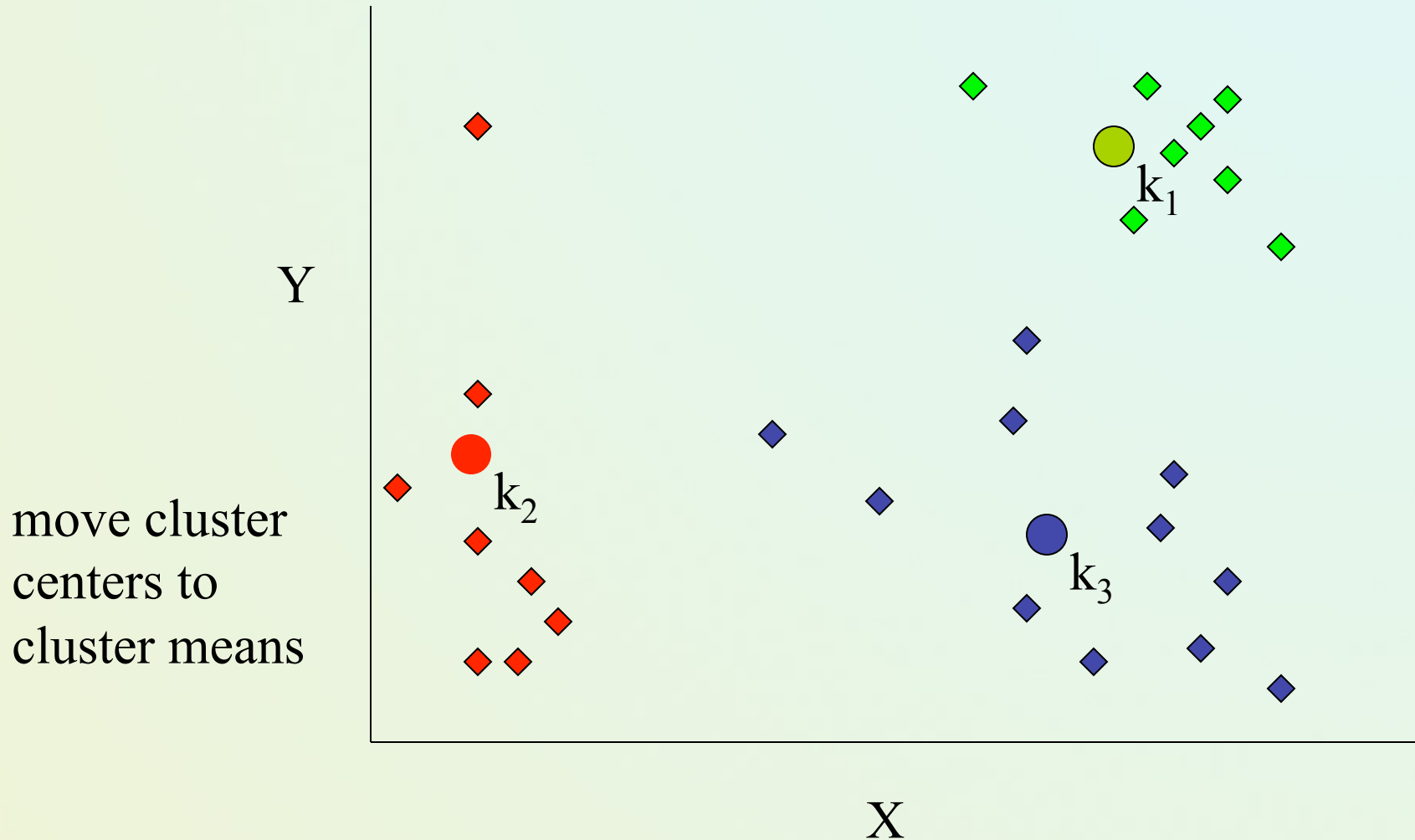
# K-means example, step 4 ...



# K-means example, step 4b



# K-means example, step 5



# Przykład (z macierzą początkową)

- Zbiór danych:  $x_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$   $x_2 = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$   $x_3 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$   $x_4 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$   $x_5 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$   
 $x_6 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$   $x_7 = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$   $x_8 = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$   $x_9 = \begin{bmatrix} 5 \\ 3 \end{bmatrix}$   $x_{10} = \begin{bmatrix} 5 \\ 4 \end{bmatrix}$

- Początkowy przydział  $B(0) = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$

- Przeliczenie centroidów  $R(0) = \begin{bmatrix} 3 & 3.2 & 2.5 \\ 2 & 2.6 & 2.5 \end{bmatrix}$

- Przeliczenie odległości

$$D(1) = \begin{bmatrix} 2 & 2.24 & 2.83 & 1.41 & 1 & 1 & 1.41 & 2 & 2.24 & 2.83 \\ 2.28 & 2.24 & 2.61 & 2 & 1.61 & 0.45 & 1.79 & 1.9 & 1.84 & 2.28 \\ 1.58 & 1.58 & 2.12 & 1.58 & 1.58 & 0.71 & 2.12 & 2.55 & 2.55 & 2.91 \end{bmatrix}$$

- Ponowny przydział do skupień:

$$B(1) = \begin{bmatrix} 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

## Przykład cd.

---

- Przeliczenie centroidów  $R(1) = \begin{bmatrix} 3 & 5 & 1.5 \\ 1 & 3 & 3 \end{bmatrix}$
- Ponowny przydział do skupień:  $B(2) = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$
- Warunek końcowy:  $B(2) = B(1)$

$$G_1 = \{x_4, x_5, x_7\} \quad G_2 = \{x_8, x_9, x_{10}\} \quad G_3 = \{x_1, x_2, x_3, x_6\}$$

# Metody optymalizacyjno-iteracyjne ( $k$ -średnich)

---

- Jednocześnie obliczana jest funkcja błędu podziału - ogólna suma kwadratów odległości wewnątrzgrupowych liczonych od środków ciężkości grup: tzn.

$$F = \sum_{j=1}^k \sum_{O_i \in S_j} d(O_i, M_j)^2$$

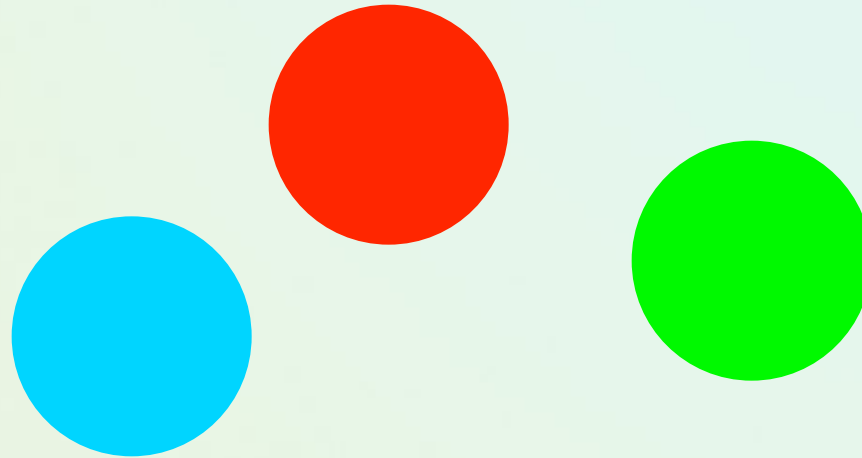
gdzie  $d$  jest odległością euklidesową.

W praktyce proces jest zbieżny po kilku lub kilkunastu iteracjach. Ponieważ w ogólności algorytm nie musi być zbieżny, ustala się maksymalną liczbę iteracji ( $L$ ).

# Pewne ukierunkowanie K-średnich

---

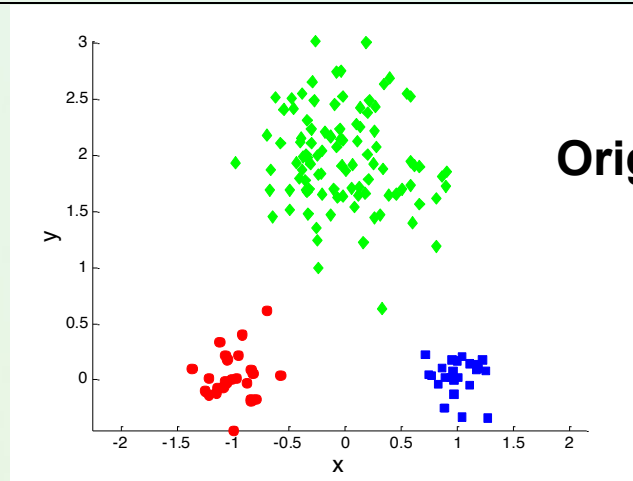
- Tworzy się kuliste kształty skupień



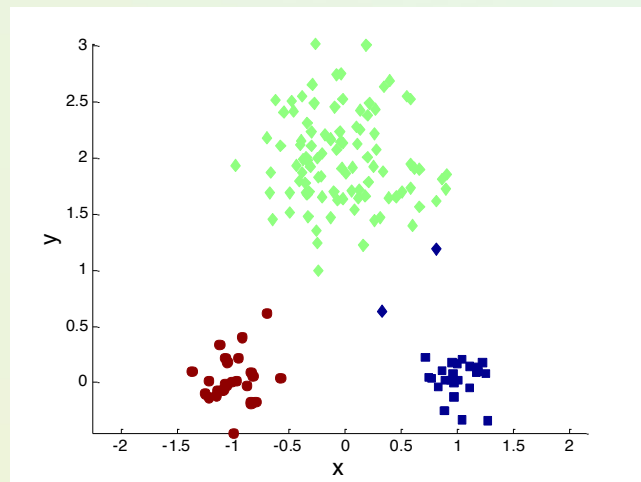
- Co z obserwacjami odstającymi i nieregularnymi kształtami skupień?



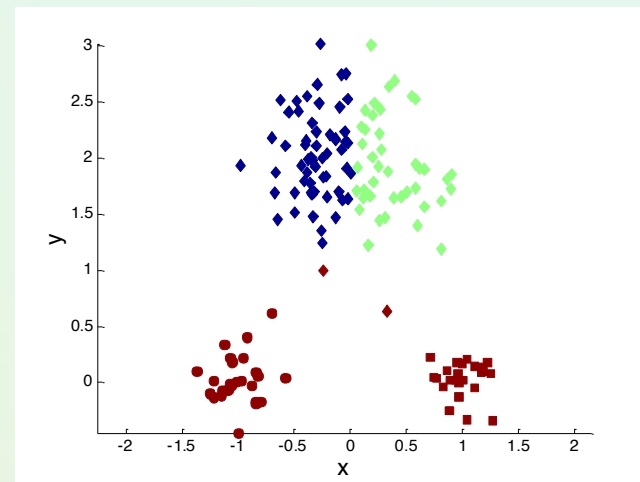
# Czy łatwo określić liczbę skupień



Original Points



Optimal Clustering



Sub-optimal Clustering

# Ustalanie liczby skupień i startowych centroidów

---

Liczbę skupień wybiera się na podstawie przesłanek merytorycznych albo szacuje się je metodami hierarchicznymi. Można dokonać obliczeń dla wszystkich wartości  $k$  z ustalonego przedziału:

$$k_{\min} \leq k \leq k_{\max}$$

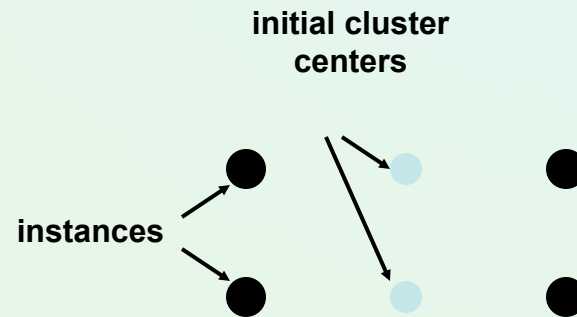
Możliwe są różne podejścia:

1. Arbitralny sposób np. przyjmuje się współrzędne pierwszych  $k$  obiektów jako załączki środków ciężkości
2. Losowy wybór środków ciężkości, przy czym może to być losowy wybór  $k$  obiektów ze zbioru danych albo losowy wybór  $k$  punktów przestrzeni niekoniecznie pokrywających się z położeniem obiektów
3. Wykorzystanie algorytmu optymalizującego w pewien sposób położenie początkowych środków ciężkości np. przez uwzględnianie  $k$  obiektów leżących daleko względem siebie
4. Przyjęcie jako początkowych środków ciężkości uzyskanych na podstawie podziału otrzymanego inna metodą, głównie jedną z metod hierarchicznych

# Uwagi nt. podziałowo- optymalizacyjnych

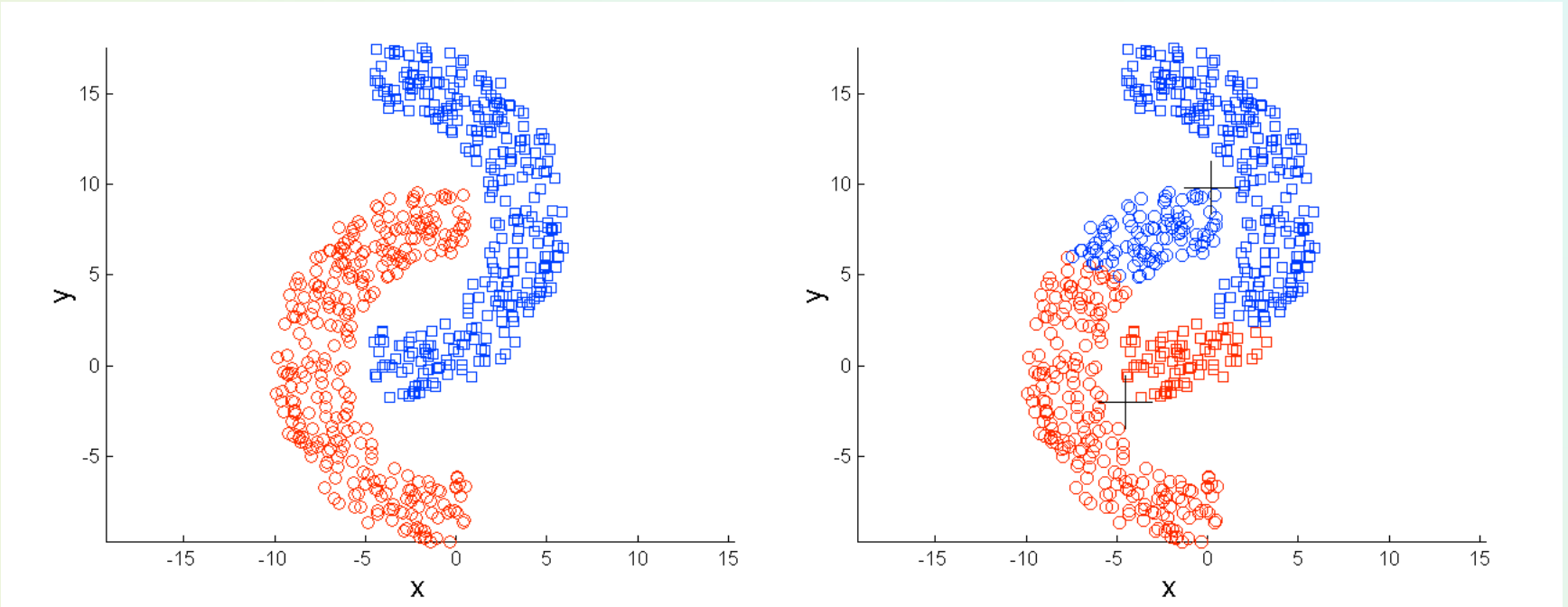
---

- Dobór parametru  $k$
- Kwestia heurystyki wyboru początkowych ziaren
- Czy jest zawsze zbieżny do „optimum globalnego”
  - Przykład:



- Możesz próbować kilka uruchomień z różnymi parametrami

# Ograniczenia K-średnich: Niesferyczne kształty



**Original Points**

**K-means (2 skupienia)**

# K-means krótkie podsumowanie

---

## Zalety

- Proste i łatwe do zrozumienia
- Reprezentacja skupień jako centroidy (centra, obserwacje centralne)

## Wady

- Jawne podanie liczby skupień
- Wszystkie przykłady muszą być przydzielone do skupień
- Problem z outliers (za duża wrażliwość)
- Ukierunkowanie na jednorodne „sferyczne” kształty skupień

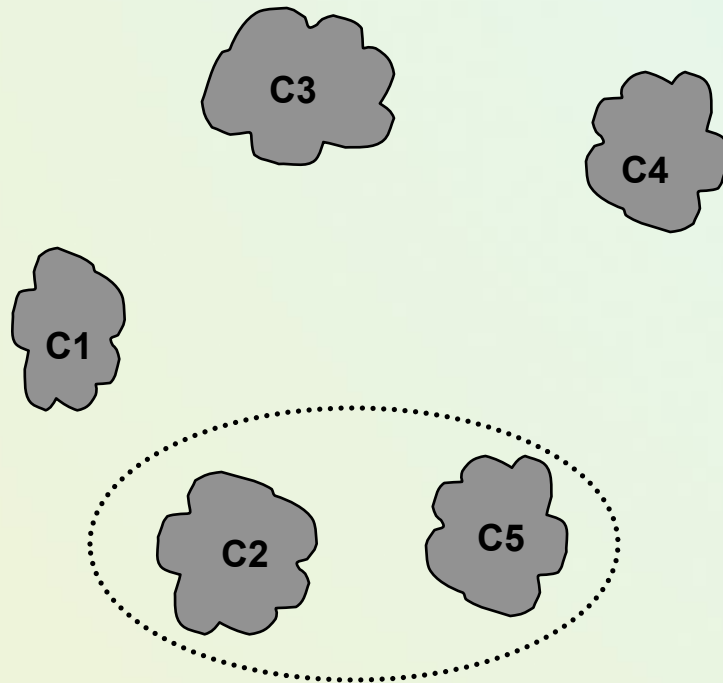
# Hierarchiczne metody aglomeracyjne - algorytm

---

1. W macierzy odległości znajduje się parę skupień najbliższych sobie.
2. Redukuje się liczbę klas łącząc znaną parę
3. Przekształca się macierz odległości metodą wybraną jako kryterium klasyfikacji
4. Powtarza się kroki 1- 3 dopóki nie powstanie jedna klasa zawierająca wszystkie skupienia.

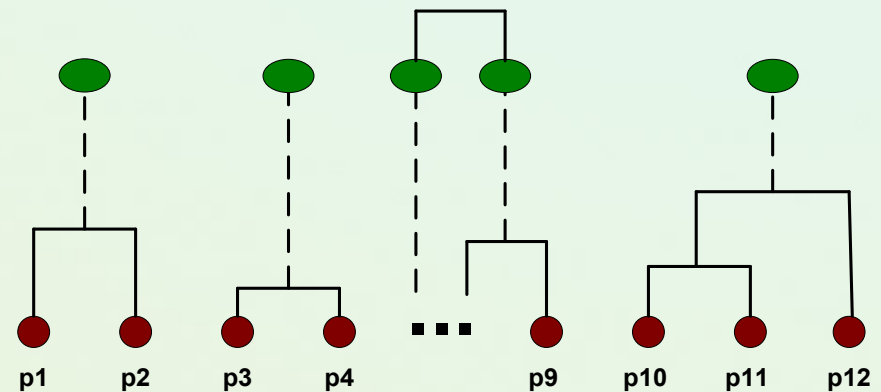
# Jak przeliczać macierz odległości?

- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.



	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



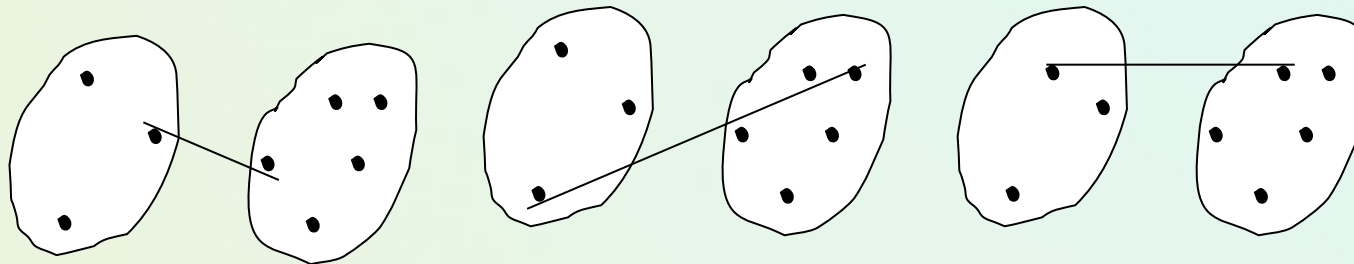
# AHC – wybór metody łączenia

---

1. Najbliższego sąsiedztwa (*Single linkage, Nearest neighbor*)
2. Najdalszego sąsiedztwa (*Complete linkage, Furthest neighbor*)
3. Mediany (*Median clustering*)
4. Środka ciężkości (*Centroid clustering*)
5. Średniej odległości wewnątrz skupień (*Average linkage within groups*)
6. Średniej odległości między skupieniami (*Average linkage between groups*)
7. Minimalnej wariancji Warda (*Ward's method*)



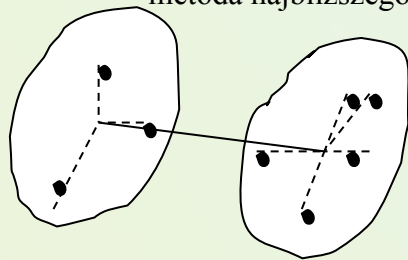
# Porównanie sposobu wyznaczania odległości między skupieniami w wybranych metodach aglomeracyjnych



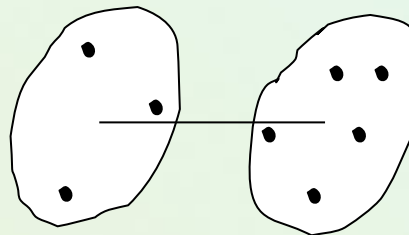
metoda najbliższego sąsiedztwa

metoda najdalszego sąsiedztwa

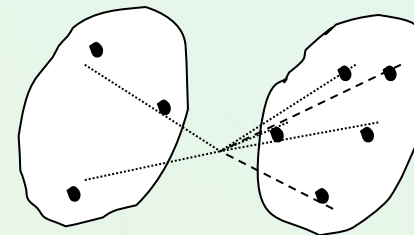
metoda mediany



metoda środka ciężkości



metoda średniej grupowej



metoda Warda

# Odległości między skupieniami

---

Single linkage  
minimum distance:

$$d_{\min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} \|p - p'\|$$

Complete linkage  
maximum distance:

$$d_{\max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} \|p - p'\|$$

mean distance:

$$d_{\text{mean}}(C_i, C_j) = \|m_i - m_j\|$$

average distance:

$$d_{\text{ave}}(C_i, C_j) = 1 / (n_i n_j) \sum_{p \in C_i} \sum_{p' \in C_j} \|p - p'\|$$

$m_i$  is the mean for cluster  $C_i$       $n_i$  is the number of points in  $C_i$

# Single Link Agglomerative Clustering

---

- Użyj maksymalnego podobieństwa dwóch obiektów:

$$\text{sim}(c_i, c_j) = \max_{x \in c_i, y \in c_j} \text{sim}(x, y)$$

- Prowadzi do „(long and thin) clusters due to *chaining effect*” (efekt łańcuchowy); prowadzi do formowania grup niejednorodnych (heterogenicznych);
  - Dogodne w specyficznych zastosowaniach
- Pozwala na wykrycie **obserwacji odstających**, nie należących do żadnej z grup, i warto przeprowadzić klasyfikację za jej pomocą na samym początku, aby wyeliminować takie obserwacje i przejść bez nich do właściwej części analizy

# Complete Link Agglomerative Clustering

---

- Użyj maksymalnej odległości – minimalnego podobieństwa

$$\text{sim}(c_i, c_j) = \min_{x \in c_i, y \in c_j} \text{sim}(x, y)$$

- Ukierunkowana do “tight,” spherical clusters
- Metoda zalecana gdy, kiedy obiekty faktycznie formują naturalnie oddzielone "kępki". Metoda ta nie jest odpowiednia, jeśli skupienia są w jakiś sposób wydłużone lub mają naturę "łańcucha".

# Wrażliwość na dobór metod łączenia skupień

Diagram dla 22 przyp.  
 Pojedyncze wiązanie  
 Odległości euklidesowe

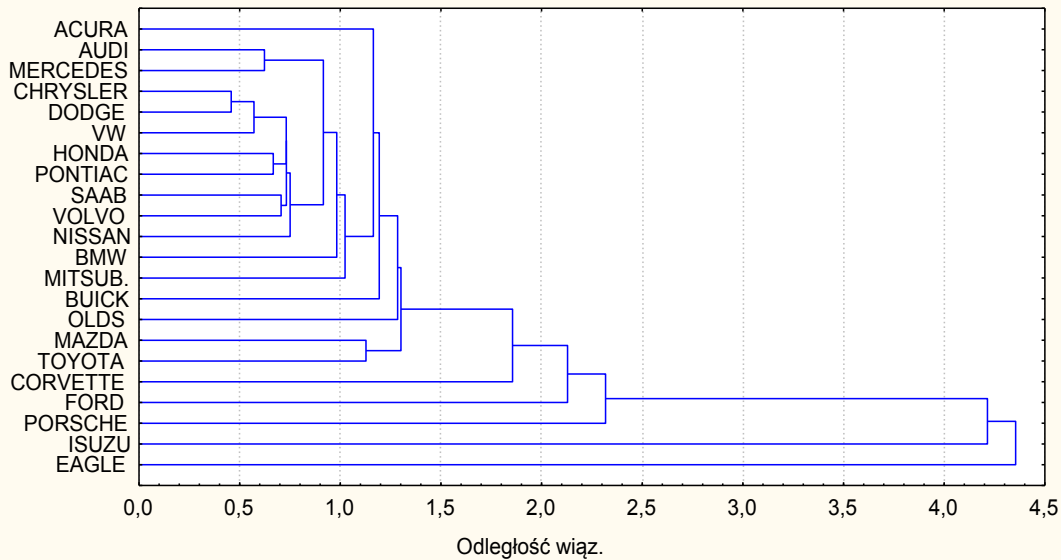
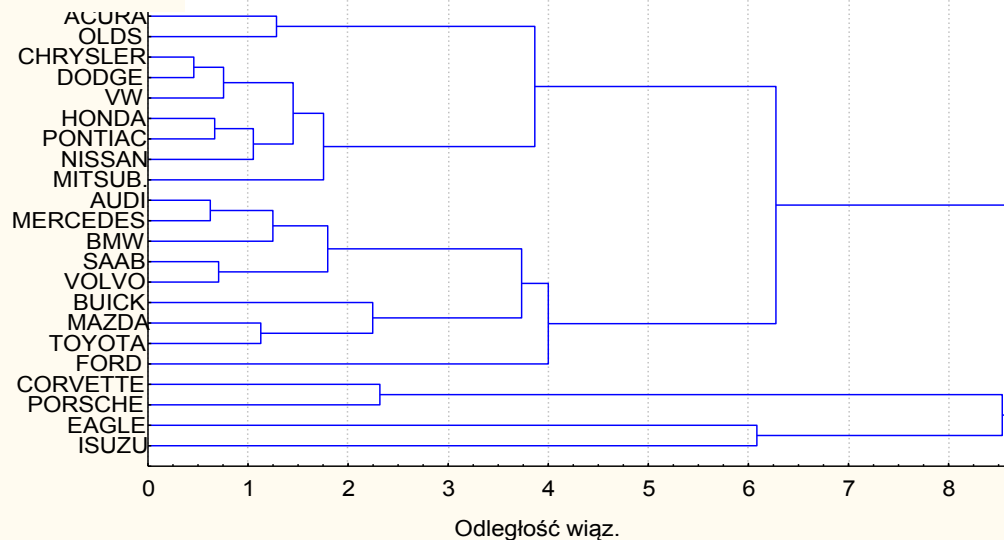


Diagram dla 22 przyp.  
 Metoda Warda  
 Odległości euklidesowe



# Metoda średnich połączeń

## [Unweighted pair-group average]

---

- W metodzie tej odległość między dwoma skupieniami oblicza się jako średnią odległość między wszystkimi parami obiektów należących do dwóch różnych skupień
- Metoda ta jest efektywna, gdy obiekty formują naturalnie oddzielone "kępki", ale zdaje także egzamin w przypadku skupień wydłużonych, mających charakter "łańcucha"

# Metoda ważonych środków ciężkości (mediany) [Weighted pair-group centroid].

---

- Jest to metoda podobna jak poprzednia, z tym wyjątkiem, że w obliczeniach wprowadza się „ważenie”, aby uwzględnić różnice między wielkościami skupień (tzn. liczbą zawartych w nich obiektów).
- Zatem, metoda ta jest lepsza od poprzedniej w sytuacji, gdy istnieją (lub podejrzewamy, że istnieją) znaczne różnice w rozmiarach (liczności) skupień

# Metody łączenia – Ward method

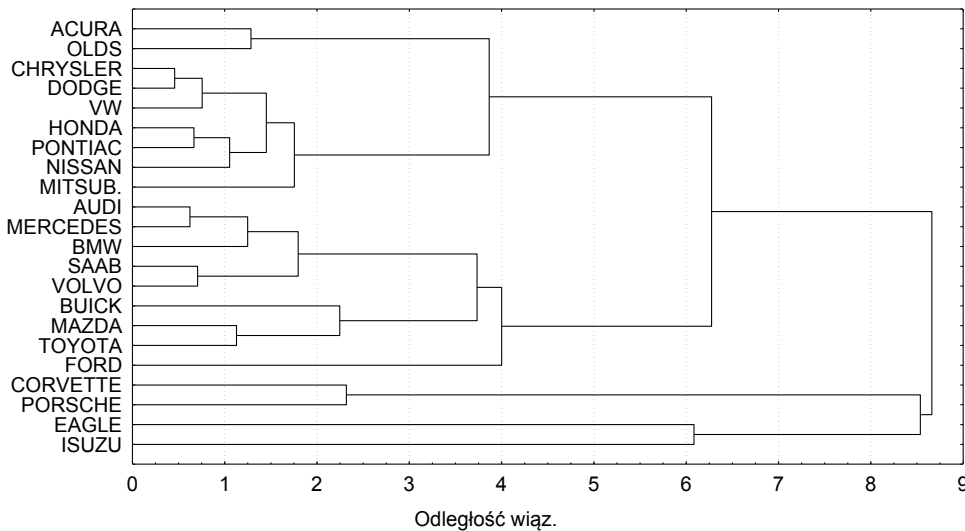
---

- Gdy powiększamy jedno ze skupień  $C_k$ , wariancja wewnątrzgrupowa (liczona przez kwadraty odchyleń od średnich w zbiorach  $C_k$ ) rośnie.
- Metoda polega na takim powiększaniu zbiorów  $C_k$ , która zapewnia **najmniejszy przyrost tej wariancji** dla danej iteracji.
- Kryterium grupowania jednostek: minimum zróżnicowania wektorów cech  $\mathbf{x}_j$  tworzących zbiór  $C_k$  ( $k = 1, \dots, K$ ) względem wartości średnich w tych zbiorach.
- Ogólnie, metoda ta jest traktowana jako bardzo efektywna, chociaż zmierza do tworzenia skupień o małej wielkości → zrównoważone drzewa o wielu elementach
- Ważne – powiązanie z miarą odległości między obiektami (Pearson vs. inne)



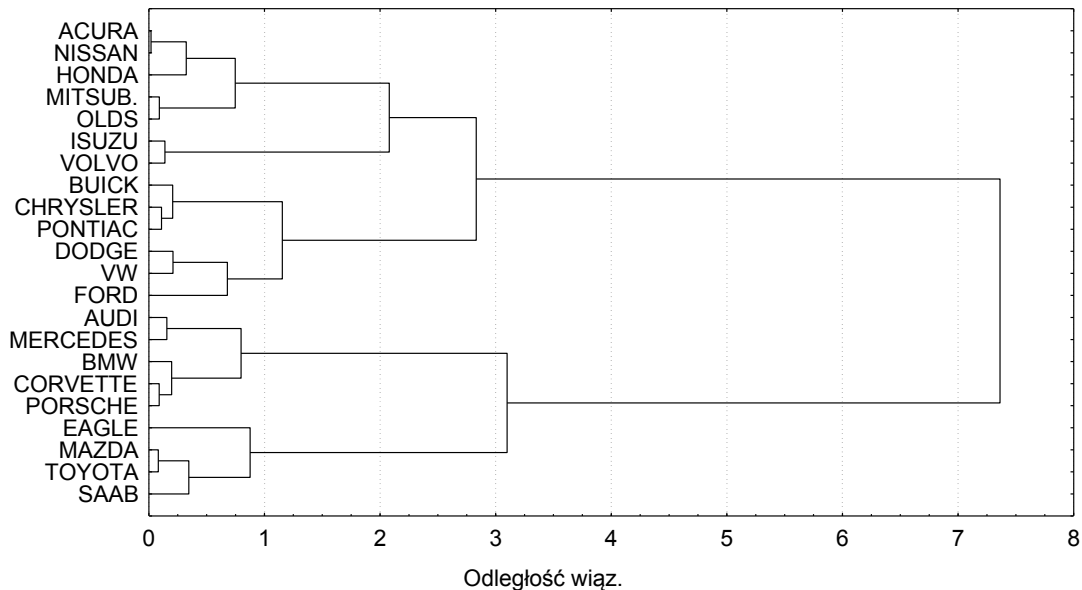
# Przykłady użycia metody Warda

Diagram dla 22 przyp.  
Metoda Warda  
Odległości euklidesowe



## Cars data

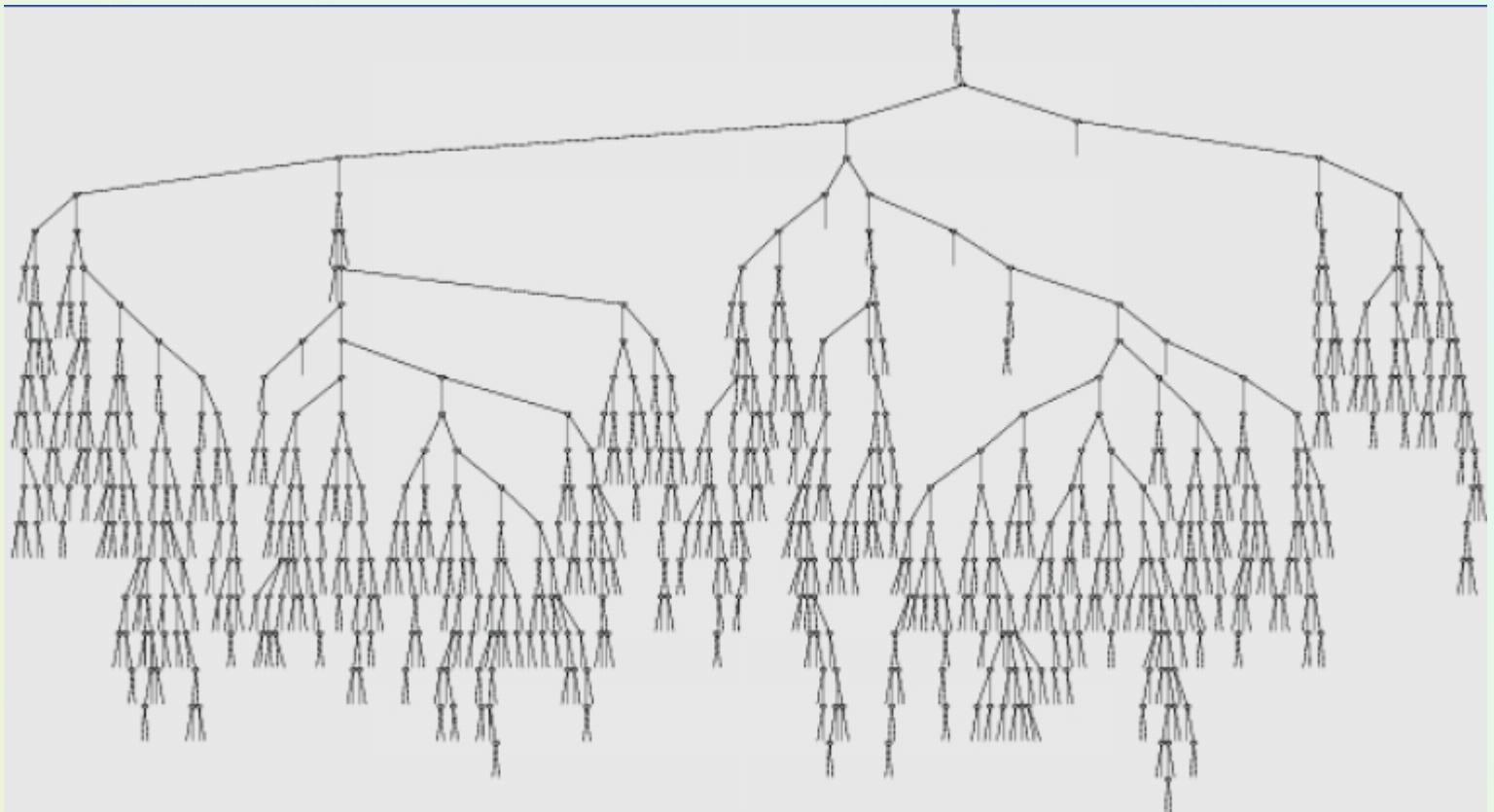
Diagram dla 22 przyp.  
Metoda Warda  
1-r Pearsona



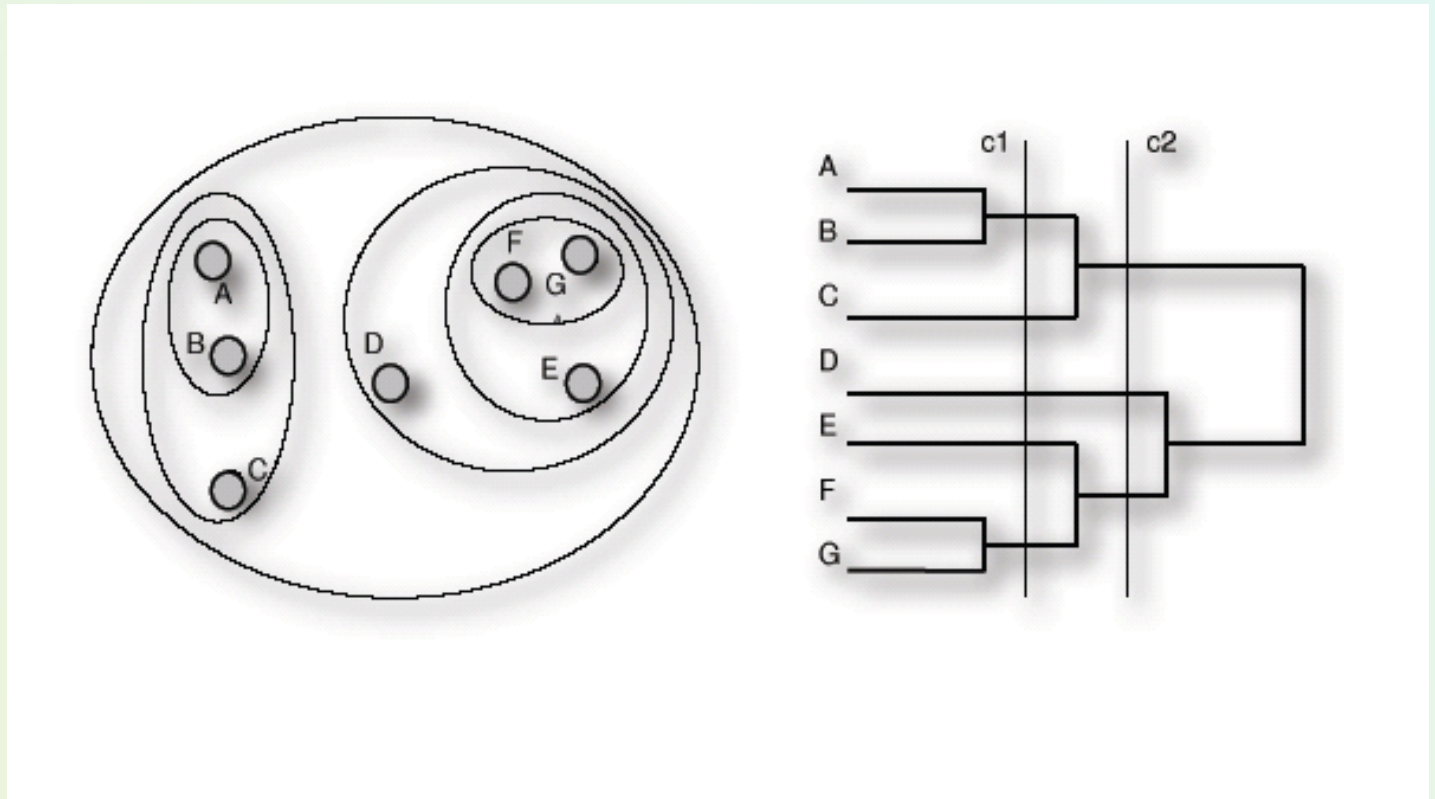
# Czy struktura drzewa skupień jest czytelna?

---

- Przykład biochemiczny

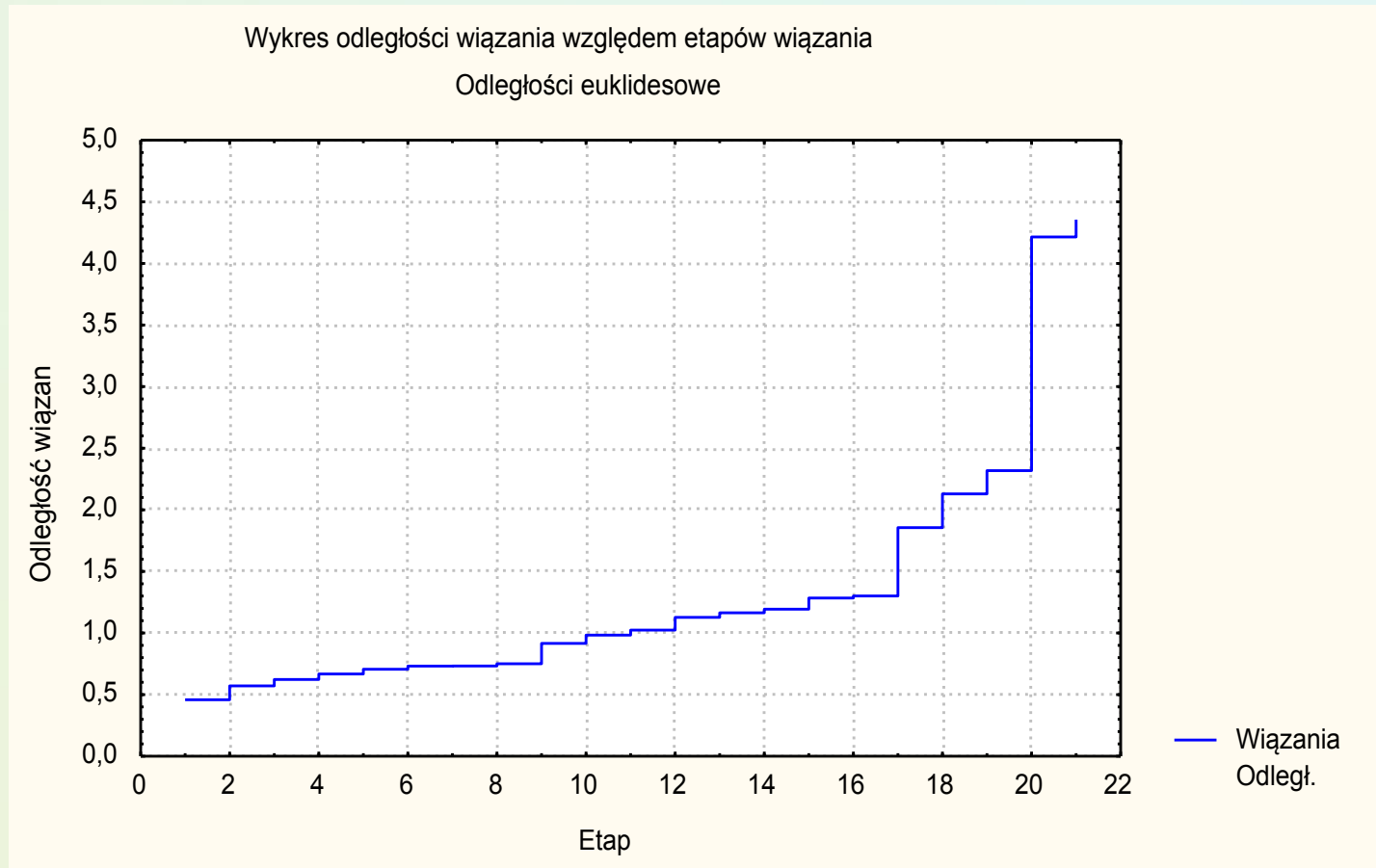


# Jak wybrać liczbę skupień?



# AHC – jak odnaleźć liczbę skupień?

- Find a cut point („kolanko” wykresu/ knee)



## **Analiza skupień w Statsoft -Statistica**

# Analiza Skupień – Statistica; więcej na [www.statsoft.com](http://www.statsoft.com). Przykład analizy danych o parametrach samochodów

STATISTICA: Analiza skupień - [Dane: CARS.STA 5v \* 22c ]

Plik Edycja Widok Analiza Wykresy Opcje Okno Pomoc

-521072362755425

Zmienne Przypadki

LICZBOWE WARTOŚCI

Cena, wydajność, trzymanie się drogi różny

	1	2	3	4	5
	CENA	PRZYSP	HAMOWAN	WSK_TRZY	ZUŻYCIE
Acura	-,521	,477	-,007	,382	2,079
Audi	,866	,208	,319	-,091	-,677
BMW	,496	-,802	,192	-,091	-,154
Buick	-,614	1,689	,933	-,210	-,154
Corvette	1,235	-1,811	-,494	,973	-,677
Chrysler	-,614	,073	,427	-,210	-,154
Dodge	-,706	-,196	,481	,145	-,154
Eagle	-,614	1,218	-4,199	-,210	-,677
Ford	-,706	-1,542	,987	,145	-1,724
Honda	-,429	,410	-,007	,027	,369
Isuzu	-,798	,410	-,061	-4,230	1,067
Mazda	,126	,679	-,133	,500	-1,724
Mercedes	1,051	,006	,120	-,091	-,154
Mitsub.	-,614	-1,003	,084	,382	,718
Nissan	-,429	,073	-,007	,263	,997
Olds	-,614	-,734	,409	,382	2,114
Pontiac	-,614	,679	,536	,145	,195
Porsche	3,454	-2,215	-,296	,618	-1,026
Saab	,588	,679	,246	,263	,021
Toyota	-,059	1,218	,228	,736	-,851
VW	-,706	-,128	,102	,382	,195
Volvo	,219	,612	,138	-,210	,369

Metoda grupowania

- Aglomeracja
- Grupowanie metodą k-średnich
- Grupowanie obiektów i cech

OK

Anuluj

Otwórz dane

SELECT CASES

W

STATYSTYKA: Analiza skupień

Widok Analiza Wykresy Opcje Okno Pomoc

Imię: CARS.STA 5v \* 22c

Cena, wydajność, trzymanie się drogi różny					
1	2	3	4	5	
CENA	PRZYSP	HANOWAN	WSK TRZY	ZUŻYCIE	
-521	477	-007	382	2,079	
866	208	319	-091	-677	
496	-802	192	-091	-154	

Przebieg aglomeracji (cars.sta)

Dalej...	Pojedyncze wiązanie Odległości euklidesowe									
połącz. odległ.	Obj. Nr 1	Obj. Nr 2	Obj. Nr 3	Obj. Nr 4	Obj. Nr 5	Obj. Nr 6	Obj. Nr 7	Obj. Nr 8	Obj. Nr 9	Obj. Nr 10
4580484	Chrysler	Dodge								
5710964	Chrysler	Dodge	VW							
6231085	Audi	Mercedes								
6670490	Honda	Pontiac								
7060042	Saab	Volvo								
7313396	Chrysler	Dodge	VW	Honda	Pontiac					
7323840	Chrysler	Dodge	VW	Honda	Pontiac	Saab	Volvo			
7506309	Chrysler	Dodge	VW	Honda	Pontiac	Saab	Volvo	Nis		
9159300	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pontiac	S		
9824548	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pontiac	S		
1.023831	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pontiac	S		
1.127473	Mazda	Toyota								
1.164055	Acura	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pont		
1.193655	Acura	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pont		
1.284603	Acura	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pont		
1.301269	Acura	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pont		
1.855838	Acura	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pont		
2.128886	Acura	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pont		
2.317976	Acura	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pont		
4.214866	Acura	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pont		
4.355048	Acura	Audi	Mercedes	Chrysler	Dodge	VW	Honda	Pont		

**Analiza skupień: Aglomeracja**

Zmienne: **WSZYSTKIE** OK

Wejście: **Dane surowe** Anuluj

Grupowanie: **Przypadki (obiekty)**

Metoda aglomeracji (wiązania): **Pojedynczego wiązania**

Miara odległości: **Odległość euklidesowa**

Braki danych: **Usuwane przypadkami**

Przetwarzanie wsadowe i drukowanie

SELECT CASES \$ W

**Wyniki aglomeracji**

Liczba zmiennych: **5**

Liczba przyp.: **22**

Łączenie przyp.

Braki danych były **usuwane przypad.**

Metoda aglomeracji: **Pojedyncze wiązanie**

Miara odległości: **Odległości euklidesowe standaryzowane**

**Poziomy hierarchiczny wykres drzewkowy** OK

**Pionowy wykres soplekowy** Anuluj

Prostokątne gałęzie

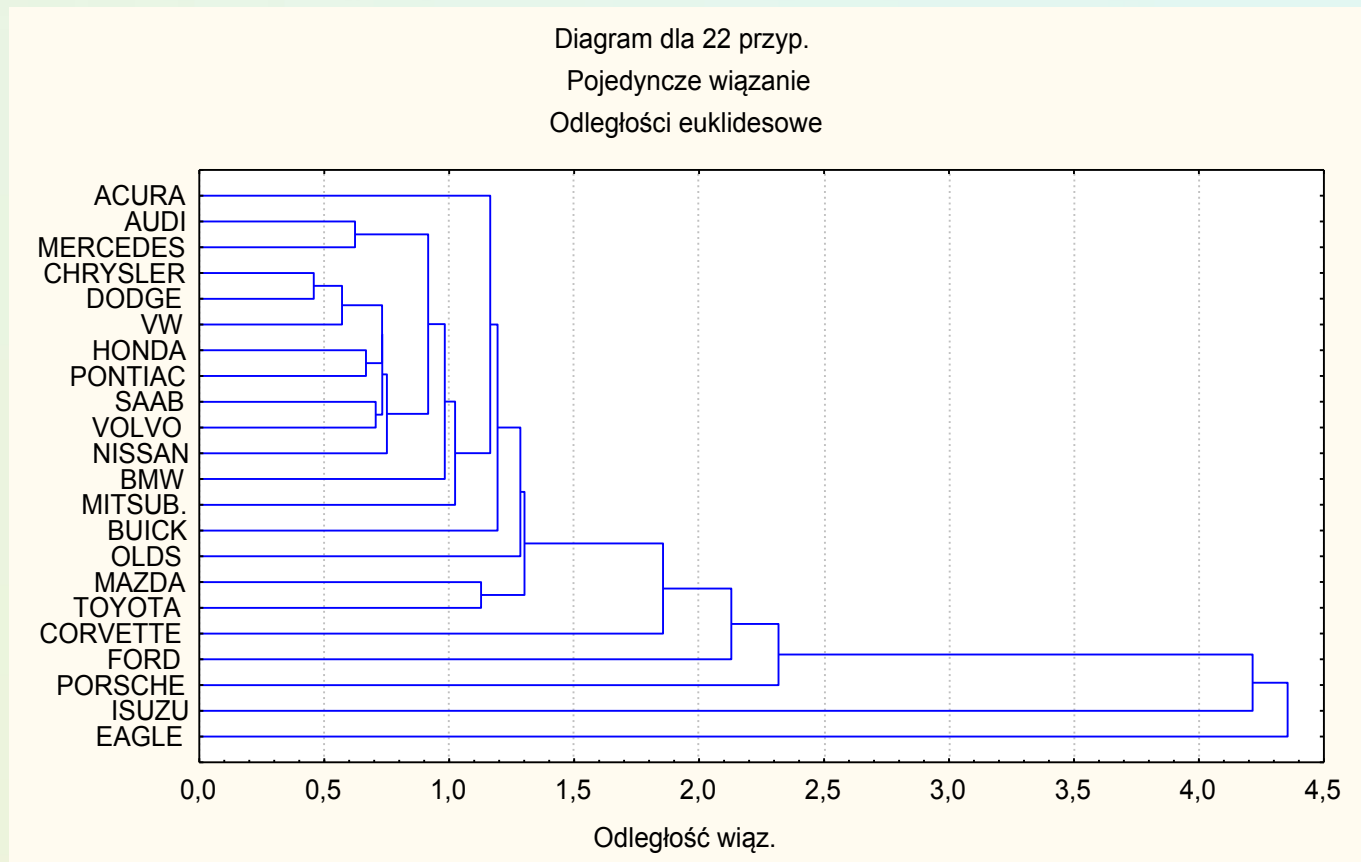
Skaluj drzewo do odl\_wiąz./odl\_maks\*100

**Macierz odległości**

**Przebieg aglomeracji** Statystyki opisowe

**Wykres przebiegu aglomeracji** Zapisz macierz odległości

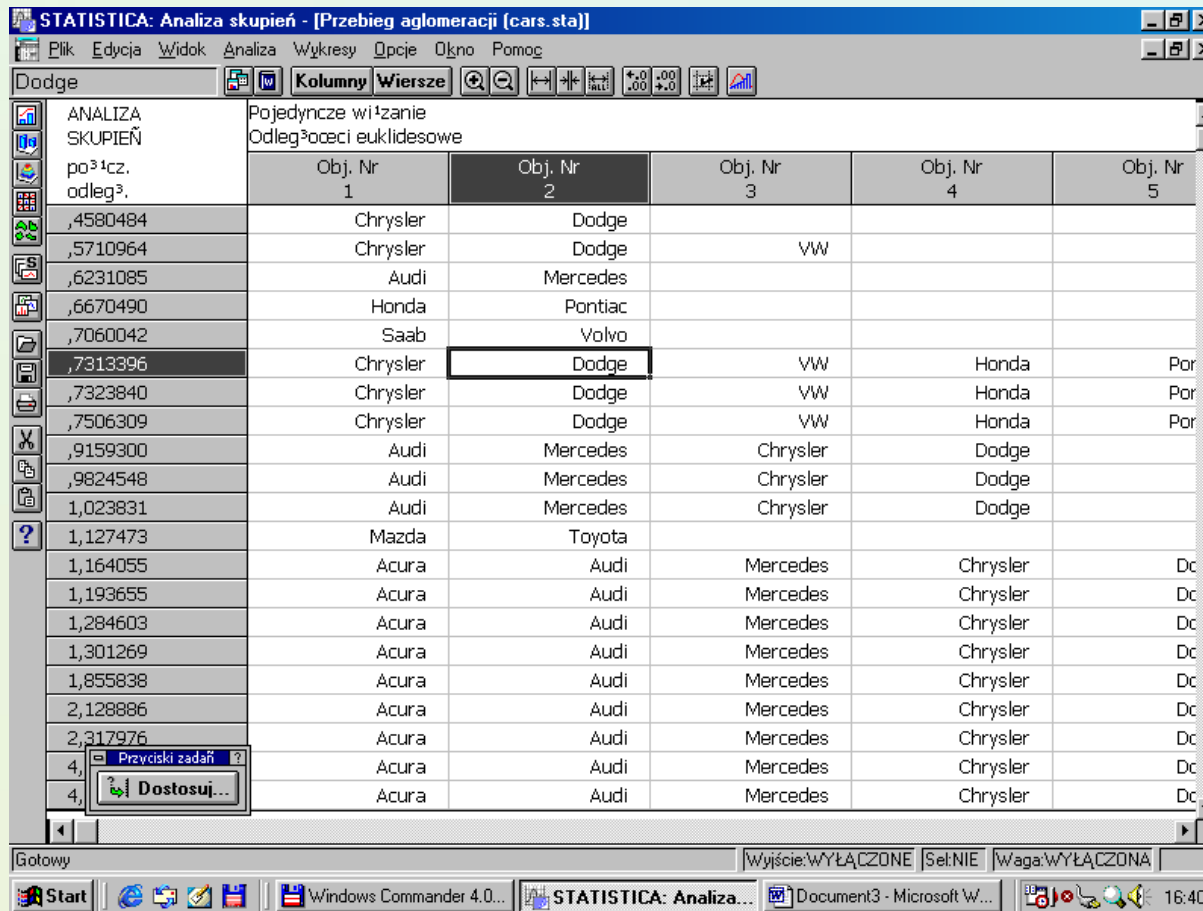
# Dendrogram for Single Linkage





# Opis tworzenia dendrogramu

- Łączenie obiektów w kolejnych krokach



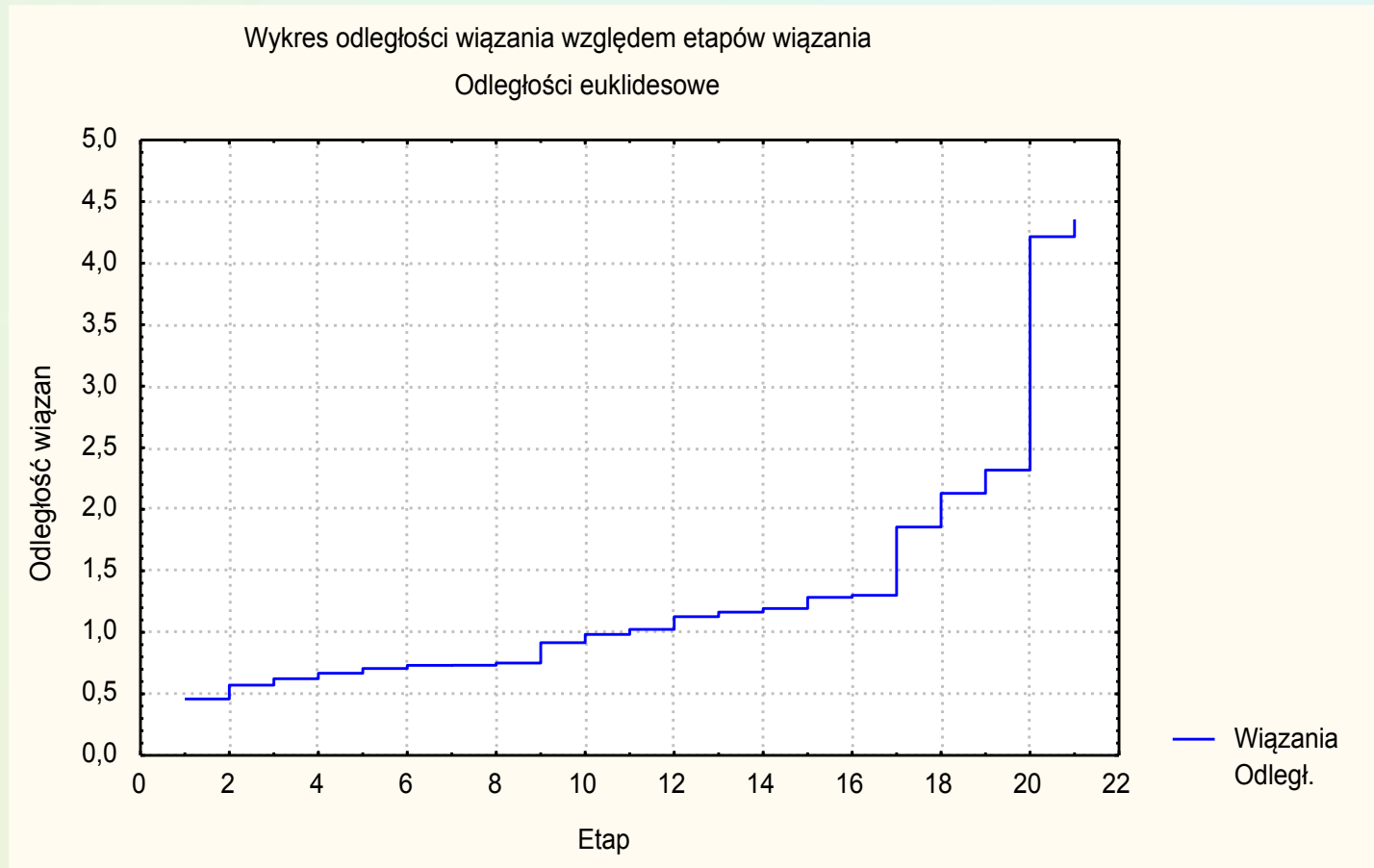
The screenshot shows the STATISTICA software interface for a clustering analysis. The window title is "STATISTICA: Analiza skupień - [Przebieg aglomeracji (cars.sta)]". The menu bar includes "Plik", "Edycja", "Widok", "Analiza", "Wykresy", "Opcje", "Okno", and "Pomoc". The toolbar shows "Kolumny" and "Wiersze" buttons. The main window displays a table with the following data:

	Obj. Nr 1	Obj. Nr 2	Obj. Nr 3	Obj. Nr 4	Obj. Nr 5
,4580484	Chrysler	Dodge			
,5710964	Chrysler	Dodge	VW		
,6231085	Audi	Mercedes			
,6670490	Honda	Pontiac			
,7060042	Saab	Volvo			
,7313396	Chrysler	Dodge	VW	Honda	Por
,7323840	Chrysler	Dodge	VW	Honda	Por
,7506309	Chrysler	Dodge	VW	Honda	Por
,9159300	Audi	Mercedes	Chrysler	Dodge	
,9824548	Audi	Mercedes	Chrysler	Dodge	
1,023831	Audi	Mercedes	Chrysler	Dodge	
1,127473	Mazda	Toyota			
1,164055	Acura	Audi	Mercedes	Chrysler	Dc
1,193655	Acura	Audi	Mercedes	Chrysler	Dc
1,284603	Acura	Audi	Mercedes	Chrysler	Dc
1,301269	Acura	Audi	Mercedes	Chrysler	Dc
1,855838	Acura	Audi	Mercedes	Chrysler	Dc
2,128886	Acura	Audi	Mercedes	Chrysler	Dc
2,317976	Acura	Audi	Mercedes	Chrysler	Dc
4,	Acura	Audi	Mercedes	Chrysler	Dc
4,	Acura	Audi	Mercedes	Chrysler	Dc

The status bar at the bottom shows "Gotowy" and "Wyjście: WYŁĄCZONE | Sel: NIE | Waga: WYŁĄCZONA". The taskbar at the bottom includes the Start button and several open applications: "Windows Commander 4.0...", "STATISTICA: Analiza...", and "Document3 - Microsoft W...". The system clock shows "16:40".

# Analiza procesu łączenia

- Wykres kolankowy – a cut point („kolanko” / knee)



**Wyniki grupowania metodą k-średnich**

Liczba zmiennych: 5  
 Liczba przyp.: 22  
 Wiązanie przypadków met.k-ś  
 Braki danych usuwano przypadkami  
 Liczba skupień: 4  
 Rozwiązanie odnaleziono po 1 iteracjach

Analiza wariacji    Anuluj

Średnie skupień i odległości euklidesowe

Wykres średnich

Statystyki opisowe każdego skupienia

Elementy każdego skupienia i odległości

Zapisz klasyfikacje i odległości

**Analiza skupień: Grupowanie metodą k-średnich**

Zmienne: **WSZYSTKIE**    OK

Grupowanie: Przypadki (obiekty)

Liczba skupień: 4

Liczba iteracji: 10

Braki danych: Usuwane przypadkami

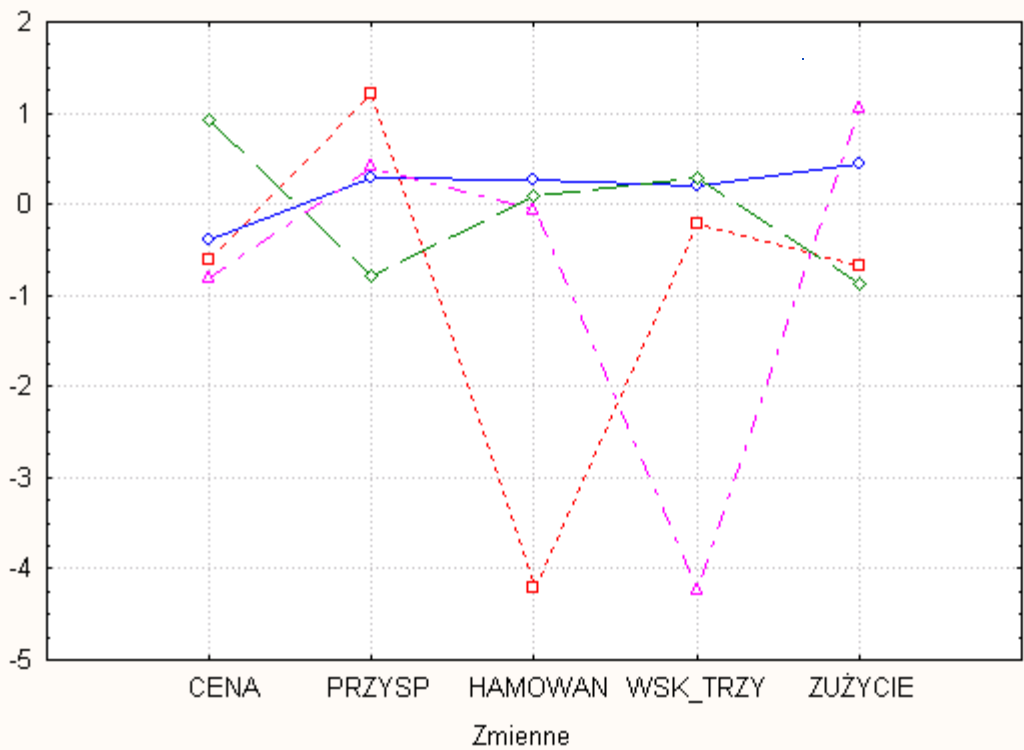
Wstępne centra skupień

- Wybierz obserwacje tak, aby zmaksymalizować odległości skupień
- Sortuj odległości i weź obserwacje przy stałym interwale
- Wybierz pierwszych N (liczba skupień) obserwacji

Przetwarzanie wsadowe i drukowanie

SELECT CASES    w

Wykres średnich każdego skupienia



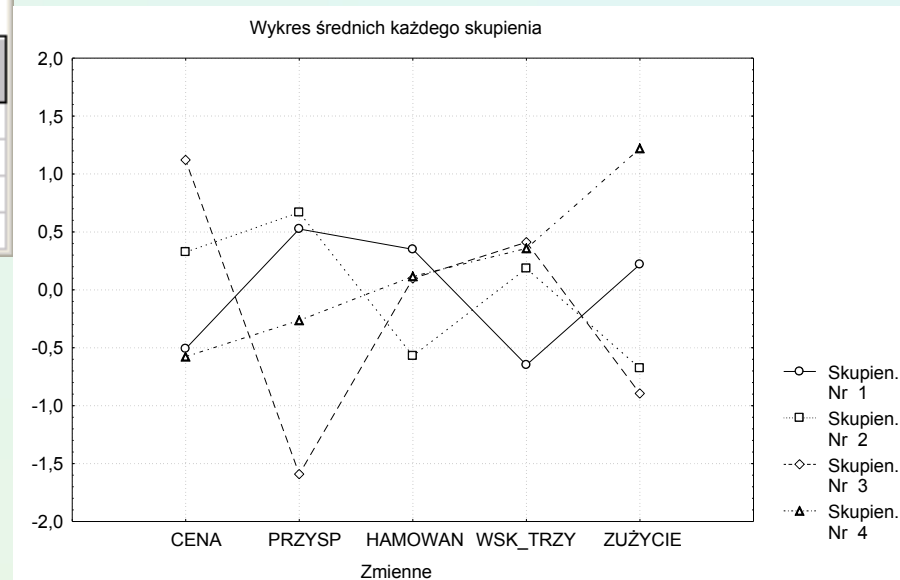
## Analiza Skupień – optymalizacja k-średnich

- ◇— Skupien. Nr 1
- - □ - - Skupien. Nr 2
- - ◇ - - Skupien. Nr 3
- - △ - - Skupien. Nr 4

# Wsparcie do charakterystyki skupień

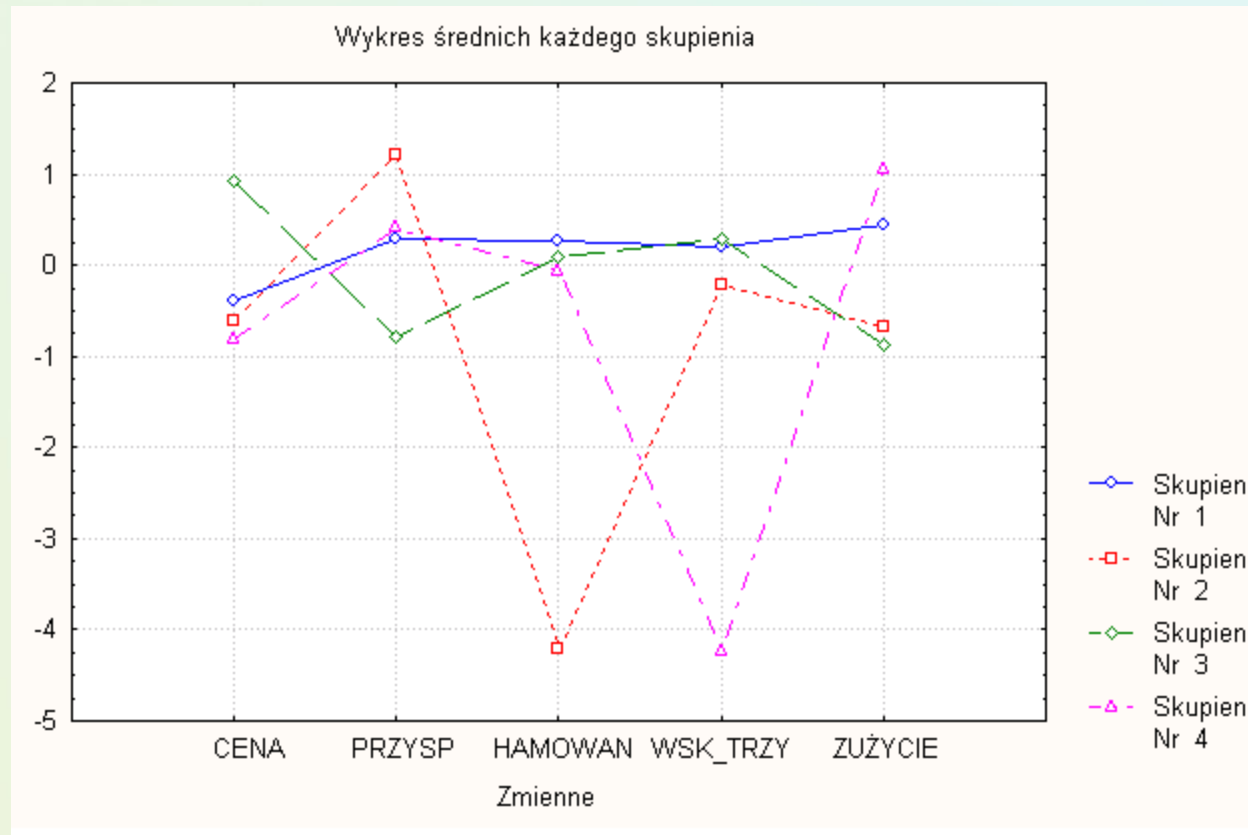
Odległości euklidesowe skupień (cars.sta)				
Dalej...	Odległości pod przekątną Kwadr. odległości nad przekątną			
Skupien. Numer	Nr 1	Nr 2	Nr 3	Nr 4
Nr 1	0,000000	,612157	1,912964	,539154
Nr 2	,782405	0,000000	1,256528	1,156810
Nr 3	1,383100	1,120950	0,000000	1,824839
Nr 4	,734271	1,075551	1,350866	0,000000

Średnie skup. (cars.sta)				
ANALIZA SKUPIEŃ	Skupien. Nr 1	Skupien. Nr 2	Skupien. Nr 3	Skupien. Nr 4
CENA		,326371	1,11989	-,576541
PRZYSP	,525328	,667950	-1,59237	-,263101
HAMOWAN	,349673	-,569764	,09733	,116307
WSK_TRZY	-,648829	,184539	,41118	,357969
ZUŻYCIE	,219949	-,677062	-,89508	1,220614



Elementy skupienia numer 2 (cars.sta)						
ANALIZA SKUPIEŃ	i odległości od środka właściwego skupienia W skupieniu jest 6 przypadków					
	Audi	Eagle	Mazda	Mercedes	Saab	Toyota
Odległ.	,523036	1,703612	,533962	,598004	,495550	,533369

# Wizualizacja centroidów



## **Analiza skupień w WEKA**

# WEKA zakładka Clustering

---

- Stopniowy przyrost implementacji
  - *k*-Means
  - EM
  - Cobweb
  - X-means
  - FarthestFirst...
  - DbScann
  - Oraz nowe
- Możliwości prostej wizualizacji i ew. porównania przydziałów do „wzorcowej klasyfikacji” – jeśli jest dostępna w pliku arff

# Exercise 1. K-means clustering in WEKA

---

- The exercise illustrates the use of the k-means algorithm.
- The example – sample of customers of the bank
  - Bank data (bank-data.csv -> bank.arff)
  - All preprocessing has been performed on cvs
  - 600 instances described by 11 attributes

```
id,age,sex,region,income,married,children,car,save_act,current_act,mortgage,pep
ID12101,48,FEMALE,INNER_CITY,17546.0,NO,1,NO,NO,NO,NO,YES
ID12102,40,MALE,TOWN,30085.1,YES,3,YES,NO,YES,YES,NO
ID12103,51,FEMALE,INNER_CITY,16575.4,YES,0,YES,YES,YES,NO,NO
ID12104,23,FEMALE,TOWN,20375.4,YES,3,NO,NO,YES,NO,NO
ID12105,57,FEMALE,RURAL,50576.3,YES,0,NO,YES,NO,NO,NO
.....
.....
```

- Cluster customers and characterize the resulting customer segments



# Loading the file and analysing the data

The screenshot shows the Weka Explorer application window. The title bar reads "Weka Explorer". The menu bar includes "Preprocess", "Classify", "Cluster", "Associate", "Select attributes", and "Visualize". Below the menu bar are buttons for "Open file...", "Open URL...", "Open DB...", "Undo", and "Save...".

The "Filter" section shows a "Choose" button and a dropdown menu set to "None", with an "Apply" button to the right.

The "Current relation" section displays "Relation: bank" and "Instances: 600". The "Attributes" section shows a list of 11 attributes:

No.	Name
1	age
2	sex
3	region
4	income
5	married
6	children
7	car
8	save_act
9	current_act
10	mortgage
11	pep

The "Selected attribute" section shows "Name: age", "Missing: 0 (0%)", "Distinct: 50", "Type: Numeric", and "Unique: 0 (0%)". Below this is a table of statistics:

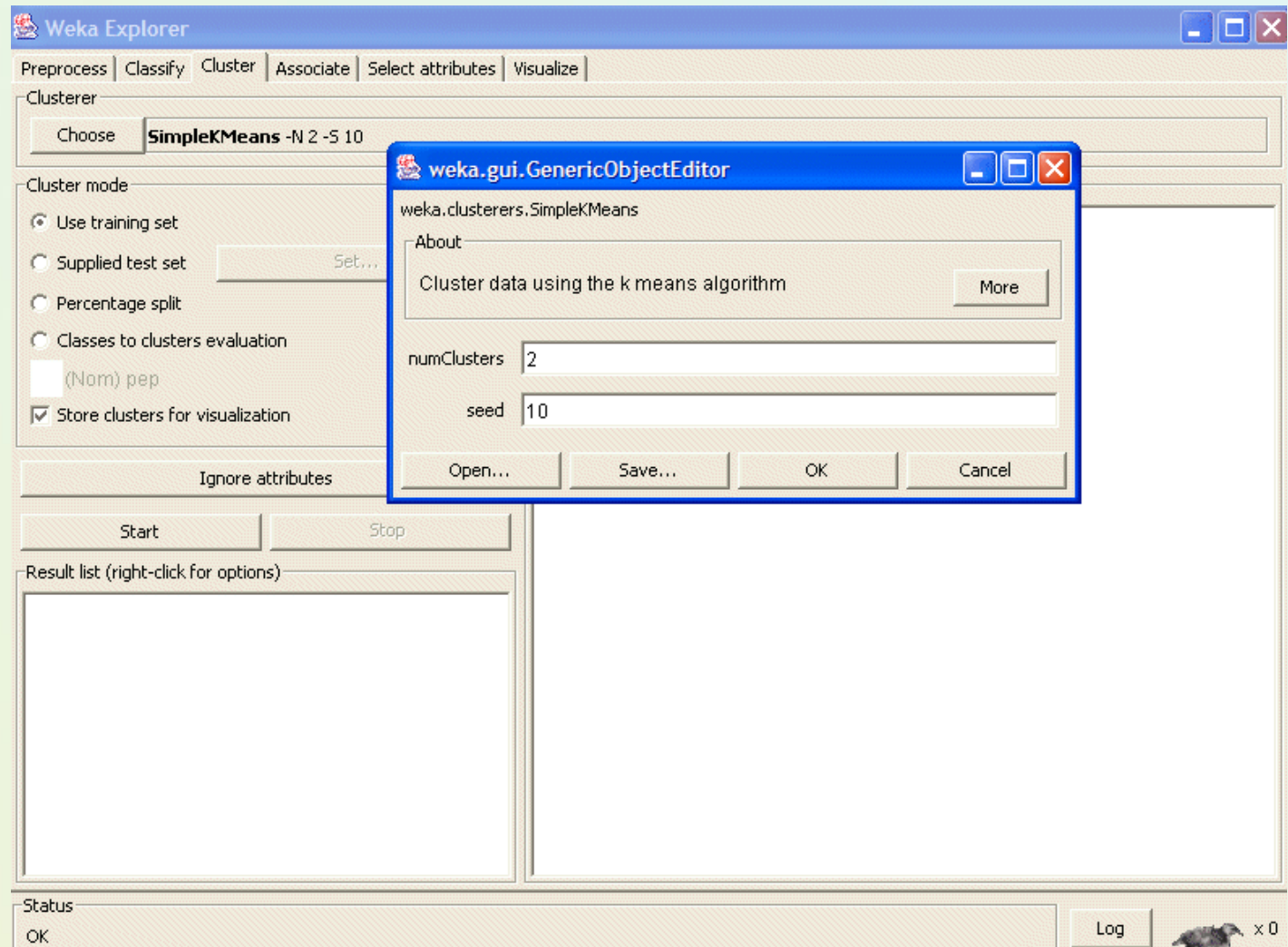
Statistic	Value
Minimum	18
Maximum	67
Mean	42.395
StdDev	14.425

The "Colour: pep (Nom)" dropdown is set to "pep (Nom)", and a "Visualize All" button is present. Below this is a histogram showing the distribution of the 'age' attribute. The x-axis is labeled with 18, 42.5, and 67. The histogram bars are colored blue and red, representing the distribution of the 'pep' variable across different age groups.

The "Status" section at the bottom left shows "OK". A "Log" button and a small animal icon with "x 0" are located at the bottom right.

# Choosing Simple k-means

- Tune proper parameters



# Clustering results

**Clusterer**  
Choose **SimpleKMeans -N 6 -S 10**

**Cluster mode**

- Use training set
- Supplied test set
- Percentage split %
- Classes to clusters evaluation
- Store clusters for visualization

**Clusterer output**

Cluster 2  
Mean/Mode: 44.0479 MALE INNER\_CITY 28547.224 YES  
Std Devs: 14.2211 N/A N/A 12696.446

Cluster 3  
Mean/Mode: 40.5068 MALE TOWN 25975.293 YES 0 YES  
Std Devs: 13.6353 N/A N/A 11111.66

Cluster 4  
Mean/Mode: 49.7843 FEMALE INNER\_CITY 33917.4538 NO  
Std Devs: 13.6872 N/A N/A 14195.168

Cluster 5  
Mean/Mode: 41.5234 FEMALE TOWN 26191.8366 YES 0 NO  
Std Devs: 13.5728 N/A N/A 11737.313

**Clustered Instances**

0	66	( 11%)
1	85	( 14%)
2	146	( 24%)
3	73	( 12%)
4	102	( 17%)
5	128	( 21%)

**Result list (right-click for options)**

- View in main window
- View in separate window**
- Save result buffer
- Load model
- Save model
- Re-evaluate model on current test set
- Visualize cluster assignments
- Visualize tree

Status: OK  x 0

- Analyse the result window

# Characterizing cluster

- How to describe clusters?
- What about descriptive statistics for centroids?

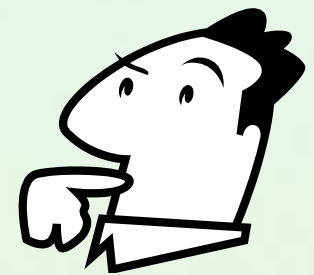
```
16:47:12 - SimpleKMeans
kMeans
=====
Number of iterations: 9
Cluster centroids:
Cluster 0
  Mean/Mode: 36.6061 FEMALE RURAL 23215.9002 NO 3 NO YES YES NO NO
  Std Devs: 14.4317 N/A N/A 12378.3336 N/A N/A N/A N/A N/A
Cluster 1
  Mean/Mode: 38.1176 FEMALE INNER_CITY 24775.7982 YES 1 NO YES YES YES YES
  Std Devs: 13.793 N/A N/A 12444.5713 N/A N/A N/A N/A N/A
Cluster 2
  Mean/Mode: 44.0479 MALE INNER_CITY 28547.224 YES 0 YES YES YES NO NO
  Std Devs: 14.2211 N/A N/A 12696.4468 N/A N/A N/A N/A N/A
Cluster 3
  Mean/Mode: 40.5068 MALE TOWN 25975.293 YES 0 YES NO YES YES YES
  Std Devs: 13.6353 N/A N/A 11111.66 N/A N/A N/A N/A N/A
Cluster 4
  Mean/Mode: 49.7843 FEMALE INNER_CITY 33917.4538 NO 0 YES YES YES NO YES
  Std Devs: 13.6872 N/A N/A 14195.1688 N/A N/A N/A N/A N/A
Cluster 5
  Mean/Mode: 41.5234 FEMALE TOWN 26191.8366 YES 0 NO YES YES NO NO
  Std Devs: 13.5728 N/A N/A 11737.3135 N/A N/A N/A N/A N/A
Clustered Instances
0 66 ( 11%)
1 85 ( 14%)
2 146 ( 24%)
3 73 ( 12%)
4 102 ( 17%)
5 128 ( 21%)
```

# Finally, cluster assignments

```
TextPad - [D:\Bamshad\CLASS\ECT584\WEKA\Cluster\bank-kmeans.arff]
File Edit Search View Tools Macros Configure Window Help
[Icons]
1 @relation bank_clustered
2
3 @attribute Instance_number numeric
4 @attribute age numeric
5 @attribute sex {FEMALE,MALE}
6 @attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
7 @attribute income numeric
8 @attribute married {NO,YES}
9 @attribute children {0,1,2,3}
10 @attribute car {NO,YES}
11 @attribute save_act {NO,YES}
12 @attribute current_act {NO,YES}
13 @attribute mortgage {NO,YES}
14 @attribute pep {YES,NO}
15 @attribute Cluster {cluster0,cluster1,cluster2,cluster3,cluster4,cluster5}
16
17 @data
18 0,48,FEMALE,INNER_CITY,17546,NO,1,NO,NO,NO,NO,YES,cluster1
19 1,40,MALE,TOWN,30085.1,YES,3,YES,NO,YES,YES,NO,cluster3
20 2,51,FEMALE,INNER_CITY,16575.4,YES,0,YES,YES,YES,NO,NO,cluster2
21 3,23,FEMALE,TOWN,20375.4,YES,3,NO,NO,YES,NO,NO,cluster5
22 4,57,FEMALE,RURAL,50576.3,YES,0,NO,YES,NO,NO,NO,cluster5
23 5,57,FEMALE,TOWN,37869.6,YES,2,NO,YES,YES,NO,YES,cluster5
24 6,22,MALE,RURAL,8877.07,NO,0,NO,NO,YES,NO,YES,cluster0
25 7,58,MALE,TOWN,24946.6,YES,0,YES,YES,YES,NO,NO,cluster2
26 8,37,FEMALE,SUBURBAN,25304.3,YES,2,YES,NO,NO,NO,NO,cluster5
27 9,54,MALE,TOWN,24212.1,YES,2,YES,YES,YES,NO,NO,cluster2
28 10,66,FEMALE,TOWN,59803.9,YES,0,NO,YES,YES,NO,NO,cluster5
29 11,52,FEMALE,INNER_CITY,26658.8,NO,0,YES,YES,YES,YES,NO,cluster4
30 12,44,FEMALE,TOWN,15735.8,YES,1,NO,YES,YES,YES,YES,cluster1
31 13,66,FEMALE,TOWN,55204.7,YES,1,YES,YES,YES,YES,YES,cluster1
32 14,36,MALE,RURAL,19474.6,YES,0,NO,YES,YES,YES,NO,cluster5
33 15,38,FEMALE,INNER_CITY,22342.1,YES,0,YES,YES,YES,YES,NO,cluster2
34 16,37,FEMALE,TOWN,17729.8,YES,2,NO,NO,NO,YES,NO,cluster5
35 17,46,FEMALE,SUBURBAN,41016,YES,0,NO,YES,NO,YES,NO,cluster5
36 18,62,FEMALE,INNER_CITY,26909.2,YES,0,NO,YES,NO,NO,YES,cluster4
37 19,31,MALE,TOWN,22522.8,YES,0,YES,YES,YES,NO,NO,cluster2
38 20,61,MALE,INNER_CITY,57880.7,YES,2,NO,YES,NO,NO,YES,cluster2
39 21,50,MALE,TOWN,16497.3,YES,2,NO,YES,YES,NO,NO,cluster5
```

---

# Ocena jakości skupień



# Dwa różne spojrzenia

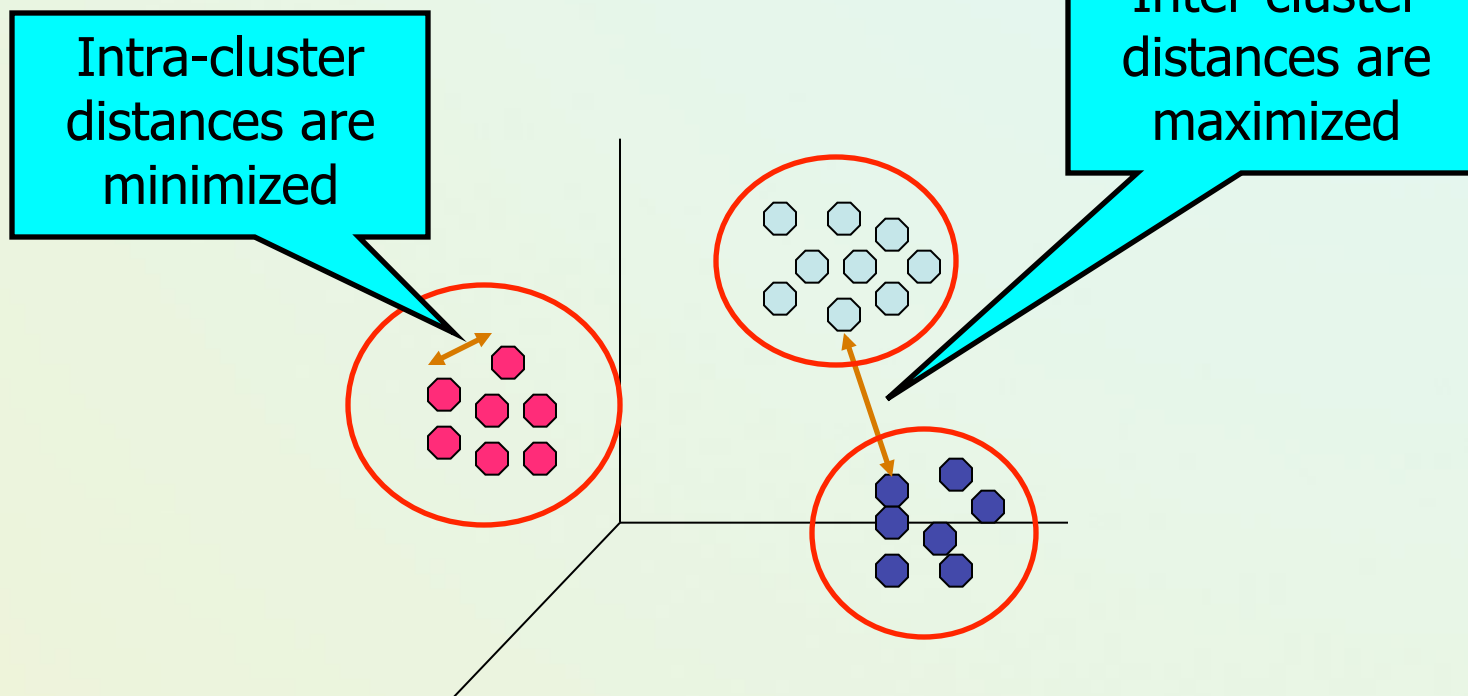
---

- Wewnętrzne (ocena tylko charakterystyki skupień)
  - Brak dodatkowych źródeł informacji, np. zbioru odniesienia etykiet
  - Miary oceny oparte na danych (internal measures)
- Zewnętrzne
  - „Benchmarking on existing labels”
  - Porównanie skupień z tzw. ground-truth categories / partitions
- Ocena ekspercka

# Ocena jakości skupień

## Miary oceny oparte na danych (internal measures)

- Oparte na odległościach lub ...
- Duże podobieństwo obiektów wewnątrz skupienia (*Compactness*)
- Same skupienia dość odległe (*Isolation*)





# Typowe miary zmienności skupień

---

- Intuicja → „zmienność wewnątrz-skupieniowa”  $wc(C)$  i „zmienność między-skupieniowa”  $bc(C)$

- Można definiować różnymi sposobami

- Wykorzystaj średni obiekt w skupieniu  $\mathbf{r}_k$  (centroids)

- Wtedy, np.  $wc(C) = \sum_{k=1}^K \sum_{\mathbf{x} \in C_k} d(\mathbf{x}, \mathbf{r}_k)$        $\mathbf{r}_k = \frac{1}{n_k} \sum_{\mathbf{x} \in C_k} \mathbf{x}$

$$bc(C) = \sum_{1 \leq j < k \leq K} d(\mathbf{r}_j, \mathbf{r}_k)$$

- Zamiast  $bc$  odległość od globalnego centrum danych (inter-class distance)

$$id = \sum_{C_j} d(\mathbf{r}_j, \mathbf{r}_{glob})$$

- Preferencja dla zwartych, jednorodnych skupień dość odległych od centrum danych

# Może pytanie lub komentarze?

---

