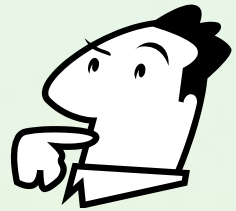
Analiza Skupień - Grupowanie

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

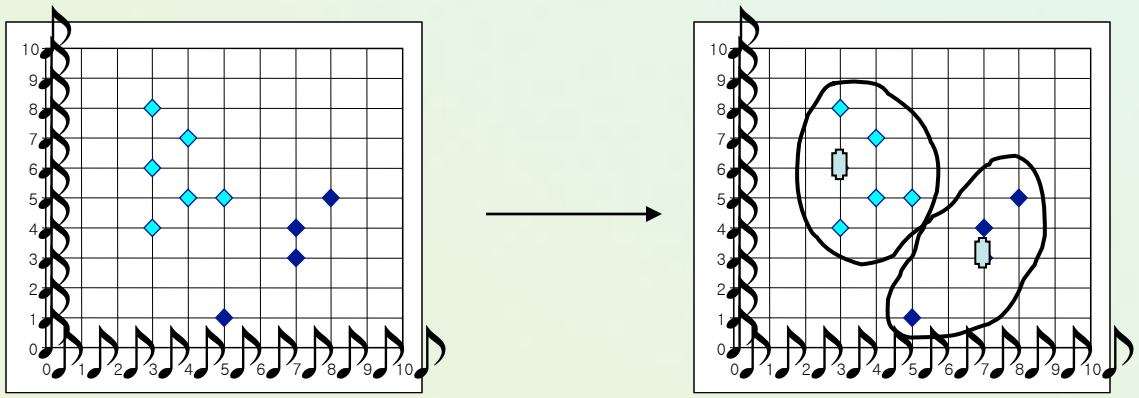
Przykłady zastosowań analizy skupień

- 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

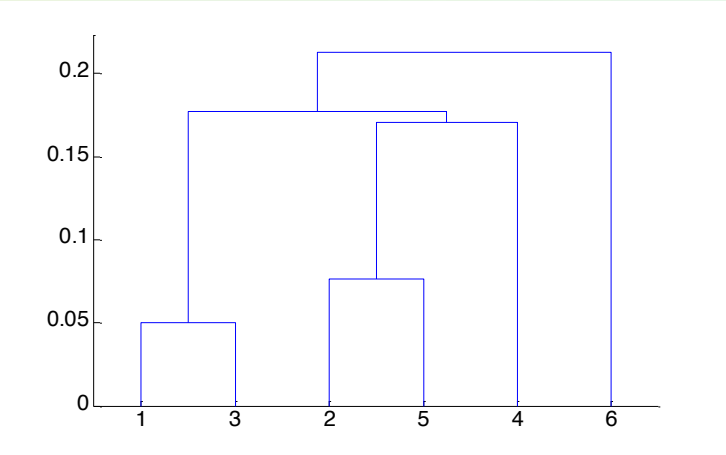
- Obiekt opisany za pomocą *n* zmiennych X_1, X_2,…X_n jest punktem x=(x_1, …,x_n) w *n*-wymiarowej przestrzeni Ω
- Cel podziału na grupy (S) → obiekty podobne (reprezentowane przez punkty znajdujące się blisko siebie w przestrzeni) przydzielone do tej samej grupy, a obiekty niepodobne (reprezentowane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

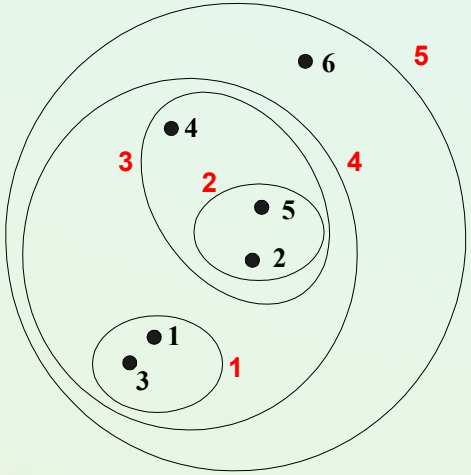
$$S_i \cap S_j = \varnothing \quad (i \neq j;\ \ i,j = 1,\ldots,p) \qquad \bigcup_{i=1}^{p} S_i = \Omega$$



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

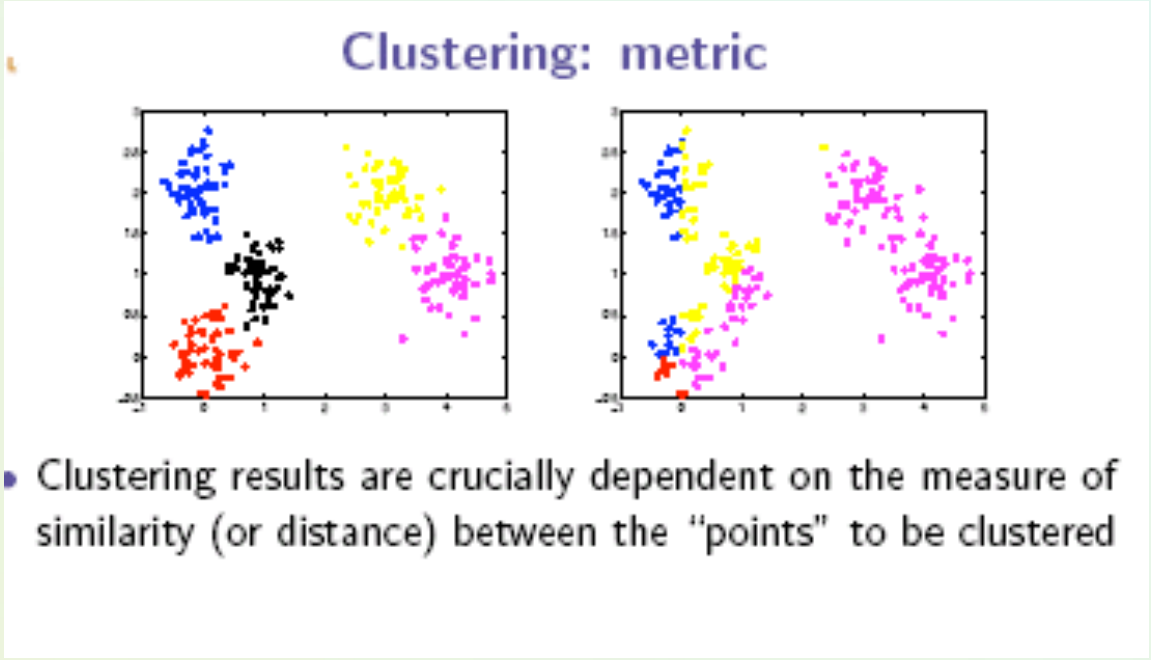
- Jak odwzorować obiekty w przestrzeni?
  - Wybór zmiennych
  - Normalizacja zmiennych
- Jak mierzyć odległości między obiektami?
- Jaką metodę grupowania zastosować?

Najczęściej stosowane miary odległości

Nazwa miary odległości	Definicja miary	Uwagi
Odległość euklidesowa	$d_{(rs)} = \left[\sum_{k=1}^{p} \left(x_{(r)k} - x_{(s)k} \right)^2 \right]^{\frac{1}{2}}$	Jest to odległość geometryczna w przestrzeni wielowymiarowej.
Kwadrat odległość euklidesowej	$d_{(rs)} = \left[\sum_{k=1}^{p} \left(x_{(r)k} - x_{(s)k} \right)^2 \right]$	Odległość euklidesową podnosi się do kwadratu, aby przypisać większą wagę obiektom, które są bardziej oddalone
Odległość miejska (Manhattan, City block)	$d_{(rs)} = \sum_{k=1}^{p} \left\| x_{(r)k} - x_{(s)k} \right\|$	Jest to przeciętna różnica mierzoną 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

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

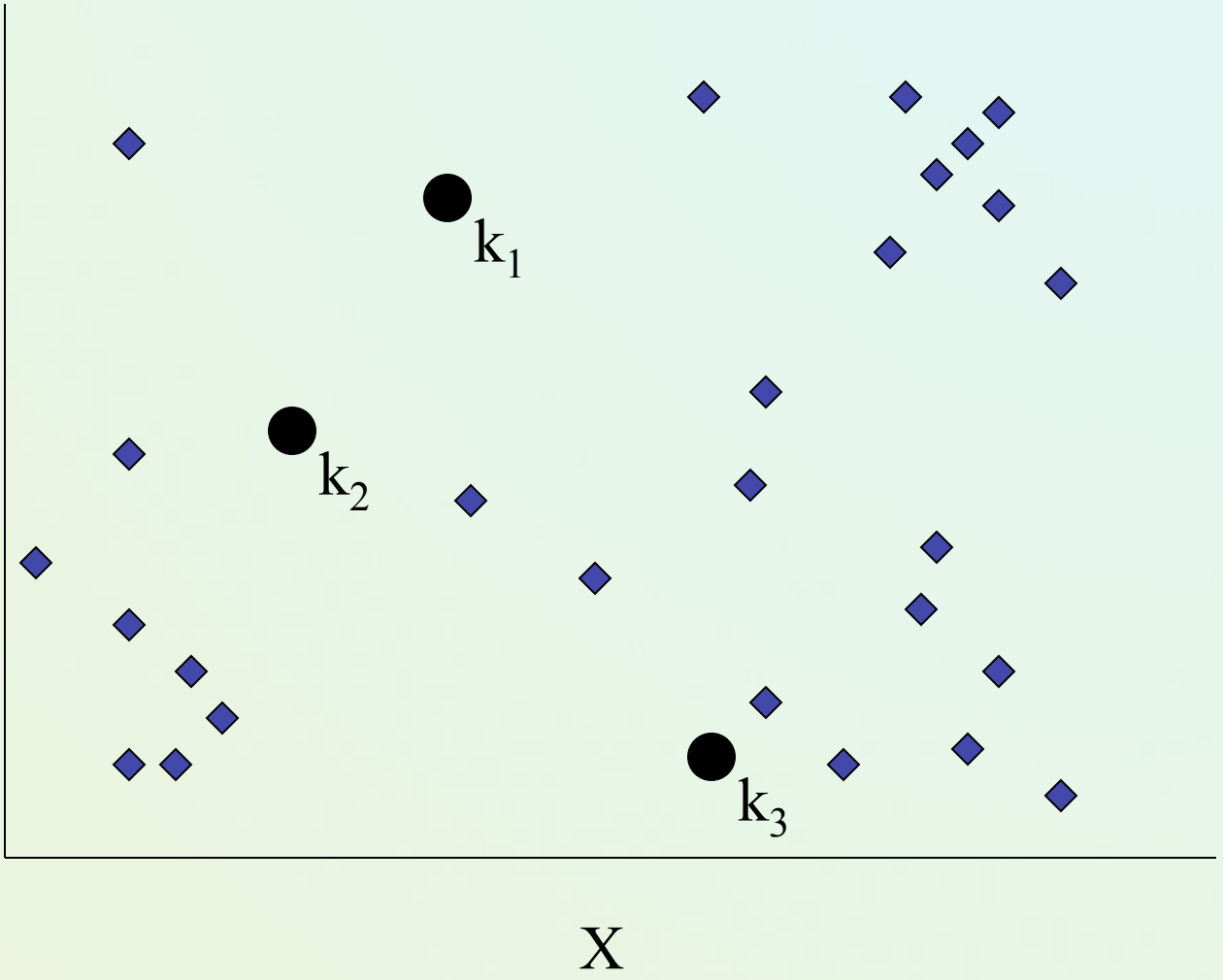
$$bc(C) = \sum_{1 \le j < k \le K} d(\mathbf{r}_j, \mathbf{r}_k)^2$$

Podstawowe algorytmy podziałowe

- Metoda *K* - średnich → minimalizacja *wc*(*C*)
- Przeszukiwanie przestrzeni możliwych przypisań → bardzo kosztowne (oszacowanie w ks. Koronackiego)
- Problem optymalizacji kombinatorycznej → 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 → proste i prowadzi do co najmniej lokalnego minimum. Różne modyfikacje, np. rozpoczynania od kilku rozwiązań startowych
- Złożoność algorytmy K - średnich → *O*(*Knl*)

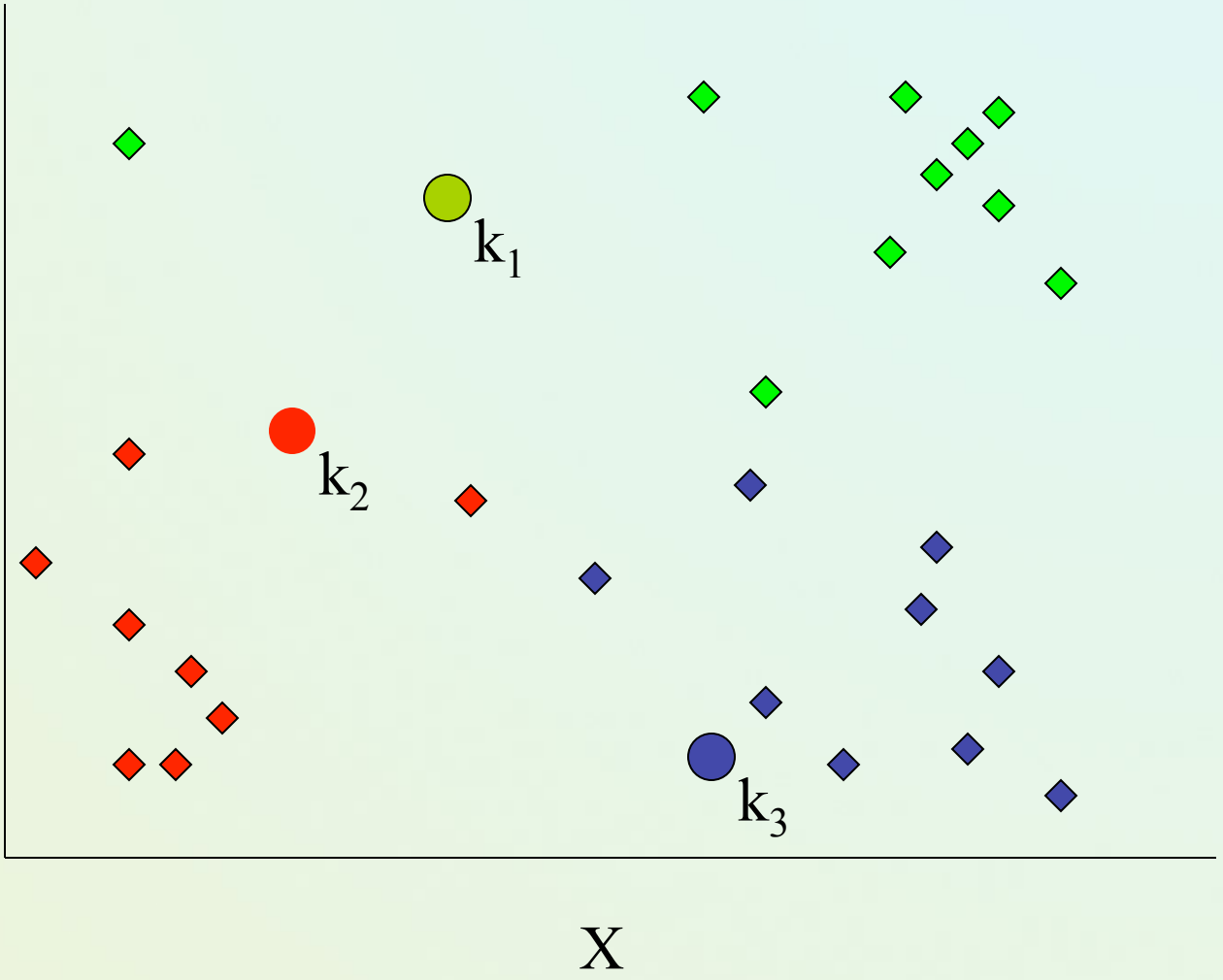
K-means example, step 1

Pick 3 initial cluster centers (randomly)



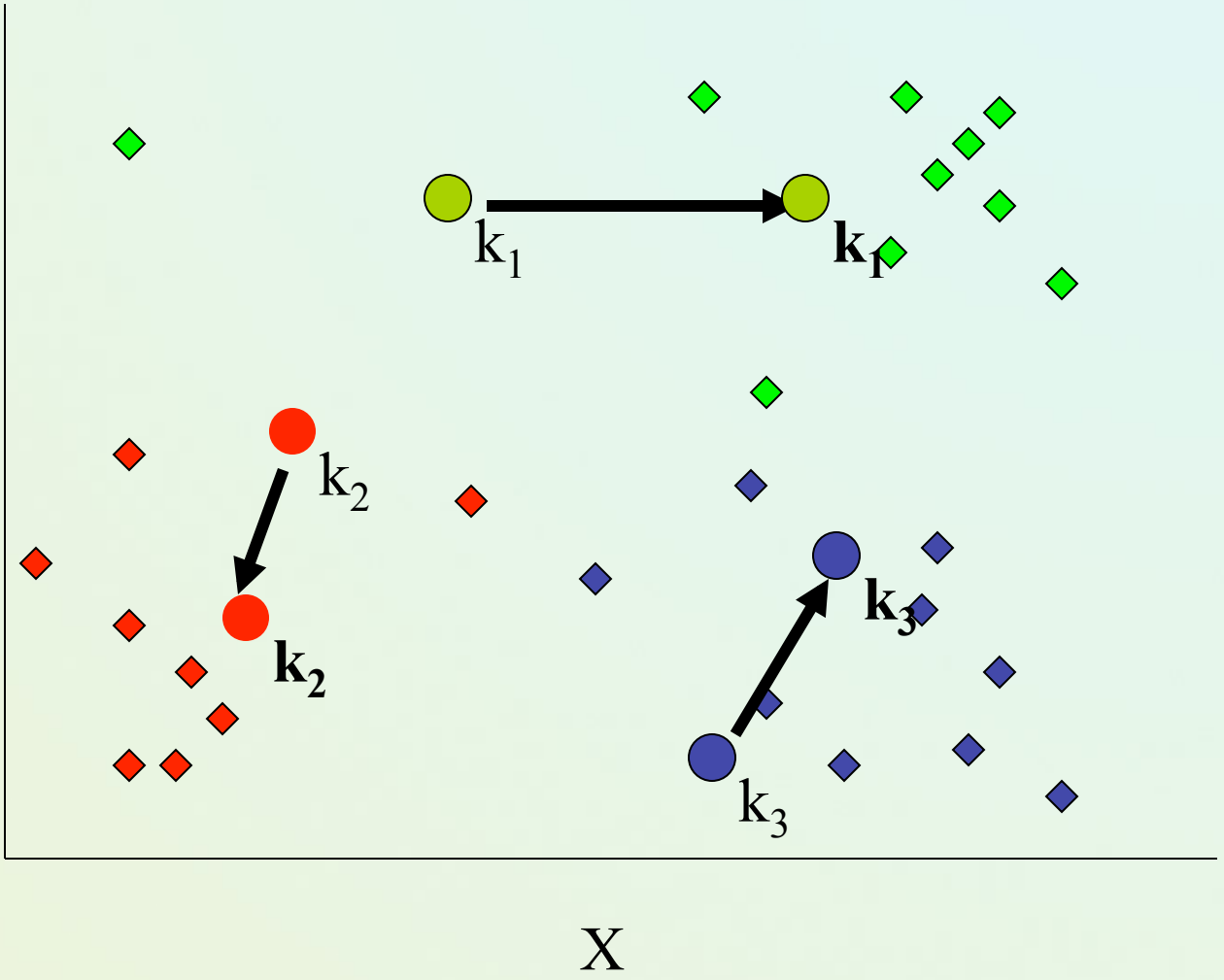
K-means example, step 2

Assign each point to the closest cluster center



K-means example, step 3

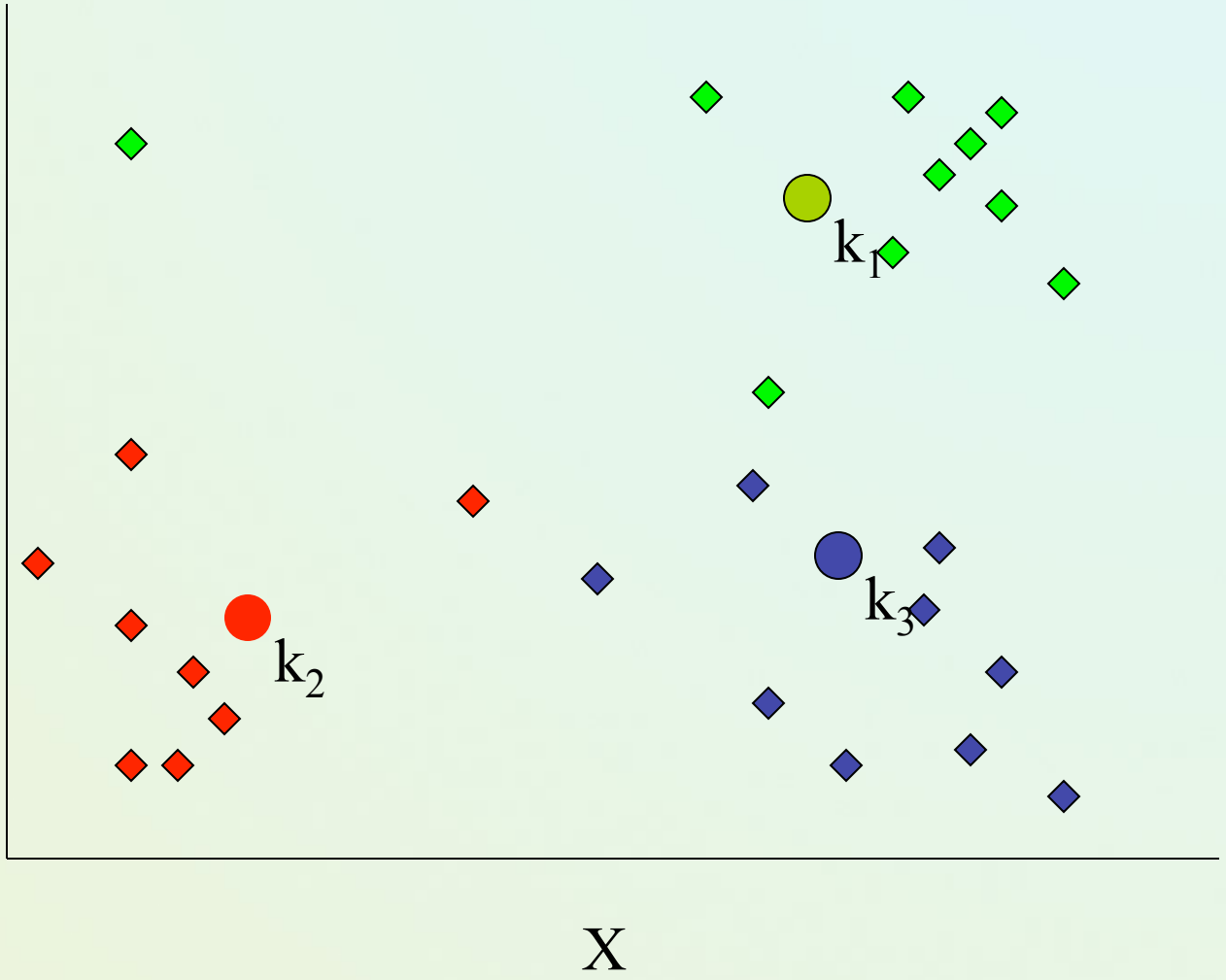
Move
each cluster
center
to the mean
of each cluster



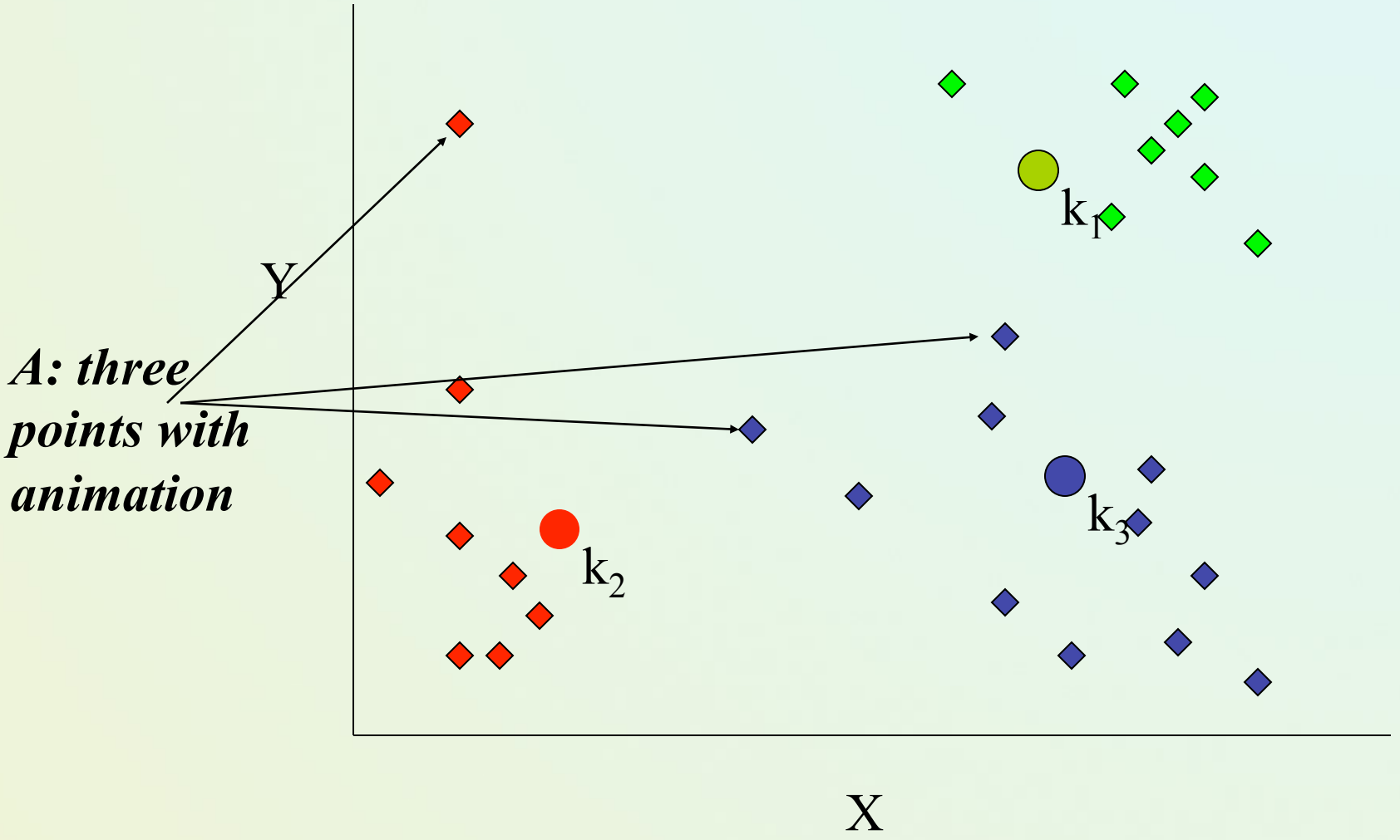
K-means example, step 4

Reassign points closest to a different new cluster center

Q: Which points are reassigned?



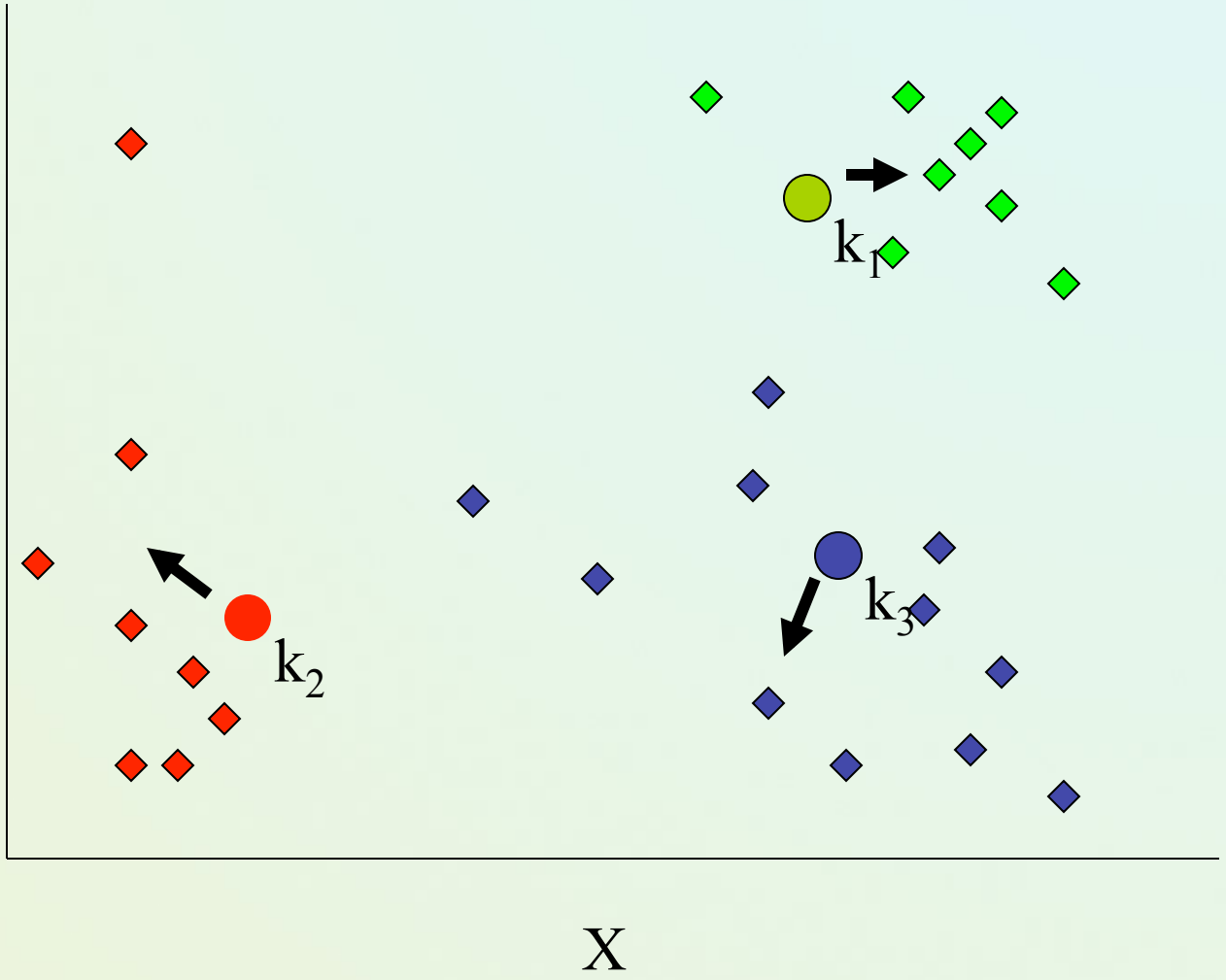
K-means example, step 4 …



A: three points with animation

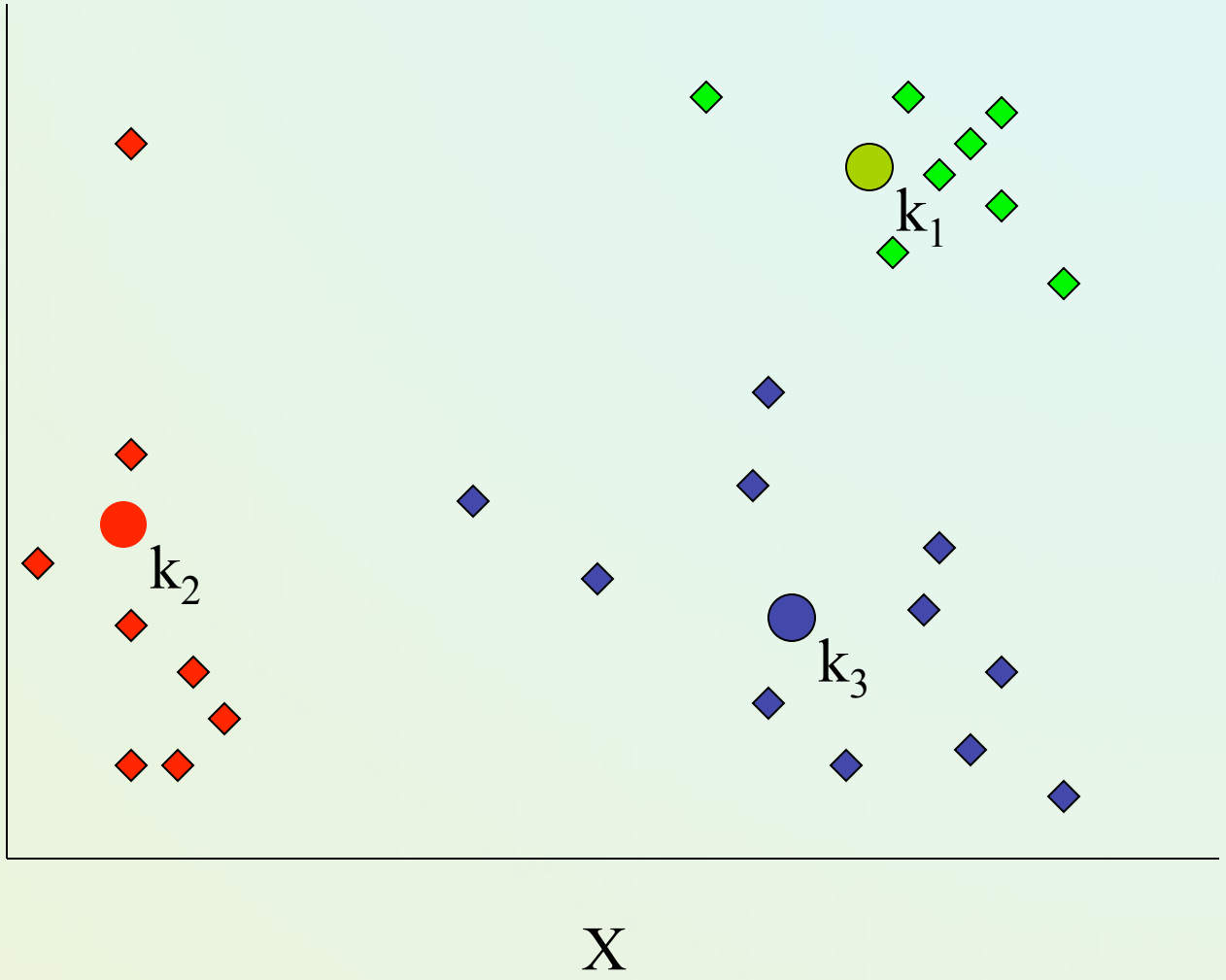
K-means example, step 4b

re-compute cluster means



K-means example, step 5

move cluster centers to cluster means



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:

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

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.

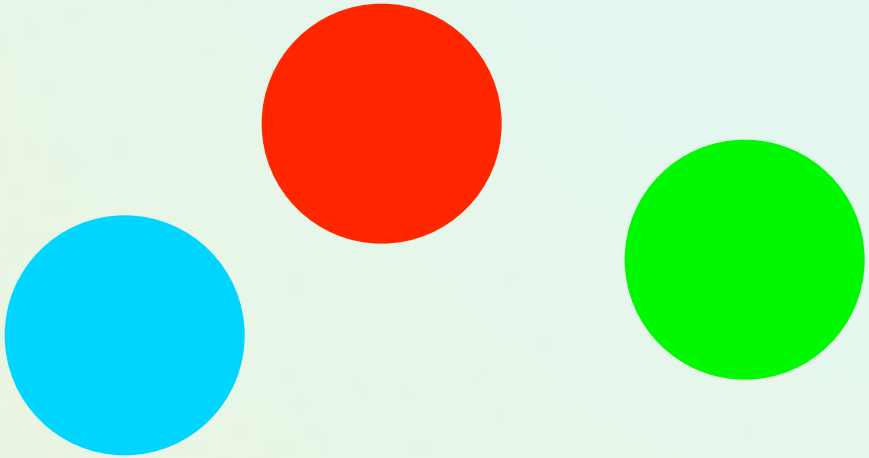
$$F = \sum_{j=1}^{k} \sum_{O_i \in S_j} d\left(O_i, M_j\right)^2$$

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*).

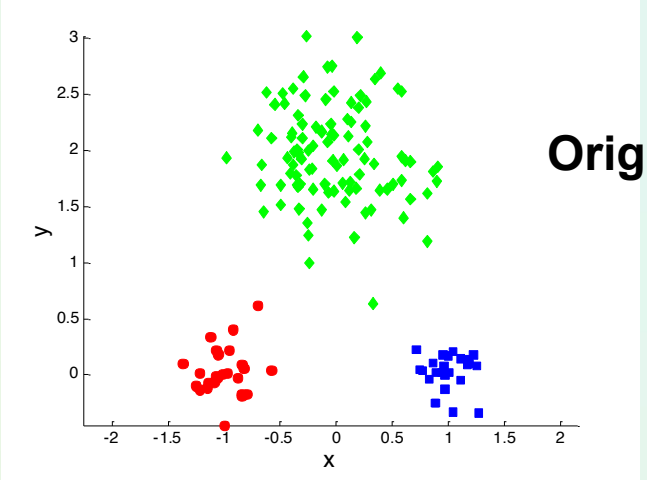
Pewne ukierunkowanie K-średnich

- Tworzy się kuliste kształty skupień

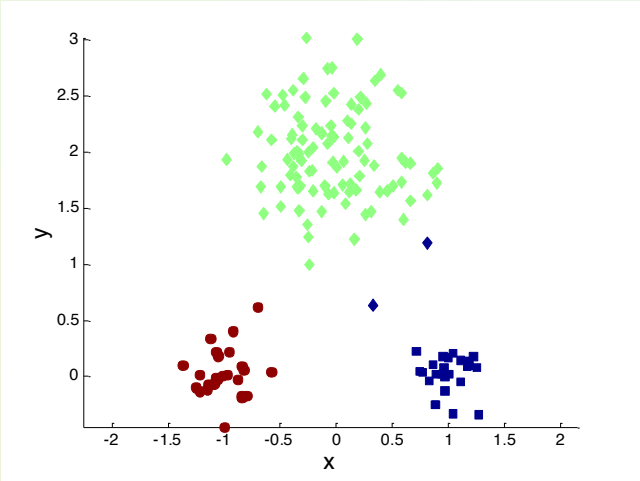


- Co z obserwacjami odstającymi i nieregularnymi kształtami skupień?

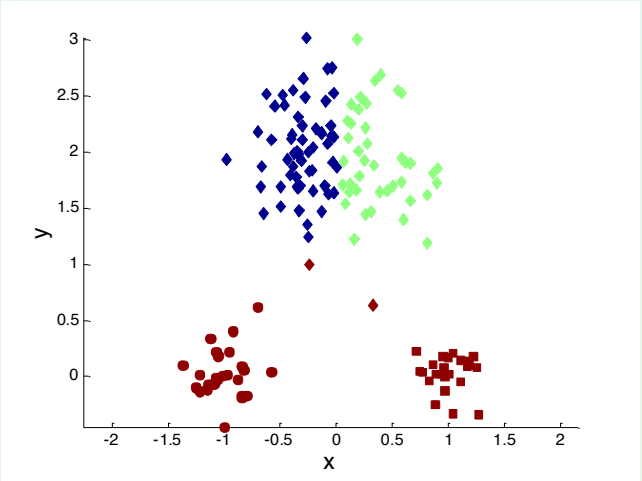
Czy łatwo określić liczbę skupień



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:

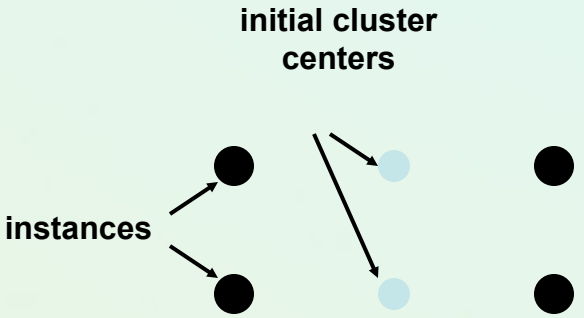
$$k_{\min} \leq k \leq k_{\max}$$

Możliwe są różne podejścia:

1. Arbitralny sposób np. przyjmuje się współrzędne pierwszych *k* obiektów jako zalążki 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

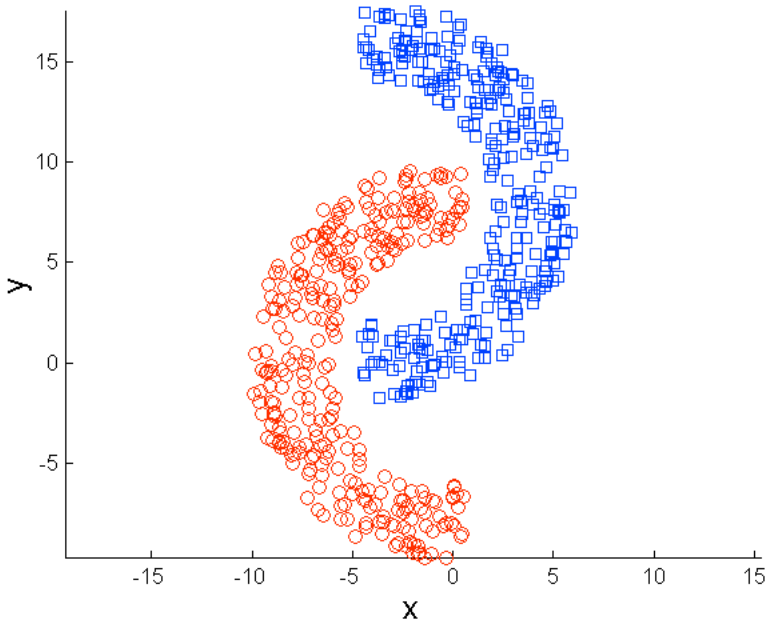
Uwagi nt. podziałowo-optymalizacyjnych

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

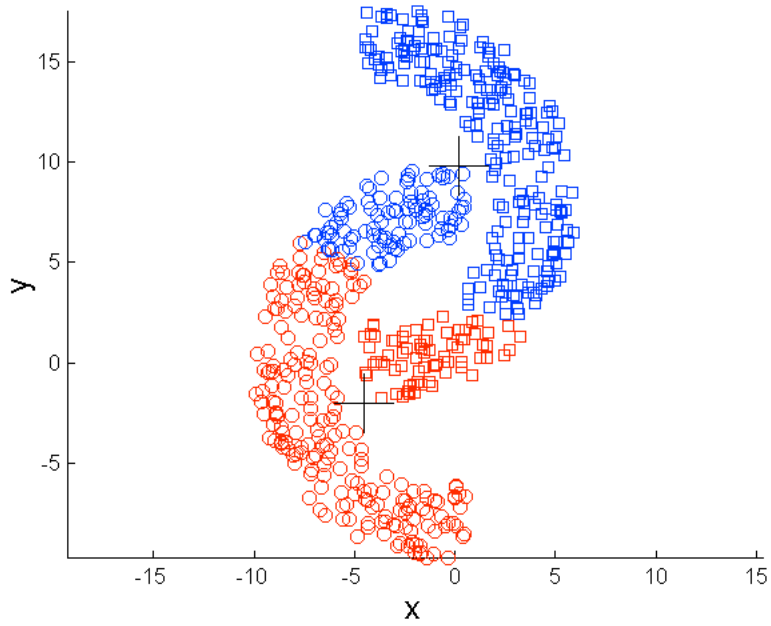


- Możesz próbować kilka uruchomień z różnymi parametrami

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

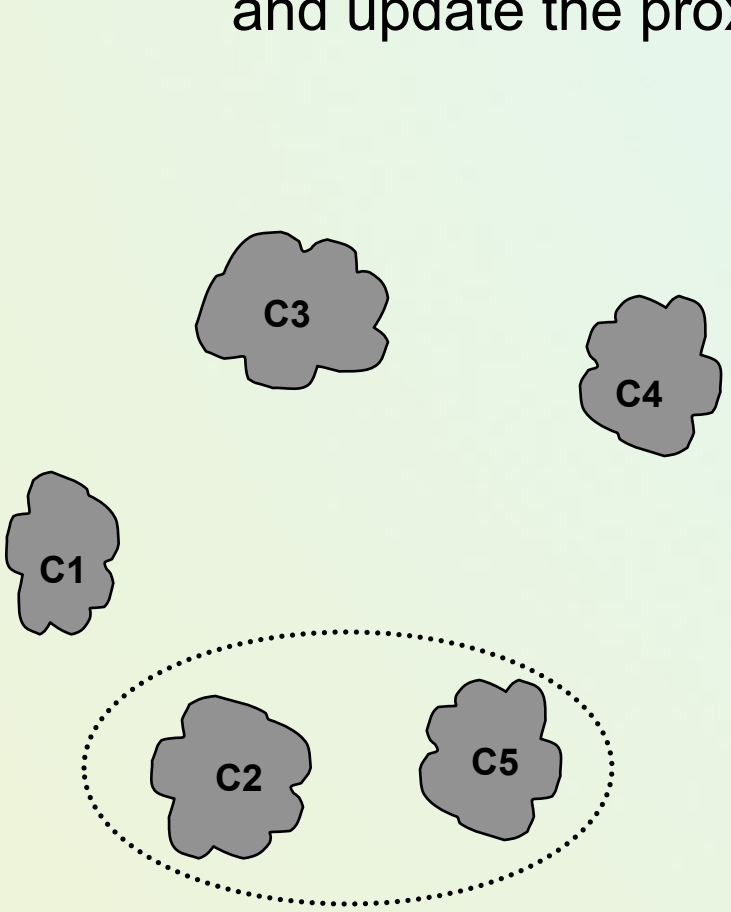
- Jawne podanie liczby skupień
- Wszystkie przykłady muszą być przydzielone do skupień
- Problem z outliers (za duża wrażliwość)
- Ukierunkowanie na jednorodne „sferyczne” kształty skupień

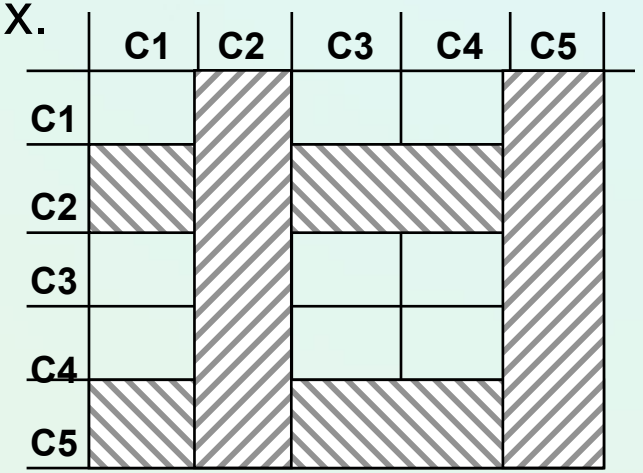
Hierarchiczne metody aglomeracyjne - algorytm

1. W macierzy odległości znajduje się parę skupień najbliższych sobie.
2. Redukuje się liczbę klas łącząc znalezioną 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.

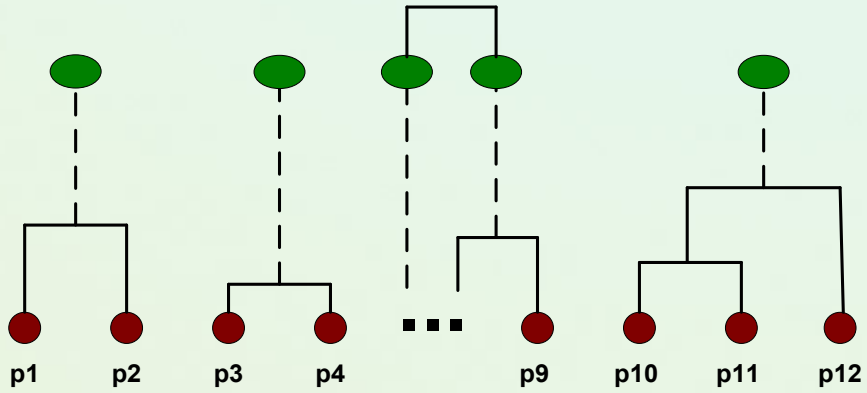
Jak przeliczać macierz odległości?

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





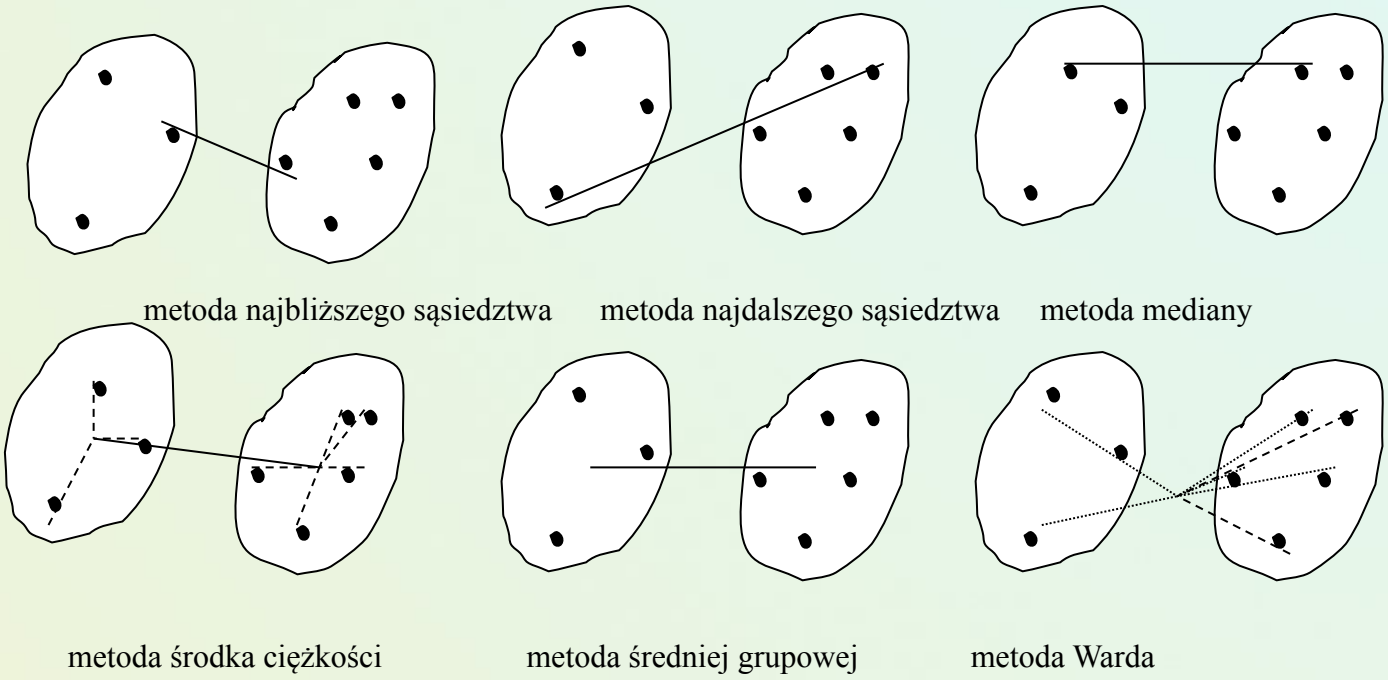
Proximity Matrix



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



Odległości między skupieniami

Single linkage
minimum distance:

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

Complete linkage
maximum distance:

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

mean distance:

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

average distance:

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

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:

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

- Prowadzi do „(long and thin) clusters due to *chaining effect*” (efekt łańcuchowy); prowadzić 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

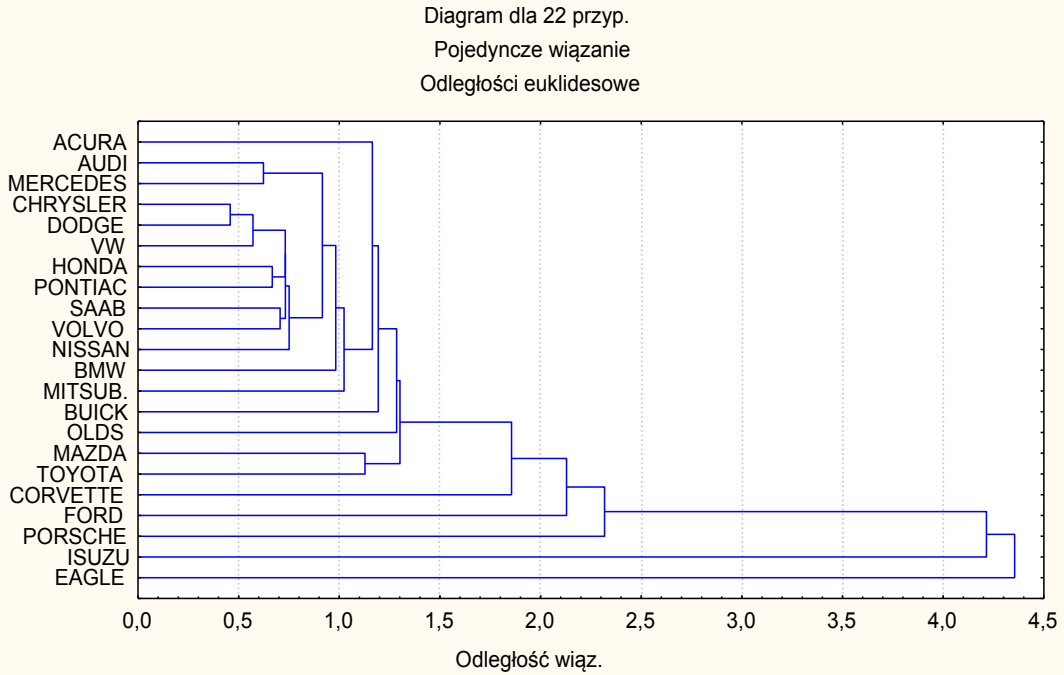
Complete Link Agglomerative Clustering

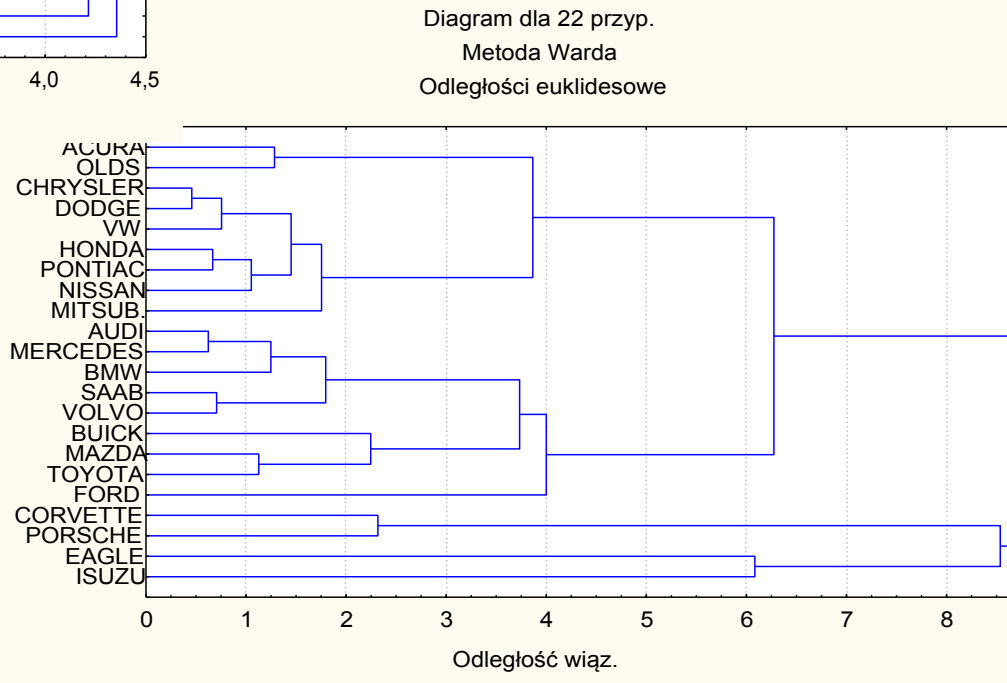
- Użyj maksymalnej odległości – minimalnego podobieństwa

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

- 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





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"

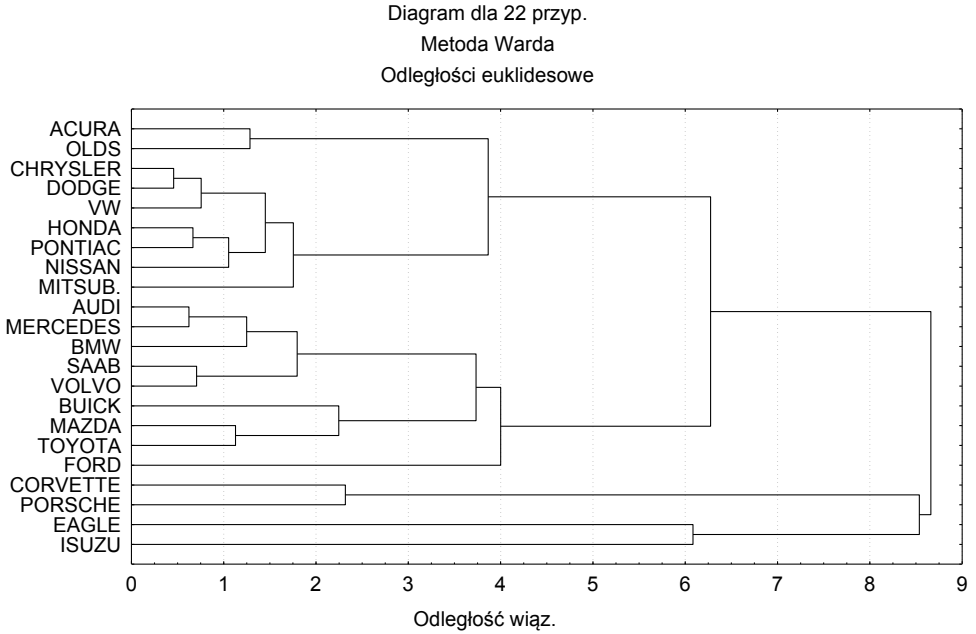
Metoda ważonych środków ciężkości (mediany) [Weighted pair-group centroid].

- 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

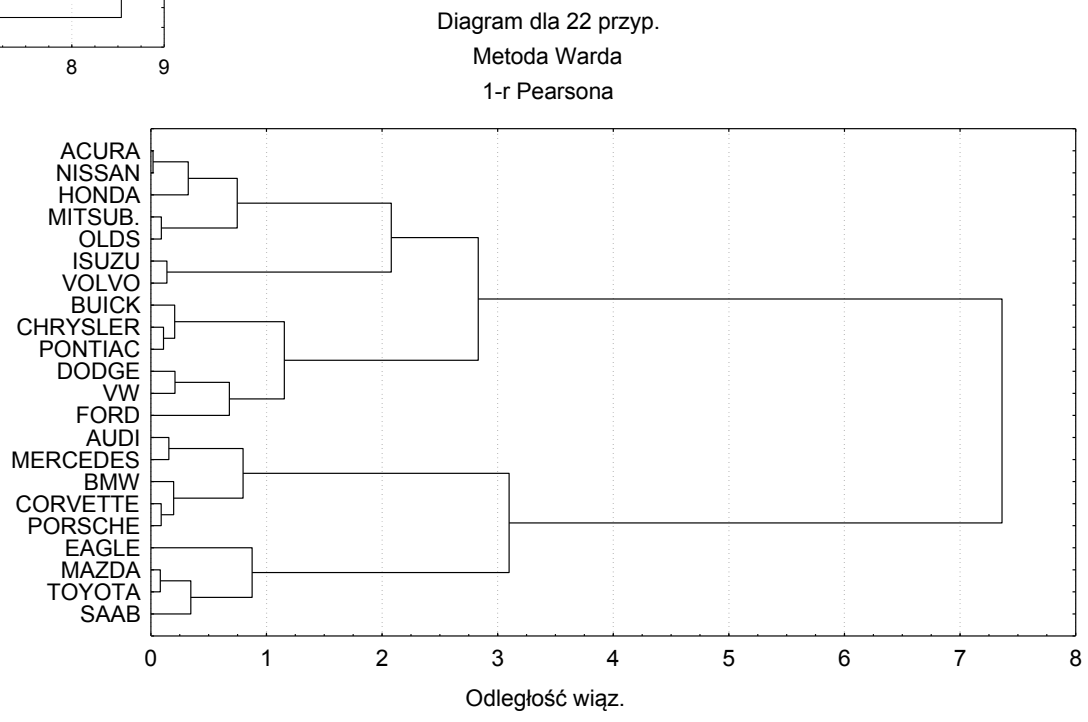
Metody łączenia – Ward method

- 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 **x**j tworzących zbiór *C*k (*k* = 1,....., *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 elemenatach
- Ważne – powiązanie z miarą odległości między obiektami (Pearson vs. inne)

Przykłady użycia metody Warda

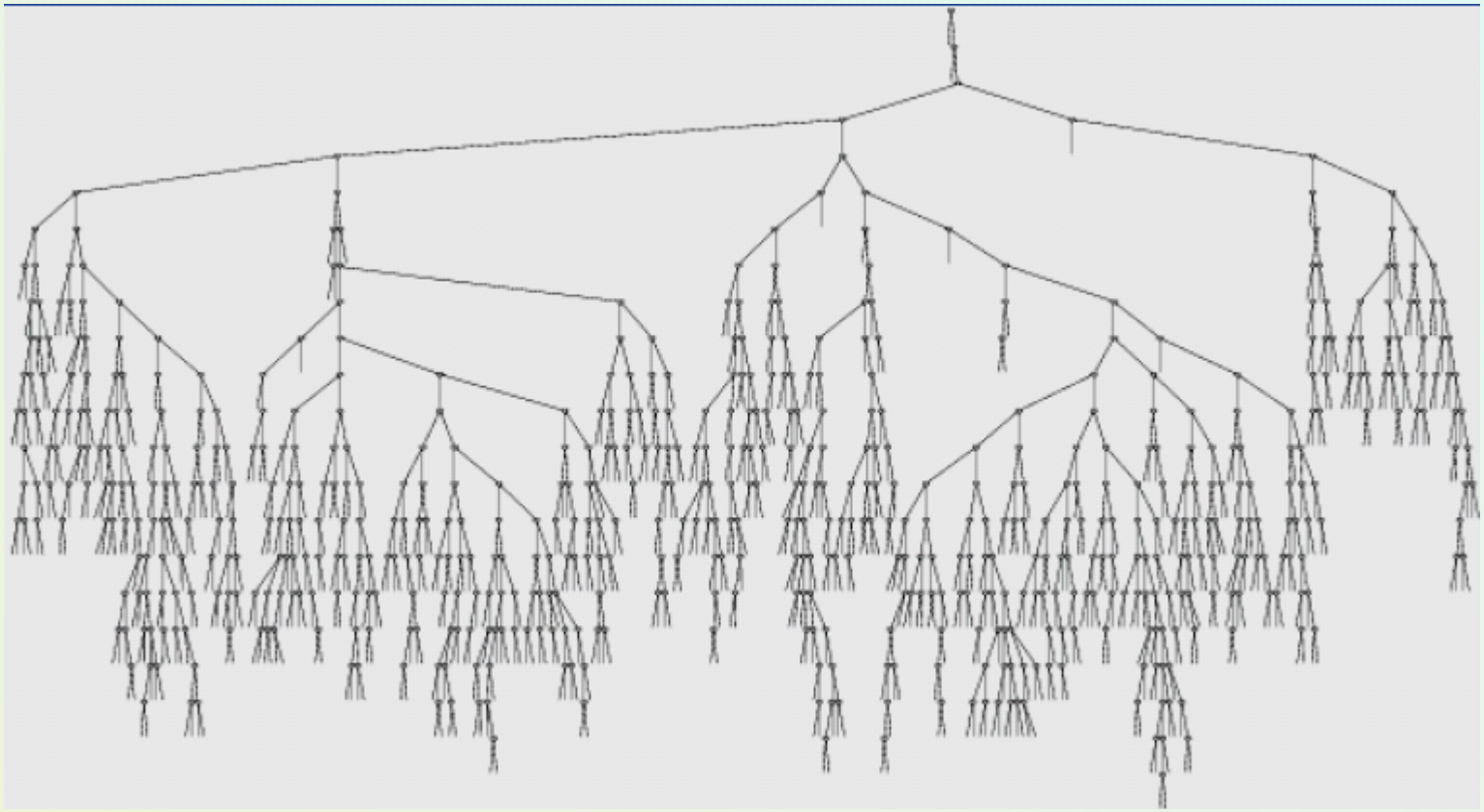


Cars data

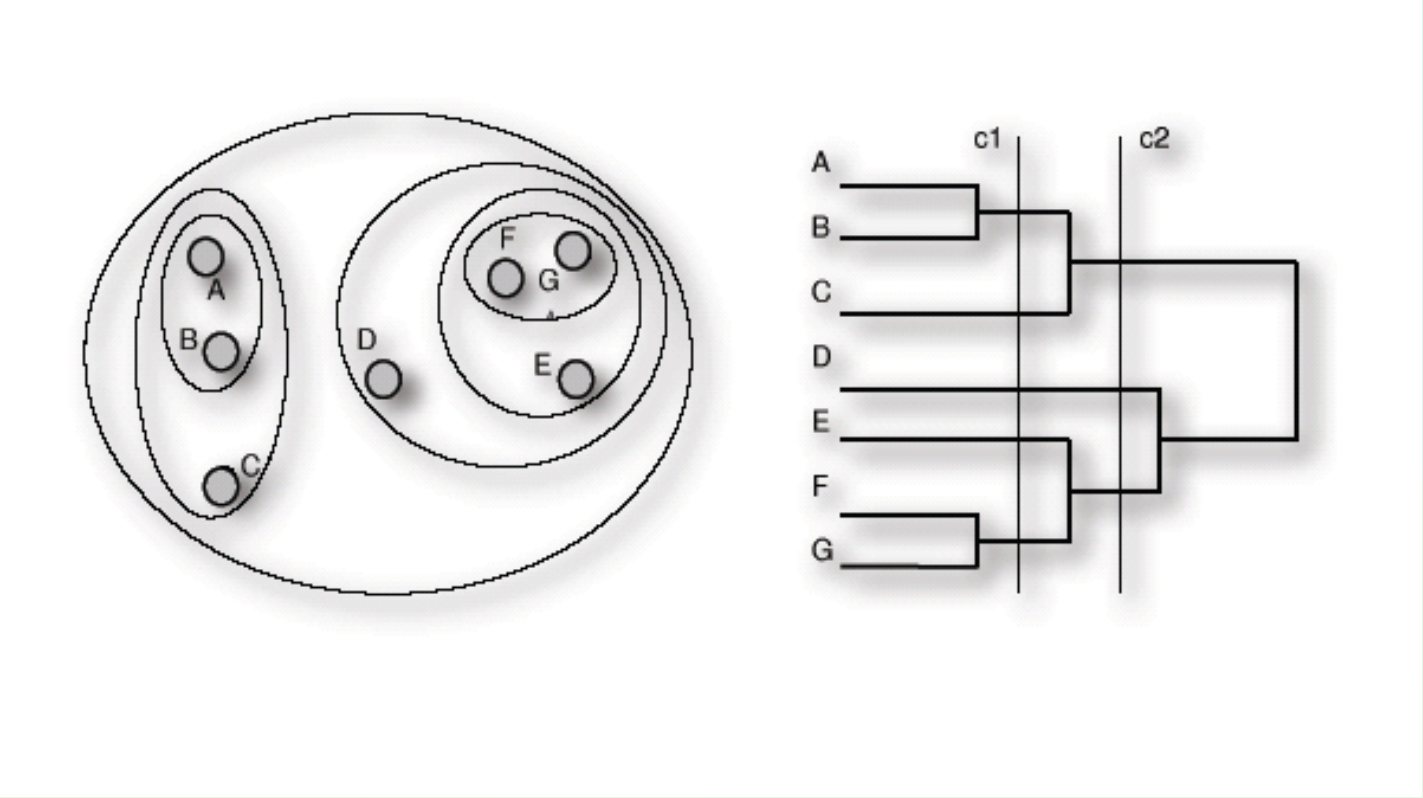


Czy struktura drzewa skupień jest czytelna?

- Przykład biochemiczny

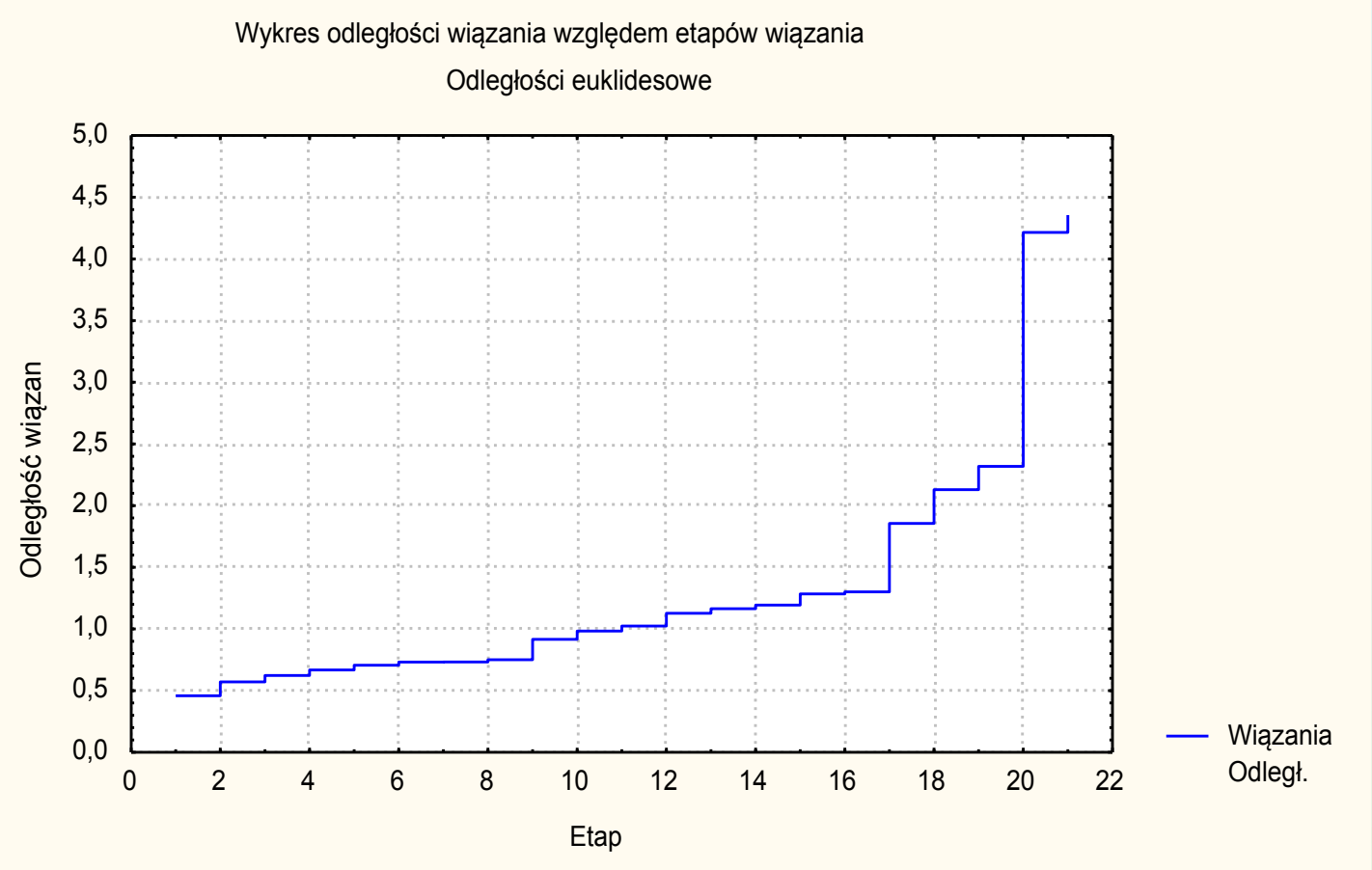


Jak wybrać liczbę skupień?



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. Przykład analizy danych o parametrach samochodów

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

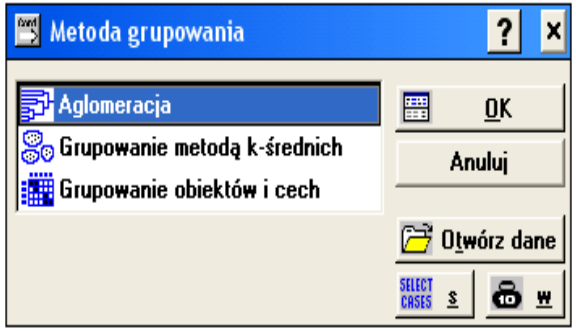
Plik Edycja Widok Analiza Wykresy Opcje Okno Pomoc

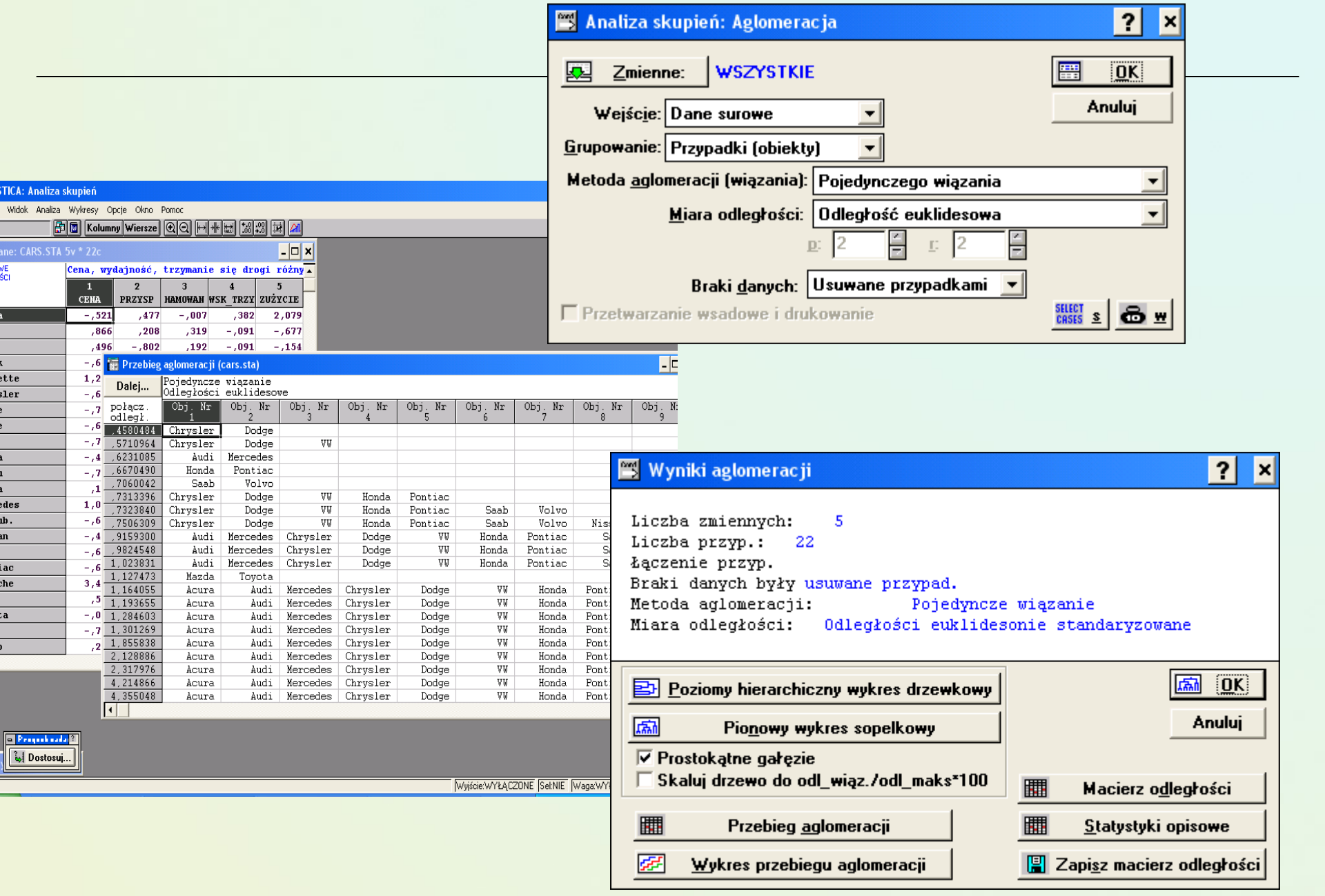
-,521072362755425 Zmienne Przypadki

LICZBOWE WARTOŚCI

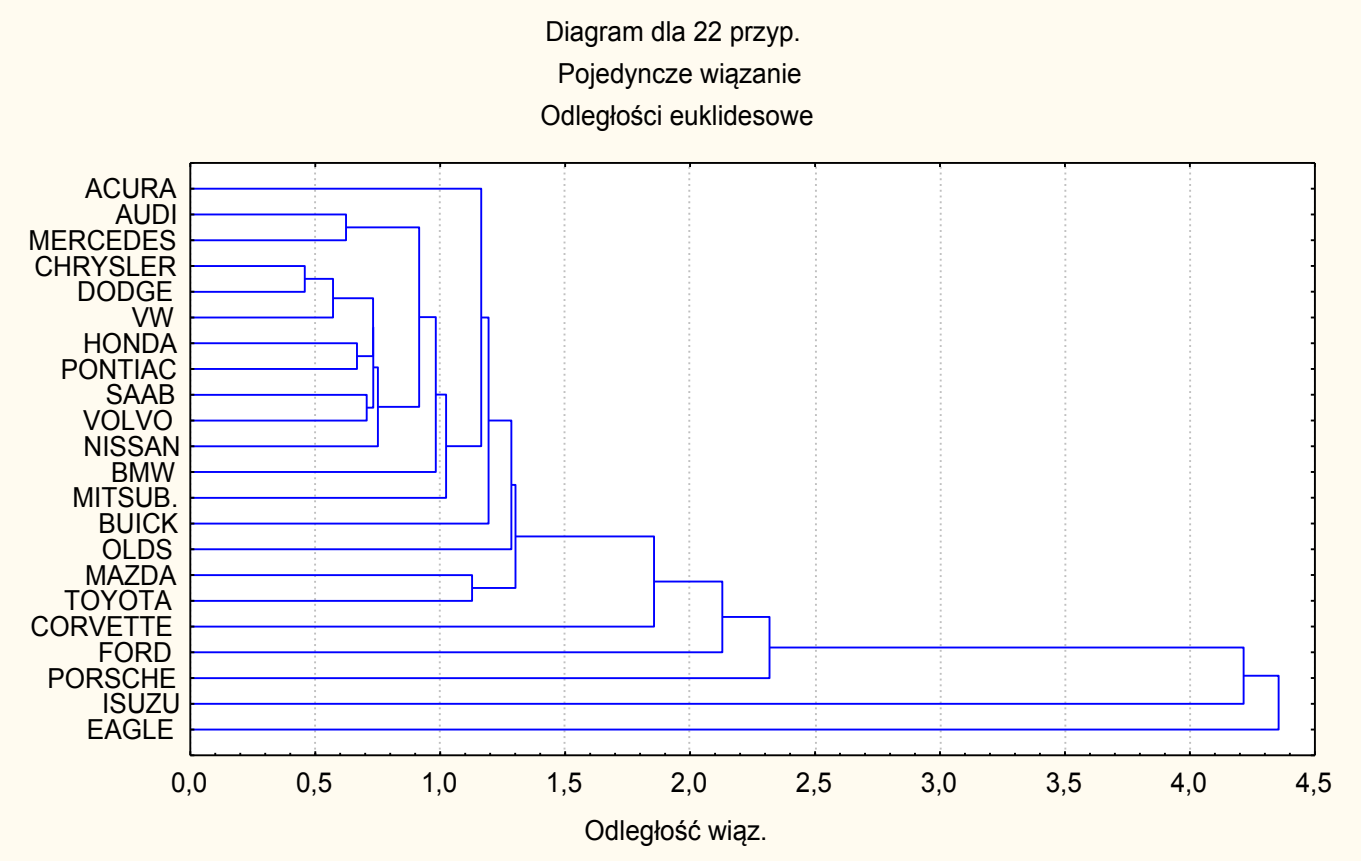
Cena, wydajność, trzymanie się drogi różny

	1 CENA	2 PRZYSP	3 HAMOWAN	4 WSK_TRZY	5 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



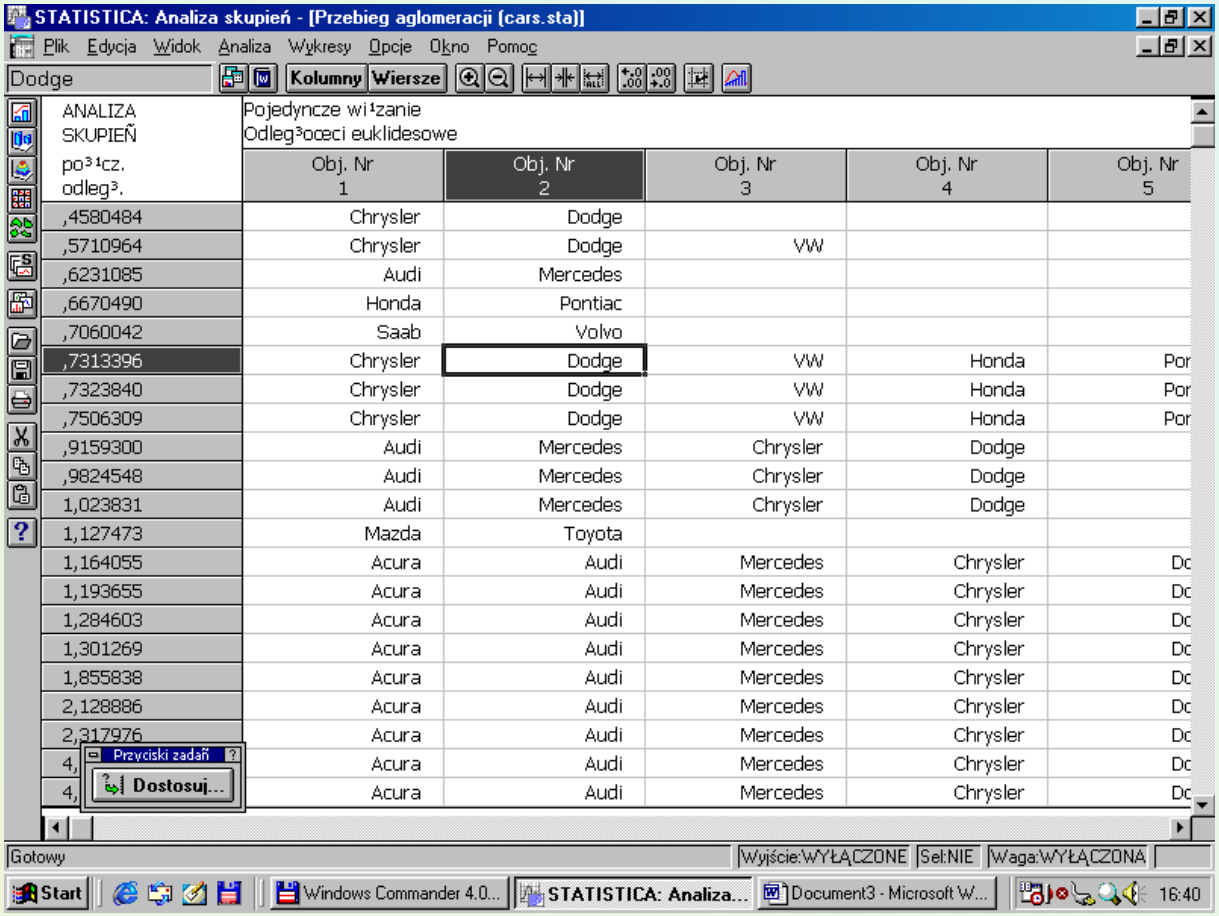


Dendrogram for Single Linkage



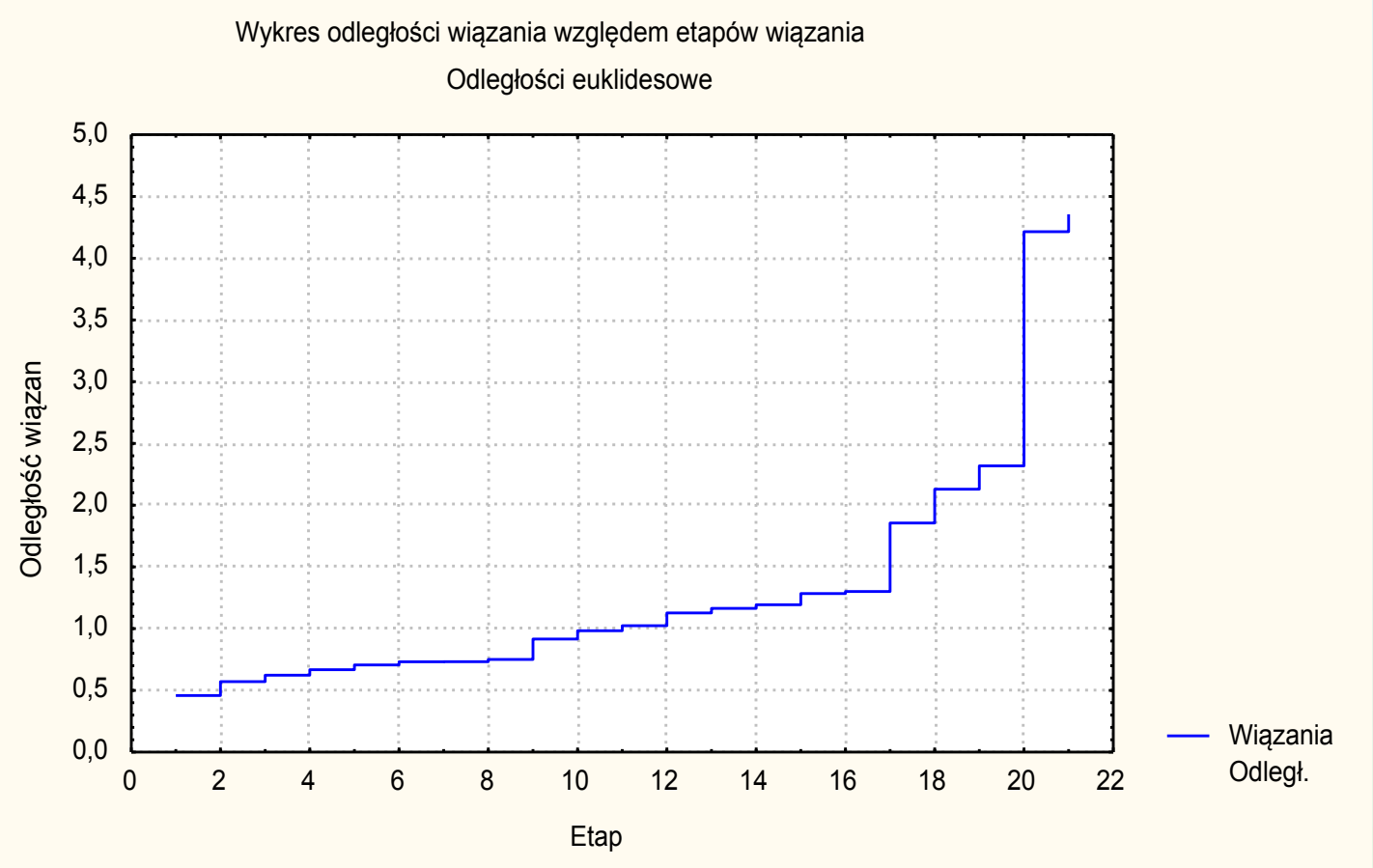
Opis tworzenia dendrogramu

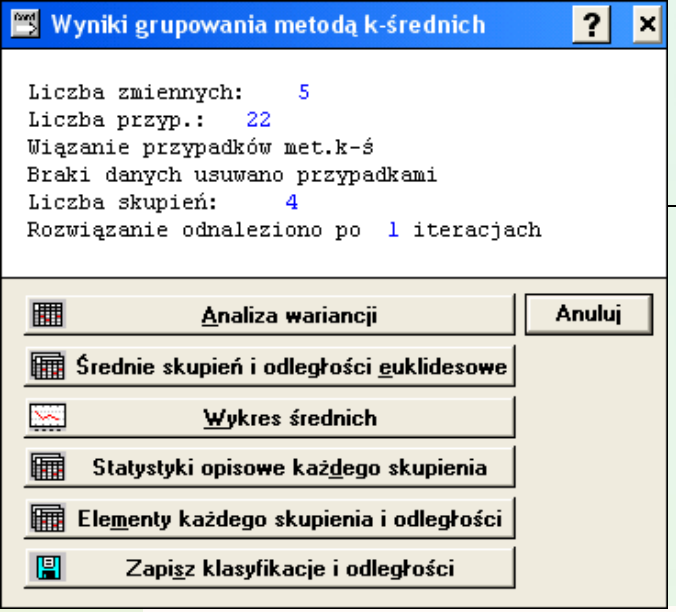
- Łączenie obiektów w kolejnych krokach

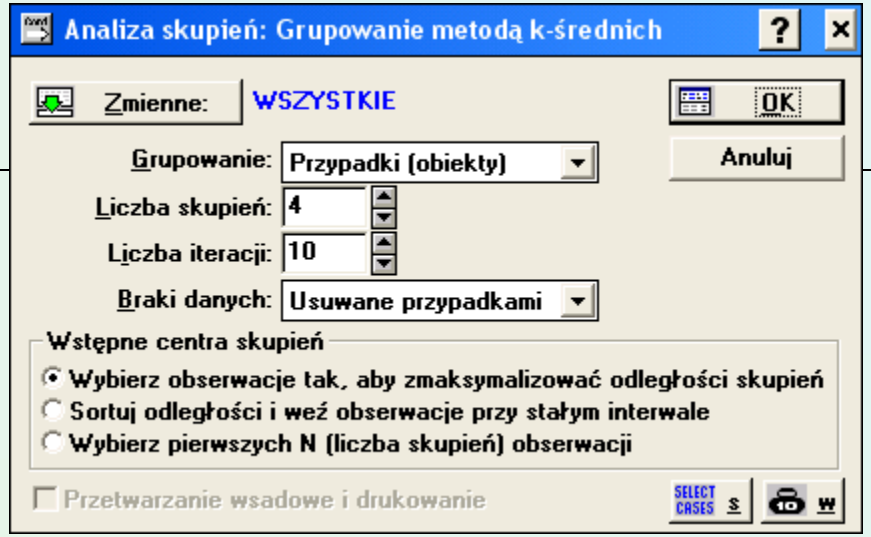


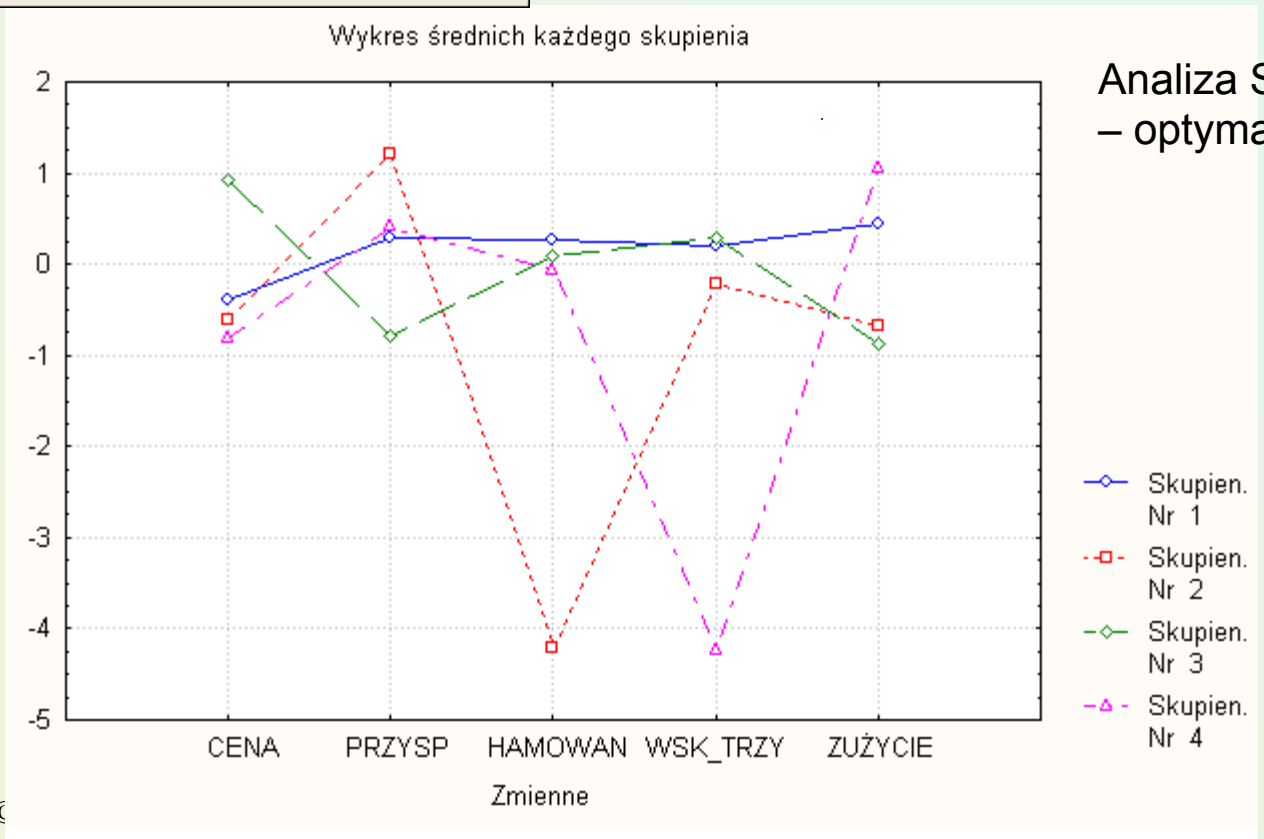
Analiza procesu łączenia

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









Analiza Skupień – optymalizacja k-średnich

Wsparcie do charakterystyki skupień

Odległości euklidesowe skupień (cars.sta)

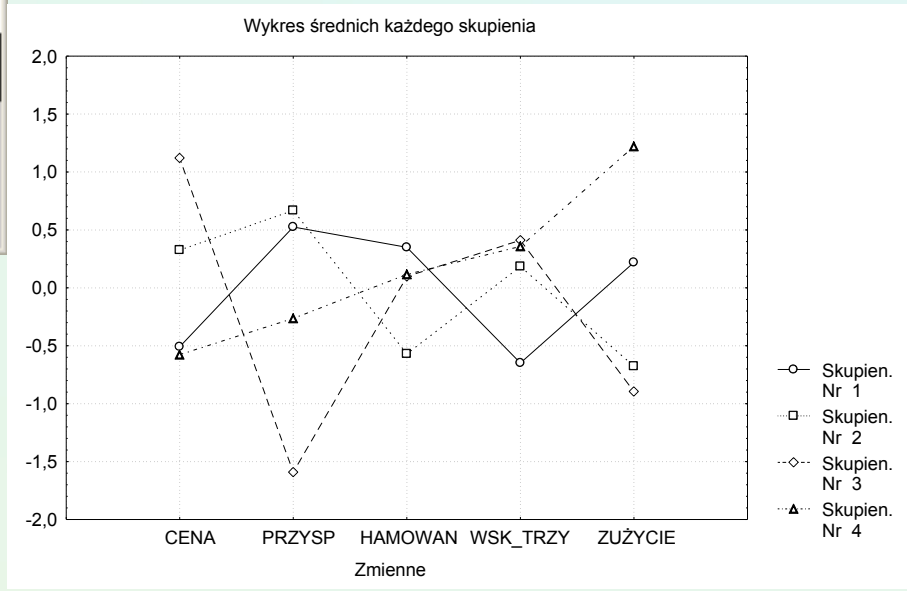
Dalej...

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



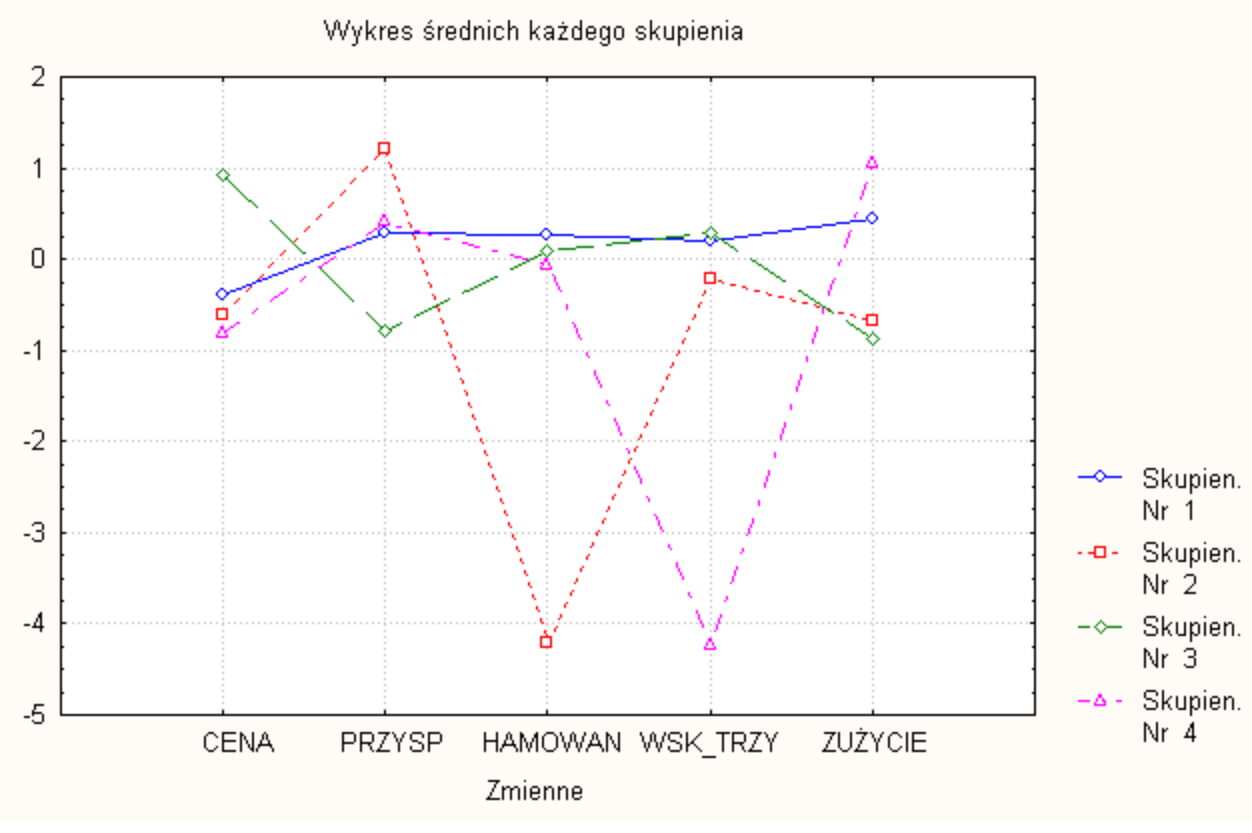
Elementy skupienia numer 2 (cars.sta)

ANALIZA SKUPIEŃ

i odległości od środka właściwego skupienia
W skupieniu jest 6 przyp

	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 dostepna w pliku arff

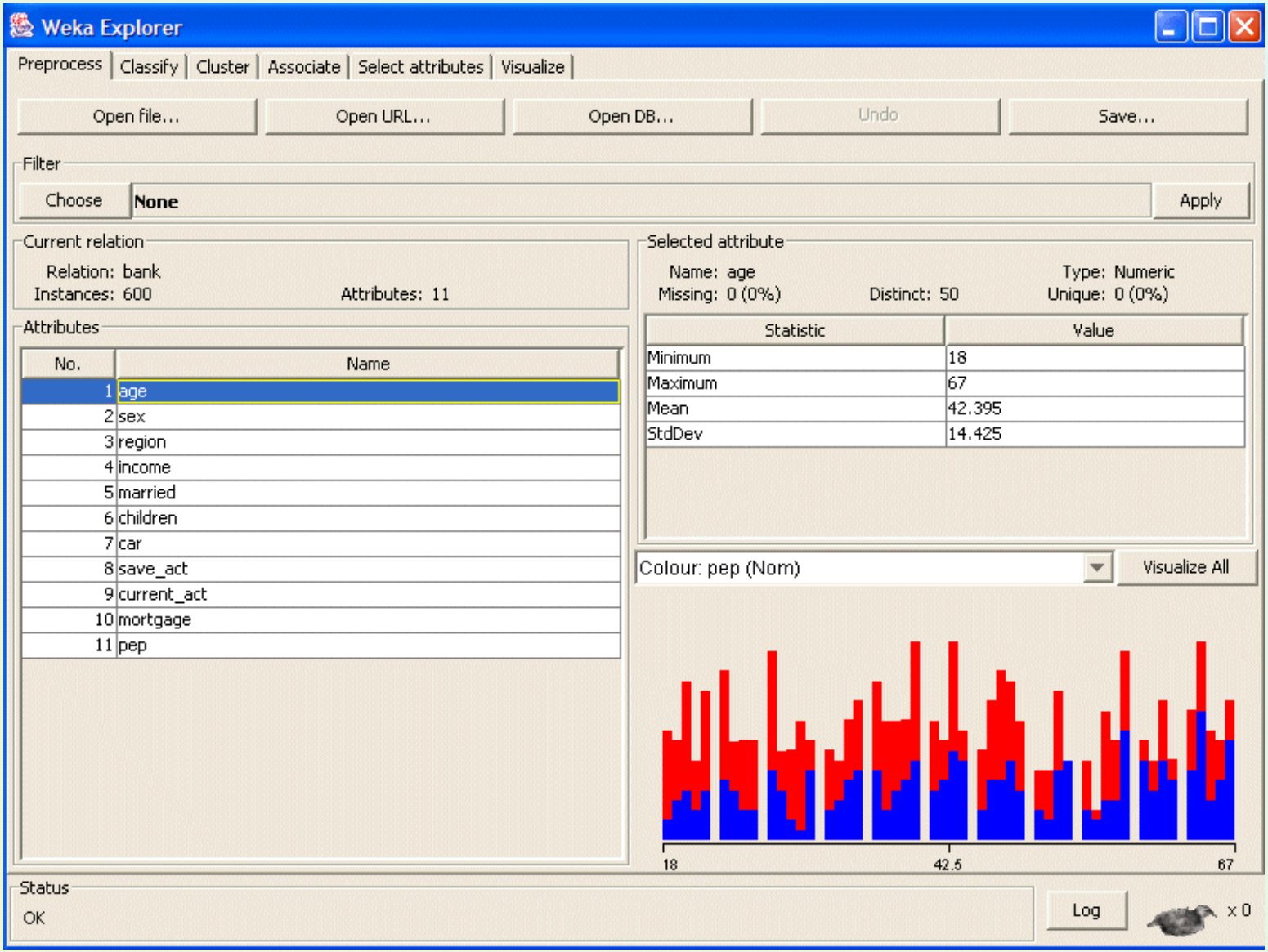
Exercise 1. K-means clustering in WEKA

- The exercise illustrates the use of the k-means algorithm.
- The example – sample of customers of the bank
  - Bank data (bank-data.cvs -> 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
……..
……..
```

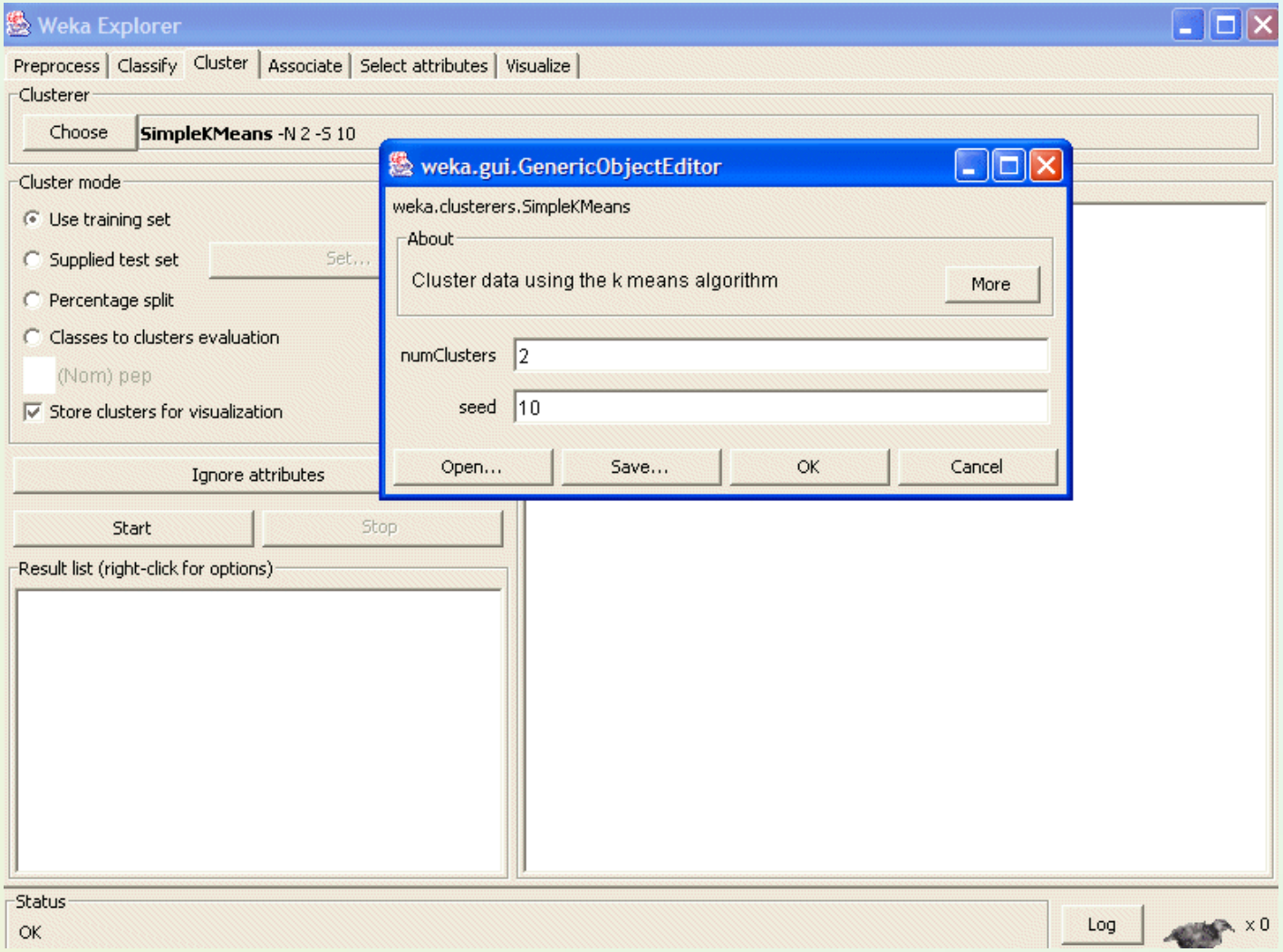
  - Cluster customers and characterize the resulting customer segments

Loading the file and analysing the data

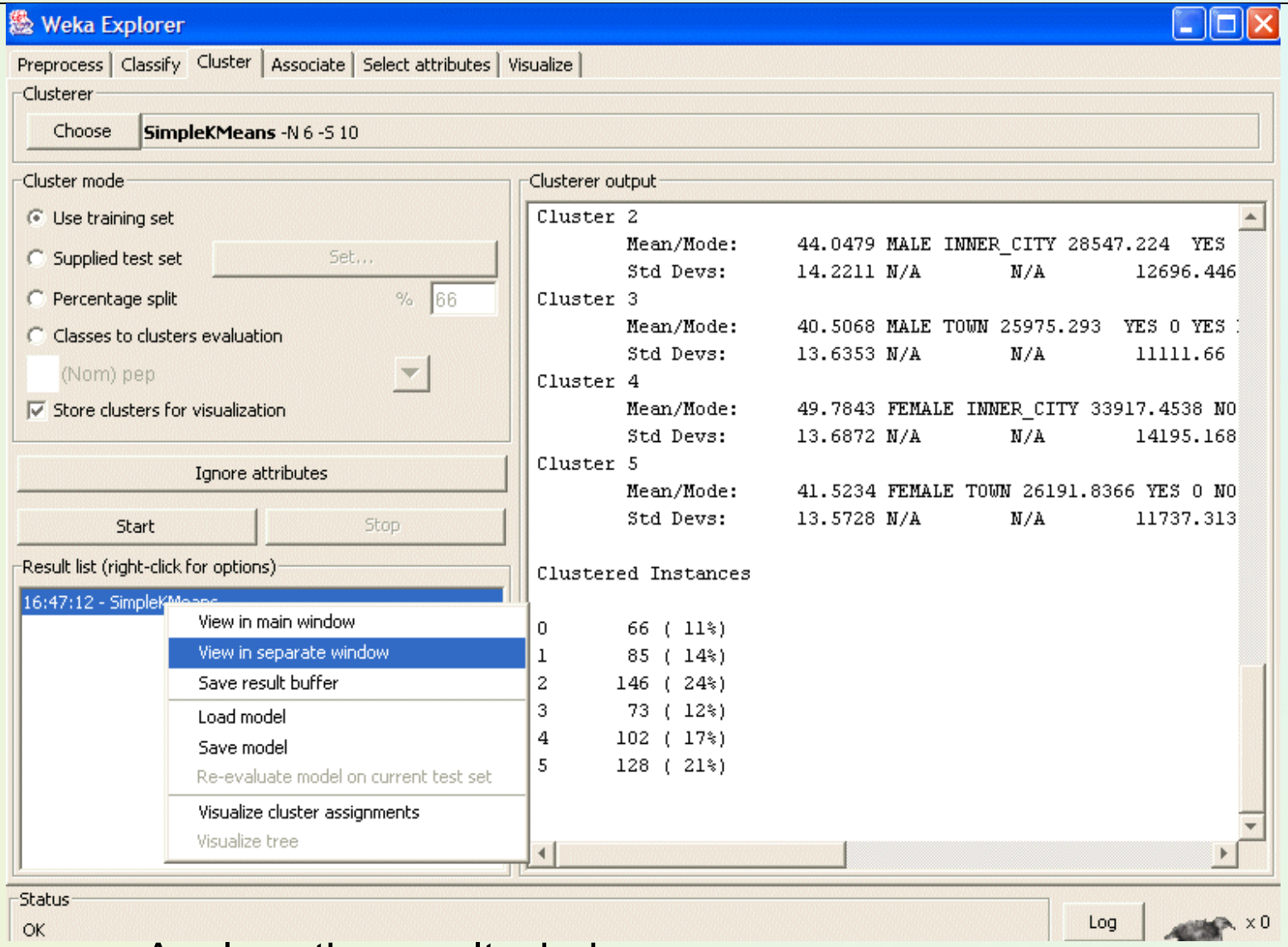


Choosing Simple k-means

- Tune proper parameters



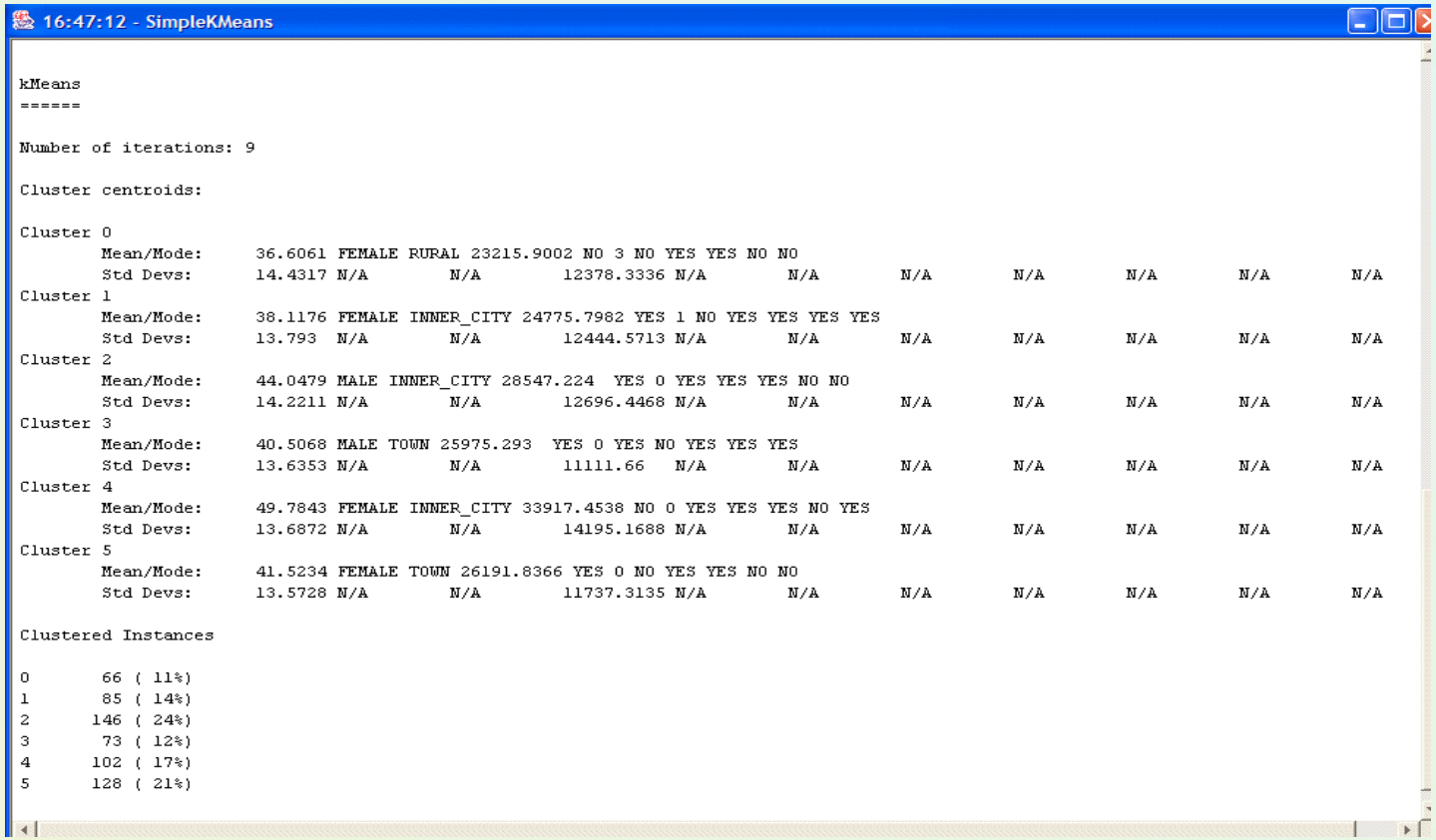
Clustering results



- Analyse the result window

Characterizing cluster

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



Finally, cluster assignments

TextPad - [D:\Bamshad\CLASS\ECT584\WEKA\Cluster\bank-kmeans.arff]

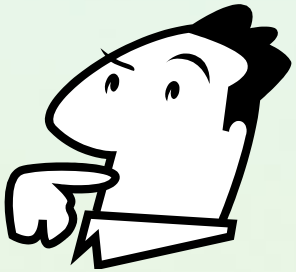
File Edit Search View Tools Macros Configure Window Help

```
@relation bank_clustered

@attribute Instance_number numeric
@attribute age numeric
@attribute sex {FEMALE,MALE}
@attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
@attribute income numeric
@attribute married {NO,YES}
@attribute children {0,1,2,3}
@attribute car {NO,YES}
@attribute save_act {NO,YES}
@attribute current_act {NO,YES}
@attribute mortgage {NO,YES}
@attribute pep {YES,NO}
@attribute Cluster {cluster0,cluster1,cluster2,cluster3,cluster4,cluster5}

@data
0,48,FEMALE,INNER_CITY,17546,NO,1,NO,NO,NO,NO,YES,cluster1
1,40,MALE,TOWN,30085.1,YES,3,YES,NO,YES,YES,NO,cluster3
2,51,FEMALE,INNER_CITY,16575.4,YES,0,YES,YES,YES,NO,NO,cluster2
3,23,FEMALE,TOWN,20375.4,YES,3,NO,NO,YES,NO,NO,cluster5
4,57,FEMALE,RURAL,50576.3,YES,0,NO,YES,NO,NO,NO,cluster5
5,57,FEMALE,TOWN,37869.6,YES,2,NO,YES,YES,NO,YES,cluster5
6,22,MALE,RURAL,8877.07,NO,0,NO,NO,YES,NO,YES,cluster0
7,58,MALE,TOWN,24946.6,YES,0,YES,YES,YES,NO,NO,cluster2
8,37,FEMALE,SUBURBAN,25304.3,YES,2,YES,NO,NO,NO,NO,cluster5
9,54,MALE,TOWN,24212.1,YES,2,YES,YES,YES,NO,NO,cluster2
10,66,FEMALE,TOWN,59803.9,YES,0,NO,YES,YES,NO,NO,cluster5
11,52,FEMALE,INNER_CITY,26658.8,NO,0,YES,YES,YES,YES,NO,cluster4
12,44,FEMALE,TOWN,15735.8,YES,1,NO,YES,YES,YES,YES,cluster1
13,66,FEMALE,TOWN,55204.7,YES,1,YES,YES,YES,YES,YES,cluster1
14,36,MALE,RURAL,19474.6,YES,0,NO,YES,YES,YES,NO,cluster5
15,38,FEMALE,INNER_CITY,22342.1,YES,0,YES,YES,YES,YES,NO,cluster2
16,37,FEMALE,TOWN,17729.8,YES,2,NO,NO,NO,YES,NO,cluster5
17,46,FEMALE,SUBURBAN,41016,YES,0,NO,YES,NO,YES,NO,cluster5
18,62,FEMALE,INNER_CITY,26909.2,YES,0,NO,YES,NO,NO,YES,cluster4
19,31,MALE,TOWN,22522.8,YES,0,YES,YES,YES,NO,NO,cluster2
20,61,MALE,INNER_CITY,57880.7,YES,2,NO,YES,NO,NO,YES,cluster2
21,50,MALE,TOWN,16497.3,YES,2,NO,YES,YES,NO,NO,cluster5
```

Ocena jakości skupień



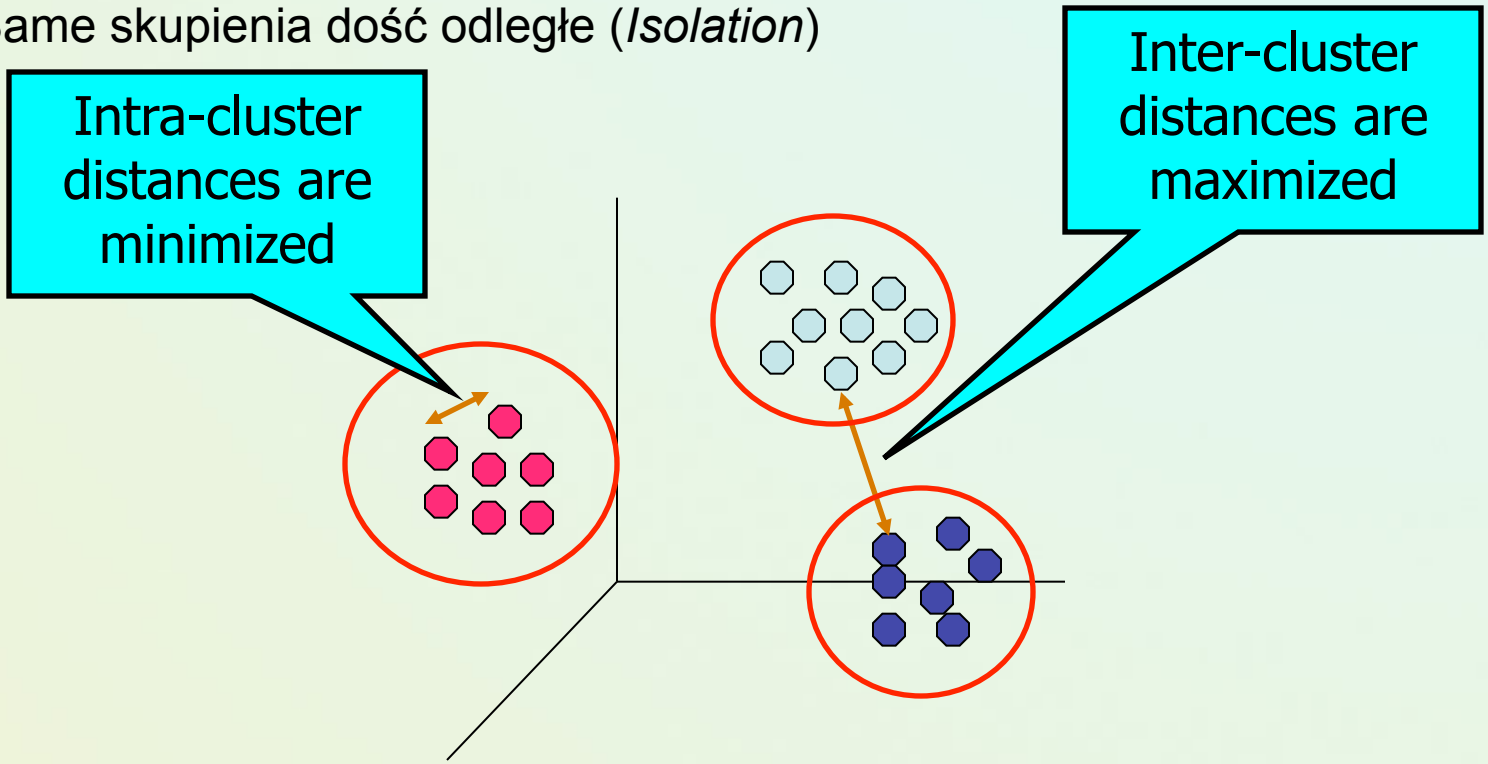
Dwa różne spojrzenia

- 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.

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

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

  - 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?