Jak radzić sobie z chaosem informacyjnym?

Mariusz Bodzioch

Wydział Matematyki i Informatyki,
Uniwersytet Warmińsko-Mazurski w Olsztynie

Co to jest eksploracja danych (data mining)?

Według szacunków firmy Cisco, w 2014 roku przez Internet przesyłanych było 61,5 miliarda gigabajtów danych miesięcznie. To odpowiednik 30 milionów filmów w technologii 3D 24 godziny na dobę. Dla porównania — ludzki mózg codziennie wchłania około 34 gigabajtów danych, co jest odpowiednikiem 100 tysięcy słów.

Chaos informacyjny plagą naszych czasów?

Do ludzkiego mózgu dociera zbyt wiele bodźców jednocześnie, by mógł je wszystkie odebrać i przeanalizować. Nadmiar informacji sprawia, że nie radzi on sobie z wymaganiami, jakie stawia przed nami otoczenie.

Co to jest eksploracja danych (data mining)?

Według szacunków firmy Cisco, w 2014 roku przez Internet przesyłanych było 61,5 miliarda gigabajtów danych miesięcznie. To odpowiednik 30 milionów filmów w technologii 3D 24 godziny na dobę. Dla porównania — ludzki mózg codziennie wchłania około 34 gigabajtów danych, co jest odpowiednikiem 100 tysięcy słów.

Chaos informacyjny plagą naszych czasów?

Do ludzkiego mózgu dociera zbyt wiele bodźców jednocześnie, by mógł je wszystkie odebrać i przeanalizować. Nadmiar informacji sprawia, że nie radzi on sobie z wymaganiami, jakie stawia przed nami otoczenie.

Co to jest eksploracja danych (data mining)?

Co miesiąc około 13 milionów klientów kontaktuje się telefonicznie z centrum obsługi klienta banku Bank of America. W przeszłości każdy klient słyszał to samo ogłoszenie marketingowe, niezależnie, czy było zgodne z zainteresowaniami dzwoniącego, czy nie. Aktualnie każdy klient jest informowany o nowych produktach i usługach banku, które mogą być dla niego najciekawsze.

Co to jest eksploracja danych (data mining)?

Były prezydent USA, Bill Clinton, wspomniał, że niedługo po wydarzeniach z 11 września 2001 roku agenci FBI przeanalizowali olbrzymie ilości danych o konsumentach i odkryli, że dane o 5 sprawcach zamachu były przechowywane w bazie. Jeden z terrorystów miał 30 kart kredytowych z łącznym saldem 250 000$ i był w kraju krócej niż 2 lata. Prowodyr terrorystów, Mohammed Atta, miał 12 różnych adresów, 2 prawdziwe domy i 10 kryjówek. Clinton podsumował, że powinniśmy aktywnie wyszukiwać dane tego typu i jeżeli *ktoś jest tutaj kilka lat lub krócej i ma 12 domów, to jest albo naprawdę bogaty, albo coś kombinuje. Nie powinno być trudno to sprawdzić*.

Eksploracja danych (data mining)

Eksploracja danych jest procesem odkrywania znaczących nowych powiązań, wzorców i trendów przez przeszukiwanie dużych ilości danych zgromdzonych w skarbnicach danych, przy wykorzystywaniu metod rozpoznawania wzorców, jak również metod statystycznych i matematycznych.

Eksploracja danych jest analizą zbiorów danych obserwacyjnych, w celu znalezienia nieoczekiwanych związków i podsumowania danych w oryginalny sposób, tak aby były zrówno zrozumiałe, jak i przydatne dla ich właściciela.

Eksploracja danych (data mining)

Eksploracja danych jest procesem odkrywania znaczących nowych powiązań, wzorców i trendów przez przeszukiwanie dużych ilości danych zgromdzonych w skarbnicach danych, przy wykorzystywaniu metod rozpoznawania wzorców, jak również metod statystycznych i matematycznych.

Eksploracja danych jest analizą zbiorów danych obserwacyjnych, w celu znalezienia nieoczekiwanych związków i podsumowania danych w oryginalny sposób, tak aby były zrówno zrozumiałe, jak i przydatne dla ich właściciela.

Po co eksploracja danych?

Problemem nie jest zbieranie i gromadzenie danych czy informacji, ale przetwarzanie ich w wiedzę. Wiedza od informacji różni się tym, że wykracza ona poza informacje i umożliwia rozwiązywanie nowych problemów.
Wiedzę utożsamia się ze zbiorem reguł (bazą wiedzy) podczas gdy informacje utożsamia się z bazą faktów.

Po co eksploracja danych?

Toniemy w informacjach, ale cierpimy na brak wiedzy.

Zadania eksploracji danych

Najczęstsze zadania eksploracji danych

- Opis.
- Szacowanie (estymacja).
- Przewidywanie (predykcja).
- Klasyfikacja.
- Grupowanie.
- Odkrywanie reguł.

Zadania eksploracji danych

Najczęstsze zadania eksploracji danych

- Opis.
- Szacowanie (estymacja).
- Przewidywanie (predykcja).
- Klasyfikacja.
- Grupowanie.
- Odkrywanie reguł.

Zadania eksploracji danych

Najczęstsze zadania eksploracji danych

- Opis.
- Szacowanie (estymacja).
- Przewidywanie (predykcja).
- Klasyfikacja.
- Grupowanie.
- Odkrywanie reguł.

Zadania eksploracji danych

Najczęstsze zadania eksploracji danych

- Opis.
- Szacowanie (estymacja).
- Przewidywanie (predykcja).
- Klasyfikacja.
- Grupowanie.
- Odkrywanie reguł.

Zadania eksploracji danych

Najczęstsze zadania eksploracji danych

- Opis.
- Szacowanie (estymacja).
- Przewidywanie (predykcja).
- Klasyfikacja.
- Grupowanie.
- Odkrywanie reguł.

Zadania eksploracji danych

Najczęstsze zadania eksploracji danych

- Opis.
- Szacowanie (estymacja).
- Przewidywanie (predykcja).
- Klasyfikacja.
- Grupowanie.
- Odkrywanie reguł.

Zadania eksploracji danych

Najczęstsze zadania eksploracji danych

- Opis.
- Szacowanie (estymacja).
- Przewidywanie (predykcja).
- Klasyfikacja.
- Grupowanie.
- Odkrywanie reguł.

Opis

Opisywanie wzorców lub trendów znajdujących się w danych jest jednym ze sposobów na ich zrozumienie. Wyniki eksploracji danych powinny opisywać wzorce jak najbardziej przejrzyste i jasne. Takie, które można w sposób intuicyjny wyjaśnić i zinterpretować.

Szacowanie

- oszacowanie średniej ocen studenta studiów drugiego stopnia na podstawie wyników ze studiów pierwszego stopnia,
- oszacowanie ryzyka zachorowania na serce w zależności od płci, wieku i rodzaju wykonywanej pracy,
- oszacowanie, ile pieniędzy wyda na zakup prezentów świątecznych losowo wybrana osoba pracująca.

Szacowanie

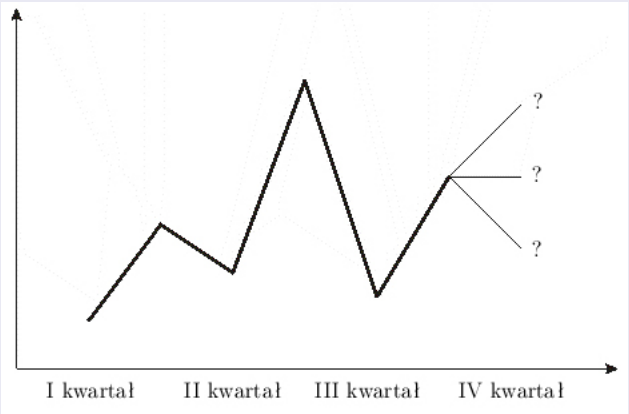
- oszacowanie średniej ocen studenta studiów drugiego stopnia na podstawie wyników ze studiów pierwszego stopnia,
- oszacowanie ryzyka zachorowania na serce w zależności od płci, wieku i rodzaju wykonywanej pracy,
- oszacowanie, ile pieniędzy wyda na zakup prezentów świątecznych losowo wybrana osoba pracująca.

Szacowanie

- oszacowanie średniej ocen studenta studiów drugiego stopnia na podstawie wyników ze studiów pierwszego stopnia,
- oszacowanie ryzyka zachorowania na serce w zależności od płci, wieku i rodzaju wykonywanej pracy,
- oszacowanie, ile pieniędzy wyda na zakup prezentów świątecznych losowo wybrana osoba pracująca.

Przewidywanie

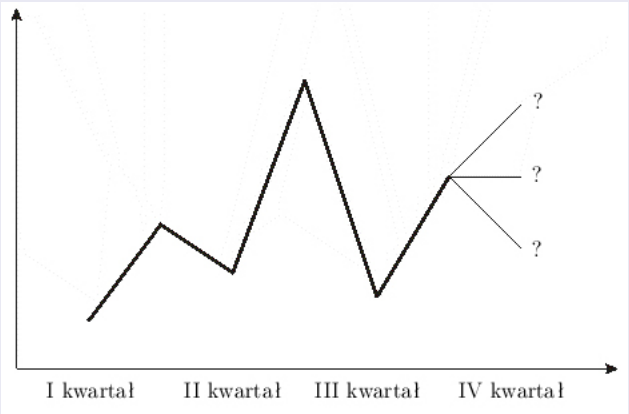
- przewidywanie ceny akcji pewnej firmy na koniec następnego kwartału,



- przewidywanie procentowego wzrostu ceny żywności po zwiększeniu podatku VAT o jeden punkt procentowy na podstawie danych statystycznych uzyskanych po poprzednim zwiększeniu stawki VAT,
- przewidywanie średniego czasu dalszego trwania życia emerytów, po podniesieniu wieku emerytalnego do 67 lat.

Przewidywanie

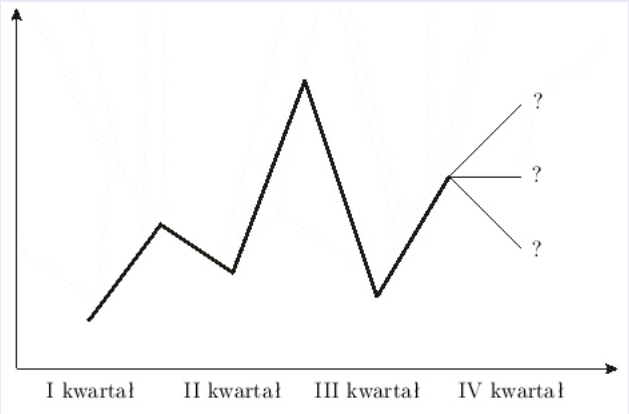
- przewidywanie ceny akcji pewnej firmy na koniec następnego kwartału,



- przewidywanie procentowego wzrostu ceny żywności po zwiększeniu podatku VAT o jeden punkt procentowy na podstawie danych statystycznych uzyskanych po poprzednim zwiększeniu stawki VAT,
- przewidywanie średniego czasu dalszego trwania życia emerytów, po podniesieniu wieku emerytalnego do 67 lat.

Przewidywanie

- przewidywanie ceny akcji pewnej firmy na koniec następnego kwartału,



- przewidywanie procentowego wzrostu ceny żywności po zwiększeniu podatku VAT o jeden punkt procentowy na podstawie danych statystycznych uzyskanych po poprzednim zwiększeniu stawki VAT,
- przewidywanie średniego czasu dalszego trwania życia emerytów, po podniesieniu wieku emerytalnego do 67 lat.

Klasyfikacja

Danymi wejściowymi jest treningowy zbiór przykładów lub obserwacji, który przy pomocy klasyfikatora służy do przydzielania każdej nowej zmiennej nieznanej wcześniej wartości atrybutu decyzyjnego w oparciu o wartości pozostałych atrybutów (deskryptorów).

Klasyfikacja

Lp.	Wiek	Płeć	Wysztalcenie	Dochód
1	27	M	średnie	niski
2	49	M	wyższe	wysoki
3	35	K	wyższe	średni
⋮				

Lp.	Wiek	Płeć	Wysztalcenie	Dochód
...	62	M	wyższe	???

Klasyfikacja

Lp.	Wiek	Płeć	Wysztalcenie	Dochód
1	27	M	średnie	niski
2	49	M	wyższe	wysoki
3	35	K	wyższe	średni
⋮				

Lp.	Wiek	Płeć	Wysztalcenie	Dochód
...	62	M	wyższe	???

Odkrywanie reguł

id	produkty	data
1	chleb, mleko	26-03-2016
2	chleb, pieluszki, piwo	29-03-2016
3	mleko, pieluszki, piwo, ser	30-03-2016
4	masło, buraki, piwo, pieluszki	12-04-2016
5	ocet, pieluszki, orzeszki	17-04-2016
⋮		

Reguła: „Jeżeli kupuje pieluszki, to kupuje piwo”
Z ufnością: $\frac{3}{4} = 75\%$.

Odkrywanie reguł

id	produkty	data
1	chleb, mleko	26-03-2016
2	chleb, pieluszki, piwo	29-03-2016
3	mleko, pieluszki, piwo, ser	30-03-2016
4	masło, buraki, piwo, pieluszki	12-04-2016
5	ocet, pieluszki, orzeszki	17-04-2016
⋮		

Reguła: „Jeżeli kupuje pieluszki, to kupuje piwo”
Z ufnością: $\frac{3}{4} = 75\%$.

Odkrywanie reguł

- szukanie produktów w markecie, które często są kupowane razem oraz takich, które razem kupowane nie są,
- zbadanie reakcji klientów banku na podwyższenie opłat za prowadzenie rachunków,
- określenie częstości występowania efektów ubocznych stosowania pewnego leku.

Odkrywanie reguł

- szukanie produktów w markecie, które często są kupowane razem oraz takich, które razem kupowane nie są,
- zbadanie reakcji klientów banku na podwyższenie opłat za prowadzenie rachunków,
- określenie częstości występowania efektów ubocznych stosowania pewnego leku.

Odkrywanie reguł

- szukanie produktów w markecie, które często są kupowane razem oraz takich, które razem kupowane nie są,
- zbadanie reakcji klientów banku na podwyższenie opłat za prowadzenie rachunków,
- określenie częstości występowania efektów ubocznych stosowania pewnego leku.

Grupowanie (clustering)

Grupowanie (clustering) oznacza grupowanie rekordów, obserwacji lub przypadków w klasy podobnych obiektów. *Grupa* (cluster) jest zbiorem rekordów, które są podobne do siebie nawzajem i niepodobne do rekordów z innych grup.

Grupowanie — przykład

Firma The Nielsen Company profesjonalnie zajmuje się grupowaniem. Firma, w ramach swoich usług, oferuje przedstawienie dla każdego z geograficznych regionów Stanów Zjednoczonych, określonego przez kod pocztowy, demograficznego profilu za pomocą jednego z 66 różnych segmentów PRIZM NE. System segmentacji PRIZM NE opisuje pod względem stylu życia każdy objęty danym kodem pocztowym region kraju. W ten sposób, wybierając określony kod pocztowy, otrzymuje się informacje o najbardziej powszechnej grupie PRIZM NE dla tego kodu.

Grupowanie — przykład

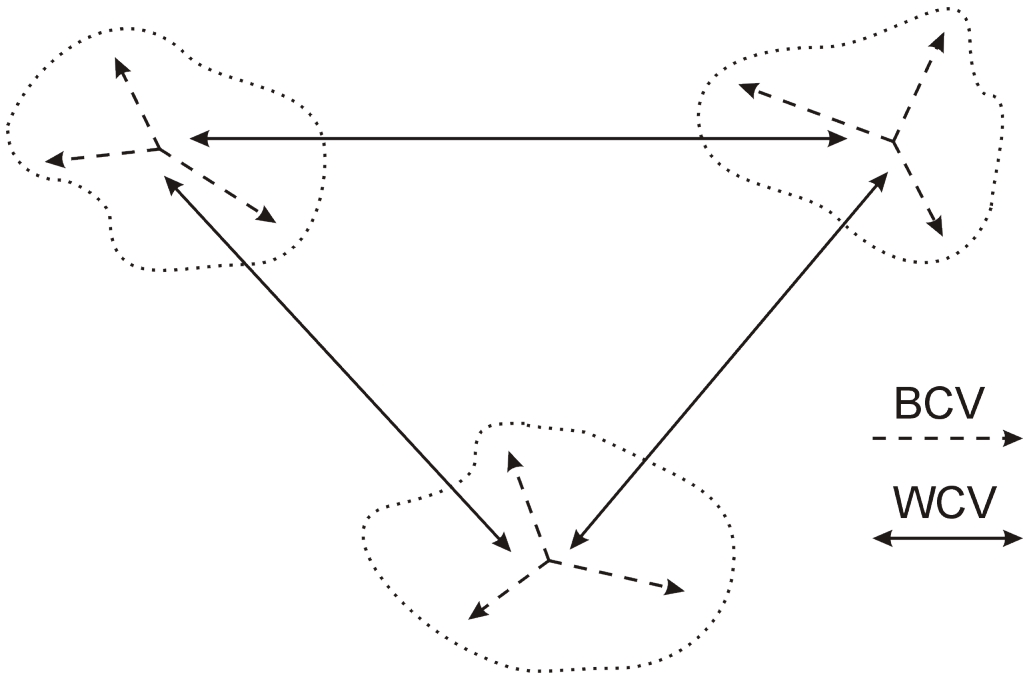
Dla kodu 90210 (Beverly Hills w Kaliforni) otrzymujemy następujące grupy:

Numer	Nazwa
01	Upper Crust
16	Bohemian Mix
07	Money & Brains
03	Movers & Shakers
04	Young Digerati

Opis grupy 01 Upper Crust jest następujący: *Najbardziej ekskluzywny adres, grupa najbardziej zamożnych Amerykanów, ludzi po 55 roku życia, którzy wychowali ju¿ dzieci, żaden z pozostałych segmentów nie posiada większej koncentracji mieszkańców zarabiających ponad 200.000 $ rocznie lub posiadających wyższe wykształcenie*. Każdy następny poziom życia uważany jest za spadek ze szczytu.

Grupowanie

Grupy powinny mieć małą zmienność wewnątrz grupy w porównaniu ze zmiennością pomiędzy grupami.



BCV - zmienność pomiędzy grupami (between-cluster variation)
WCV - zmienność wewnątrz grupy (within-cluster variation)

Grupowanie

Grupowanie przeprowadza się w celach:

- zrozumienia struktury danych,
- wykrywania skupisk,
- wykrywania punktów osobliwych,
- redukcji dużej liczby danych i zastąpienia ich przez grupy (kategorie).

Grupowanie

Grupowanie przeprowadza się w celach:

- zrozumienia struktury danych,
- wykrywania skupisk,
- wykrywania punktów osobliwych,
- redukcji dużej liczby danych i zastąpienia ich przez grupy (kategorie).

Grupowanie

Grupowanie przeprowadza się w celach:

- zrozumienia struktury danych,
- wykrywania skupisk,
- wykrywania punktów osobliwych,
- redukcji dużej liczby danych i zastąpienia ich przez grupy (kategorie).

Grupowanie

Grupowanie przeprowadza się w celach:

- zrozumienia struktury danych,
- wykrywania skupisk,
- wykrywania punktów osobliwych,
- redukcji dużej liczby danych i zastąpienia ich przez grupy (kategorie).

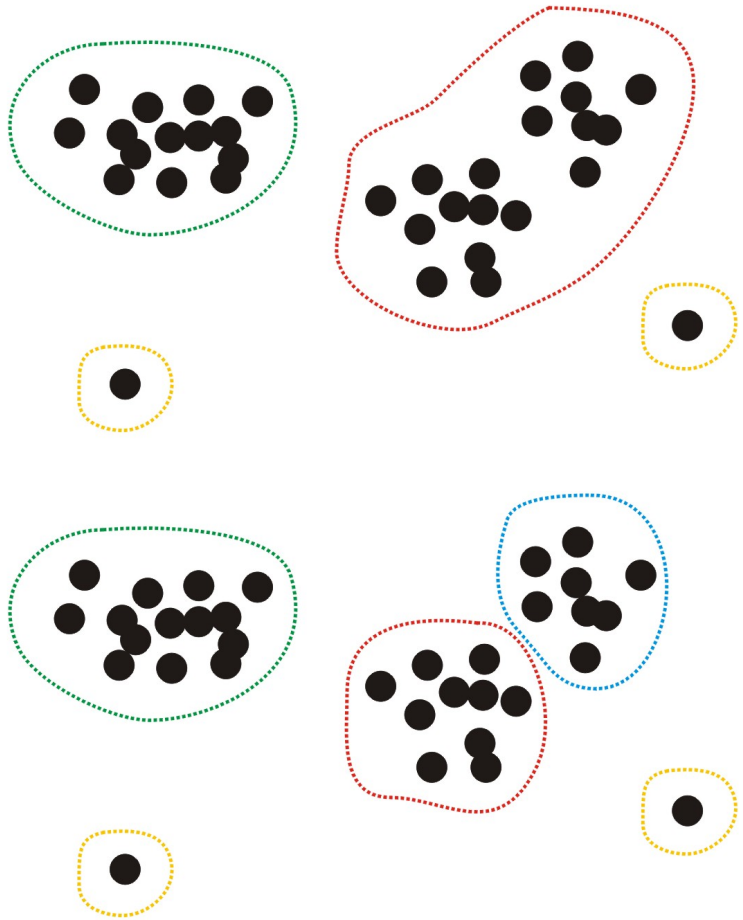
Grupowanie

Grupowanie przeprowadza się w celach:

- zrozumienia struktury danych,
- wykrywania skupisk,
- wykrywania punktów osobliwych,
- redukcji dużej liczby danych i zastąpienia ich przez grupy (kategorie).

Grupowanie

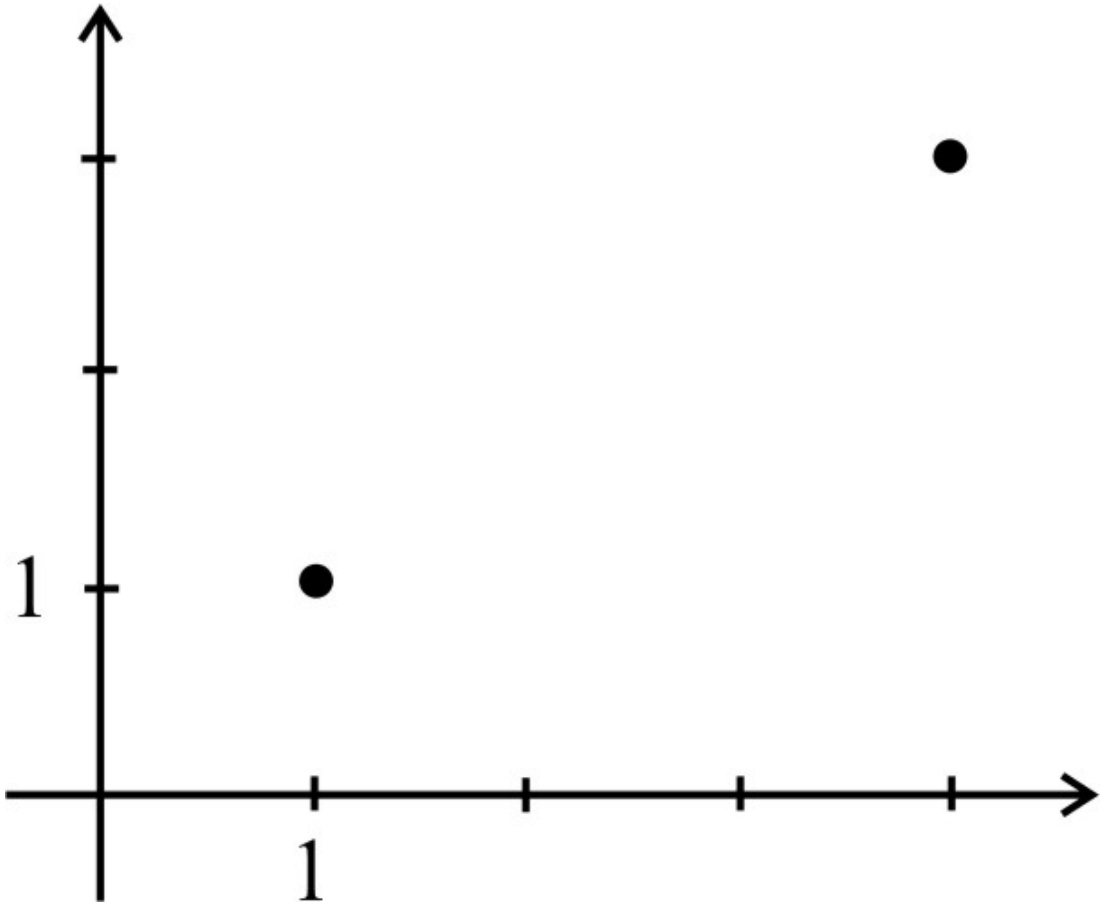
Efekt przykładowego grupowania wraz z detekcją punktów osobliwych



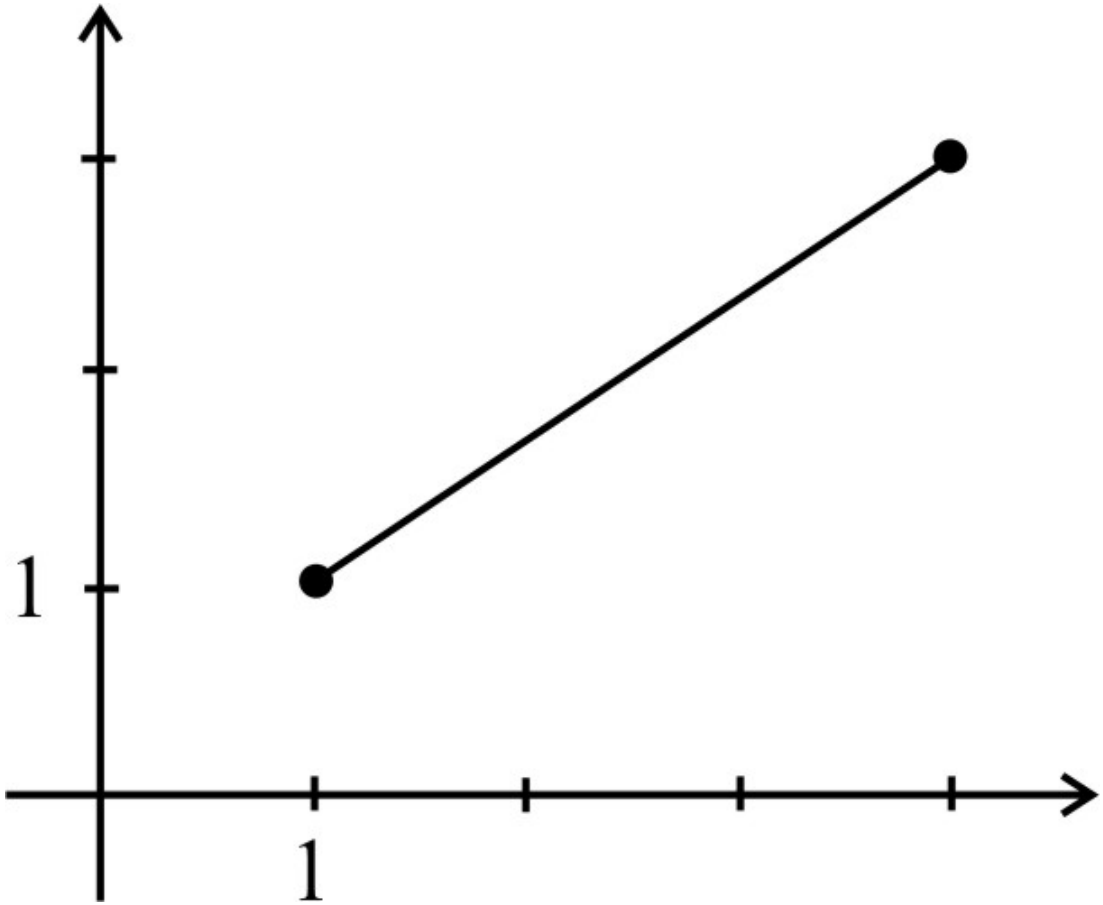
Miary odległości

Co to jest odległość i jak ją mierzyć?

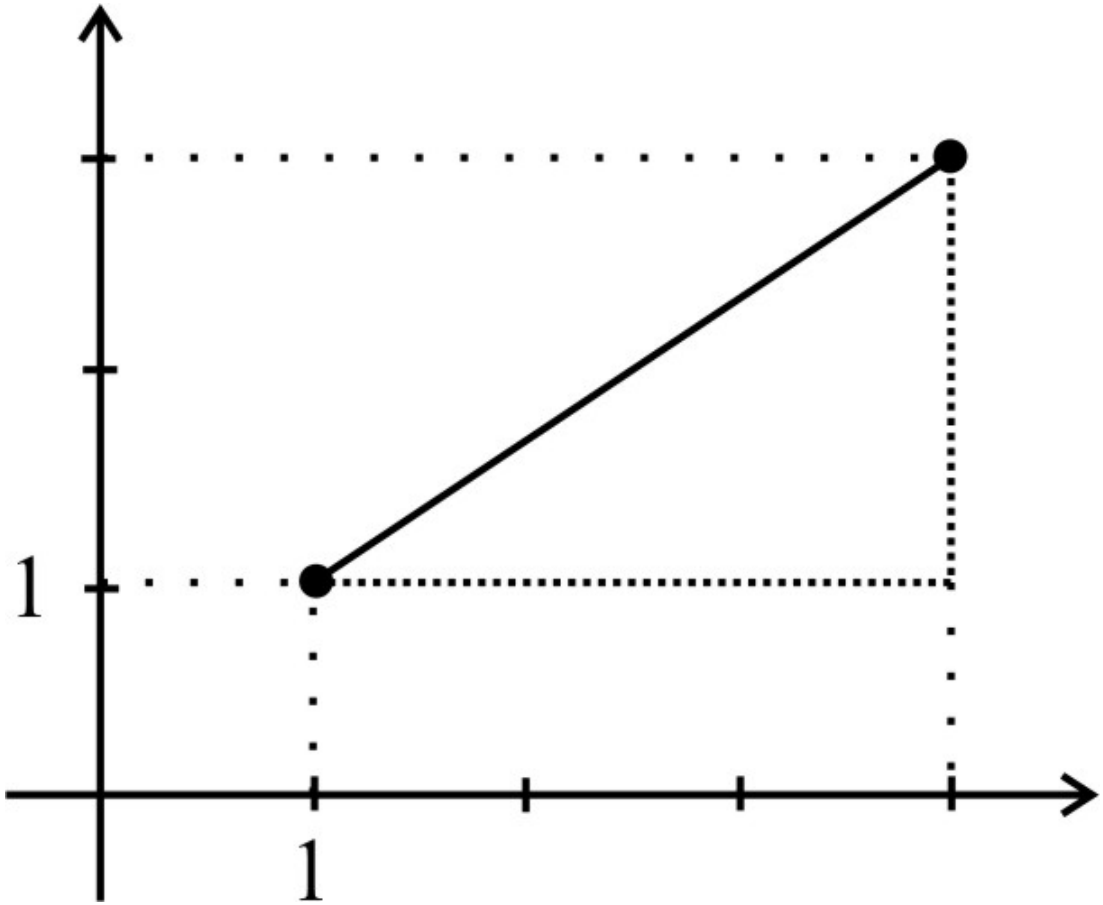
Miary odległości



Miary odległości



Miary odległości



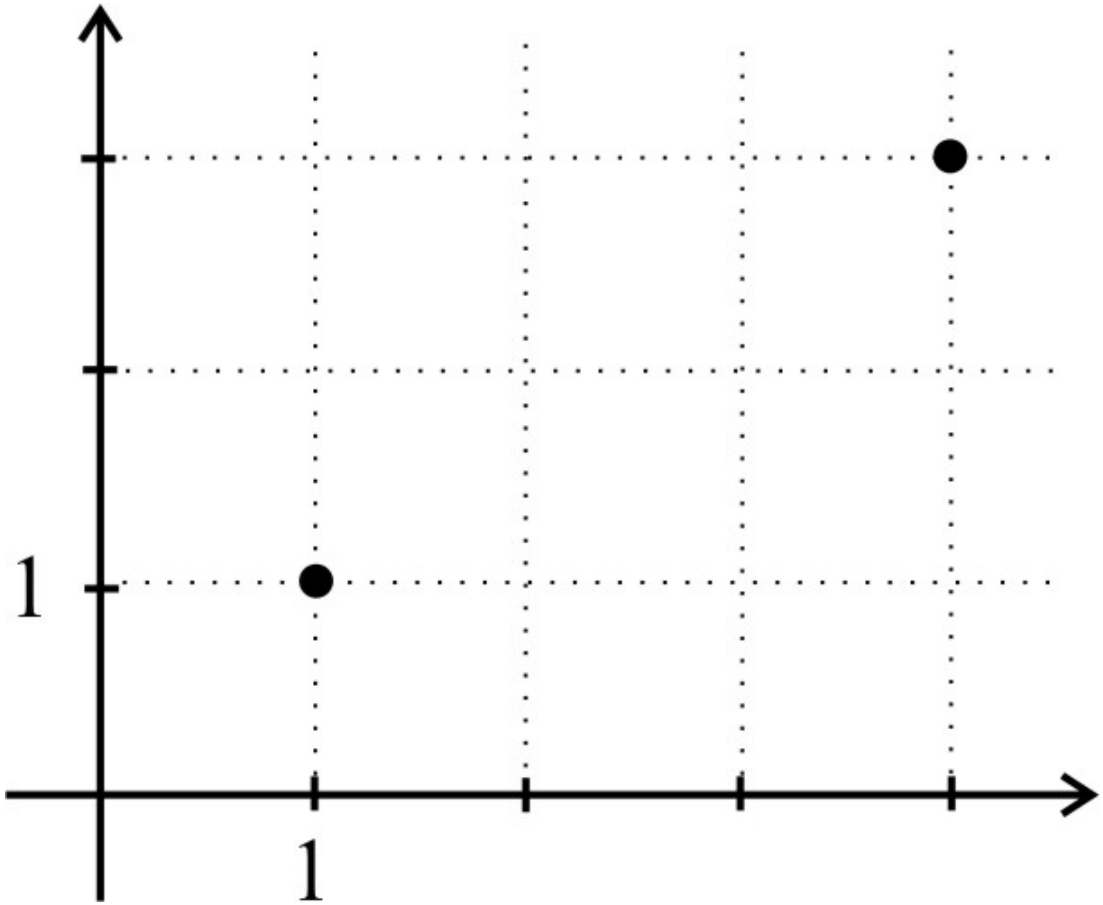
Miary odległości

- Odległość euklidesowa pomiędzy rekordami:

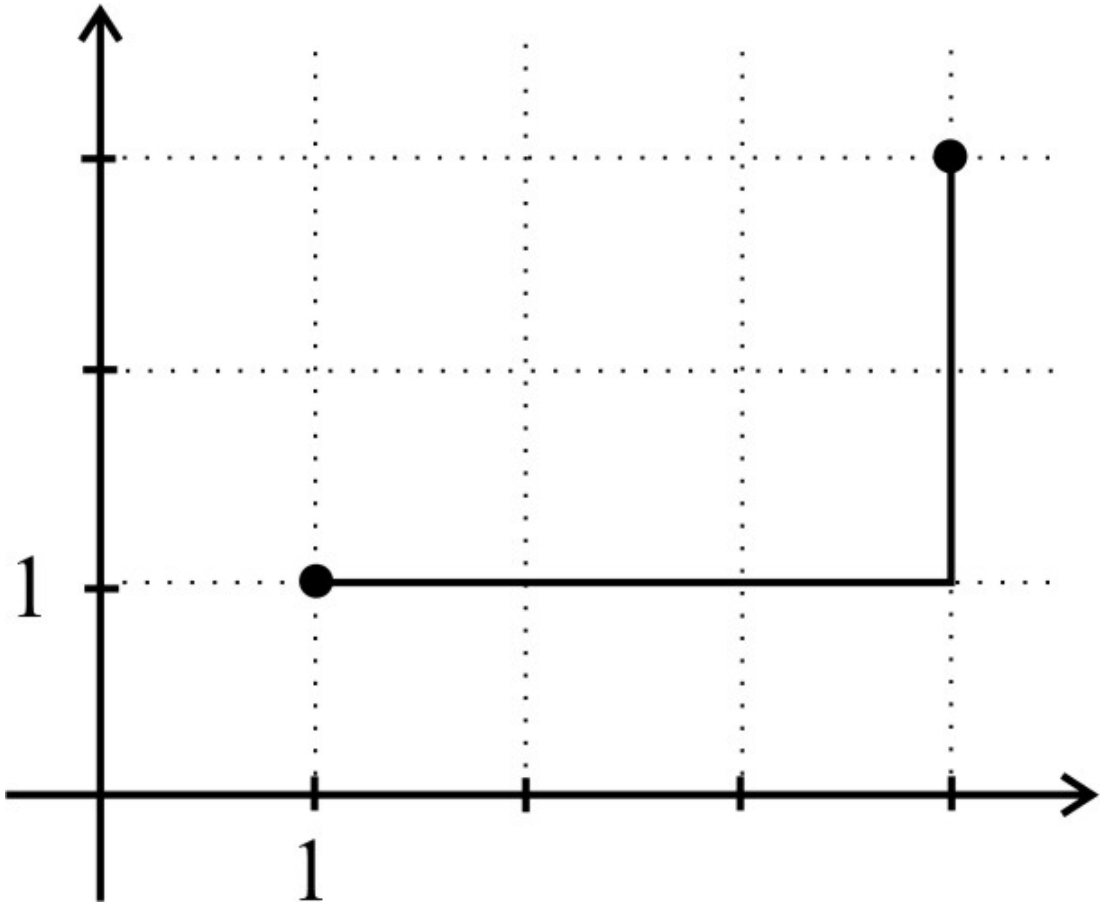
$$d_{Euklides}(x, y) = \sqrt{\sum_i (x_i - y_i)^2},$$

gdzie $x = x_1, x_2, ..., x_m$ i $y = y_1, y_2, ..., y_m$ reprezentują wartości m atrybutów dwóch wzorców.

Miary odległości



Miary odległości



Miary odległości

- Odległość Manhattan:

$$d_{Manhattan}(x, y) = \sum_i |x_i - y_i| ,$$

gdzie $x = x_1, x_2, ..., x_m$ i $y = y_1, y_2, ..., y_m$ reprezentują wartości m atrybutów dwóch wzorców.

Miary odległości

Jak interpretować odległość dla cech jakościowych?

$$d(x_i, y_i) = \begin{cases} 0, & \text{gdy} \quad x_i = y_i, \\ 1, & \text{gdy} \quad x_i \neq y_i. \end{cases}$$

Miary odległości

Jak interpretować odległość dla cech jakościowych?

$$d(x_i, y_i) = \begin{cases} 0, & \text{gdy} \quad x_i = y_i, \\ 1, & \text{gdy} \quad x_i \neq y_i. \end{cases}$$

Miary odległości

Pacjent	Wiek	Płeć
A	50	M
B	20	M
C	50	K

$$d(A,B) = \sqrt{(50-20)^2 + 0^2} = 30$$

$$d(A,C) = \sqrt{(50-50)^2 + 1^2} = 1$$

Kto jest bardziej „podobny” do kogo?

Miary odległości

Pacjent	Wiek	Płeć
A	50	M
B	20	M
C	50	K

$$d(A,B) = \sqrt{(50-20)^2 + 0^2} = 30$$

$$d(A,C) = \sqrt{(50-50)^2 + 1^2} = 1$$

Kto jest bardziej „podobny” do kogo?

Miary odległości

Pacjent	Wiek	Płeć
A	50	M
B	20	M
C	50	K

$$d(A,B) = \sqrt{(50-20)^2 + 0^2} = 30$$

$$d(A,C) = \sqrt{(50-50)^2 + 1^2} = 1$$

Kto jest bardziej „podobny” do kogo?

Miary odległości

Pacjent	Wiek	Płeć
A	50	M
B	20	M
C	50	K

$$d(A,B) = \sqrt{(50-20)^2 + 0^2} = 30$$

$$d(A,C) = \sqrt{(50-50)^2 + 1^2} = 1$$

Kto jest bardziej „podobny” do kogo?

Normalizacja

Do optymalnego działania algorytmy grupowania, tak jak algorytmy klasyfikacyjne, wymagają normalizacji danych, tak żeby żadna zmienna ani podzbiór zmiennych nie zdominowała analizy. Można użyć np.:

- *normalizacji min-max*

$$X^* = \frac{X - min(X)}{zakres(X)} = \frac{X - min(X)}{max(X) - min(X)}.$$

Normalizacja

Do optymalnego działania algorytmy grupowania, tak jak algorytmy klasyfikacyjne, wymagają normalizacji danych, tak żeby żadna zmienna ani podzbiór zmiennych nie zdominowała analizy. Można użyć np.:

- *normalizacji min-max*

$$X^* = \frac{X - min(X)}{zakres(X)} = \frac{X - min(X)}{max(X) - min(X)}.$$

Normalizacja

Załóżmy, że wartość minimalna dla zmiennej *Wiek* to 10 lat, a maksymalna 90:

Pacjent	Wiek	Wiek (min-max)	Płeć
A	50	$\frac{50-10}{80} = 0,5$	M
B	20	$\frac{20-10}{80} = 0,125$	M
C	50	$\frac{50-10}{80} = 0,5$	K

$$d(A,B) = \sqrt{(0,5-0,125)^2 + 0^2} = 0,375$$

$$d(A,C) = \sqrt{(0,5-0,5)^2 + 1^2} = 1$$

Kto jest teraz bardziej „podobny” do kogo?

Normalizacja

Załóżmy, że wartość minimalna dla zmiennej *Wiek* to 10 lat, a maksymalna 90:

Pacjent	Wiek	Wiek (min-max)	Płeć
A	50	$\frac{50-10}{80} = 0,5$	M
B	20	$\frac{20-10}{80} = 0,125$	M
C	50	$\frac{50-10}{80} = 0,5$	K

$$d(A,B) = \sqrt{(0,5-0,125)^2 + 0^2} = 0,375$$

$$d(A,C) = \sqrt{(0,5-0,5)^2 + 1^2} = 1$$

Kto jest teraz bardziej „podobny” do kogo?

Normalizacja

Załóżmy, że wartość minimalna dla zmiennej *Wiek* to 10 lat, a maksymalna 90:

Pacjent	Wiek	Wiek (min-max)	Płeć
A	50	$\frac{50-10}{80} = 0,5$	M
B	20	$\frac{20-10}{80} = 0,125$	M
C	50	$\frac{50-10}{80} = 0,5$	K

$$d(A,B) = \sqrt{(0,5-0,125)^2 + 0^2} = 0,375$$

$$d(A,C) = \sqrt{(0,5-0,5)^2 + 1^2} = 1$$

Kto jest teraz bardziej „podobny” do kogo?

Normalizacja

Załóżmy, że wartość minimalna dla zmiennej *Wiek* to 10 lat, a maksymalna 90:

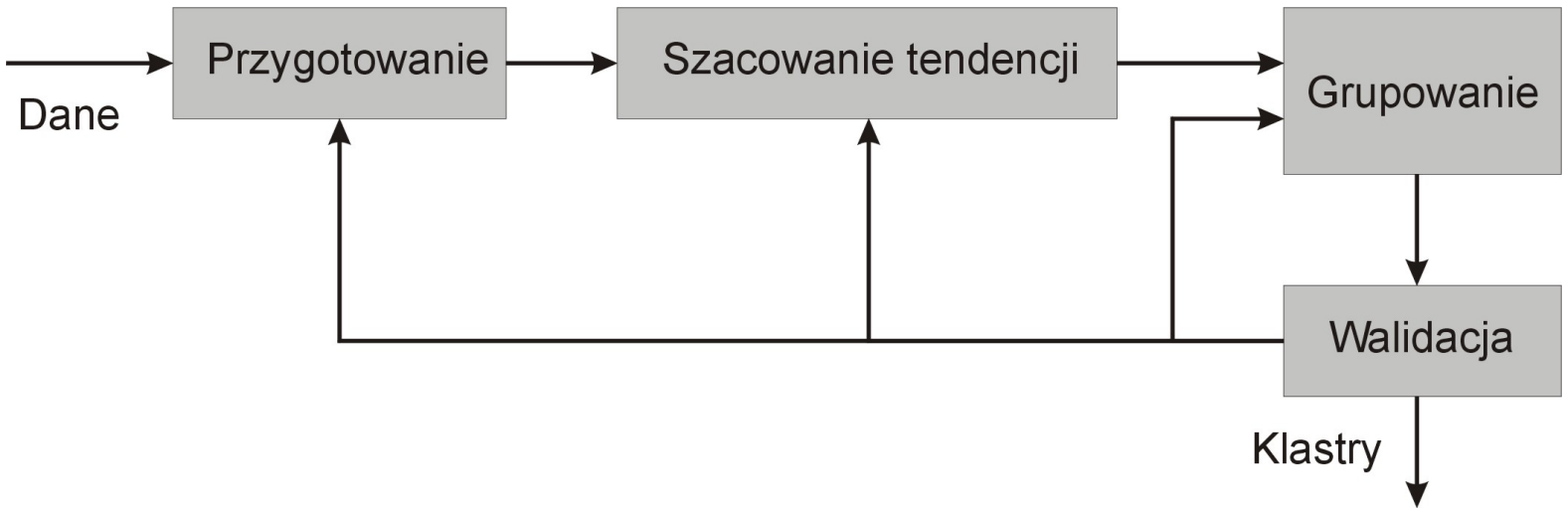
Pacjent	Wiek	Wiek (min-max)	Płeć
A	50	$\frac{50-10}{80} = 0,5$	M
B	20	$\frac{20-10}{80} = 0,125$	M
C	50	$\frac{50-10}{80} = 0,5$	K

$$d(A,B) = \sqrt{(0,5-0,125)^2 + 0^2} = 0,375$$

$$d(A,C) = \sqrt{(0,5-0,5)^2 + 1^2} = 1$$

Kto jest teraz bardziej „podobny” do kogo?

Składowe procesu grupowania



Algorytmy grupowania

Rodzaje algorytmów grupowania

- hierarchiczne,
- niehierarchiczne.

Algorytmy hierarchiczne

- aglomeracyjne,
- rozdzielające.

Algorytmy grupowania

Rodzaje algorytmów grupowania

- hierarchiczne,
- niehierarchiczne.

Algorytmy hierarchiczne

- aglomeracyjne,
- rozdzielające.

Algorytmy grupowania

Rodzaje algorytmów grupowania

- hierarchiczne,
- niehierarchiczne.

Algorytmy hierarchiczne

- aglomeracyjne,
- rozdzielające.

Algorytmy grupowania

Rodzaje algorytmów grupowania

- hierarchiczne,
- niehierarchiczne.

Algorytmy hierarchiczne

- aglomeracyjne,
- rozdzielające.

Algorytmy grupowania

Rodzaje algorytmów grupowania

- hierarchiczne,
- niehierarchiczne.

Algorytmy hierarchiczne

- aglomeracyjne,
- rozdzielające.

Algorytmy grupowania

Rodzaje algorytmów grupowania

- hierarchiczne,
- niehierarchiczne.

Algorytmy hierarchiczne

- aglomeracyjne,
- rozdzielające.

Grupowanie hierarchiczne

W *grupowaniu hierarchicznym* tworzona jest struktura drzewiasta (*dendrogram*) poprzez rekurencyjne dzielenie (metody rozdzielające) lub łączenie (metody aglomeracyjne) istniejących grup.

Metody aglomeracyjne

Pojęcie odległości między rekordami jest raczej jasne. Ale w jaki sposób określić odległość między grupami? Czy powinniśmy uważać dwie grupy za bliskie, jeżeli ich najbliżsi sąsiedzi są blisko siebie, czy ich najdalsi sąsiedzi są blisko siebie?

Kryteria odległości pomiędzy dowolnymi grupami

- *Metoda najbliższego sąsiedztwa*,
- *Metoda najdalszego sąsiedztwa*,
- *Metoda średniego połączenia*.

Metody aglomeracyjne

Pojęcie odległości między rekordami jest raczej jasne. Ale w jaki sposób określić odległość między grupami? Czy powinniśmy uważać dwie grupy za bliskie, jeżeli ich najbliżsi sąsiedzi są blisko siebie, czy ich najdalsi sąsiedzi są blisko siebie?

Kryteria odległości pomiędzy dowolnymi grupami

- *Metoda najbliższego sąsiedztwa*,
- *Metoda najdalszego sąsiedztwa*,
- *Metoda średniego połączenia*.

Metody aglomeracyjne

Pojęcie odległości między rekordami jest raczej jasne. Ale w jaki sposób określić odległość między grupami? Czy powinniśmy uważać dwie grupy za bliskie, jeżeli ich najbliżsi sąsiedzi są blisko siebie, czy ich najdalsi sąsiedzi są blisko siebie?

Kryteria odległości pomiędzy dowolnymi grupami

- *Metoda najbliższego sąsiedztwa*,
- *Metoda najdalszego sąsiedztwa*,
- *Metoda średniego połączenia*.

Metody aglomeracyjne

Pojęcie odległości między rekordami jest raczej jasne. Ale w jaki sposób określić odległość między grupami? Czy powinniśmy uważać dwie grupy za bliskie, jeżeli ich najbliżsi sąsiedzi są blisko siebie, czy ich najdalsi sąsiedzi są blisko siebie?

Kryteria odległości pomiędzy dowolnymi grupami

- *Metoda najbliższego sąsiedztwa*,
- *Metoda najdalszego sąsiedztwa*,
- *Metoda średniego połączenia*.

Metody aglomeracyjne

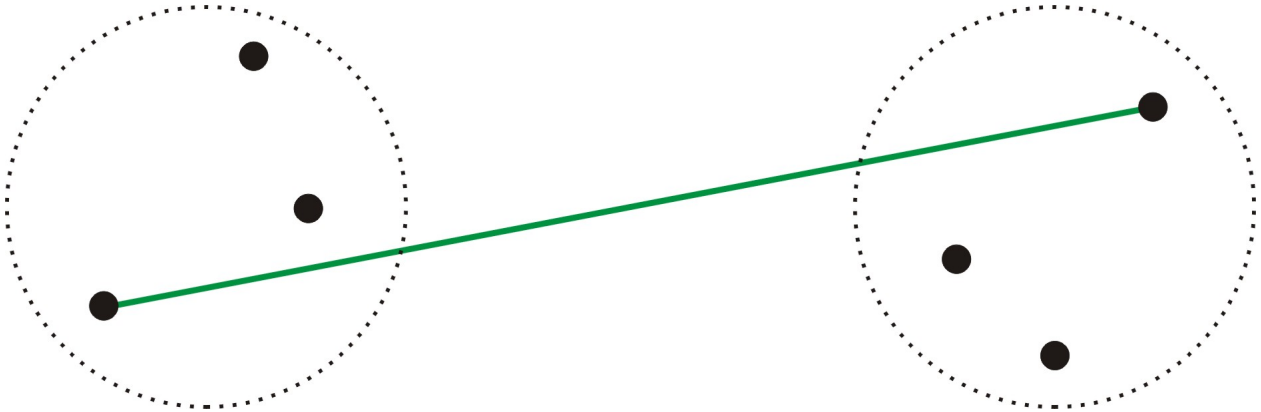
Pojęcie odległości między rekordami jest raczej jasne. Ale w jaki sposób określić odległość między grupami? Czy powinniśmy uważać dwie grupy za bliskie, jeżeli ich najbliżsi sąsiedzi są blisko siebie, czy ich najdalsi sąsiedzi są blisko siebie?

Kryteria odległości pomiędzy dowolnymi grupami

- *Metoda najbliższego sąsiedztwa*,
- *Metoda najdalszego sąsiedztwa*,
- *Metoda średniego połączenia*.

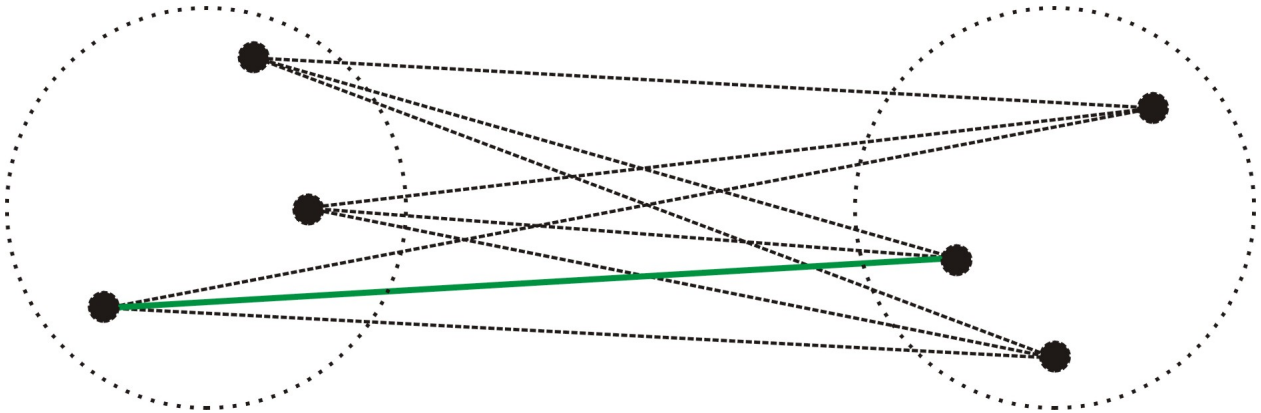
Metody aglomeracyjne

metoda najbliższego sąsiada



metoda najdalszego sąsiada

Metody aglomeracyjne



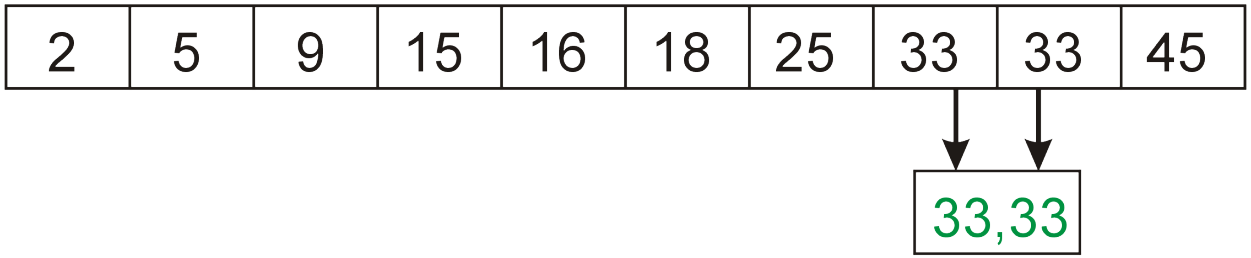
metoda mediany

metoda środka ciężkości

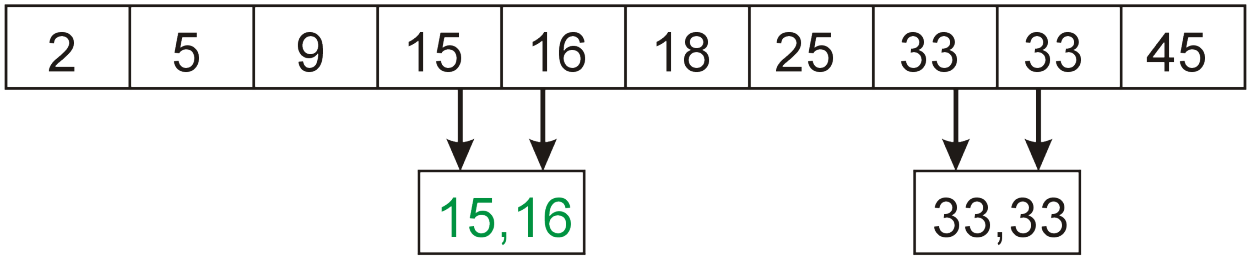
Metoda najbliższego sąsiedztwa (1)

2	5	9	15	16	18	25	33	33	45

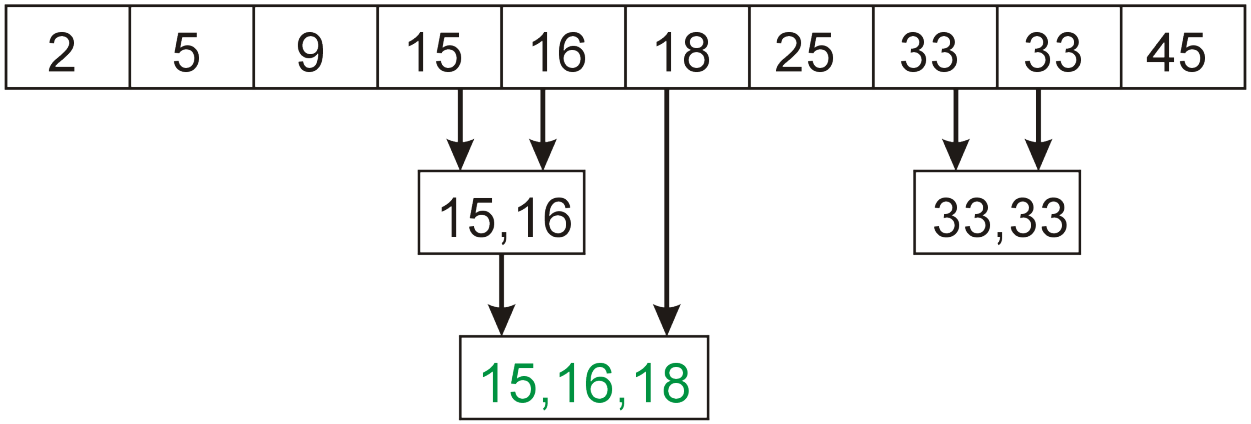
Metoda najbliższego sąsiedztwa (2)



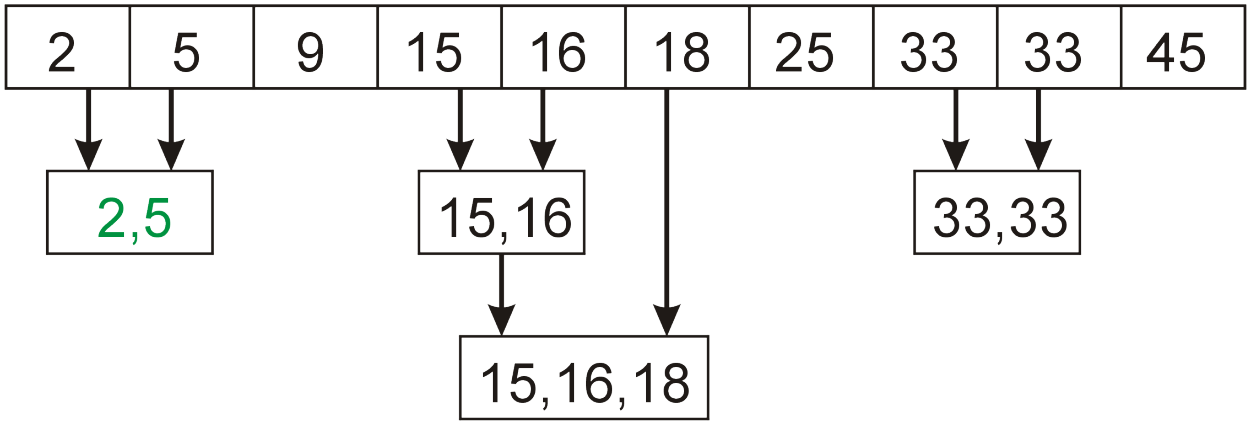
Metoda najbliższego sąsiedztwa (3)



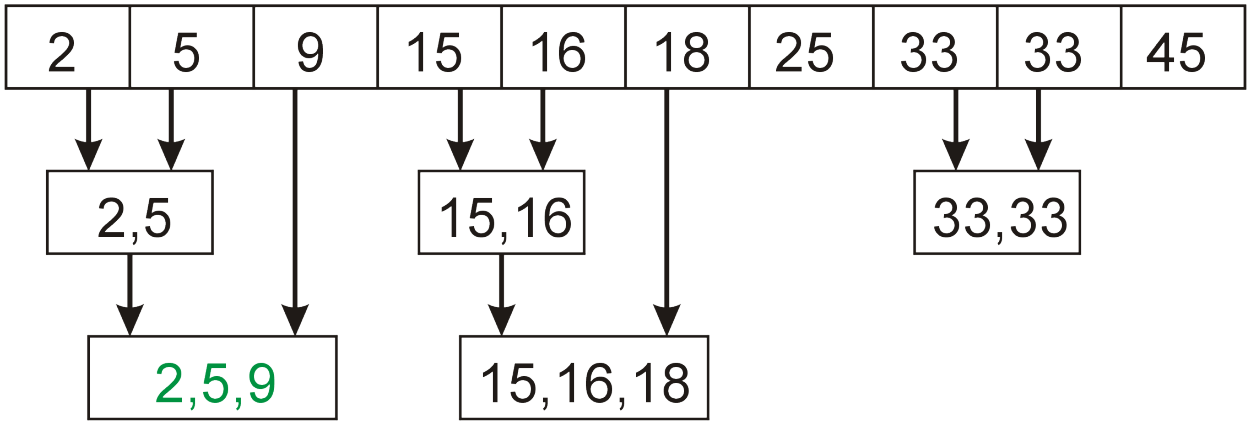
Metoda najbliższego sąsiedztwa (4)



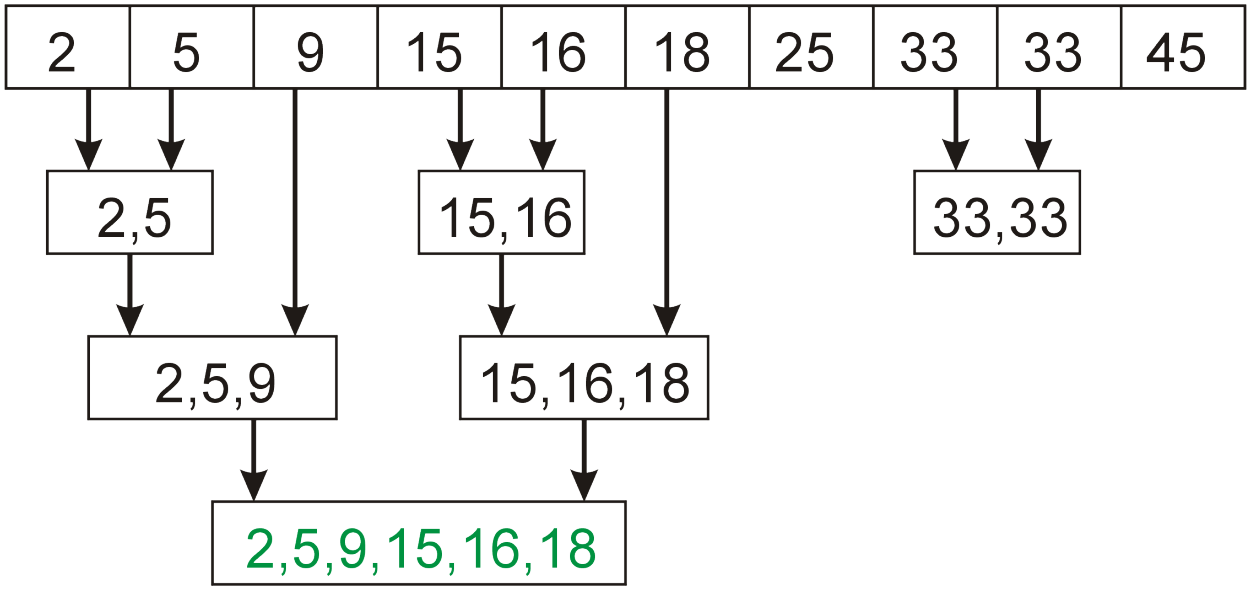
Metoda najbliższego sąsiedztwa (5)



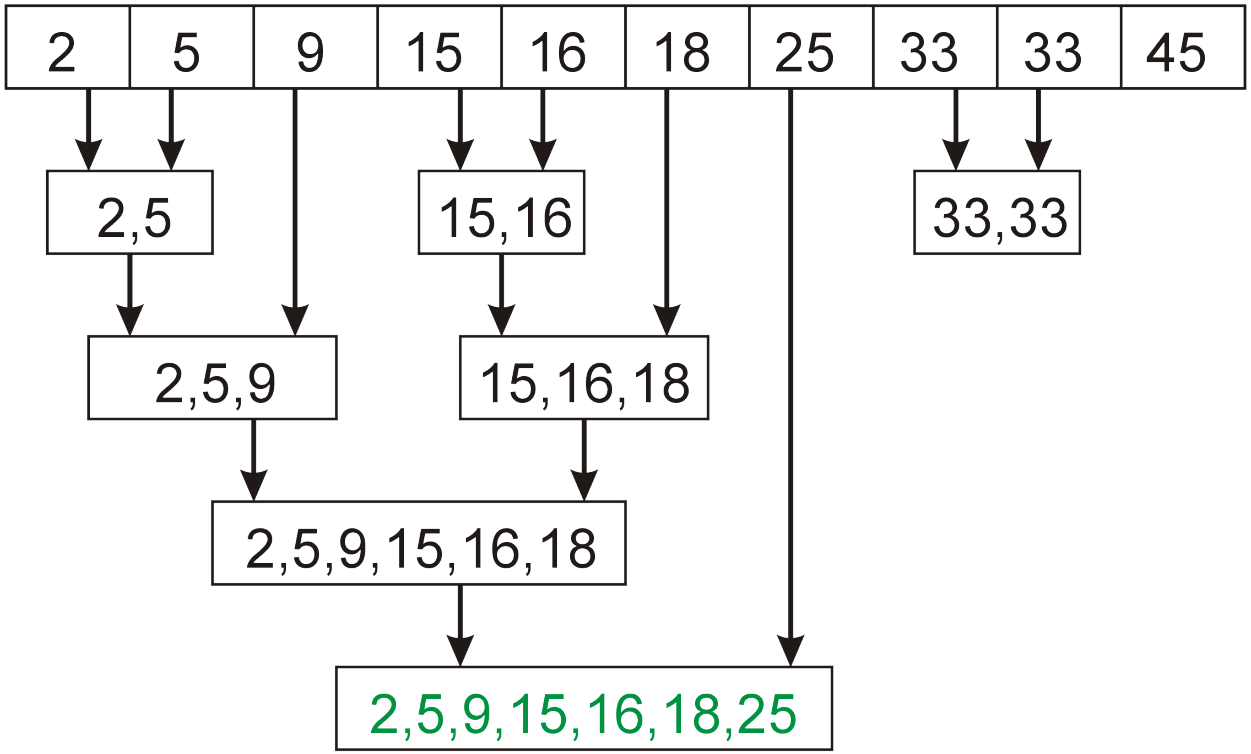
Metoda najbliższego sąsiedztwa (6)



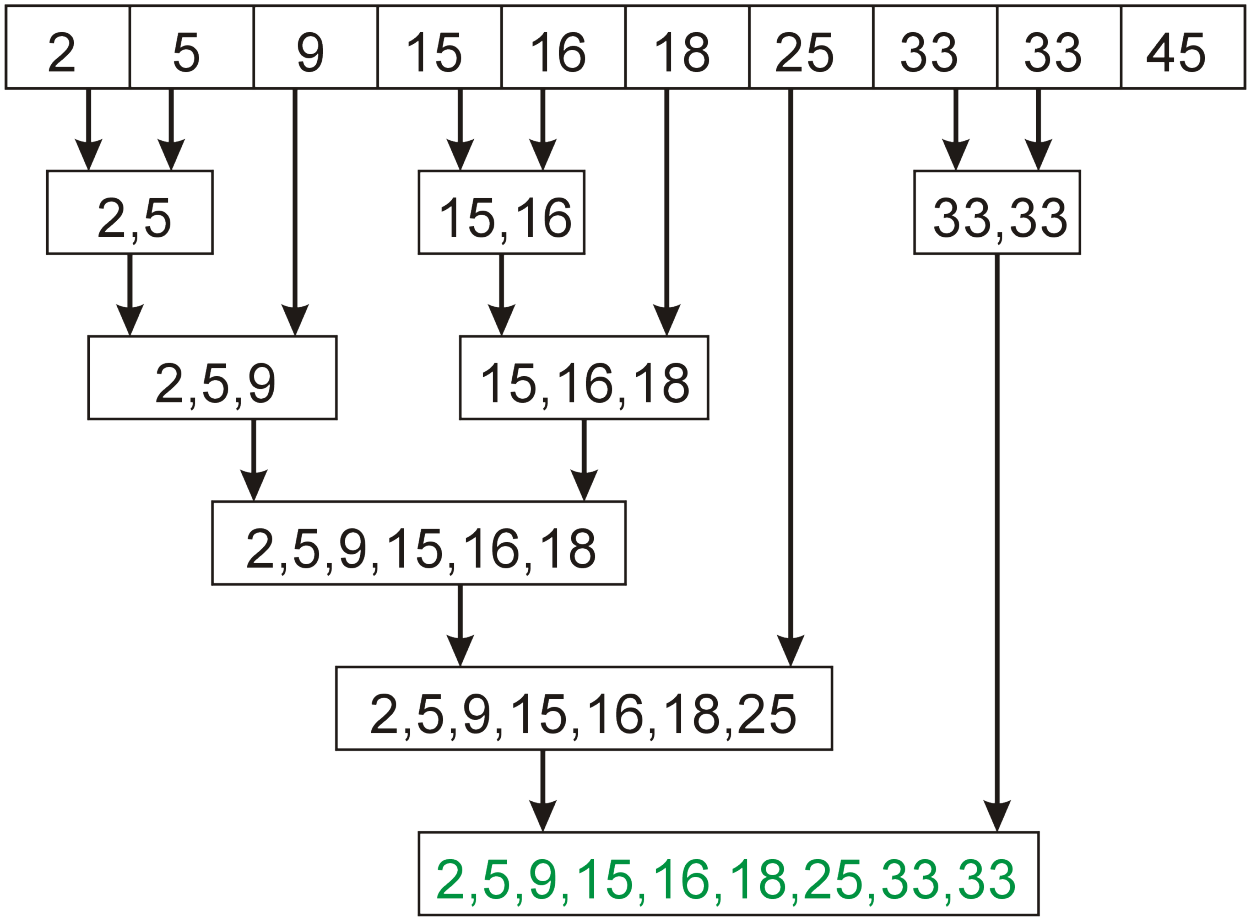
Metoda najbliższego sąsiedztwa (7)



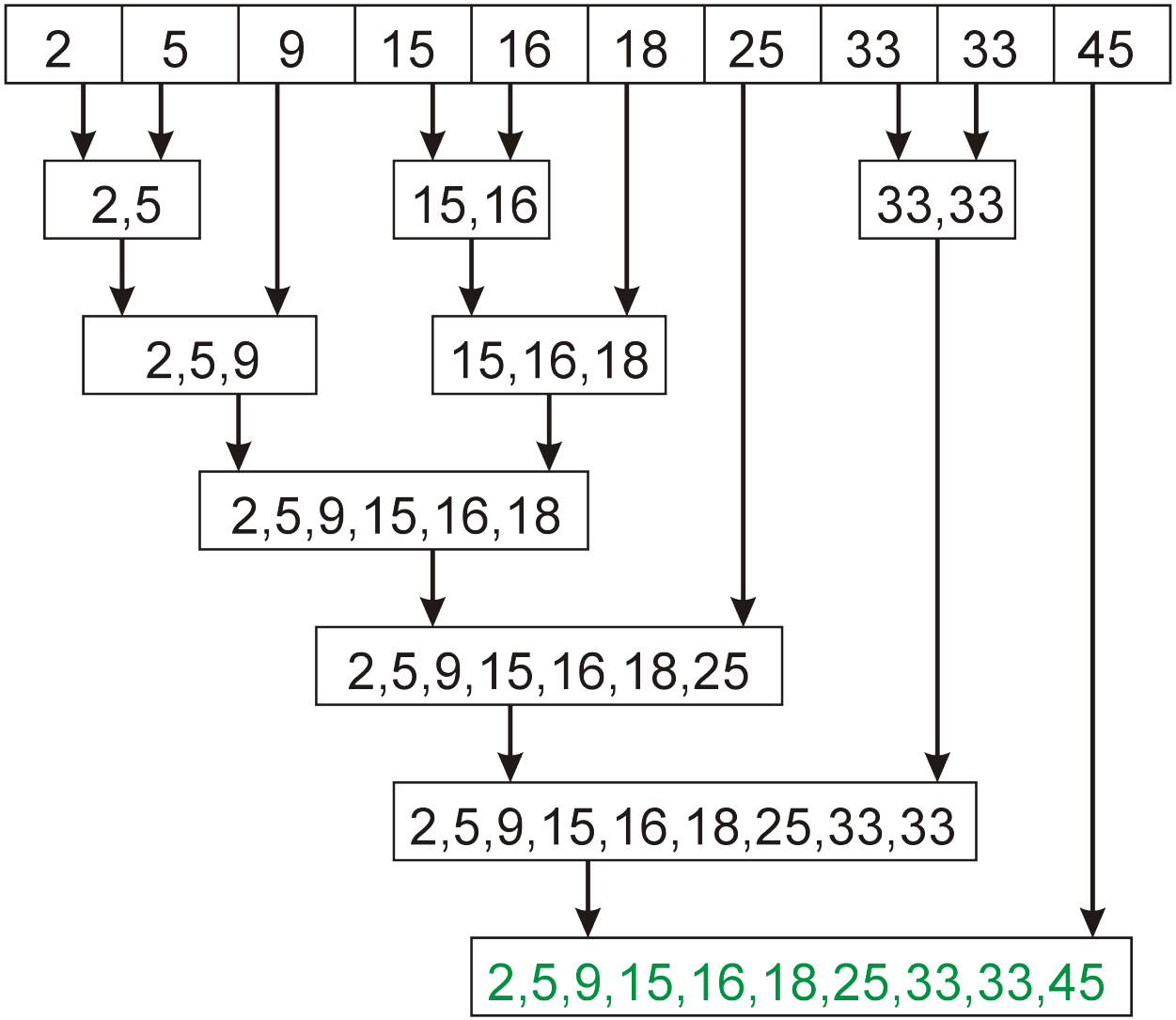
Metoda najbliższego sąsiedztwa (8)



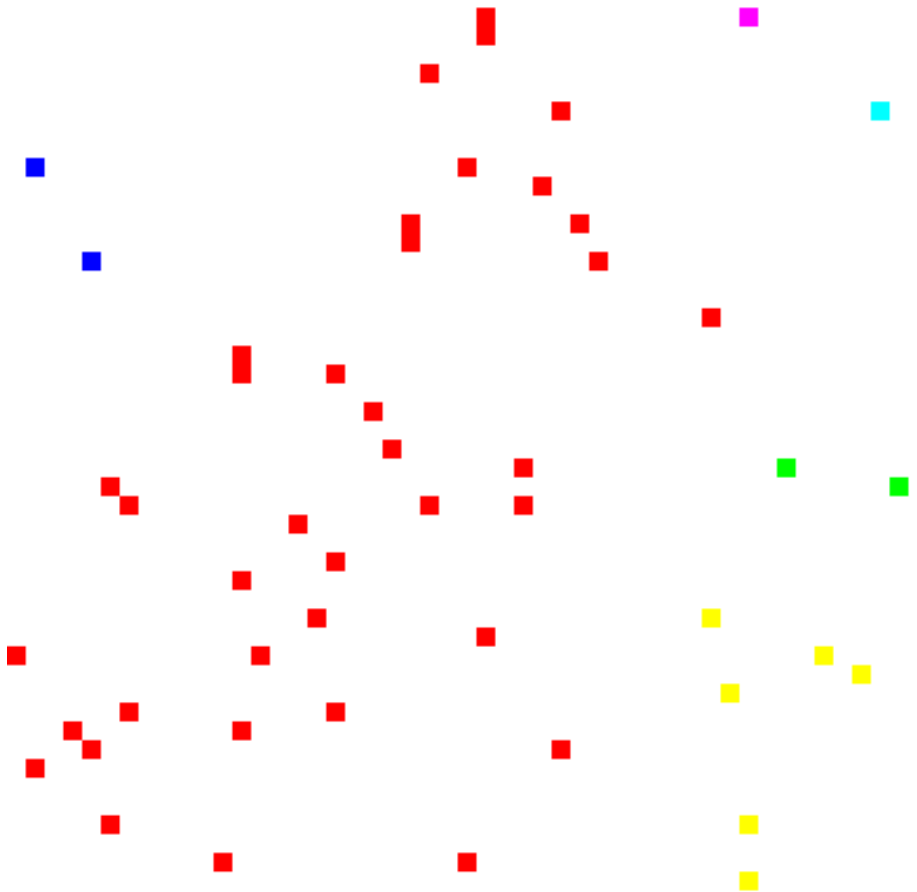
Metoda najbliższego sąsiedztwa (9)



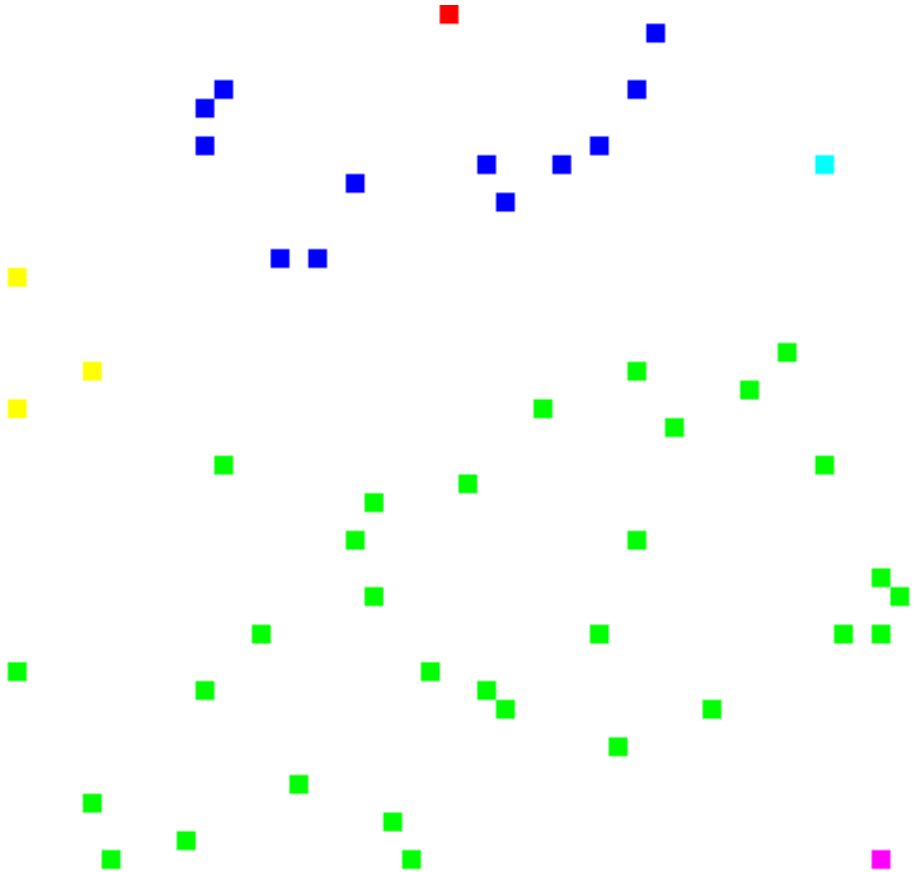
Metoda najbliższego sąsiedztwa (10)



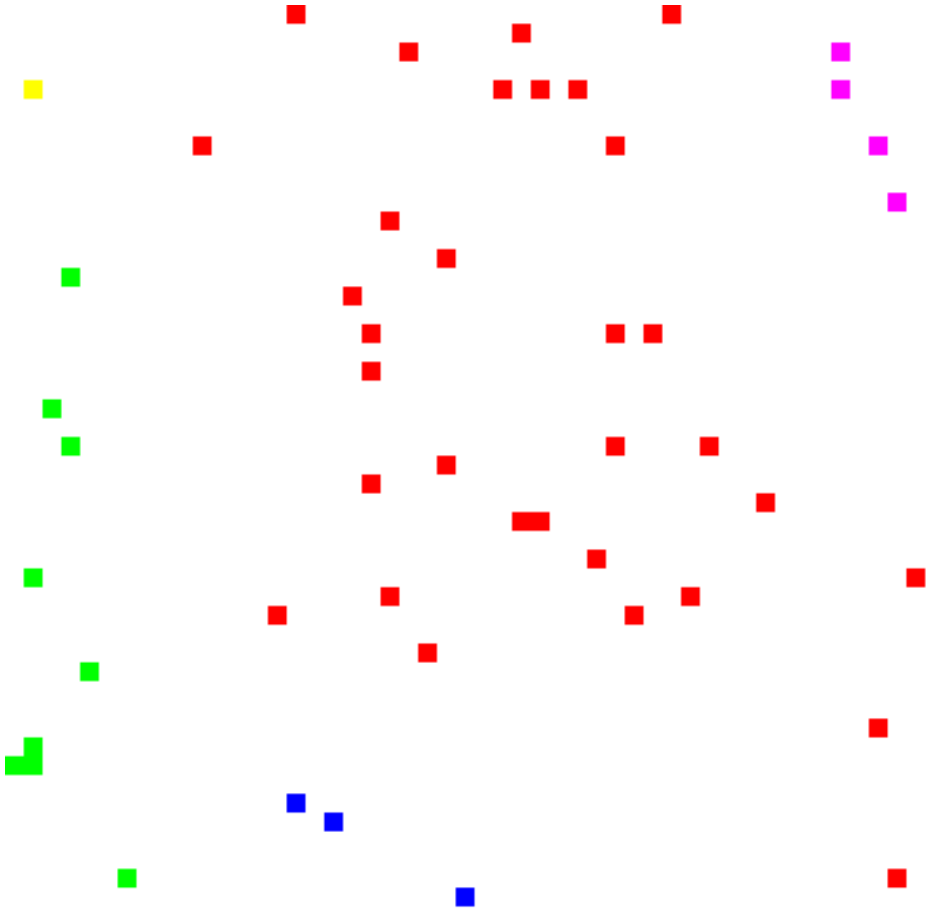
Metoda najbliższego sąsiedztwa - 50 pkt., 6 grup (1)



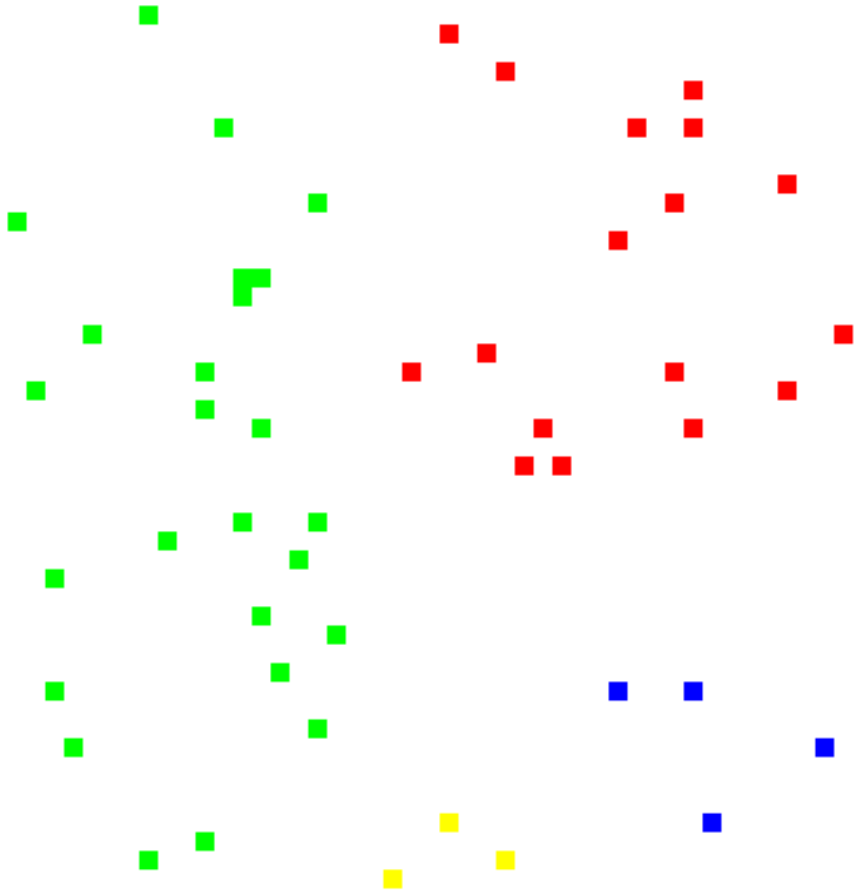
Metoda najbliższego sąsiedztwa - 50 pkt., 6 grup (2)



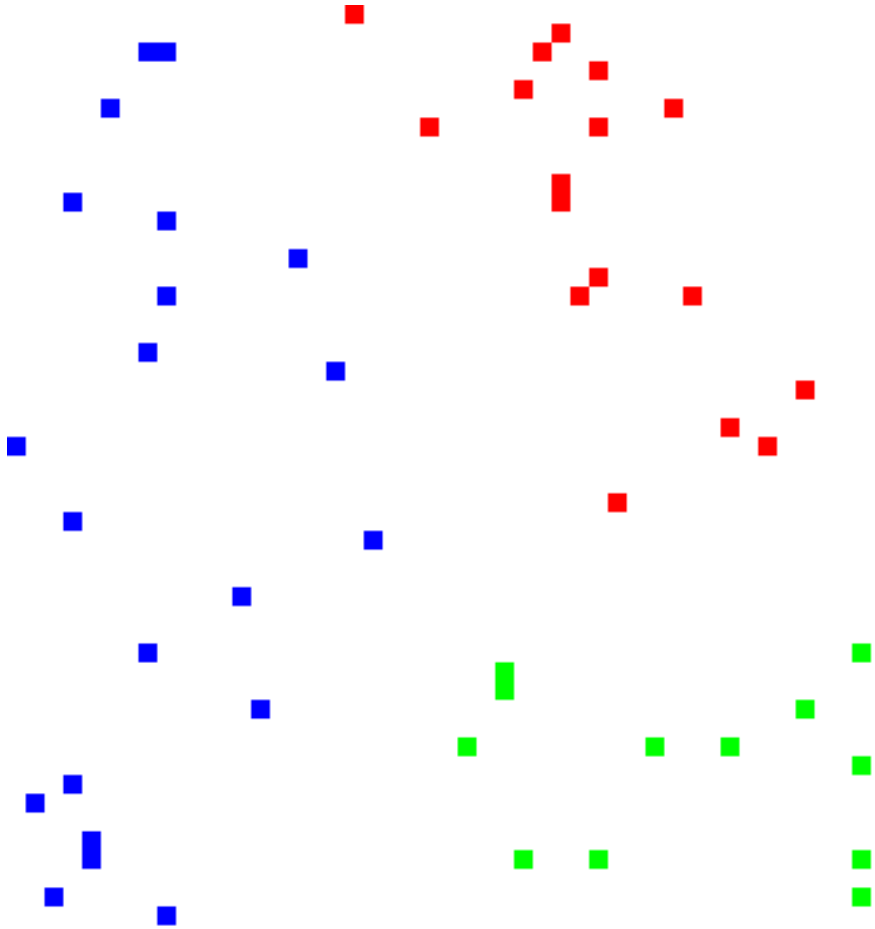
Metoda najbliższego sąsiedztwa - 50 pkt., 5 grup (3)



Metoda najbliższego sąsiedztwa - 50 pkt., 4 grupy (4)



Metoda najbliższego sąsiedztwa - 50 pkt., 3 grupy (5)



Algorytm *k*-średnich (*k*-means) (1)

Algorytm *k-średnich* jest prostym i efektywnym algorytmem znajdowania grupy w danych. Algorytm postępuje w następujący posób:

- *Krok 1:* Zapytaj użytkownika, na ile grup (k) zbiór danych powinien zostać podzielony.
- *Krok 2:* Losowo przypisz k rekordów jako początkowe środki grup.
- *Krok 3:* Dla każdego rekordu znajdź najbliższy środek grupy.
- *Krok 4:* Dla każdej z k grup znajdź *centroid* grupy i uaktualnij położenie każdego środka grupy jako nową wartość centroidu.
- *Krok 5:* Powtarzaj kroki od 3 do 5 do zbieżności lub zakończenia.

Algorytm *k*-średnich (*k*-means) (1)

Algorytm *k-średnich* jest prostym i efektywnym algorytmem znajdowania grupy w danych. Algorytm postępuje w następujący posób:

- *Krok 1:* Zapytaj użytkownika, na ile grup (*k*) zbiór danych powinien zostać podzielony.
- *Krok 2:* Losowo przypisz *k* rekordów jako początkowe środki grup.
- *Krok 3:* Dla każdego rekordu znajdź najbliższy środek grupy.
- *Krok 4:* Dla każdej z *k* grup znajdź *centroid* grupy i uaktualnij położenie każdego środka grupy jako nową wartość centroidu.
- *Krok 5:* Powtarzaj kroki od 3 do 5 do zbieżności lub zakończenia.

Algorytm *k*-średnich (*k*-means) (1)

Algorytm *k-średnich* jest prostym i efektywnym algorytmem znajdowania grupy w danych. Algorytm postępuje w następujący posób:

- *Krok 1:* Zapytaj użytkownika, na ile grup (*k*) zbiór danych powinien zostać podzielony.
- *Krok 2:* Losowo przypisz *k* rekordów jako początkowe środki grup.
- *Krok 3:* Dla każdego rekordu znajdź najbliższy środek grupy.
- *Krok 4:* Dla każdej z *k* grup znajdź *centroid* grupy i uaktualnij położenie każdego środka grupy jako nową wartość centroidu.
- *Krok 5:* Powtarzaj kroki od 3 do 5 do zbieżności lub zakończenia.

Algorytm *k*-średnich (*k*-means) (1)

Algorytm *k-średnich* jest prostym i efektywnym algorytmem znajdowania grupy w danych. Algorytm postępuje w następujący posób:

- *Krok 1:* Zapytaj użytkownika, na ile grup (*k*) zbiór danych powinien zostać podzielony.
- *Krok 2:* Losowo przypisz *k* rekordów jako początkowe środki grup.
- *Krok 3:* Dla każdego rekordu znajdź najbliższy środek grupy.
- *Krok 4:* Dla każdej z *k* grup znajdź *centroid* grupy i uaktualnij położenie każdego środka grupy jako nową wartość centroidu.
- *Krok 5:* Powtarzaj kroki od 3 do 5 do zbieżności lub zakończenia.

Algorytm *k*-średnich (*k*-means) (1)

Algorytm *k-średnich* jest prostym i efektywnym algorytmem znajdowania grupy w danych. Algorytm postępuje w następujący posób:

- *Krok 1:* Zapytaj użytkownika, na ile grup (*k*) zbiór danych powinien zostać podzielony.
- *Krok 2:* Losowo przypisz *k* rekordów jako początkowe środki grup.
- *Krok 3:* Dla każdego rekordu znajdź najbliższy środek grupy.
- *Krok 4:* Dla każdej z *k* grup znajdź *centroid* grupy i uaktualnij położenie każdego środka grupy jako nową wartość centroidu.
- *Krok 5:* Powtarzaj kroki od 3 do 5 do zbieżności lub zakończenia.

Algorytm *k*-średnich (*k*-means) (1)

Algorytm *k-średnich* jest prostym i efektywnym algorytmem znajdowania grupy w danych. Algorytm postępuje w następujący posób:

- *Krok 1:* Zapytaj użytkownika, na ile grup (k) zbiór danych powinien zostać podzielony.
- *Krok 2:* Losowo przypisz k rekordów jako początkowe środki grup.
- *Krok 3:* Dla każdego rekordu znajdź najbliższy środek grupy.
- *Krok 4:* Dla każdej z k grup znajdź *centroid* grupy i uaktualnij położenie każdego środka grupy jako nową wartość centroidu.
- *Krok 5:* Powtarzaj kroki od 3 do 5 do zbieżności lub zakończenia.

Algorytm k-średnich (2)

Odległość euklidesowa jest zwykle miarą ”bliskości” z punktu 3, chociaż i inne miary mogą być również stosowane.

Centroid grupy w kroku 4 oblicza się w następujący sposób. Załóżmy, że mamy n punktów danych (a_1, b_1, c_1),...,(a_n, b_n, c_n). *Centroid* tych punktów jest środkiem ciężkości tych punktów i znajduje się w punkcie

$$\left(\sum_{i=1}^{n} \frac{a_i}{n}, \sum_{i=1}^{n} \frac{b_i}{n}, \sum_{i=1}^{n} \frac{c_i}{n} \right).$$

Algorytm k-średnich (2)

Odległość euklidesowa jest zwykle miarą ”bliskości” z punktu 3, chociaż i inne miary mogą być również stosowane.

Centroid grupy w kroku 4 oblicza się w następujący sposób. Załóżmy, że mamy n punktów danych (a_1, b_1, c_1),...,(a_n, b_n, c_n). *Centroid* tych punktów jest środkiem ciężkości tych punktów i znajduje się w punkcie

$$\left(\sum_{i=1}^{n} \frac{a_i}{n}, \sum_{i=1}^{n} \frac{b_i}{n}, \sum_{i=1}^{n} \frac{c_i}{n} \right).$$

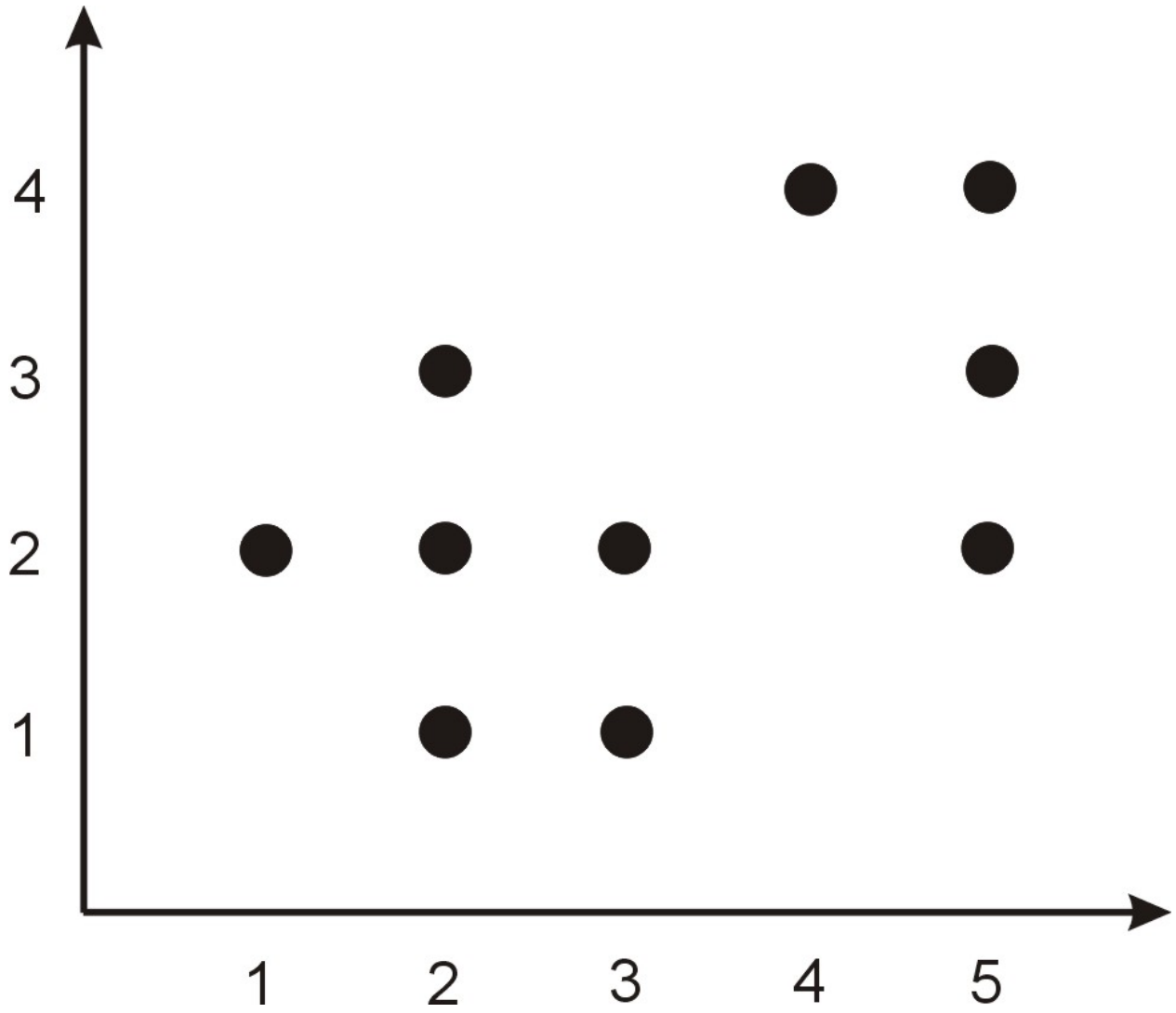
Algorytm k-średnich (3)

Algorytm kończy działanie, gdy centroidy już się nie zmieniają lub gdy spełnione jest pewne kryterium zbieżności, takie jak brak istotnego zmniejszania *sumarycznego błędu kwadratowego*:

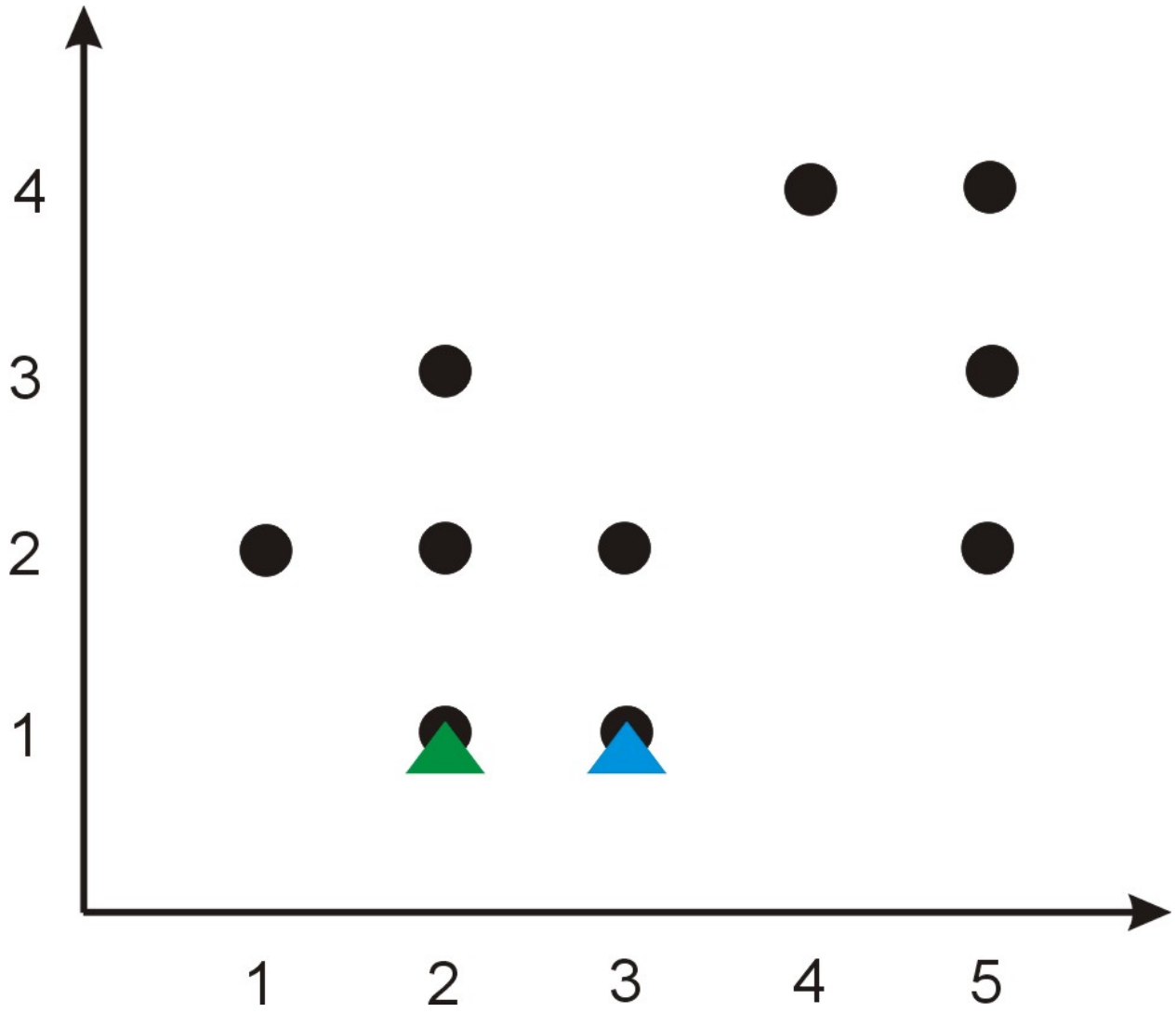
$$SSE = \sum_{i=1}^{n} \sum_{p \in C_i} d(p, m_i)^2,$$

gdzie $p \in C_i$ reprezentuje każdy punkt danych w i-tej grupie, a m_i jest centroidem i-tej grupy.

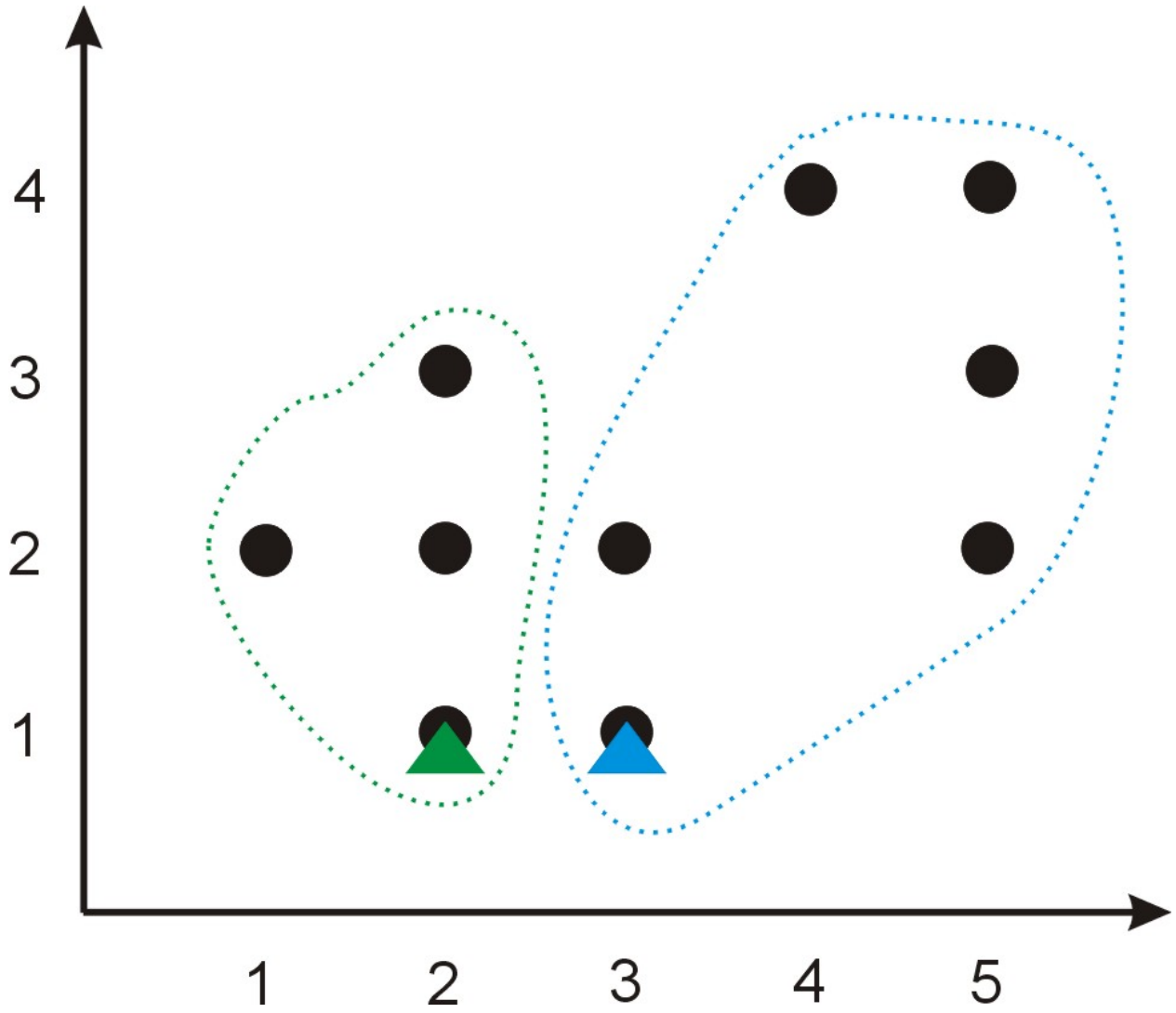
Algorytm k-średnich - przykład (1)



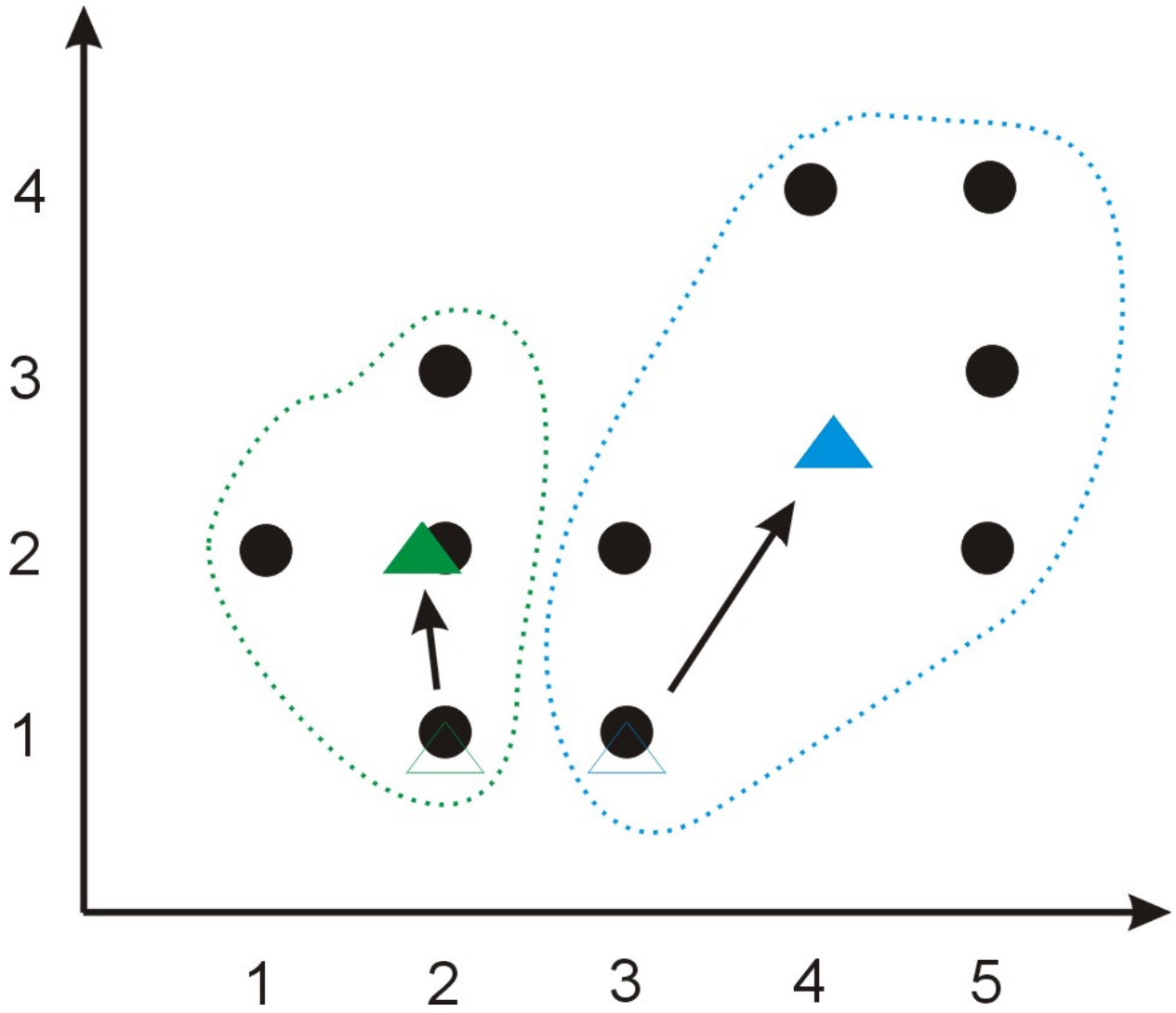
Algorytm k-średnich - przykład (2)



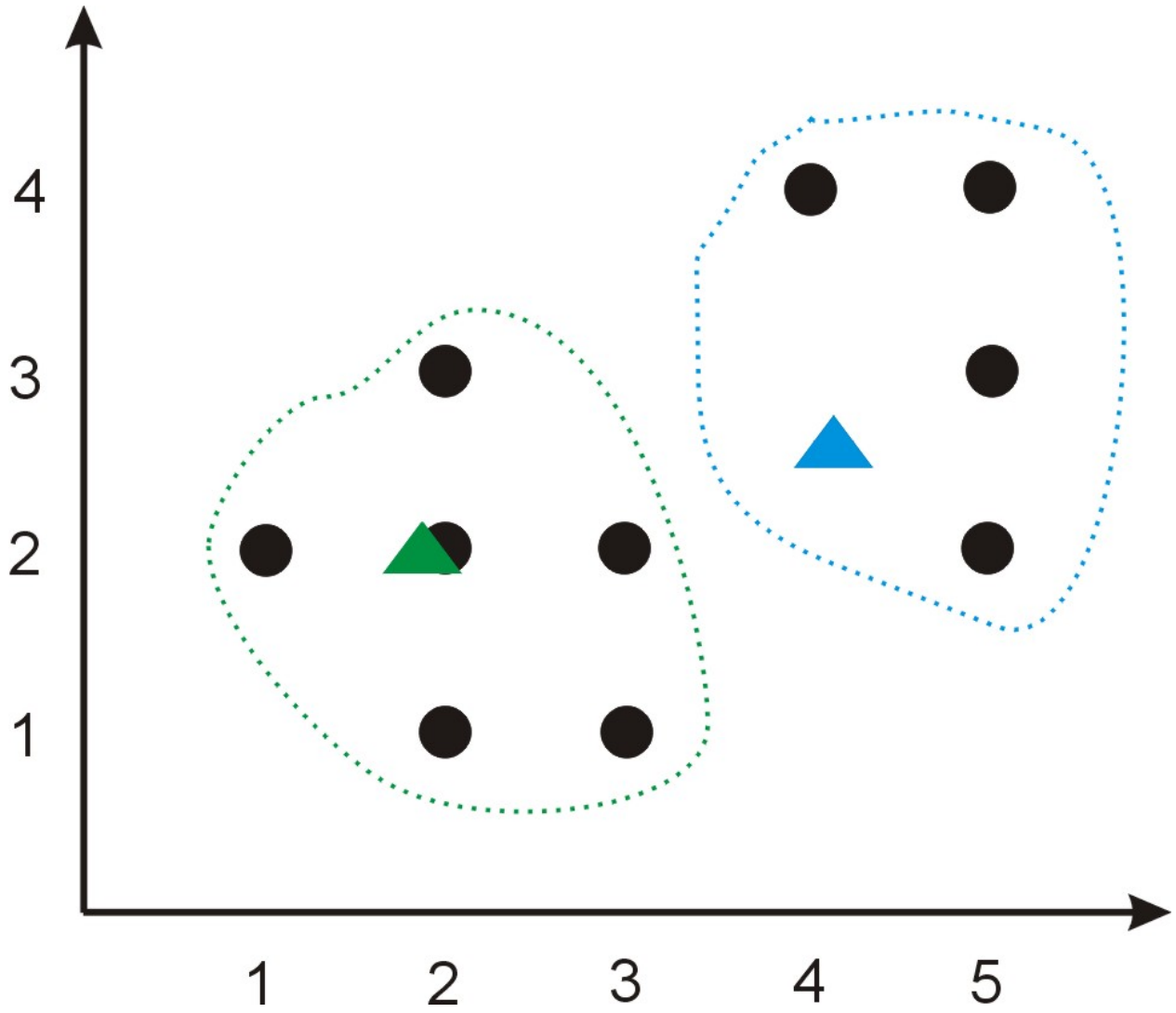
Algorytm k-średnich - przykład (3)



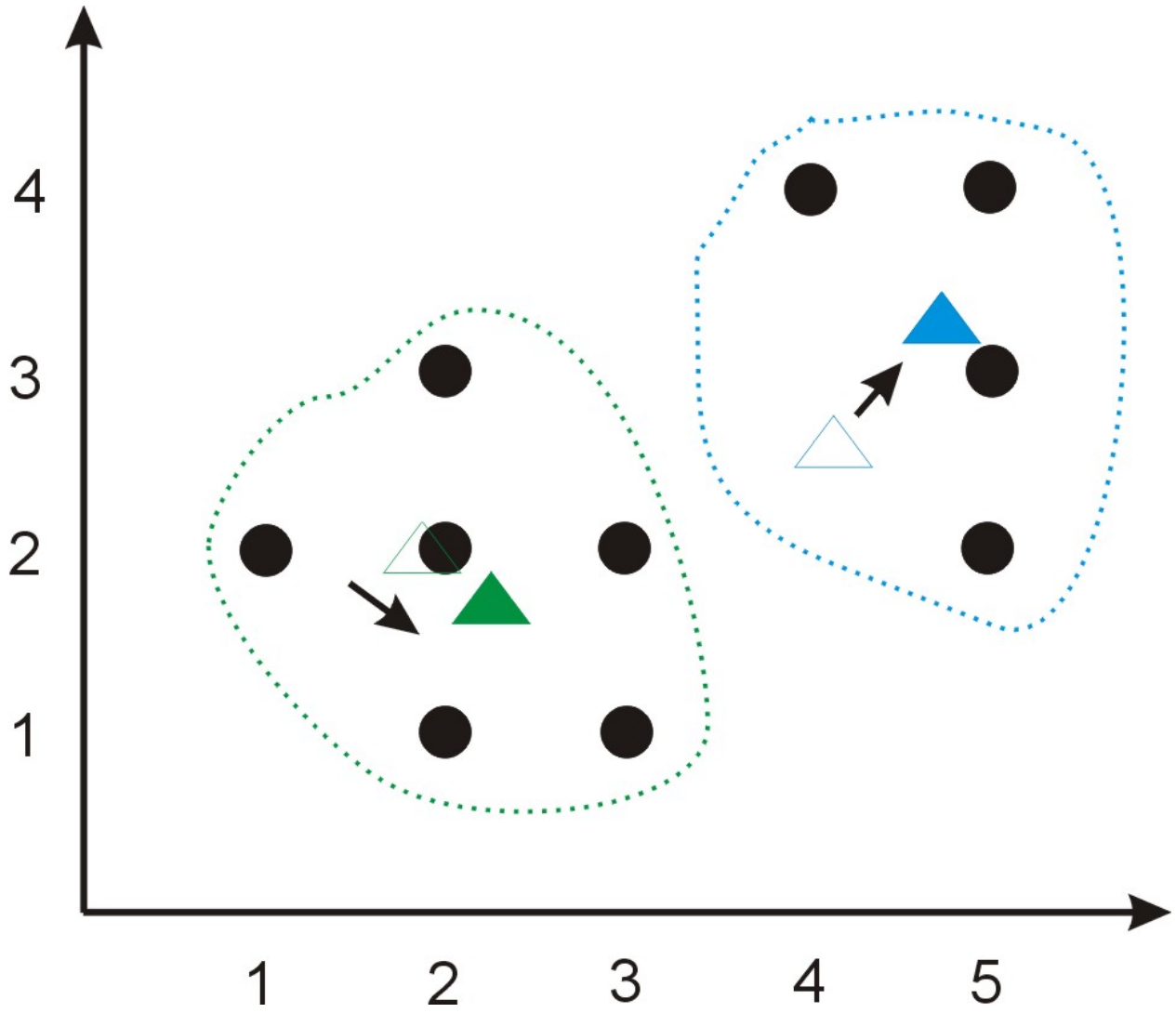
Algorytm k-średnich - przykład (4)



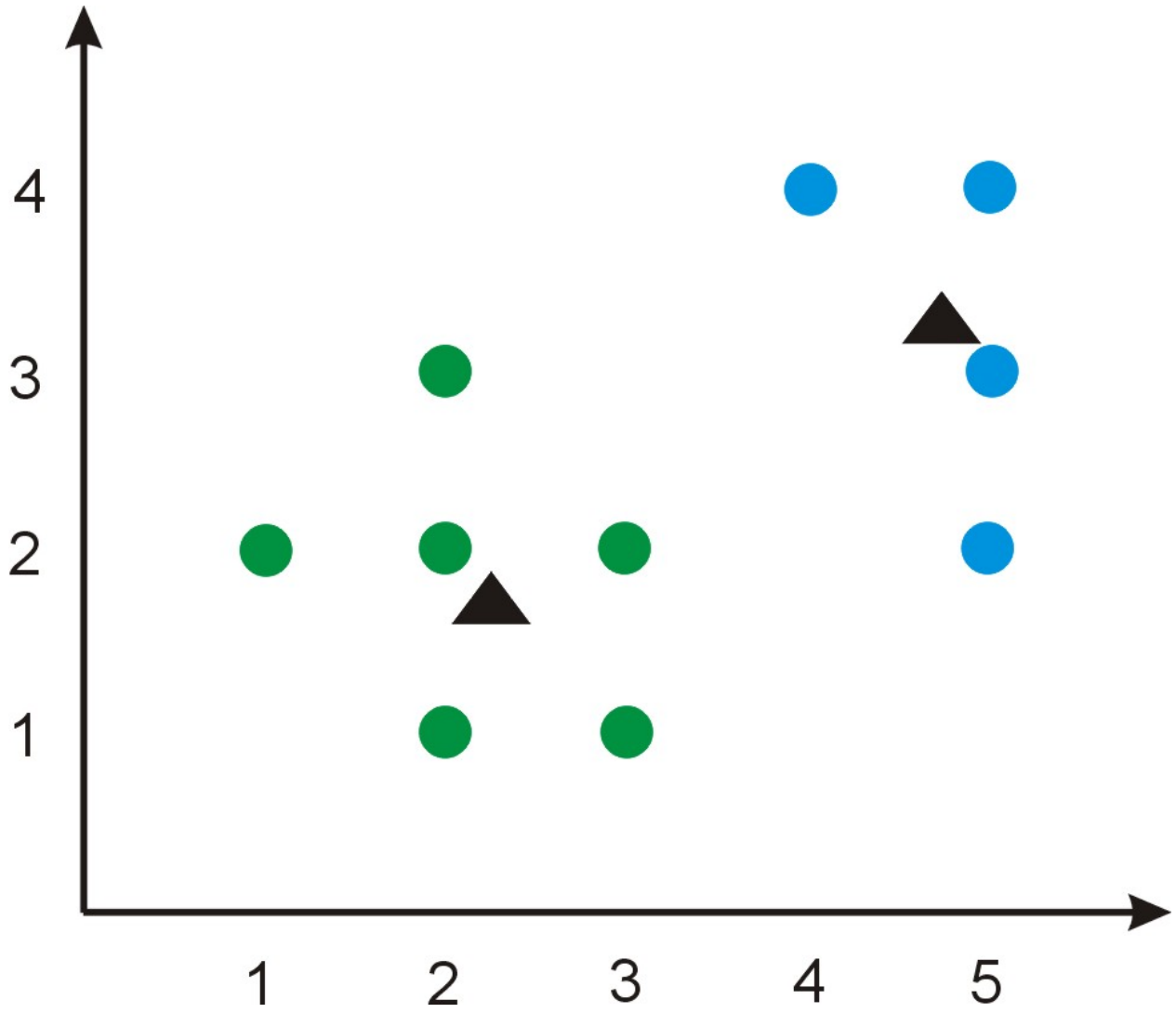
Algorytm k-średnich - przykład (5)



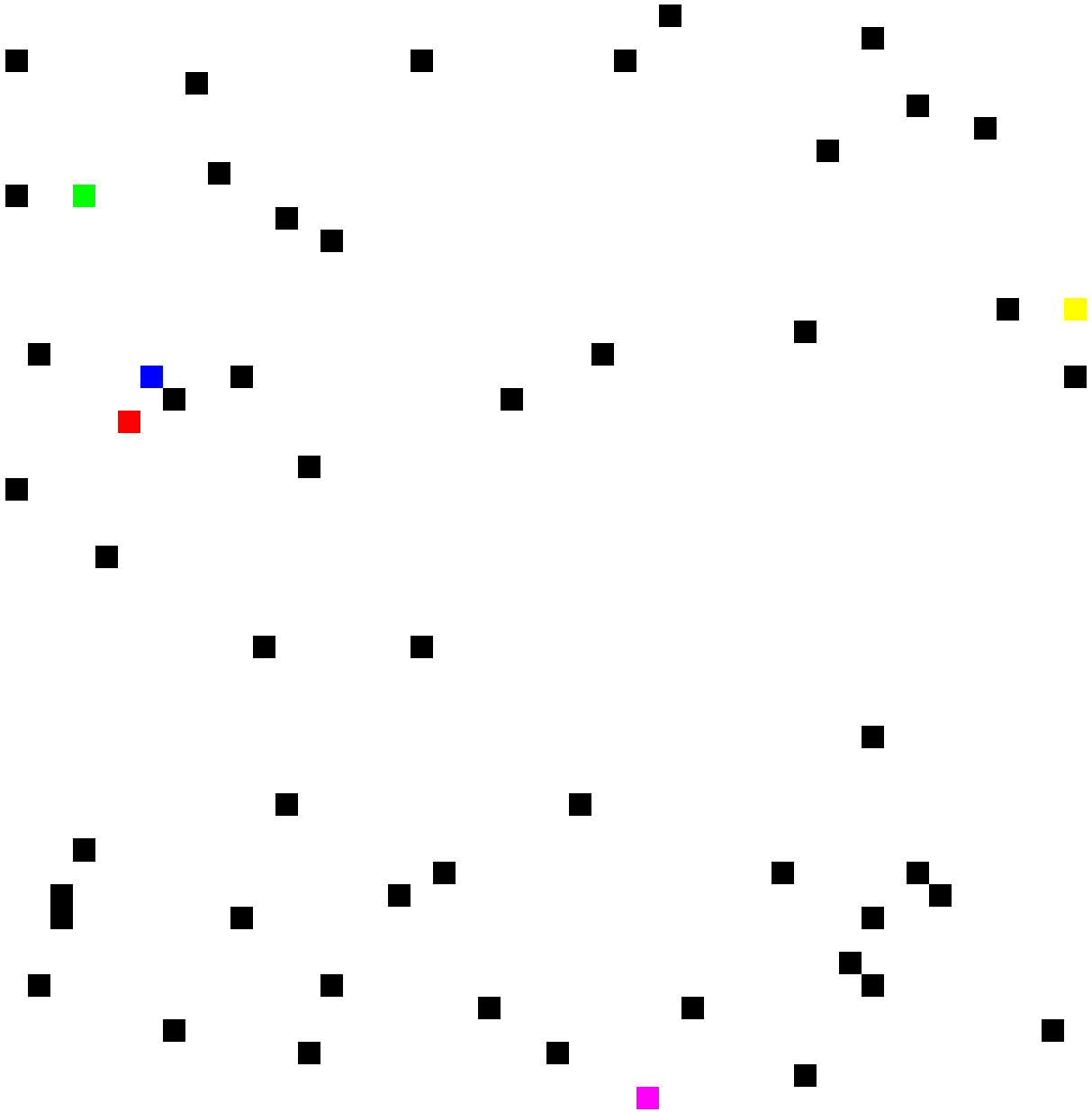
Algorytm k-średnich - przykład (6)



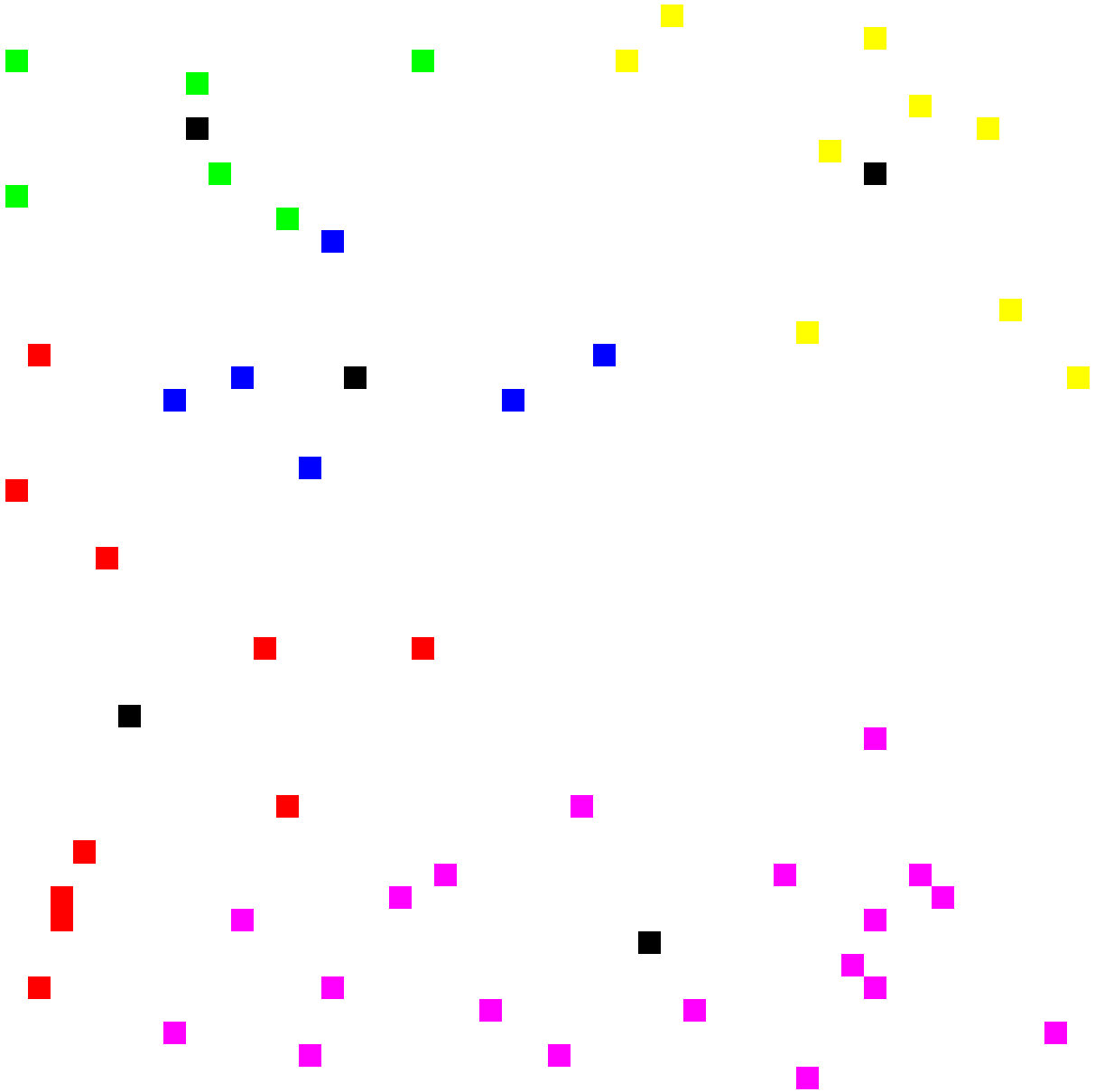
Algorytm k-średnich - przykład (7)



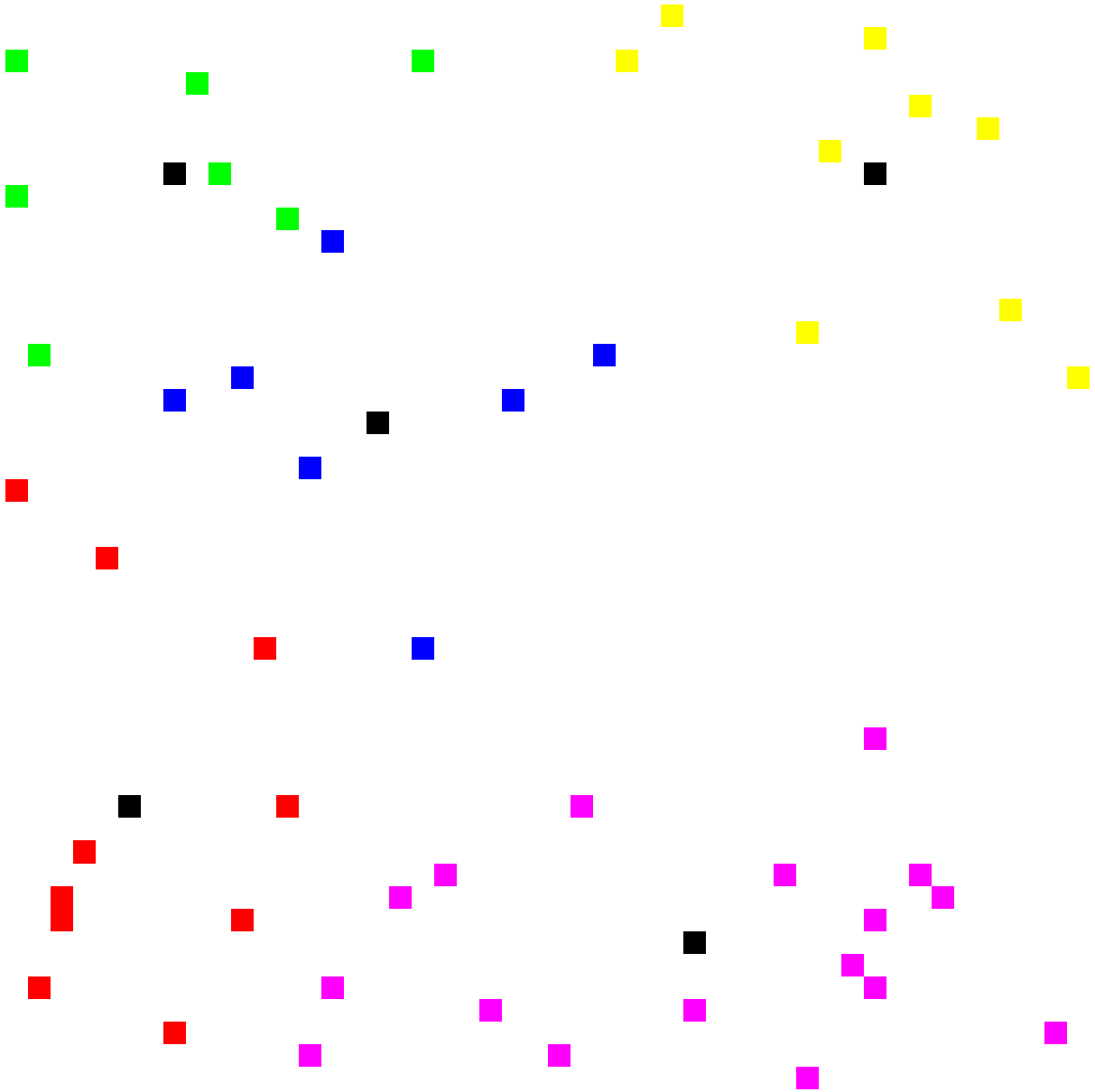
50 punktów, 5 grup, przed grupowaniem



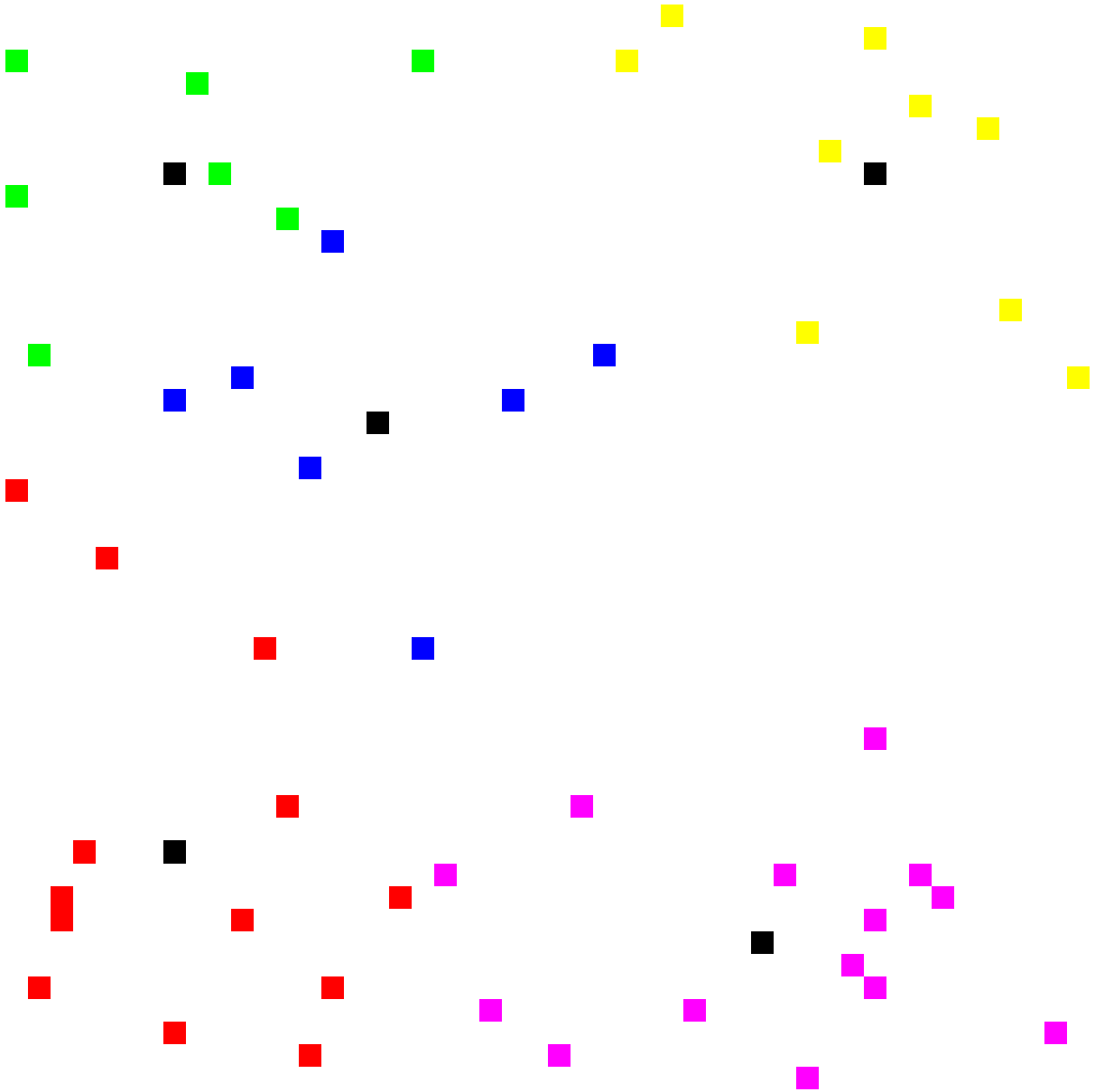
50 punktów, 5 grup, pierwsza iteracja



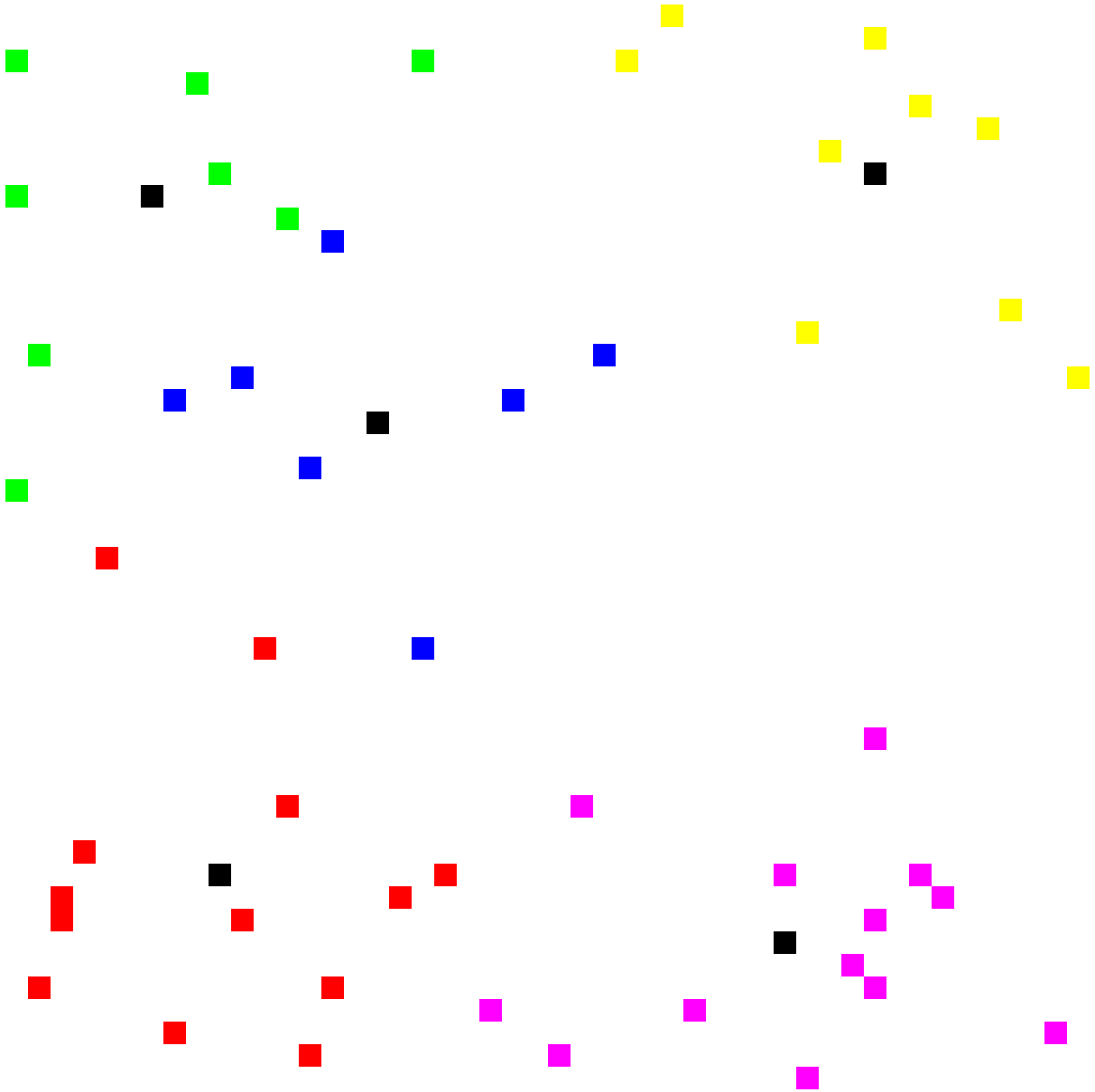
50 punktów, 5 grup, druga iteracja



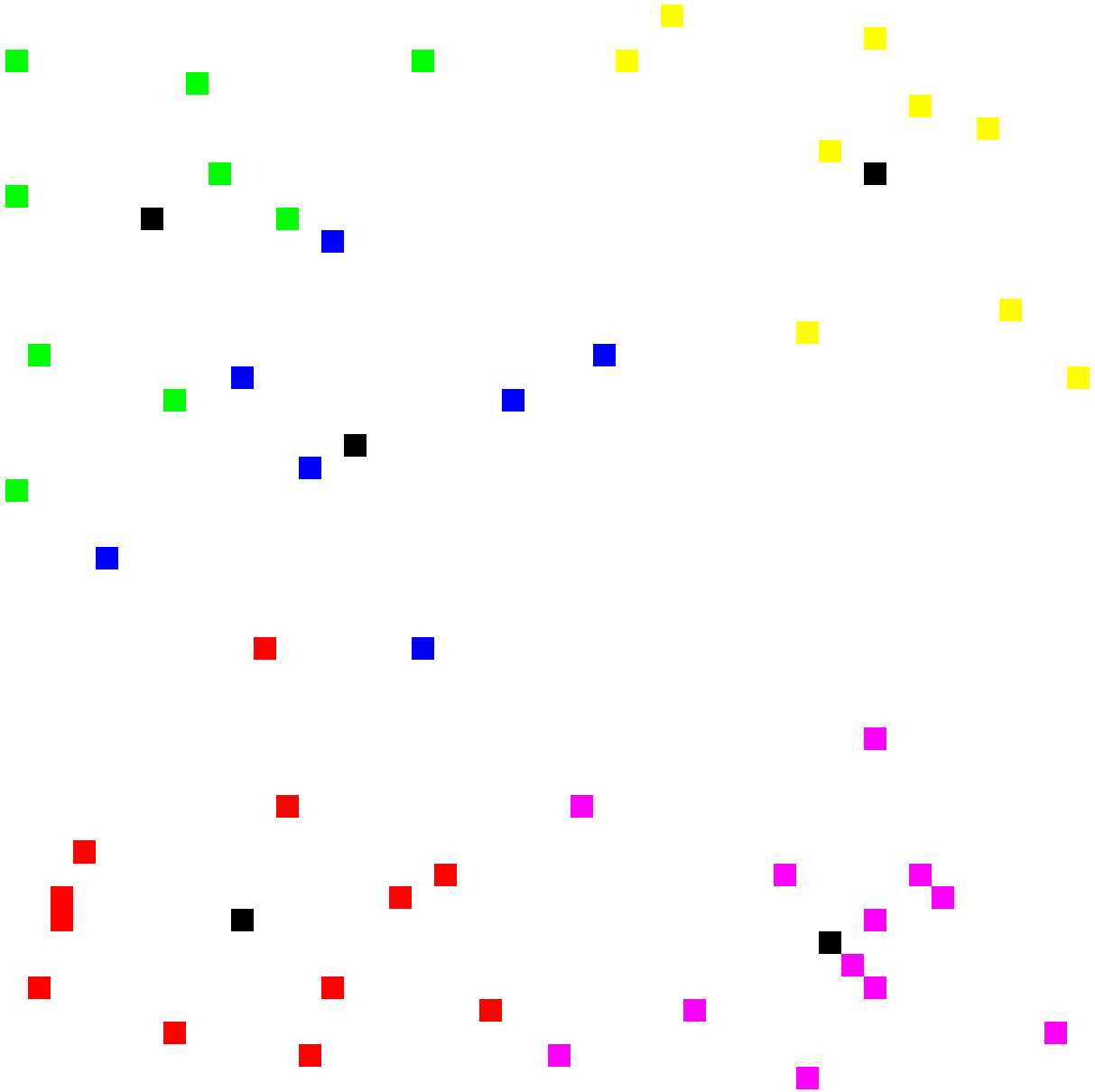
50 punktów, 5 grup, trzecia iteracja



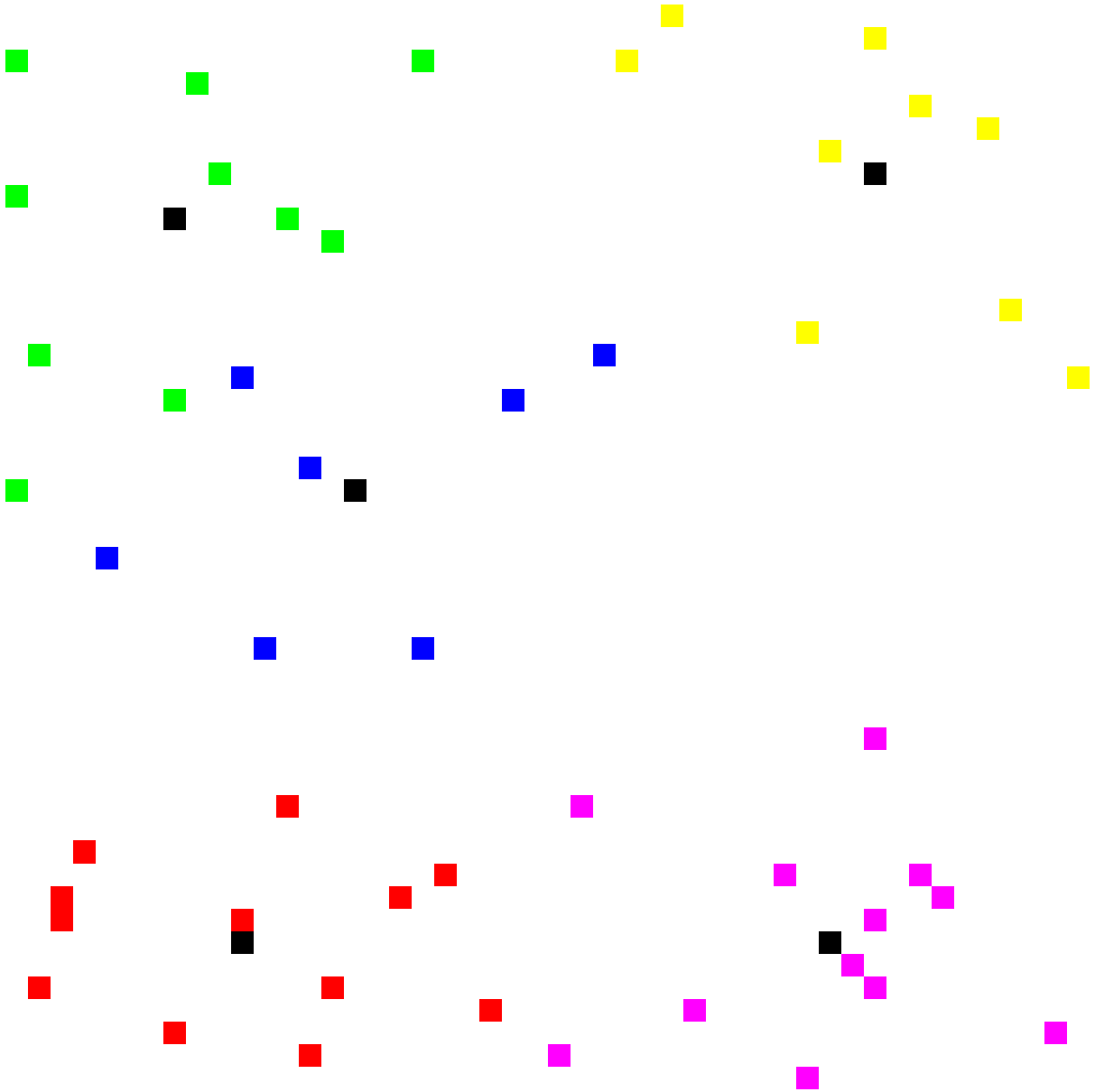
50 punktów, 5 grup, czwarta iteracja



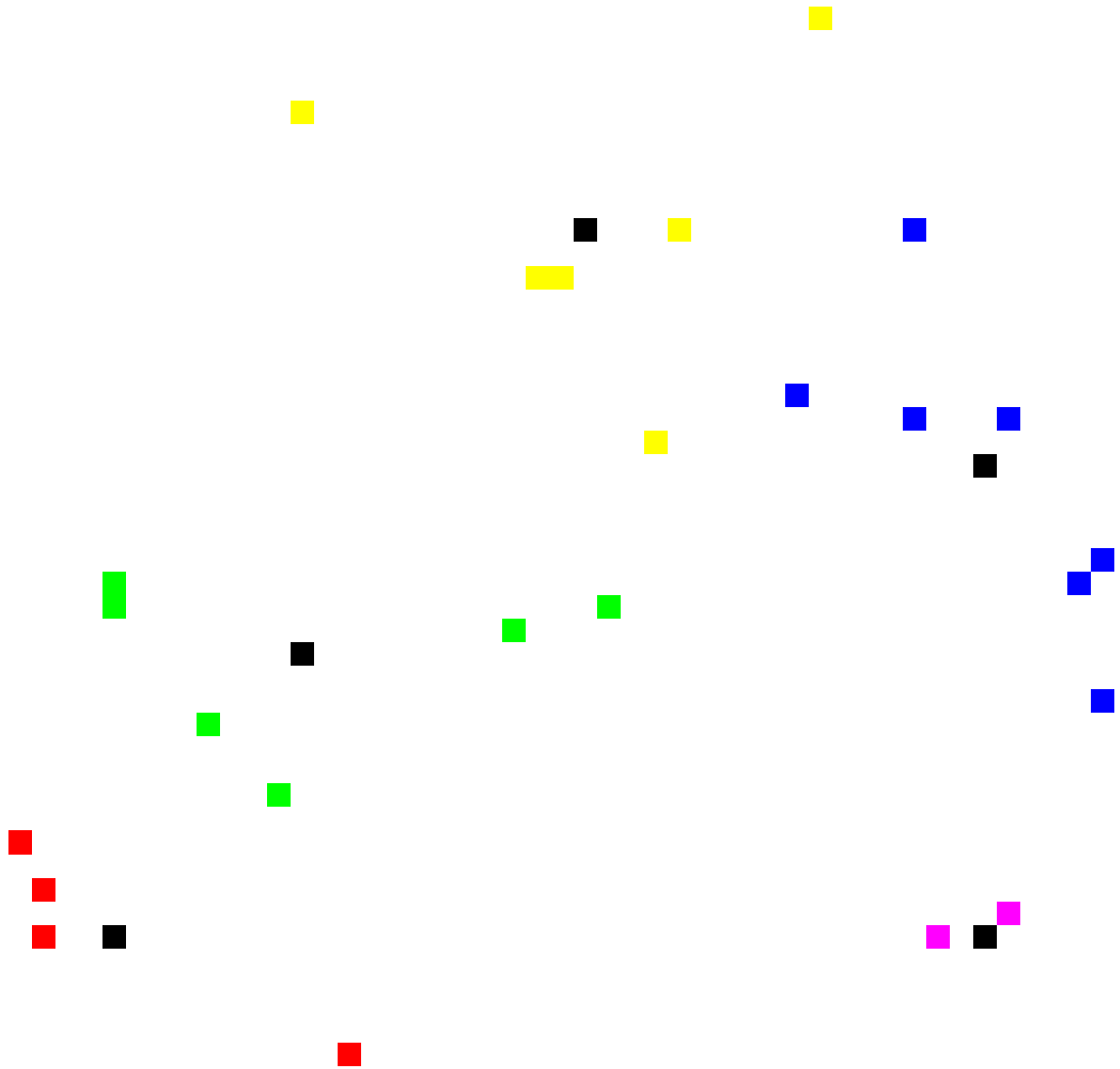
50 punktów, 5 grup, piąta iteracja



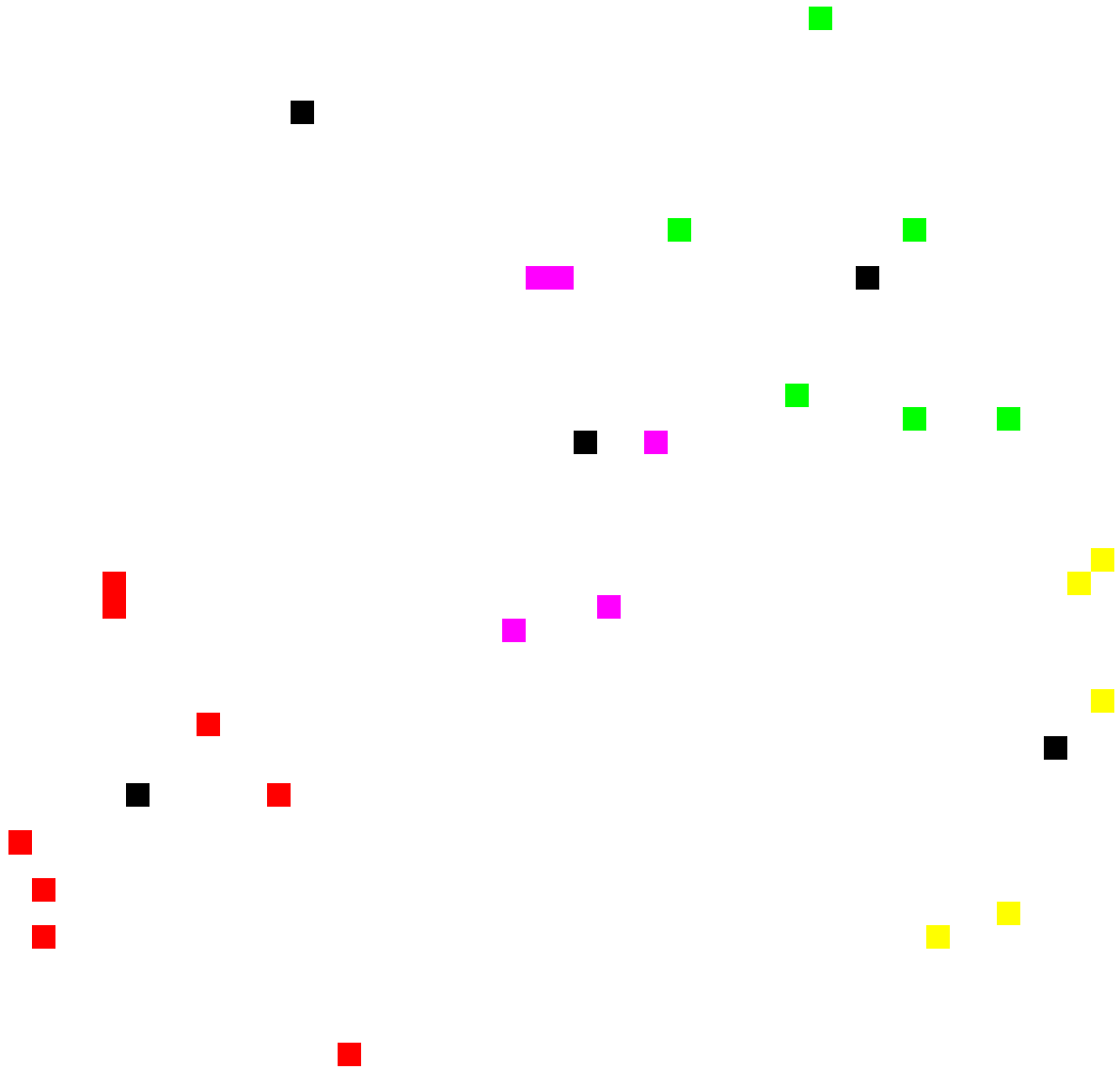
50 punktów, 5 grup, szósta iteracja



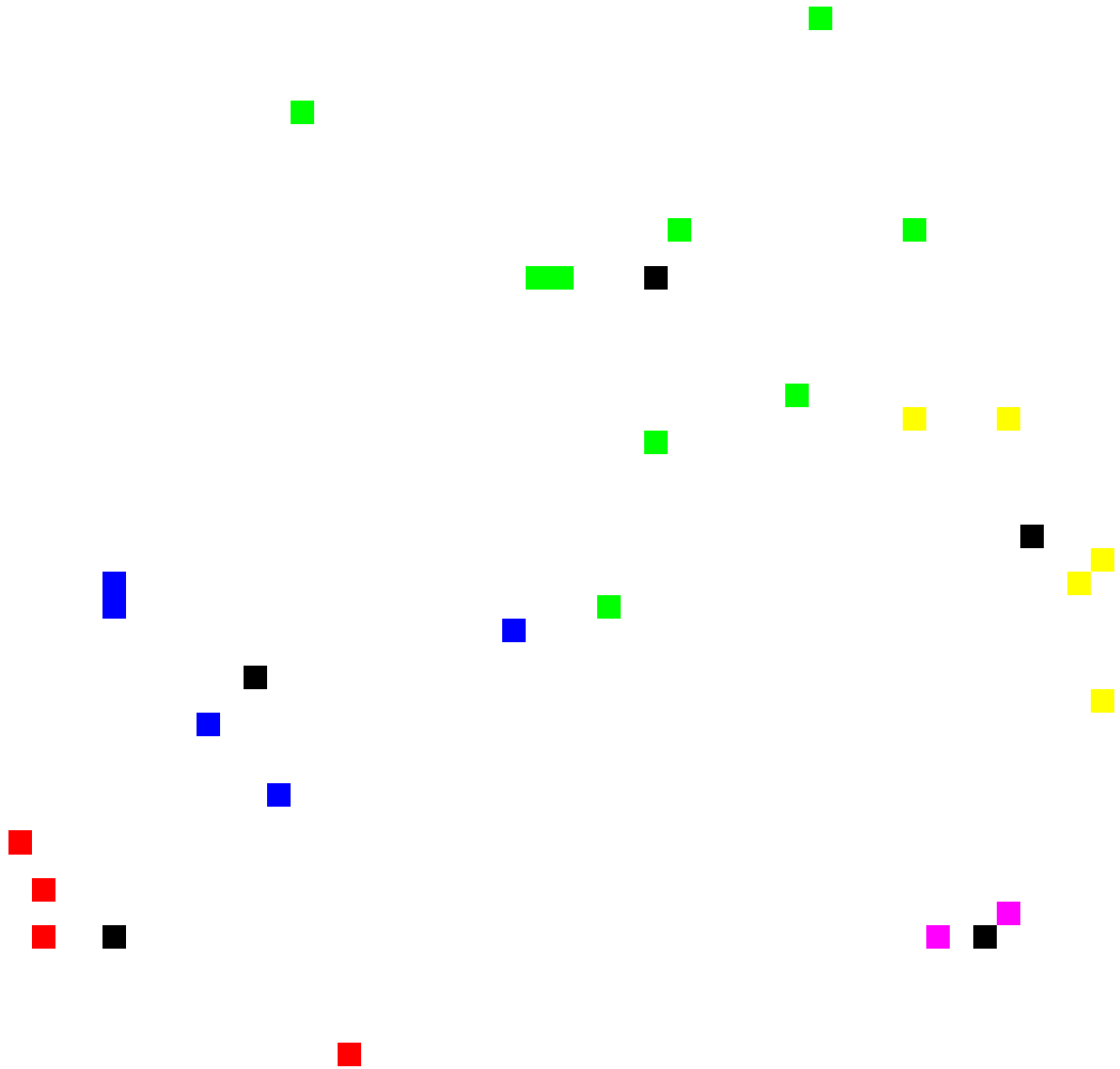
25 punktów, 5 grup, zmienne położenie środków (1)



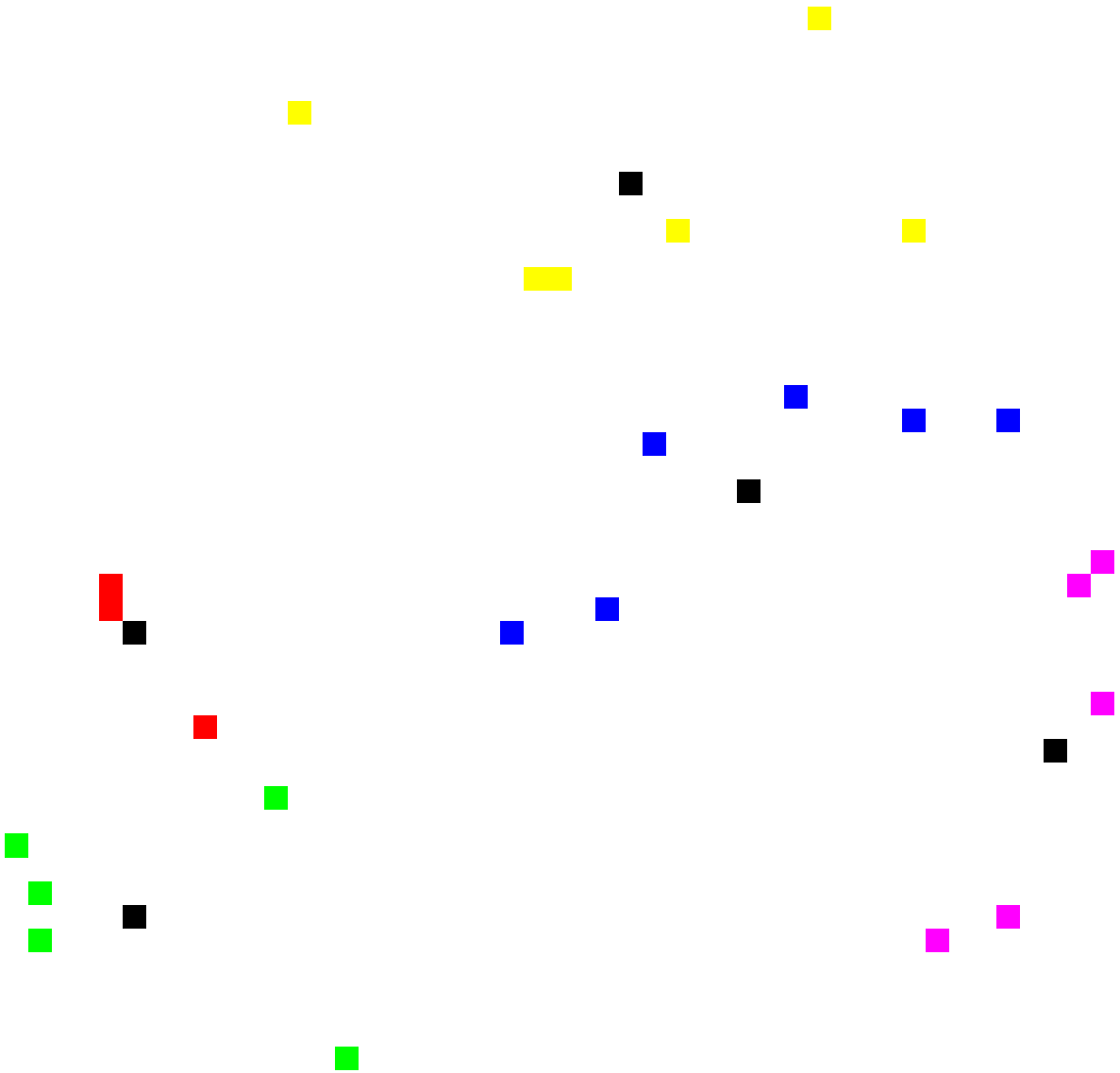
25 punktów, 5 grup, zmienne położenie środków (2)



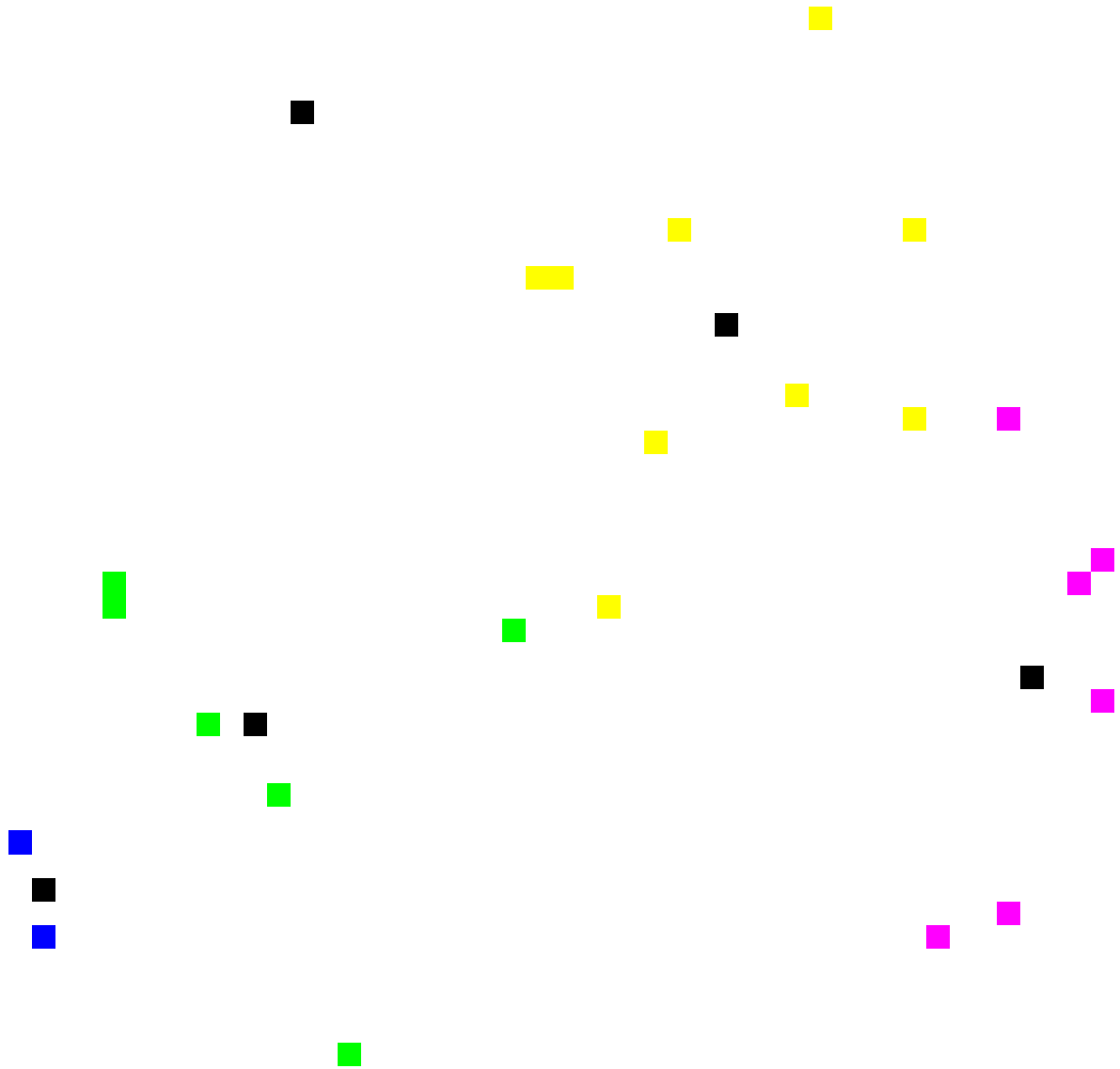
25 punktów, 5 grup, zmienne położenie środków (3)



25 punktów, 5 grup, zmienne położenie środków (4)



25 punktów, 5 grup, zmienne położenie środków (5)



Ograniczenia

Algorytm k-średnich nie może zagwarantować znalezienia minimum globalnego SSE, zamiast tego często zatrzymuje się w minimum lokalnym.

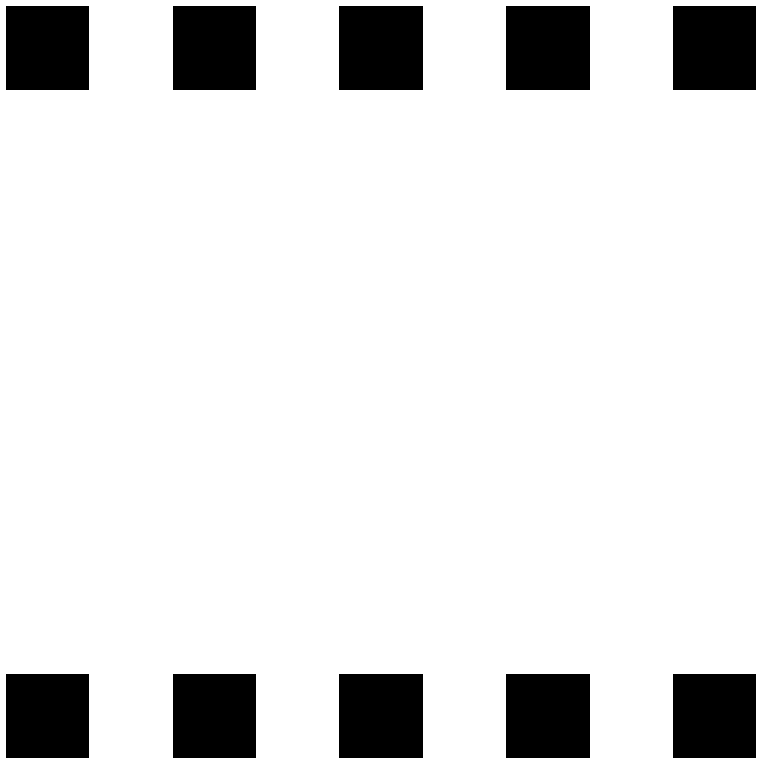
Innym możliwym problemem algorytmu k-średnich jest: *Kto decyduje, ile grup należy szukać?* Czyli, kto określa wartość k?

Ograniczenia

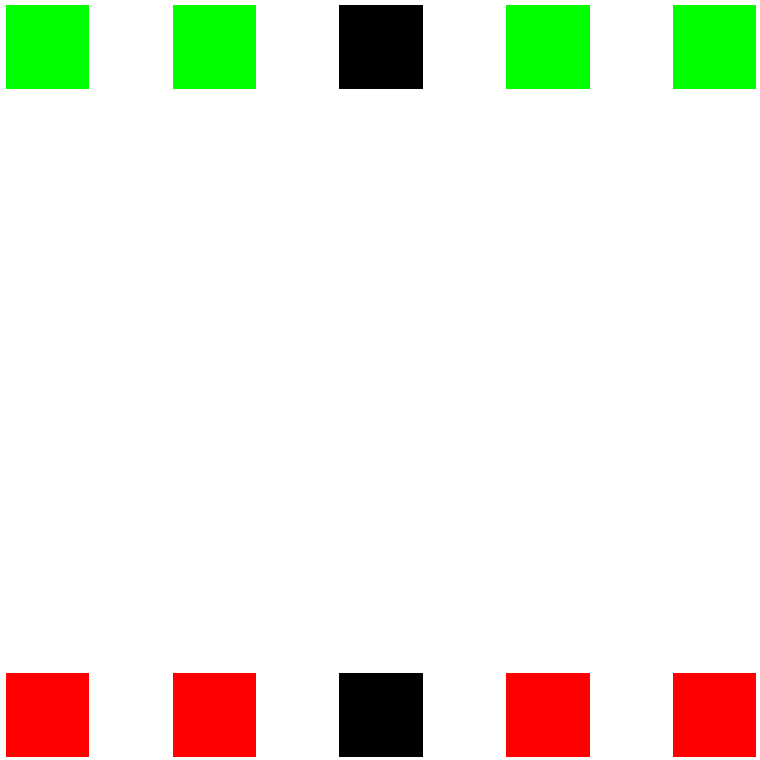
Algorytm k-średnich nie może zagwarantować znalezienia minimum globalnego SSE, zamiast tego często zatrzymuje się w minimum lokalnym.

Innym możliwym problemem algorytmu k-średnich jest: *Kto decyduje, ile grup należy szukać?* Czyli, kto określa wartość k?

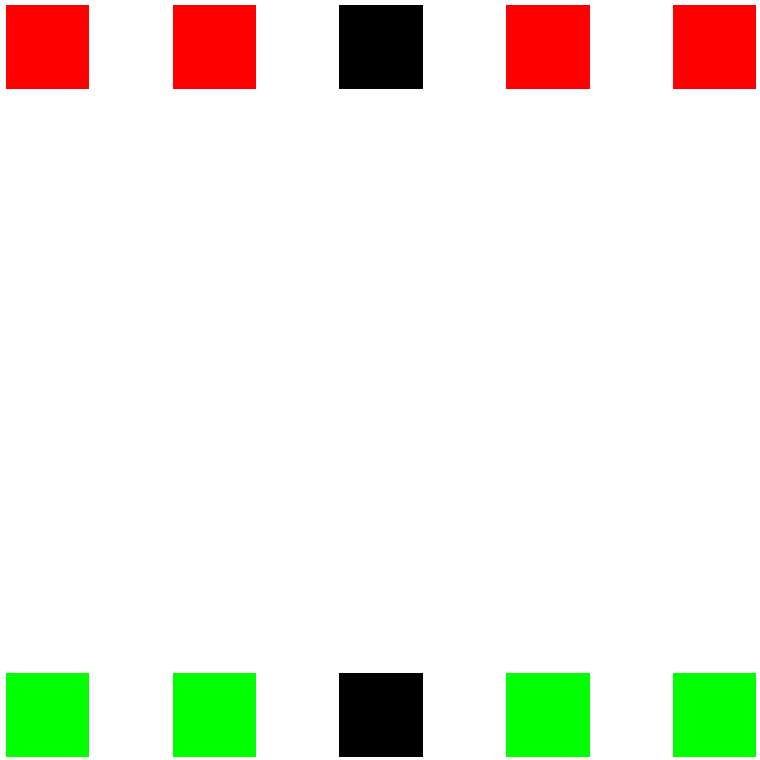
Algorytm k-średnich - minimum lokalne SSE (1)



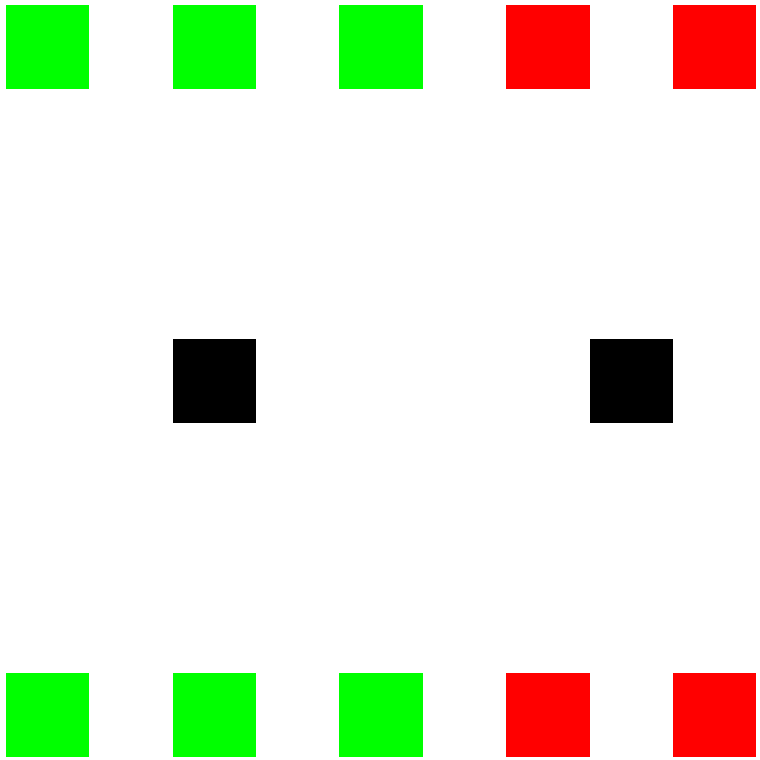
Algorytm k-średnich - minimum lokalne SSE (2)



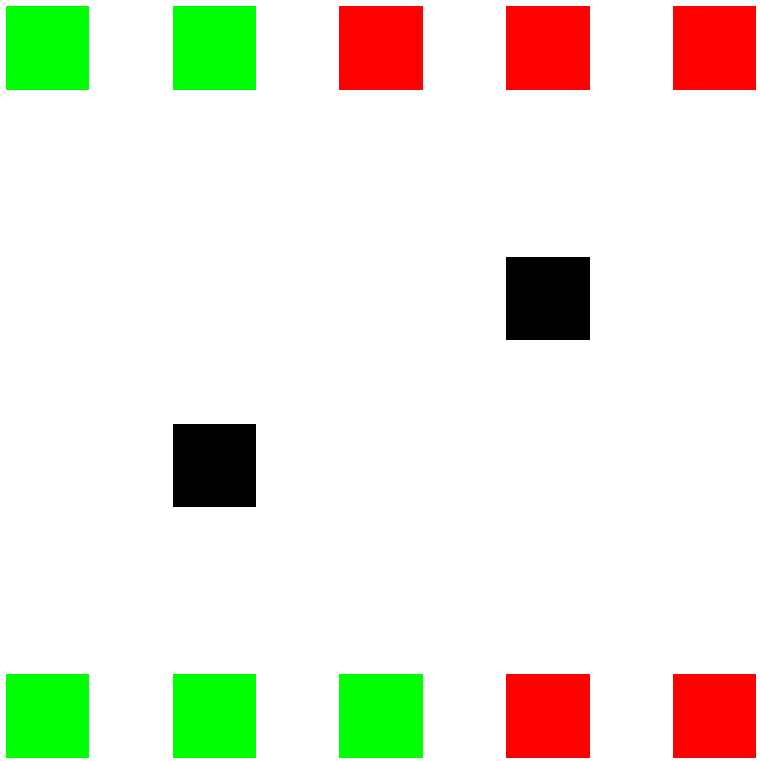
Algorytm k-średnich - minimum lokalne SSE (3)



Algorytm k-średnich - minimum lokalne SSE (4)



Algorytm k-średnich - minimum lokalne SSE (5)



Algorytm *k*-medoidów

- *Krok 1:* Zapytaj użytkownika, na ile grup (k) zbiór danych powinien zostać podzielony.
- *Krok 2:* Losowo przypisz k rekordów jako początkowe środki grup.
- *Krok 3:* Dla każdego rekordu znajdź najbliższy środek grupy.
- *Krok 4:* Dla każdej z k grup znajdź *medoid* grupy i uaktualnij położenie każdego środka grupy jako nową wartość medoidu.
- *Krok 5:* Powtarzaj kroki od 3 do 5 do zbieżności lub zakończenia.

Algorytm k-medoidów

- *Krok 1:* Zapytaj użytkownika, na ile grup (k) zbiór danych powinien zostać podzielony.
- *Krok 2:* Losowo przypisz k rekordów jako początkowe środki grup.
- *Krok 3:* Dla każdego rekordu znajdź najbliższy środek grupy.
- *Krok 4:* Dla każdej z k grup znajdź *medoid* grupy i uaktualnij położenie każdego środka grupy jako nową wartość medoidu.
- *Krok 5:* Powtarzaj kroki od 3 do 5 do zbieżności lub zakończenia.

Algorytm k-medoidów

- *Krok 1:* Zapytaj użytkownika, na ile grup (k) zbiór danych powinien zostać podzielony.
- *Krok 2:* Losowo przypisz k rekordów jako początkowe środki grup.
- *Krok 3:* Dla każdego rekordu znajdź najbliższy środek grupy.
- *Krok 4:* Dla każdej z k grup znajdź *medoid* grupy i uaktualnij położenie każdego środka grupy jako nową wartość medoidu.
- *Krok 5:* Powtarzaj kroki od 3 do 5 do zbieżności lub zakończenia.

Algorytm k-medoidów

- *Krok 1:* Zapytaj użytkownika, na ile grup (k) zbiór danych powinien zostać podzielony.
- *Krok 2:* Losowo przypisz k rekordów jako początkowe środki grup.
- *Krok 3:* Dla każdego rekordu znajdź najbliższy środek grupy.
- *Krok 4:* Dla każdej z k grup znajdź *medoid* grupy i uaktualnij położenie każdego środka grupy jako nową wartość medoidu.
- *Krok 5:* Powtarzaj kroki od 3 do 5 do zbieżności lub zakończenia.

Algorytm k-medoidów

- *Krok 1:* Zapytaj użytkownika, na ile grup (k) zbiór danych powinien zostać podzielony.
- *Krok 2:* Losowo przypisz k rekordów jako początkowe środki grup.
- *Krok 3:* Dla każdego rekordu znajdź najbliższy środek grupy.
- *Krok 4:* Dla każdej z k grup znajdź *medoid* grupy i uaktualnij położenie każdego środka grupy jako nową wartość medoidu.
- *Krok 5:* Powtarzaj kroki od 3 do 5 do zbieżności lub zakończenia.

Algorytm k-medoidów - motywacja

W celu uodpornienia algorytmu k-średnich na występowanie punktów osobliwych, zamiast średnich klastrów jako środków tych klastrów, możemy przyjąć medoidy - najbardziej centralnie ulokowane punkty klastrów:

- średnia zbioru $1, 3, 5, 7, 1009$ wynosi **205**;
- medoid zbioru $1, 3, 5, 7, 1009$ wynosi **5**.

Algorytm k-medoidów - motywacja

W celu uodpornienia algorytmu k-średnich na występowanie punktów osobliwych, zamiast średnich klastrów jako środków tych klastrów, możemy przyjąć medoidy - najbardziej centralnie ulokowane punkty klastrów:

- średnia zbioru $1, 3, 5, 7, 1009$ wynosi **205**;
- medoid zbioru $1, 3, 5, 7, 1009$ wynosi **5**.

Algorytm k-medoidów - motywacja

W celu uodpornienia algorytmu k-średnich na występowanie punktów osobliwych, zamiast średnich klastrów jako środków tych klastrów, możemy przyjąć medoidy - najbardziej centralnie ulokowane punkty klastrów:

- średnia zbioru $1, 3, 5, 7, 1009$ wynosi **205**;
- medoid zbioru $1, 3, 5, 7, 1009$ wynosi **5**.

Zbiór danych przedstawiamy jako graf $G = (V, E)$, gdzie zbiór wierzchołków jest tożsamy ze zbiorem przykładów $V \equiv X$. Każde dwa przykłady łączy krawędź o wadze równej odległości między tymi przykładami.

Algorytm minimalnego drzewa rozpinającego *MST*

- w pierwszym kroku algorytmu wyznaczany jest graf będący minimalnym drzewem rozpinającym zbioru przykładów;
- następnie porządkuje się malejąco pod względem wag krawędzie tego grafu;
- usuwając kolejne krawędzie otrzymujemy spójne części tego grafu, będące kolejnymi hierarchicznymi grupowaniami zbioru danych.

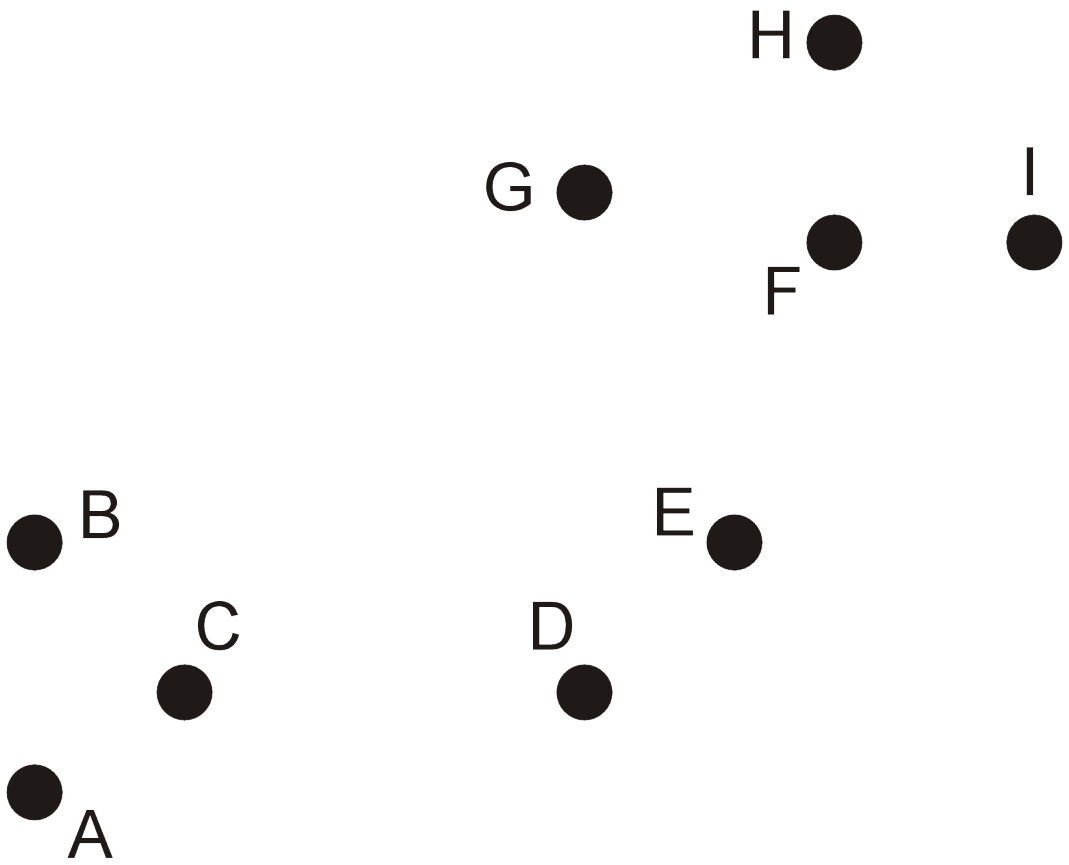
Algorytm minimalnego drzewa rozpinającego *MST*

- w pierwszym kroku algorytmu wyznaczany jest graf będący minimalnym drzewem rozpinającym zbioru przykładów;
- następnie porządkuje się malejąco pod względem wag krawędzie tego grafu;
- usuwając kolejne krawędzie otrzymujemy spójne części tego grafu, będące kolejnymi hierarchicznymi grupowaniami zbioru danych.

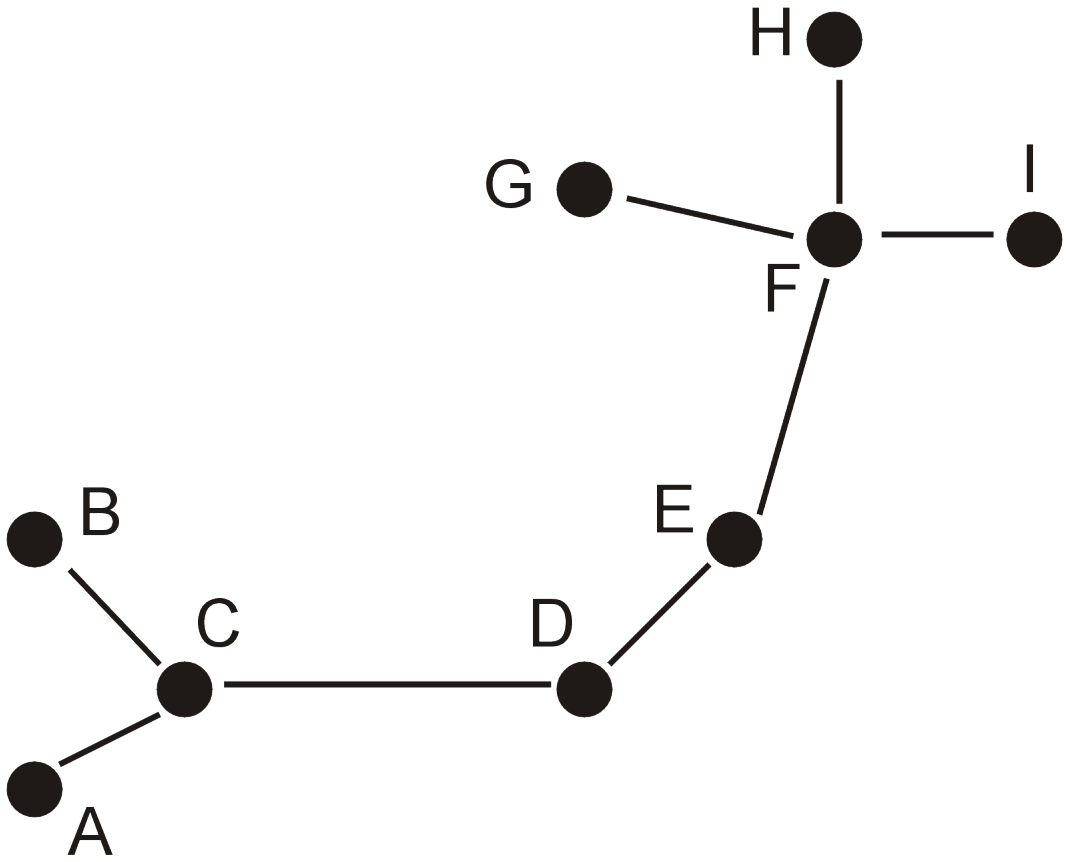
Algorytm minimalnego drzewa rozpinającego *MST*

- w pierwszym kroku algorytmu wyznaczany jest graf będący minimalnym drzewem rozpinającym zbioru przykładów;
- następnie porządkuje się malejąco pod względem wag krawędzie tego grafu;
- usuwając kolejne krawędzie otrzymujemy spójne części tego grafu, będące kolejnymi hierarchicznymi grupowaniami zbioru danych.

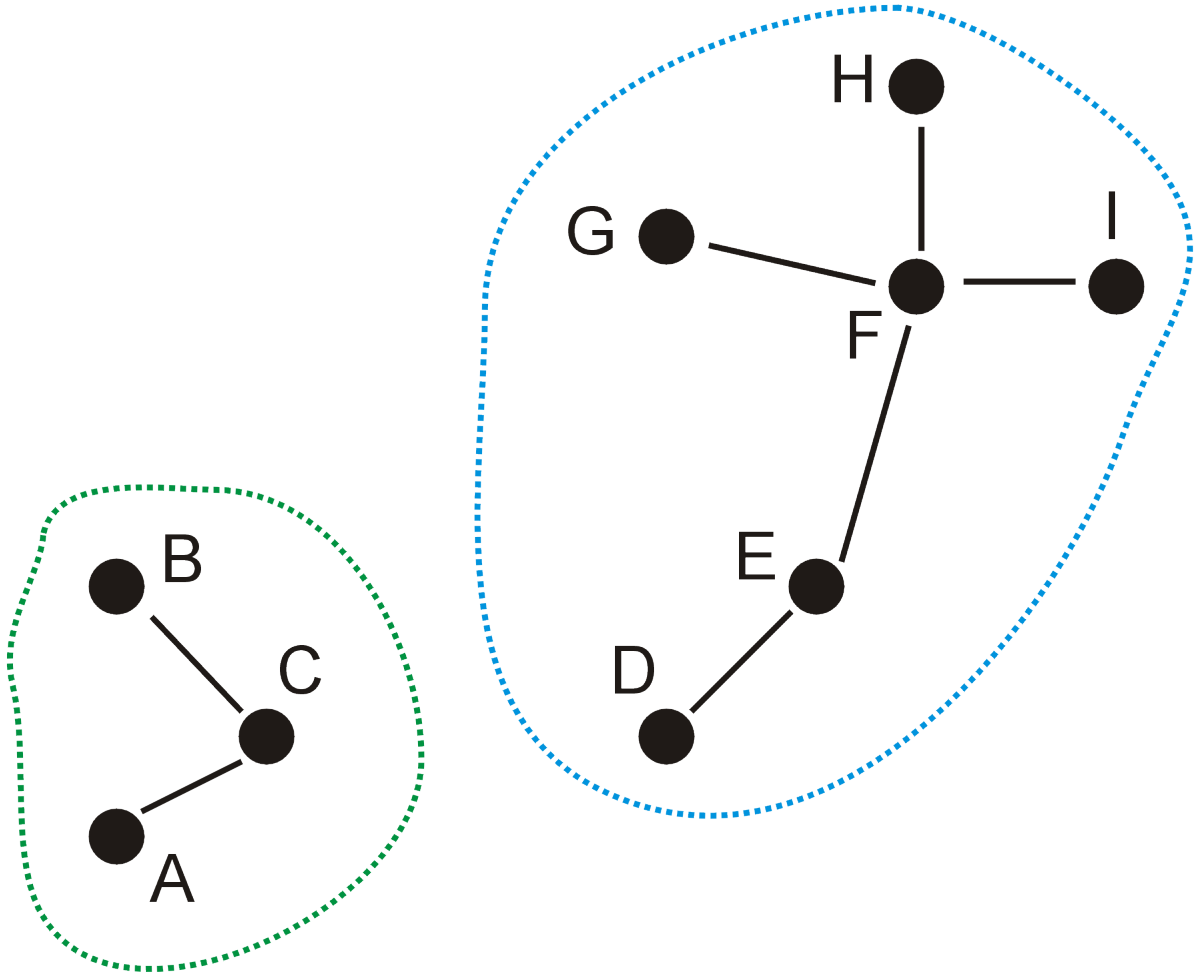
Algorytm *MST* (1)



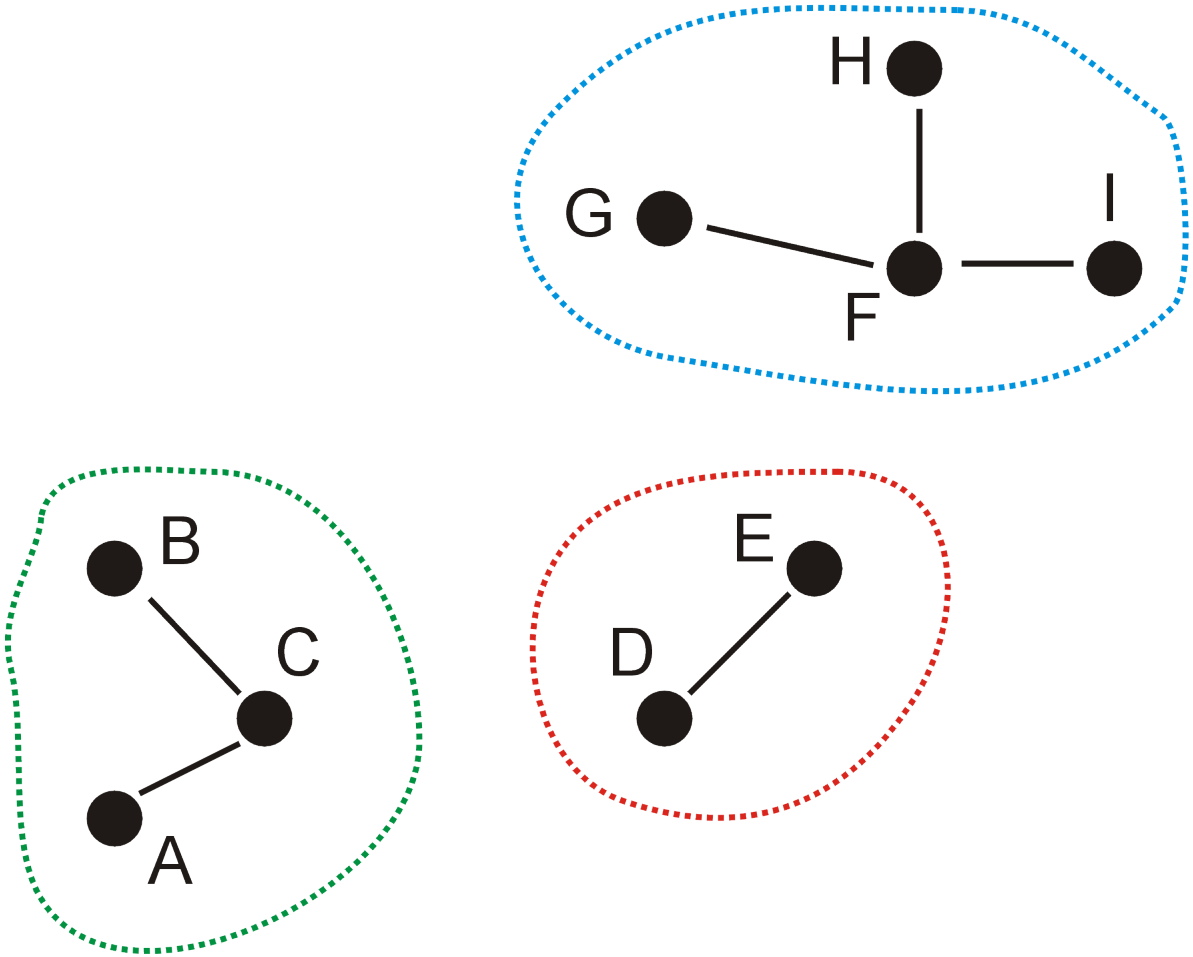
Algorytm *MST* (2)



Algorytm *MST* (3)



Algorytm *MST* (4)



Analiza tendencji gupowania

Algorytmy używane do grupowania mają zazwyczaj wiele parametrów, wśród których można wyznaczyć parametr kluczowy, od którego wynik zależy w największym stopniu.

Analizą tendencji grupowania można nazwać badanie działania algorytmu w zależności od wartości kluczowego parametru, przy ustalonych wartościach innych parametrów.

Na tej podstawie można dużo lepiej zrozumieć zależności w zbiorze danych, który jest badany. Możliwe jest również wyznaczenie najwłaściwszej wartości kluczowego parametru.

Analiza tendencji gupowania

Algorytmy używane do grupowania mają zazwyczaj wiele parametrów, wśród których można wyznaczyć parametr kluczowy, od którego wynik zależy w największym stopniu.

Analizą tendencji grupowania można nazwać badanie działania algorytmu w zależności od wartości kluczowego parametru, przy ustalonych wartościach innych parametrów.

Na tej podstawie można dużo lepiej zrozumieć zależności w zbiorze danych, który jest badany. Możliwe jest również wyznaczenie najwłaściwszej wartości kluczowego parametru.

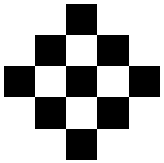
Analiza tendencji gupowania

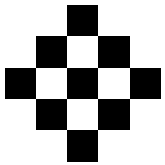
Algorytmy używane do grupowania mają zazwyczaj wiele parametrów, wśród których można wyznaczyć parametr kluczowy, od którego wynik zależy w największym stopniu.

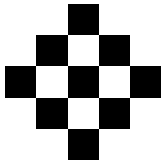
Analizą tendencji grupowania można nazwać badanie działania algorytmu w zależności od wartości kluczowego parametru, przy ustalonych wartościach innych parametrów.

Na tej podstawie można dużo lepiej zrozumieć zależności w zbiorze danych, który jest badany. Możliwe jest również wyznaczenie najwłaściwszej wartości kluczowego parametru.

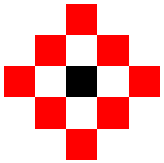
Algorytm k-średnich

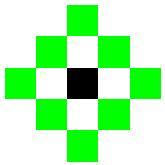


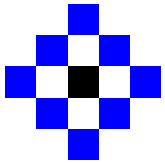




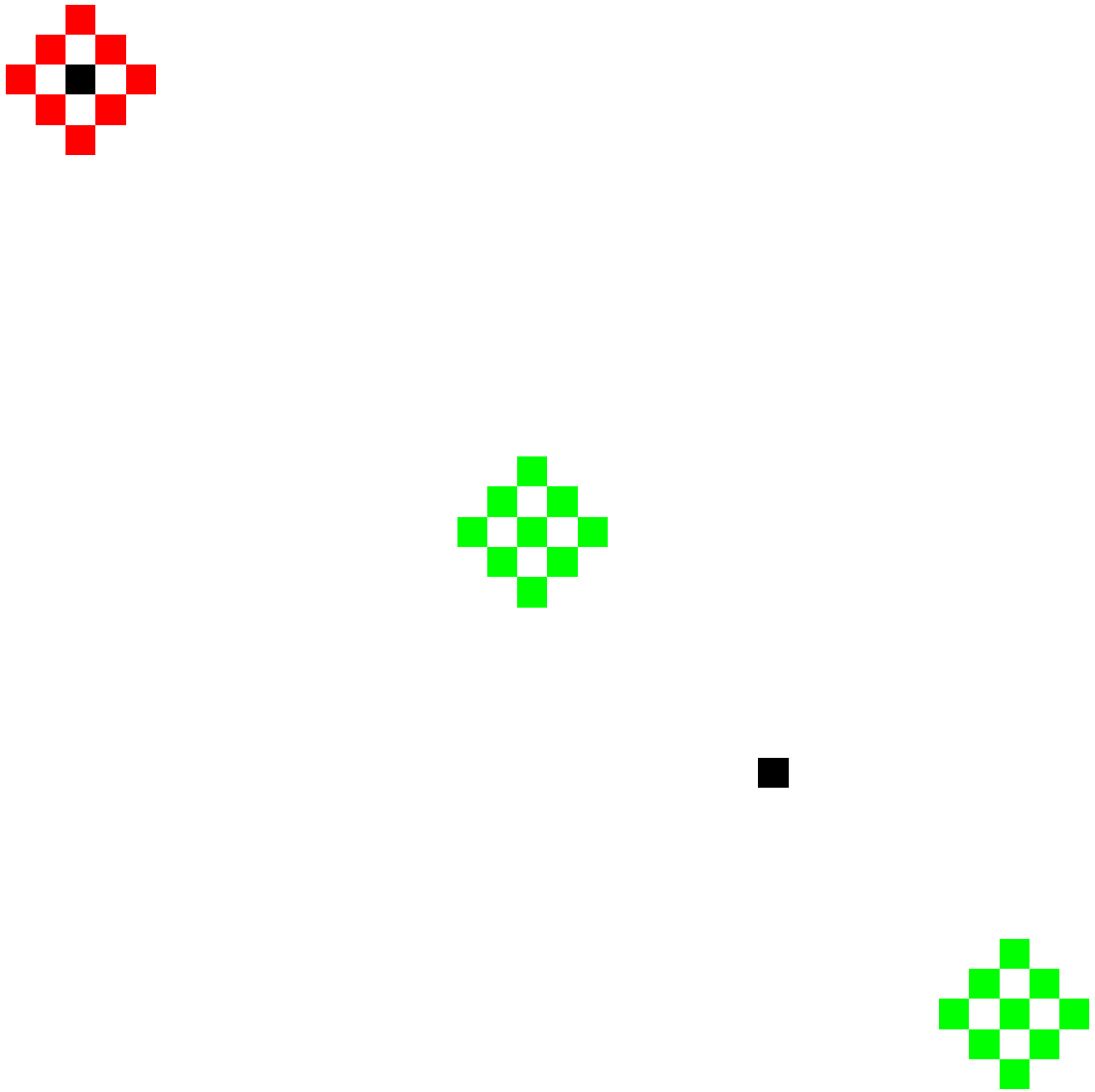
Algorytm k-średnich, 3 klastry



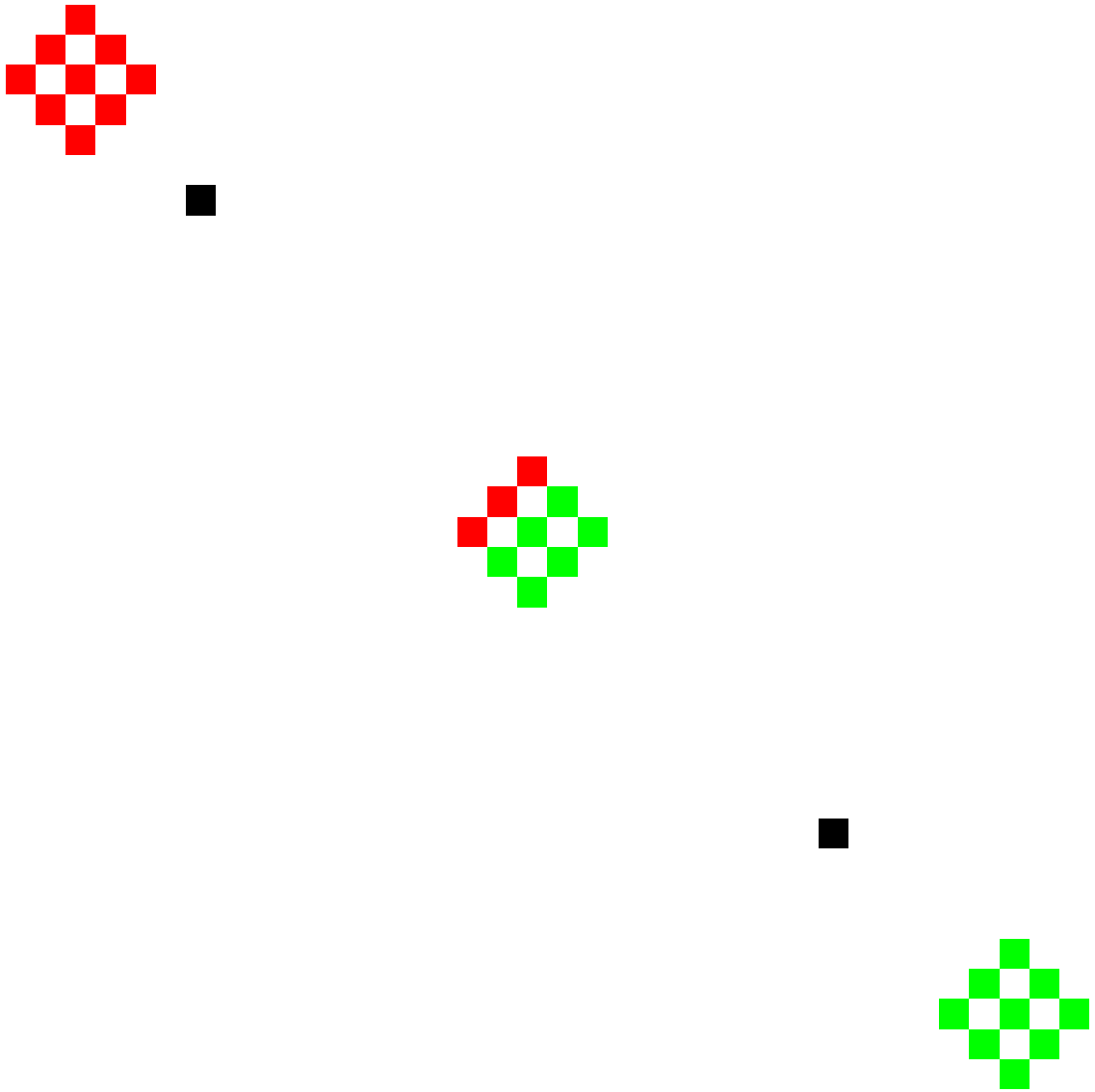




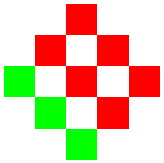
Algorytm k-średnich, 2 klastry (1)

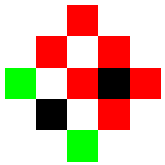


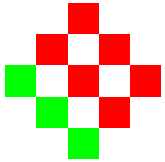
Algorytm k-średnich, 2 klastry (2)



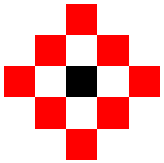
Algorytm k-średnich, 2 klastry (3)

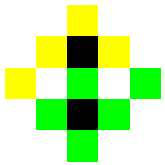


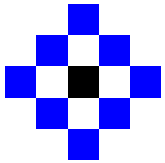




Algorytm k-średnich, 4 klastry







Analiza tendencji gupowania

Problem zaczyna się, gdy dane mają nie dwa czy trzy wymiary, ale cztery i więcej. Przy niemożliwości pełnego przedstawienia graficznego takiego zbioru danych, trudno jest zauważć klastry nawet tak dobrze oddzielone, jak te rozważane, a co za tym idzie, właściwie ocenić rezultaty grupowania.

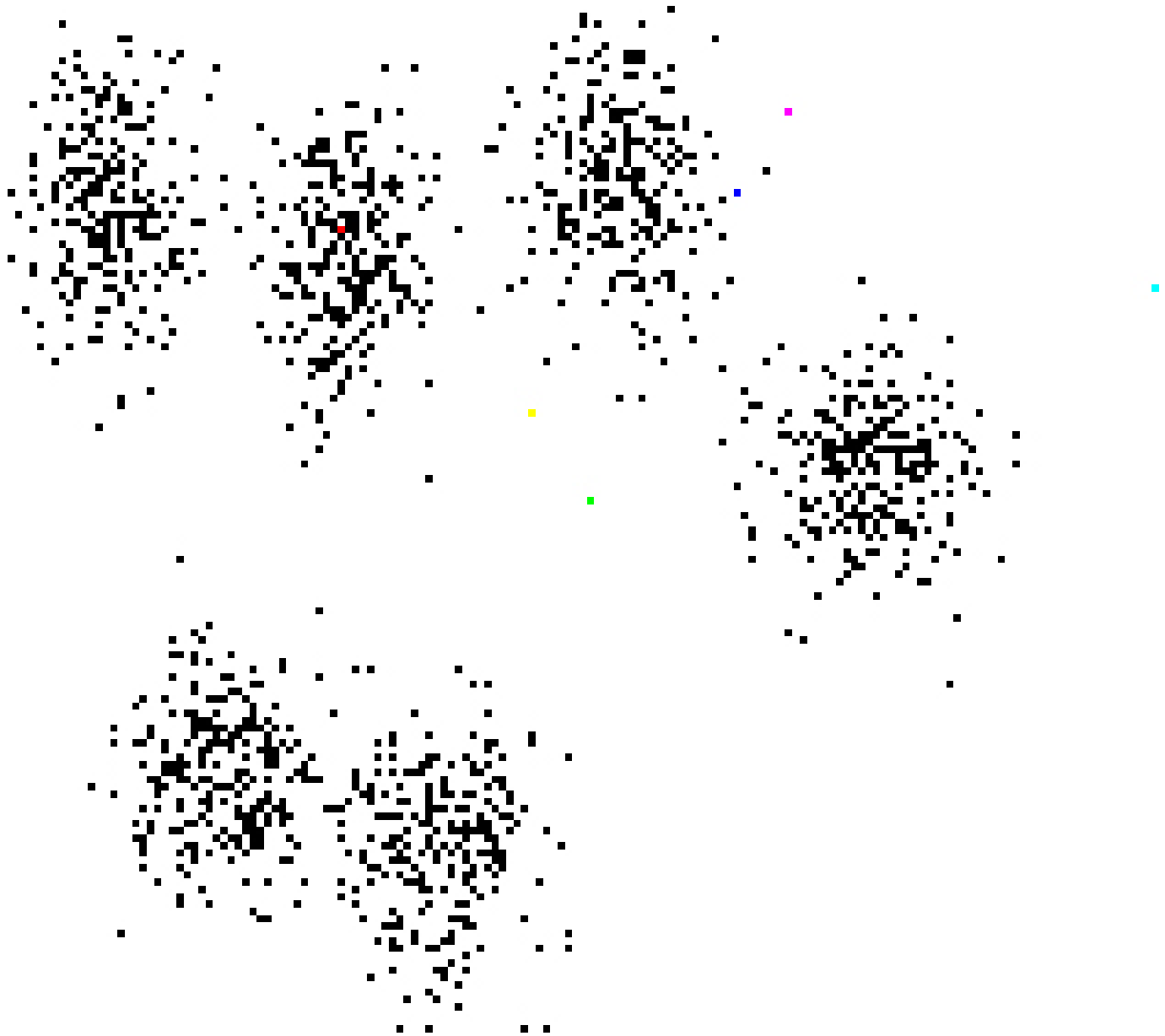
Funkcje oceny grupowania

By przeprowadzić analizę tendencji, trzeba umieć porównać wyniki działania algorytmu dla różnych parametrów. Najprościej jest to zrobić wybierając pewną funkcję oceny grupowania, która danemu grupowaniu przypisuje pewną ocenę - liczbę rzeczywistą. Najlepiej jest, gdy funkcja ta ma globalne ekstremum dla najlepszej wartości parametru.

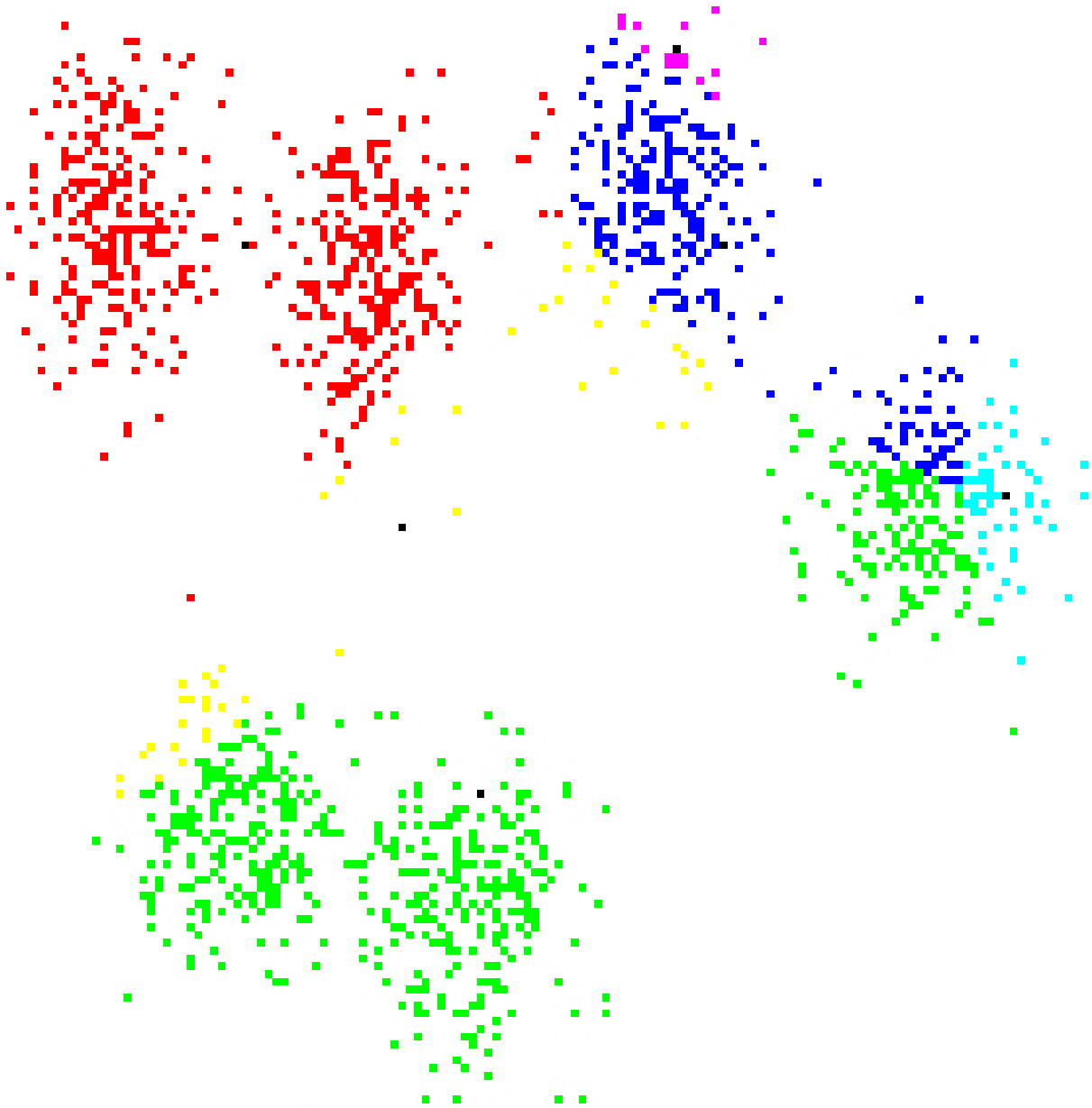
Problemy z badaniem tendencji

Badanie analizy tendencji jest zazwyczaj bardzo kosztowne obliczeniowo, wiąże się bowiem z wielokrotnym uruchamianiem algorytmu dla nieznacznie tylko zmienionego parametru.

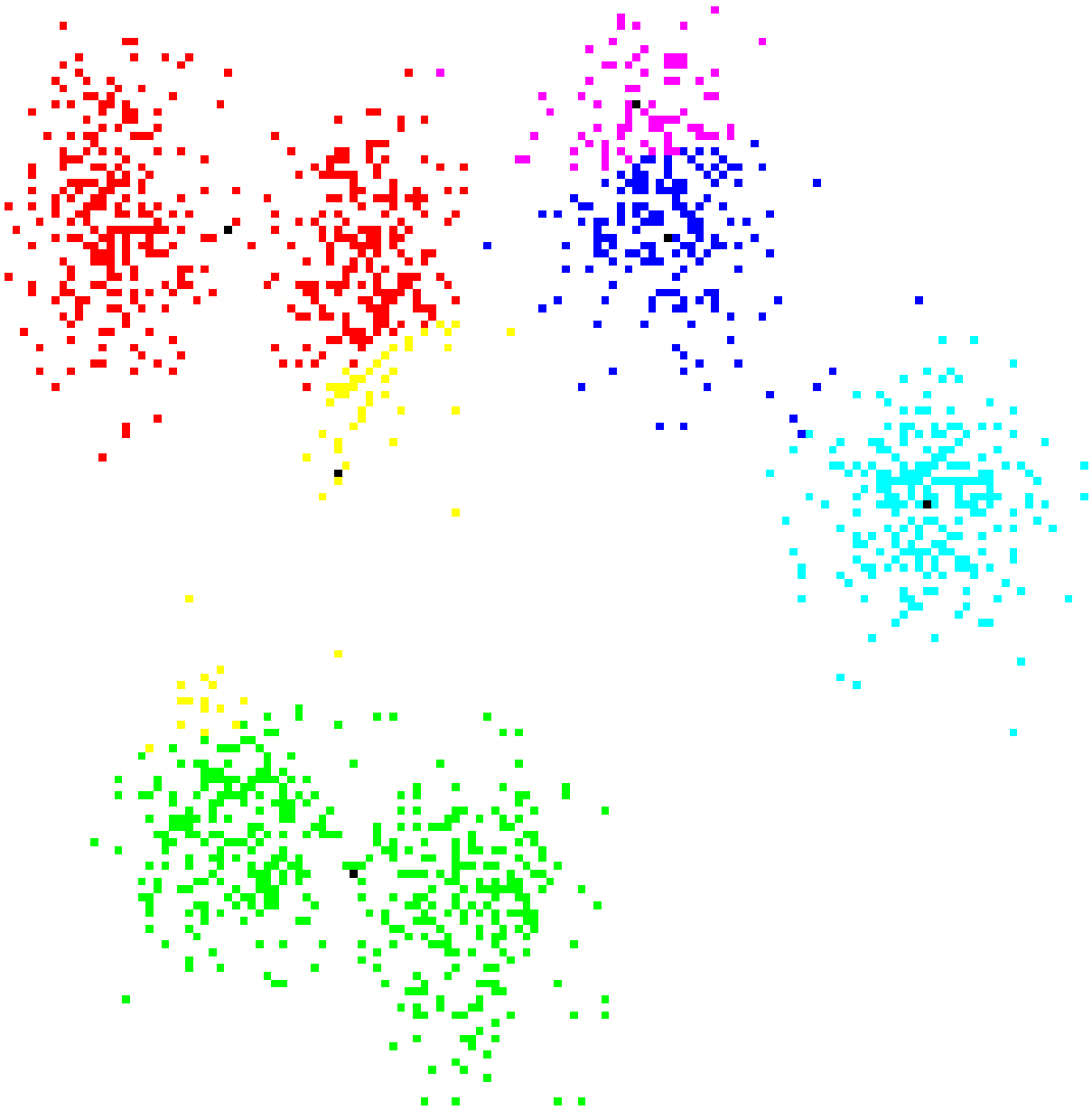
6 x 250 punktów



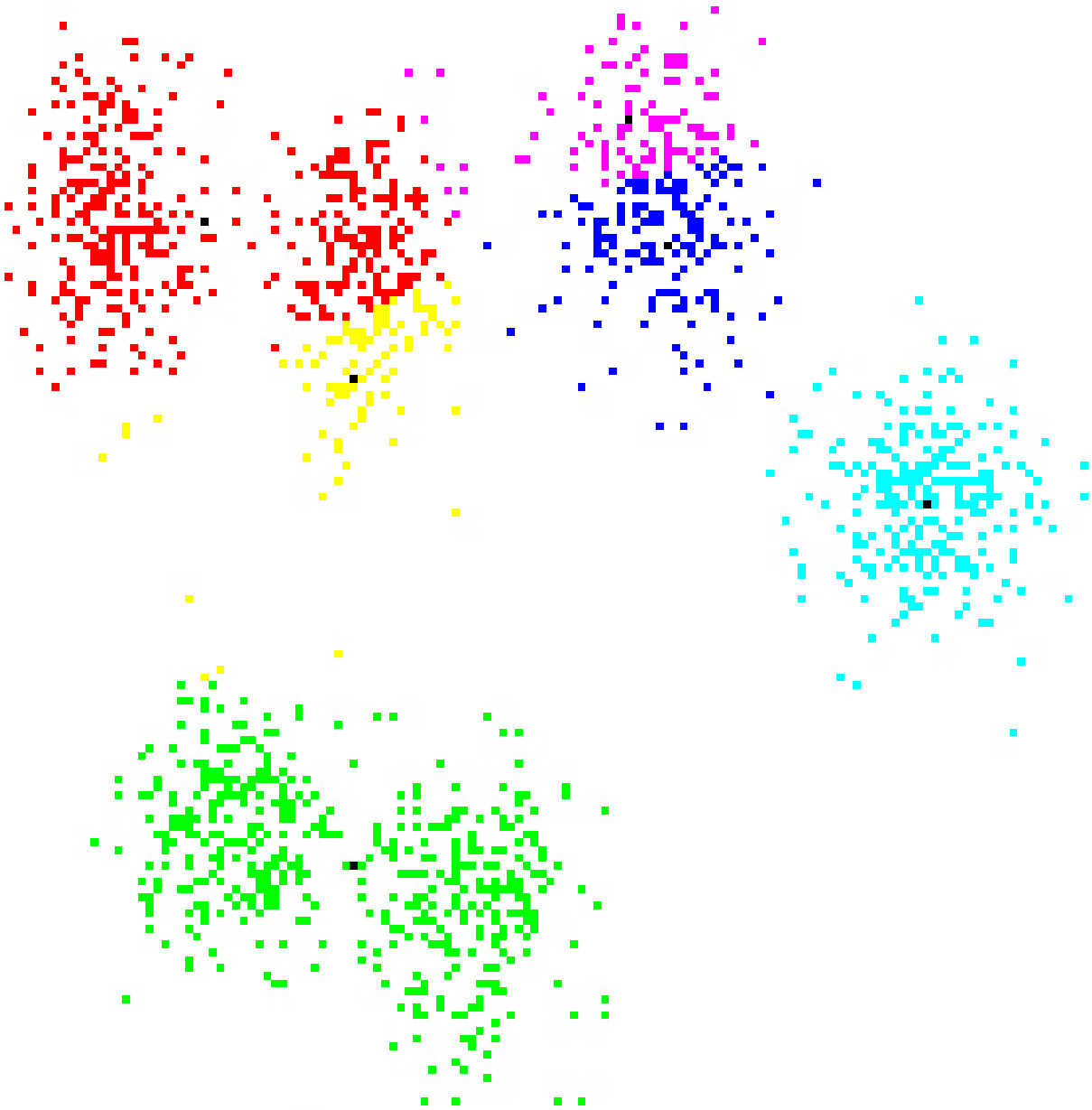
6 x 250 punktów, iteracja 1



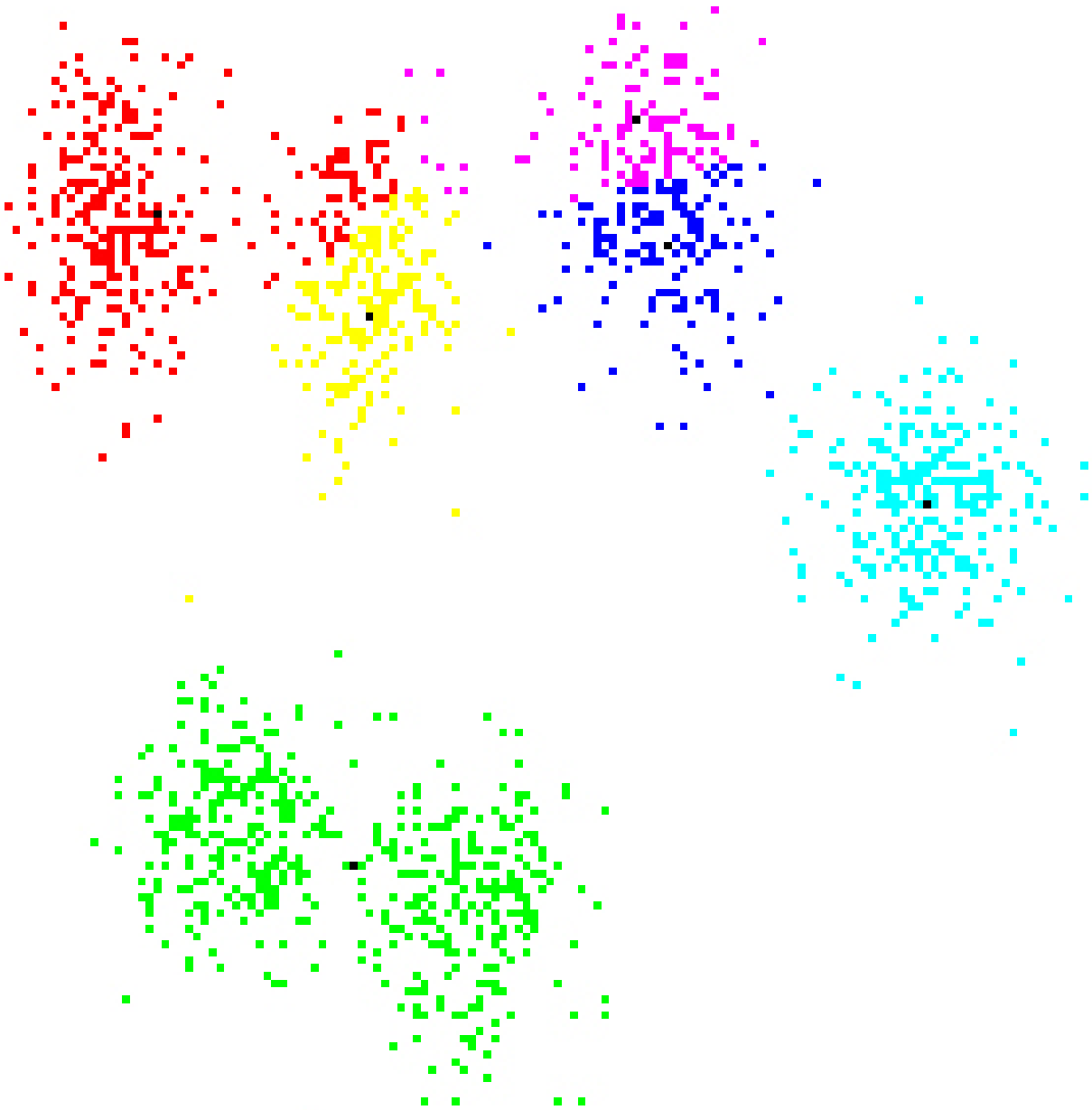
6 x 250 punktów, iteracja 2



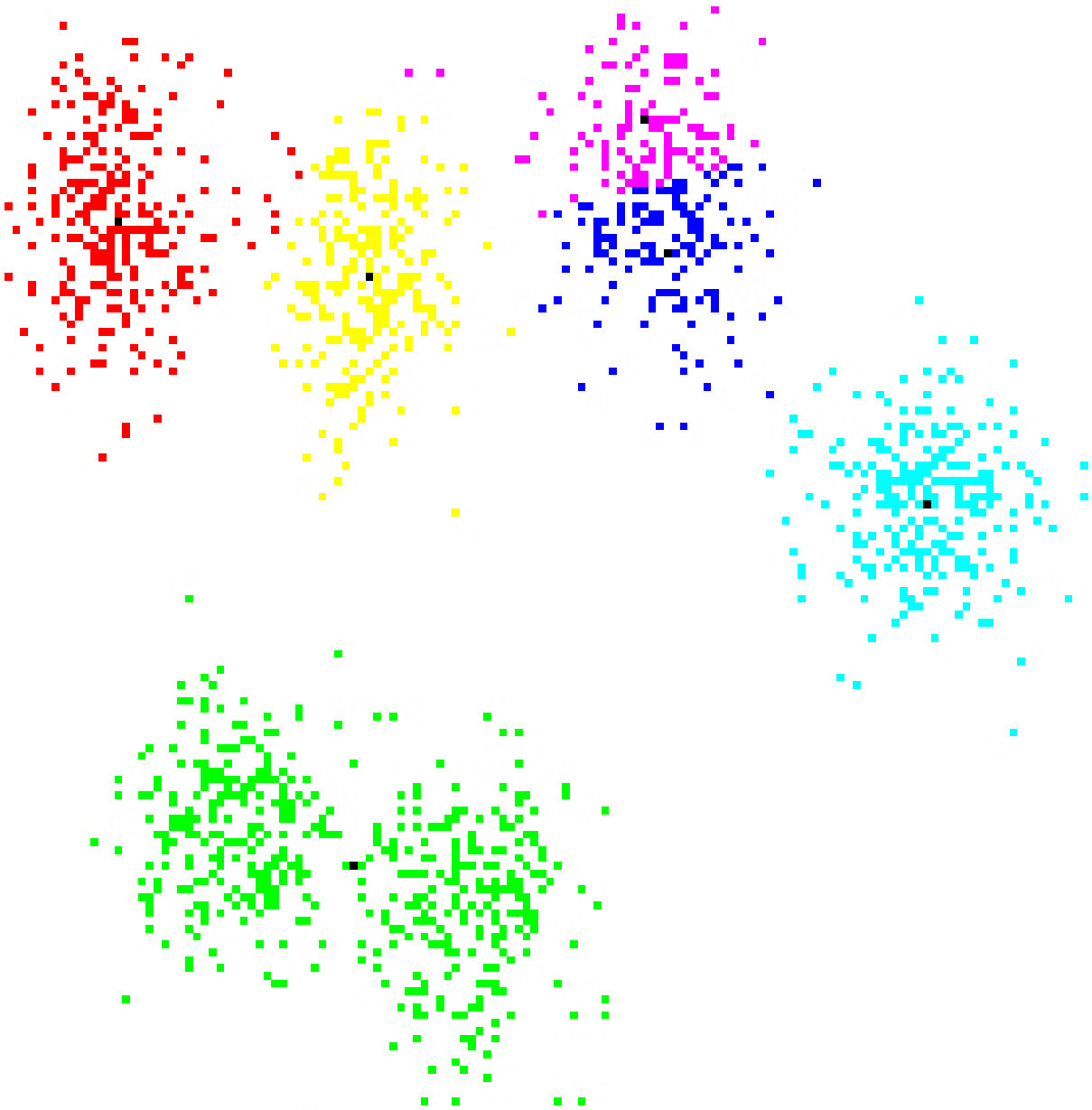
6 x 250 punktów, iteracja 3



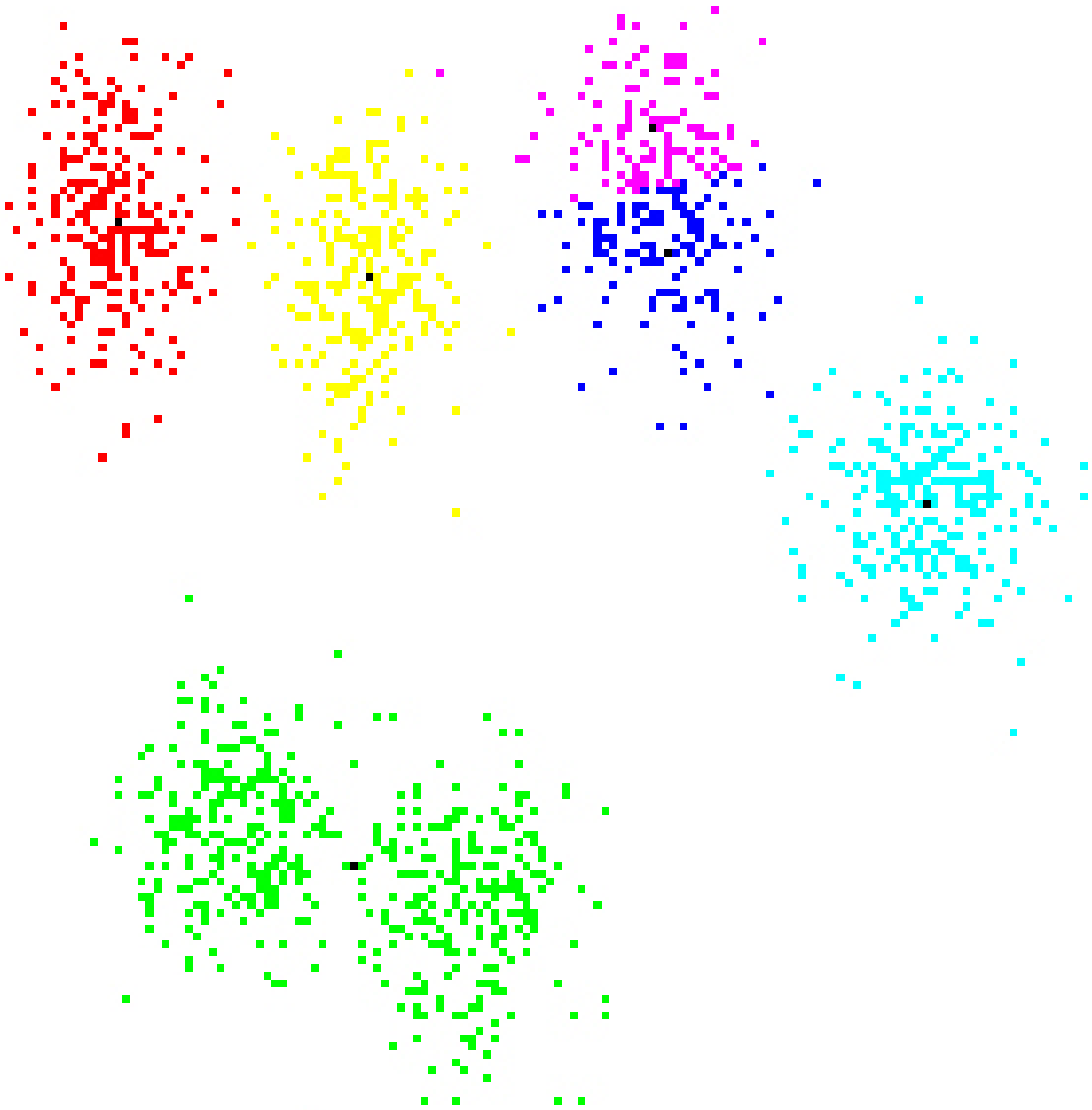
6 x 250 punktów, iteraja 4



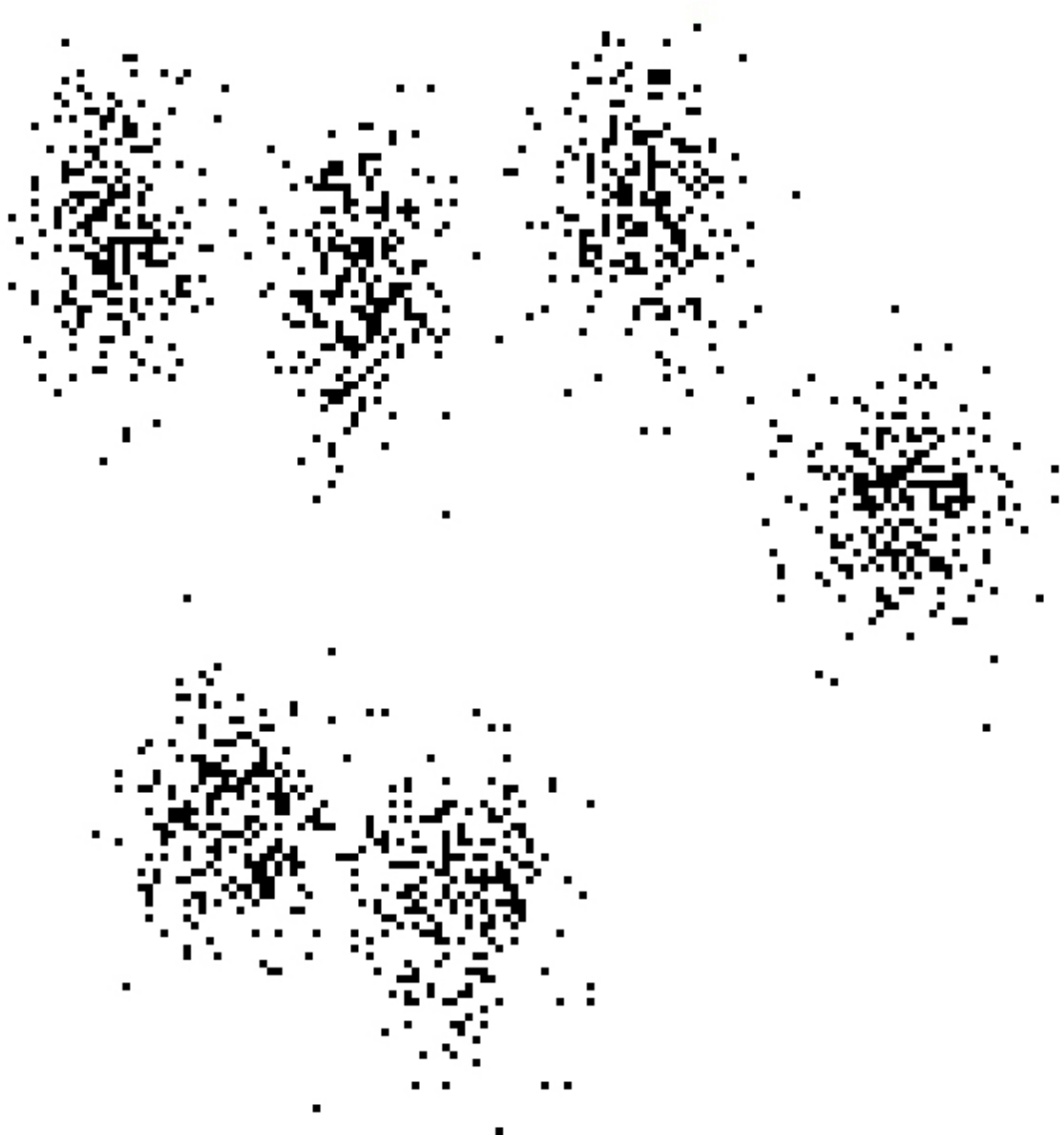
6 x 250 punktów, iteracja 5



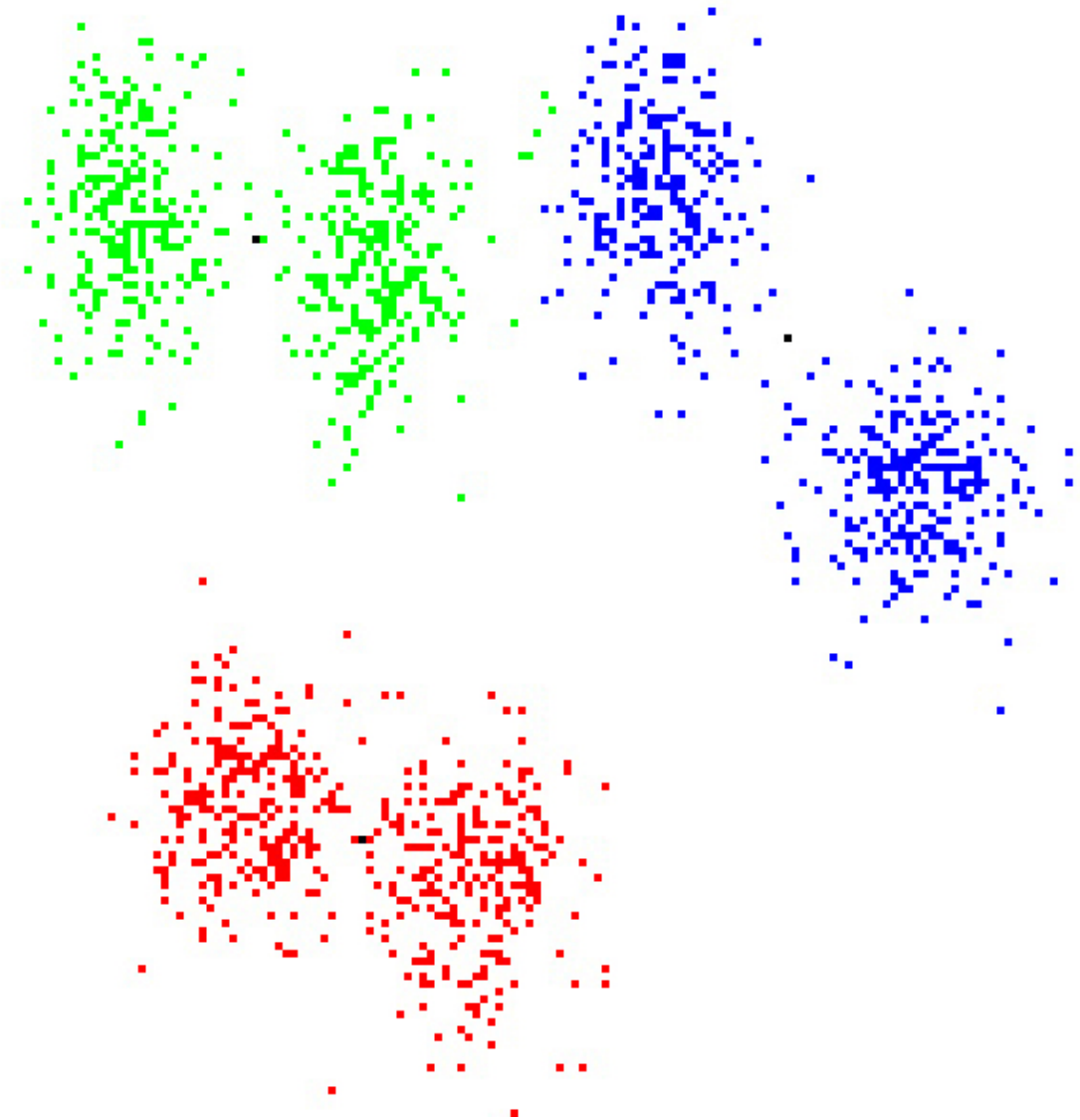
6 x 250 punktów, iteracja 6



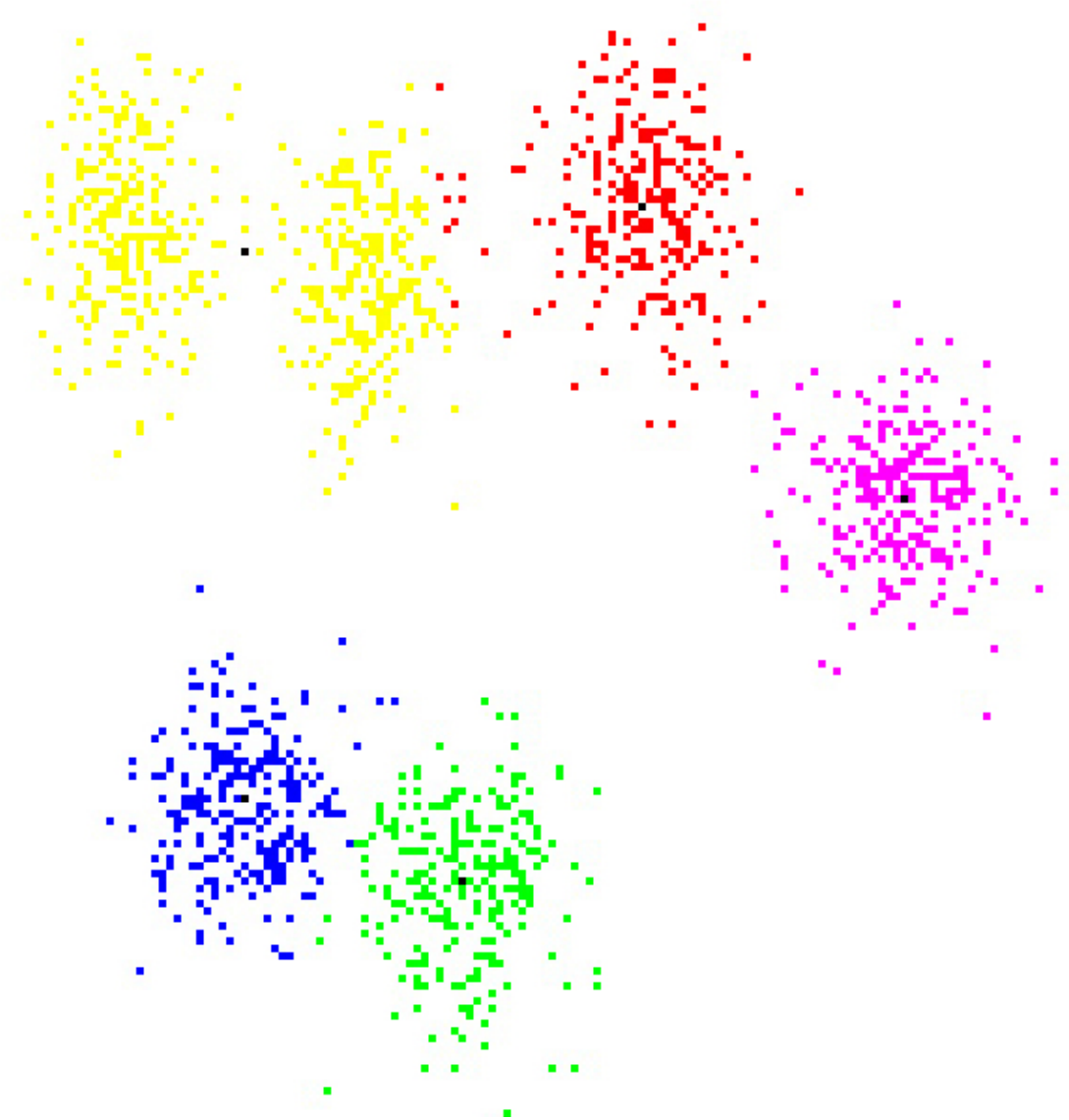
6 x 250 punktów



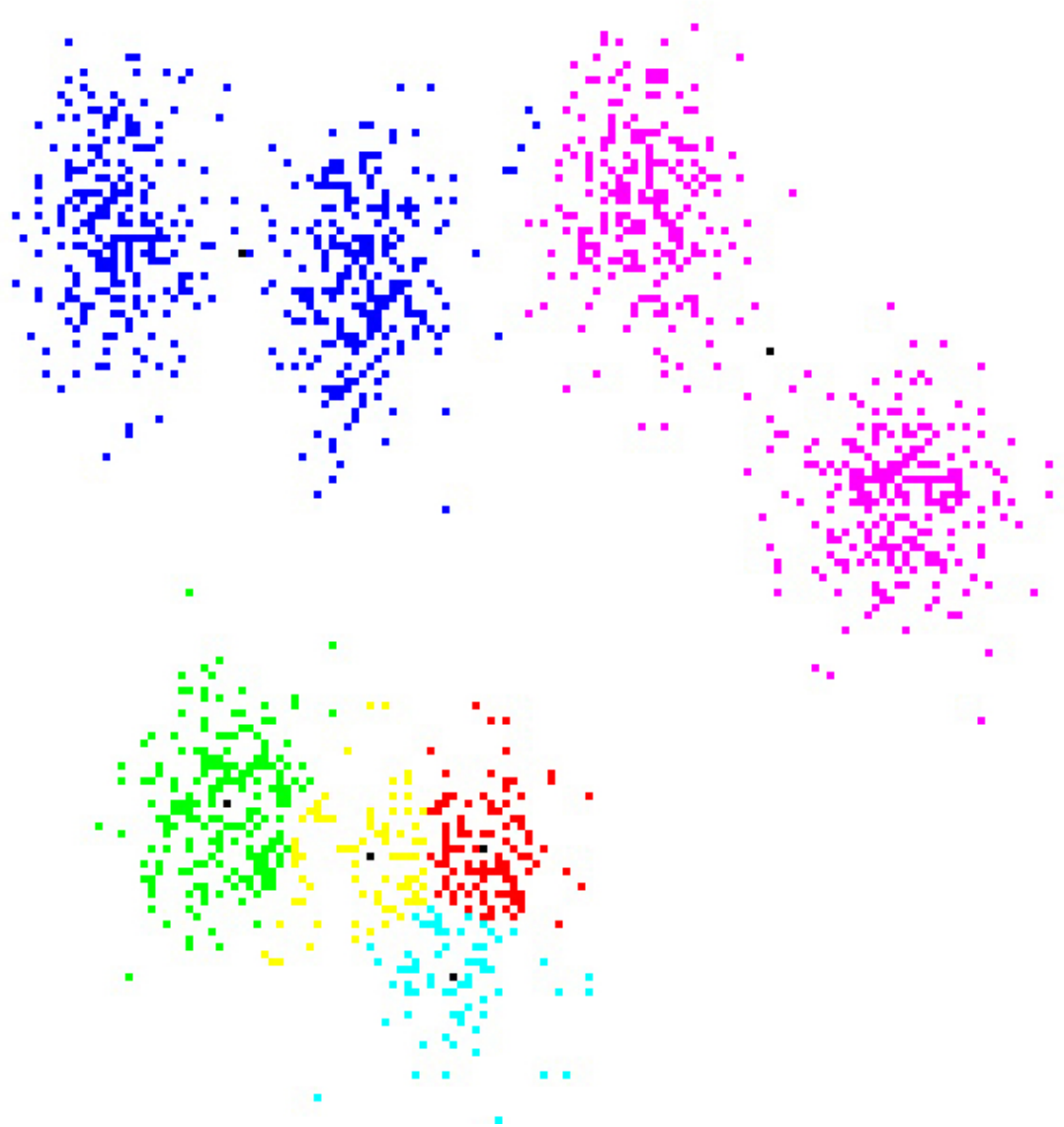
6 x 250 punktów, 3 skupiska



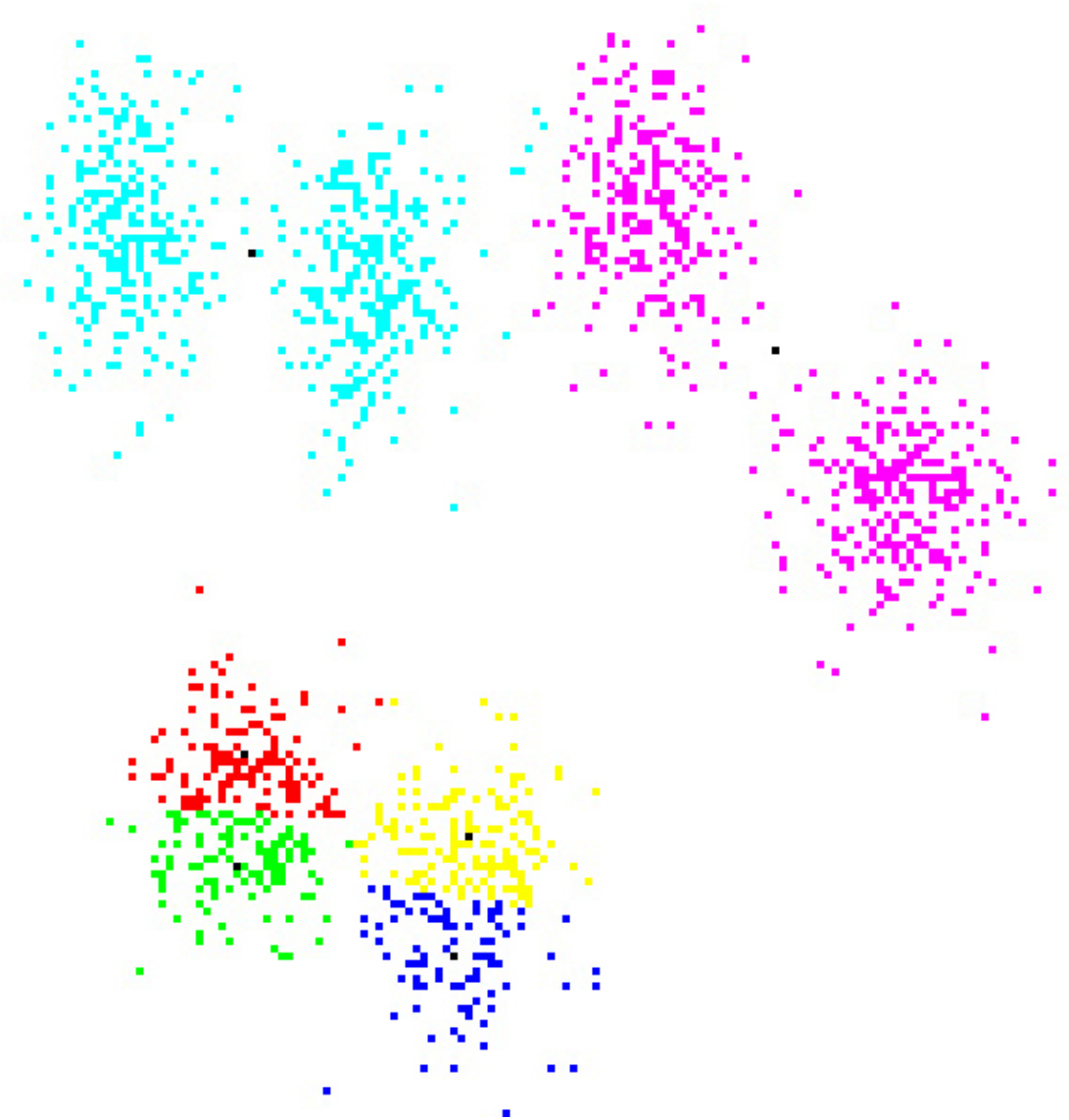
6 x 250 punktów, 5 skupisk



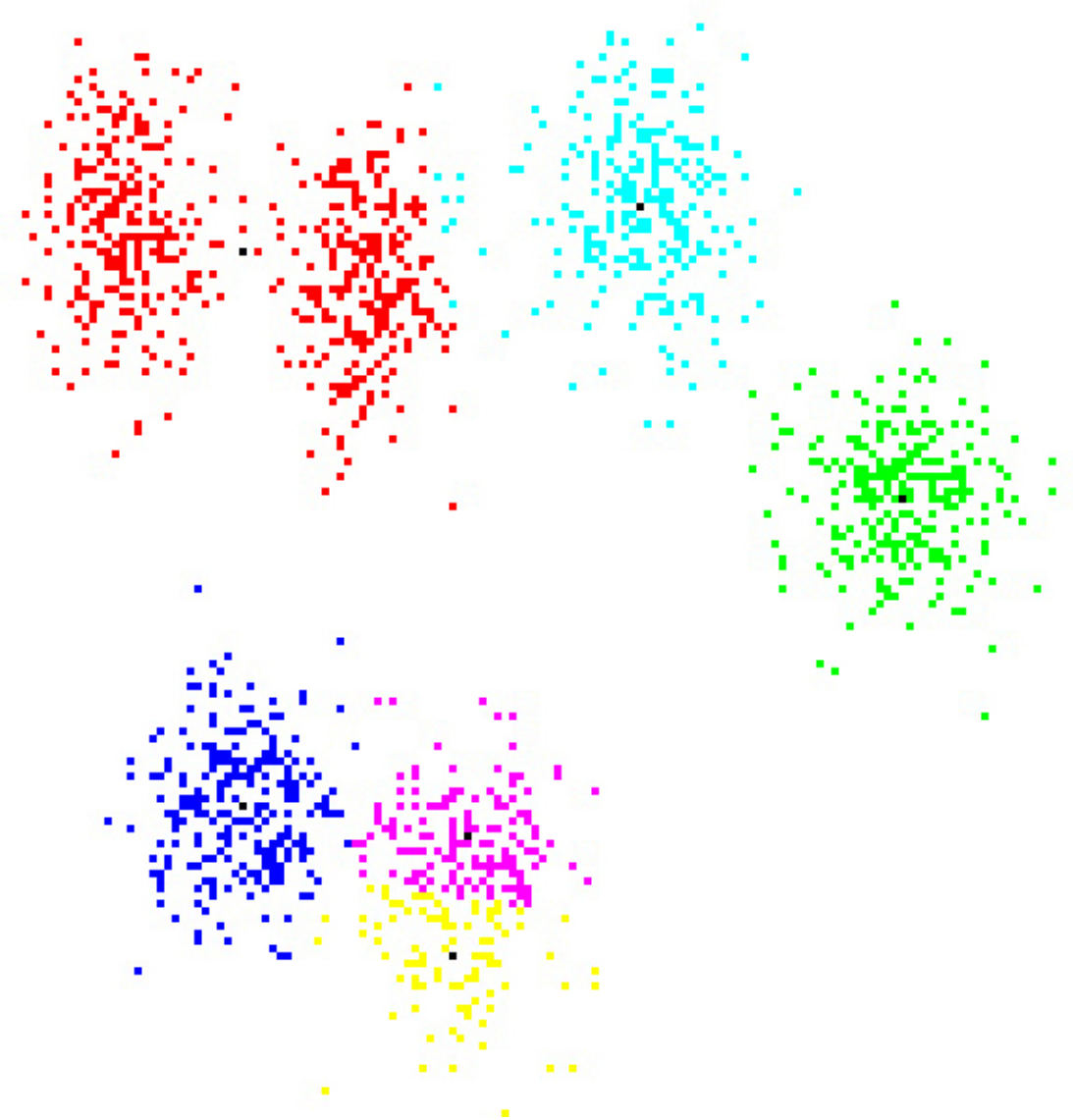
6 x 250 punktów, 6 skupisk (1)



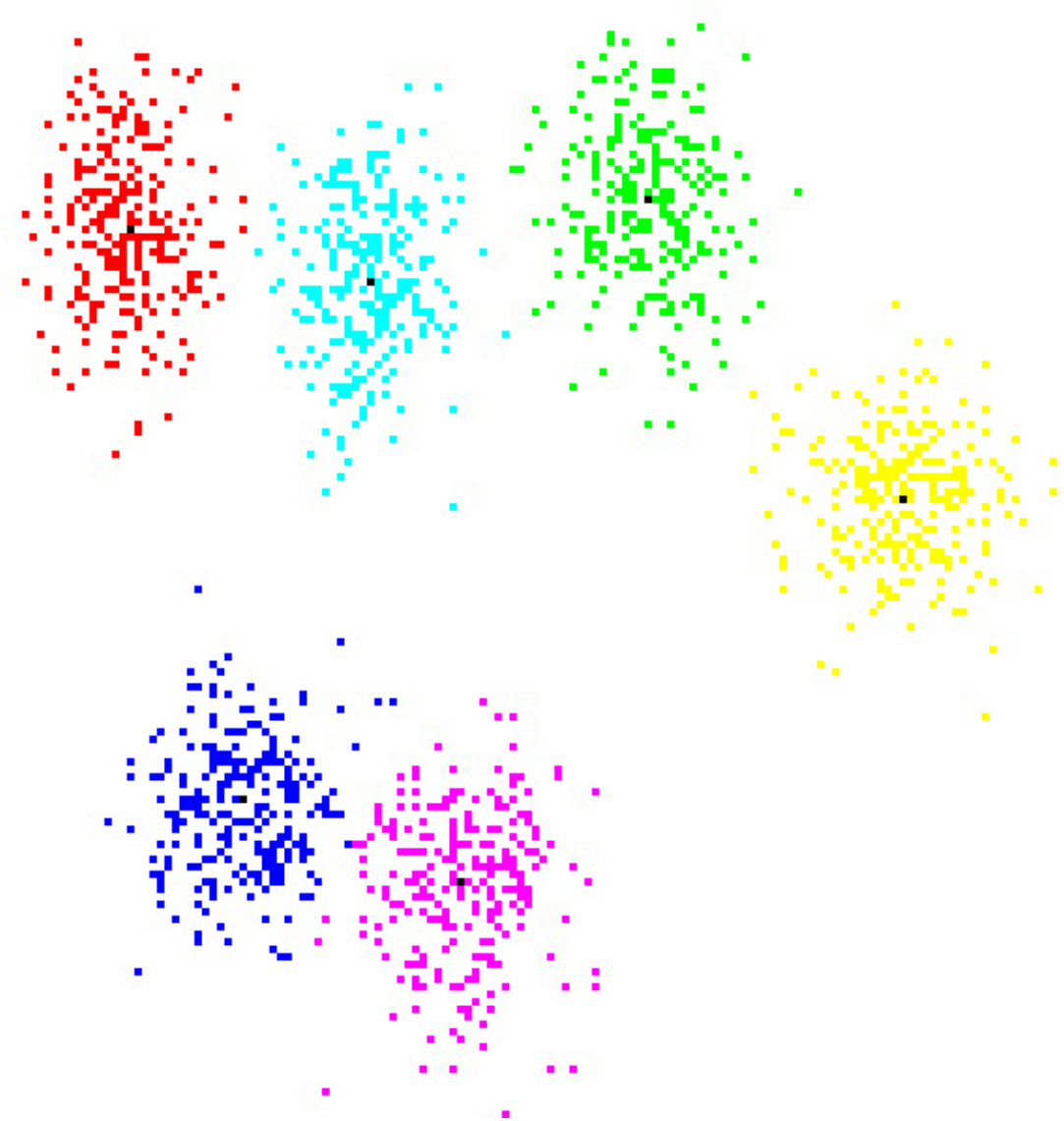
6 x 250 punktów, 6 skupisk (2)



6 x 250 punktów, 6 skupisk (3)

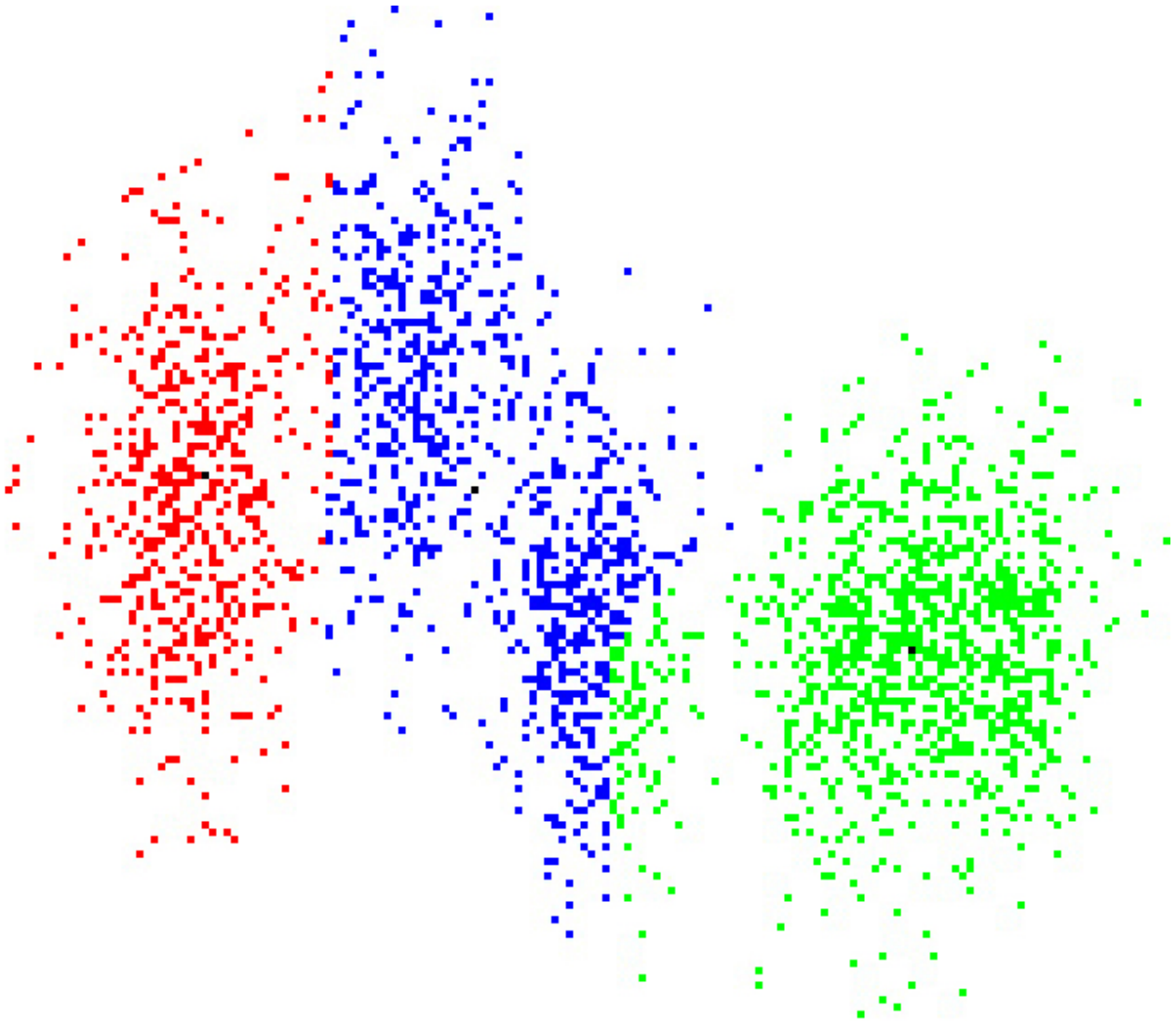


6 x 250 punktów, 6 skupisk (4)

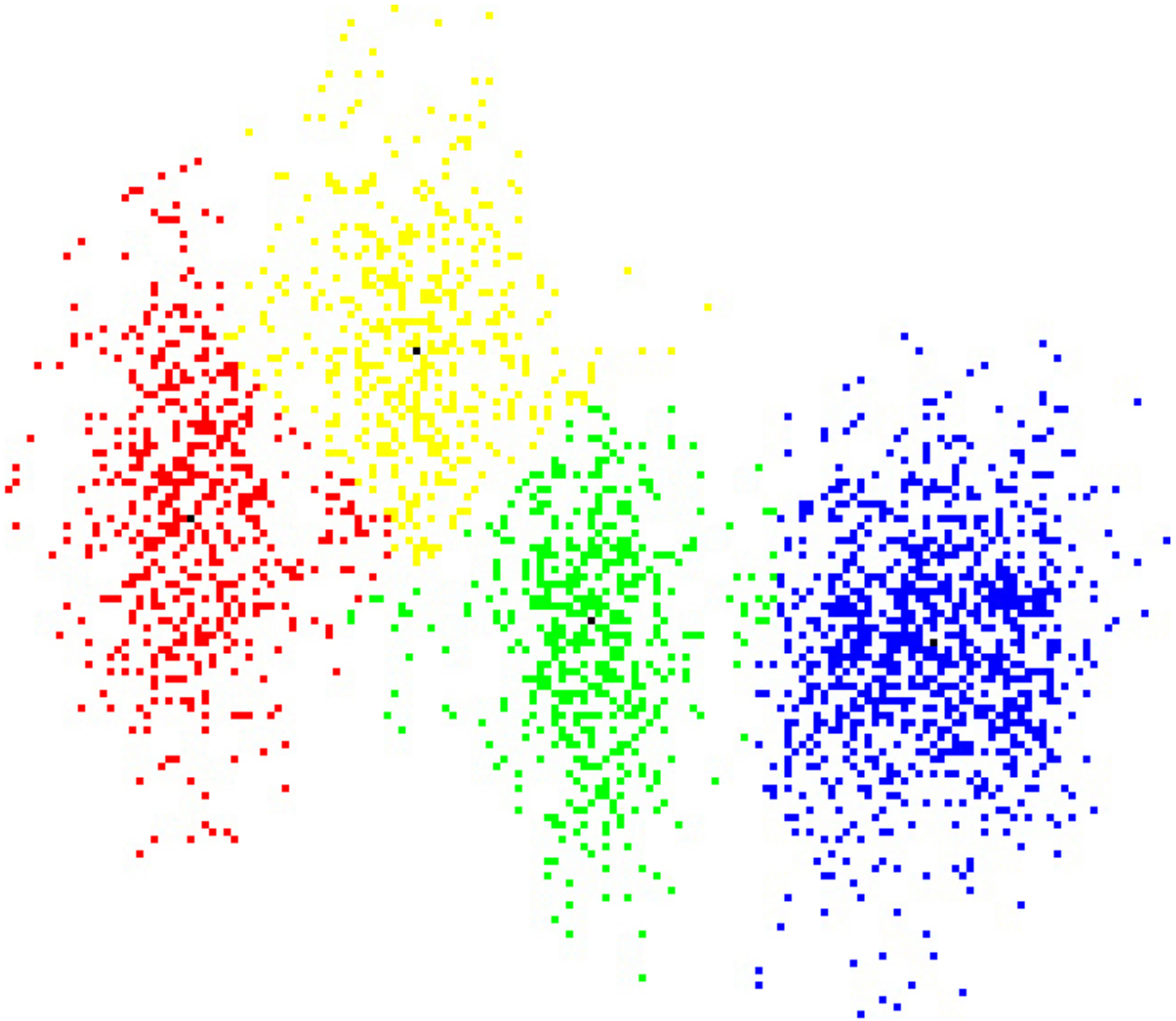


5 x 500 punktów

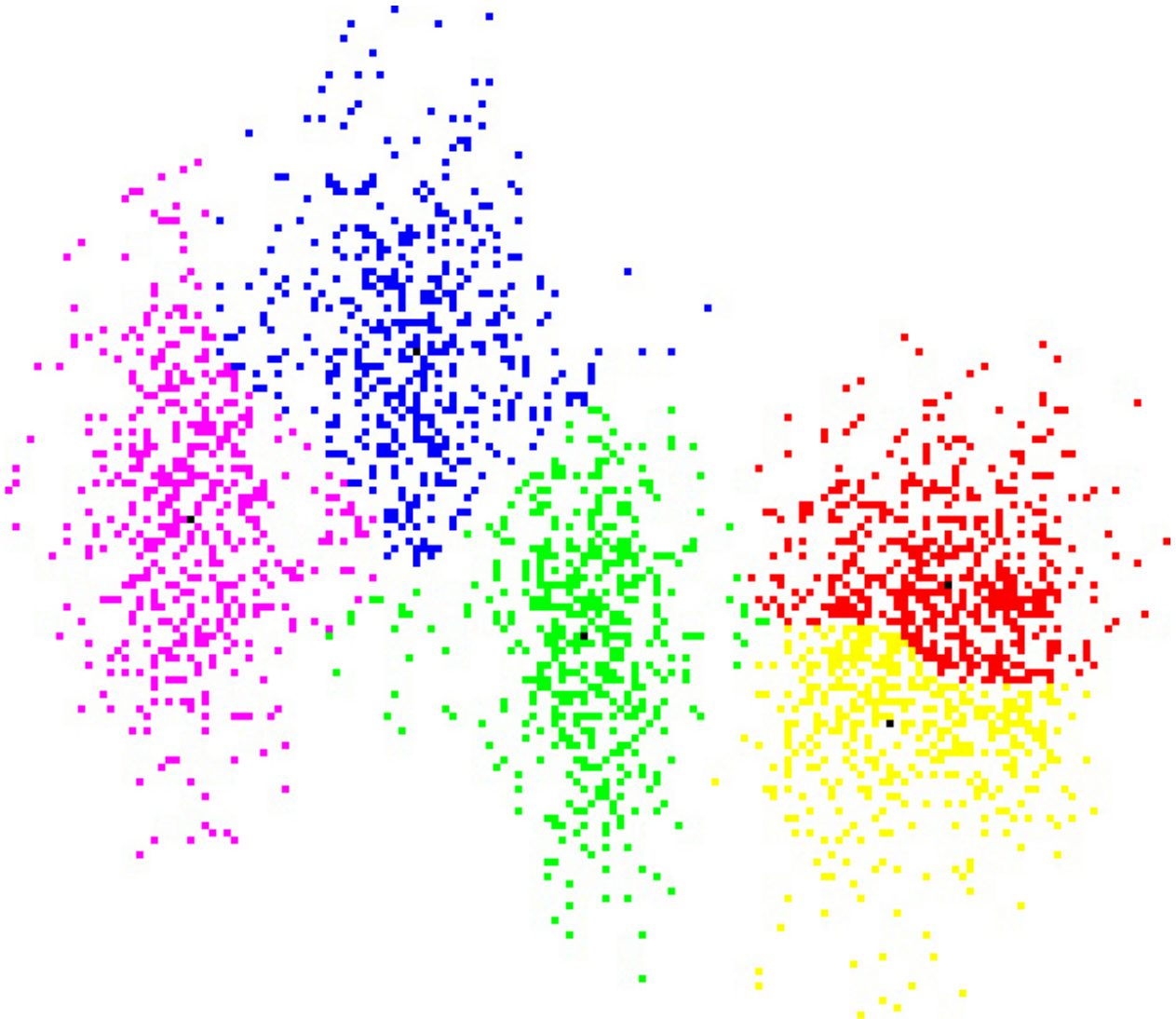
5 x 500 punktów, 3 skupiska



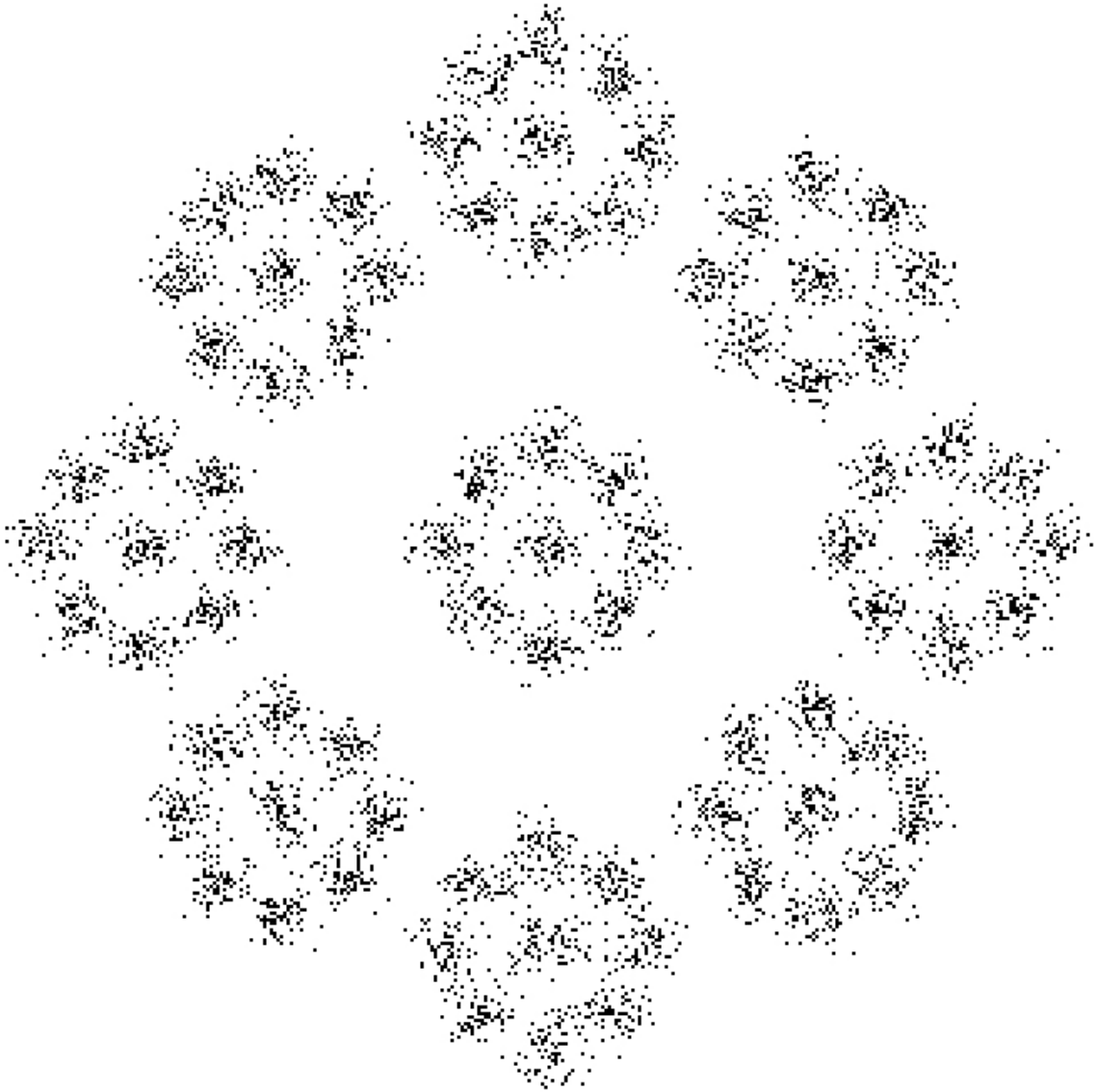
5 x 500 punktów, 4 skupiska



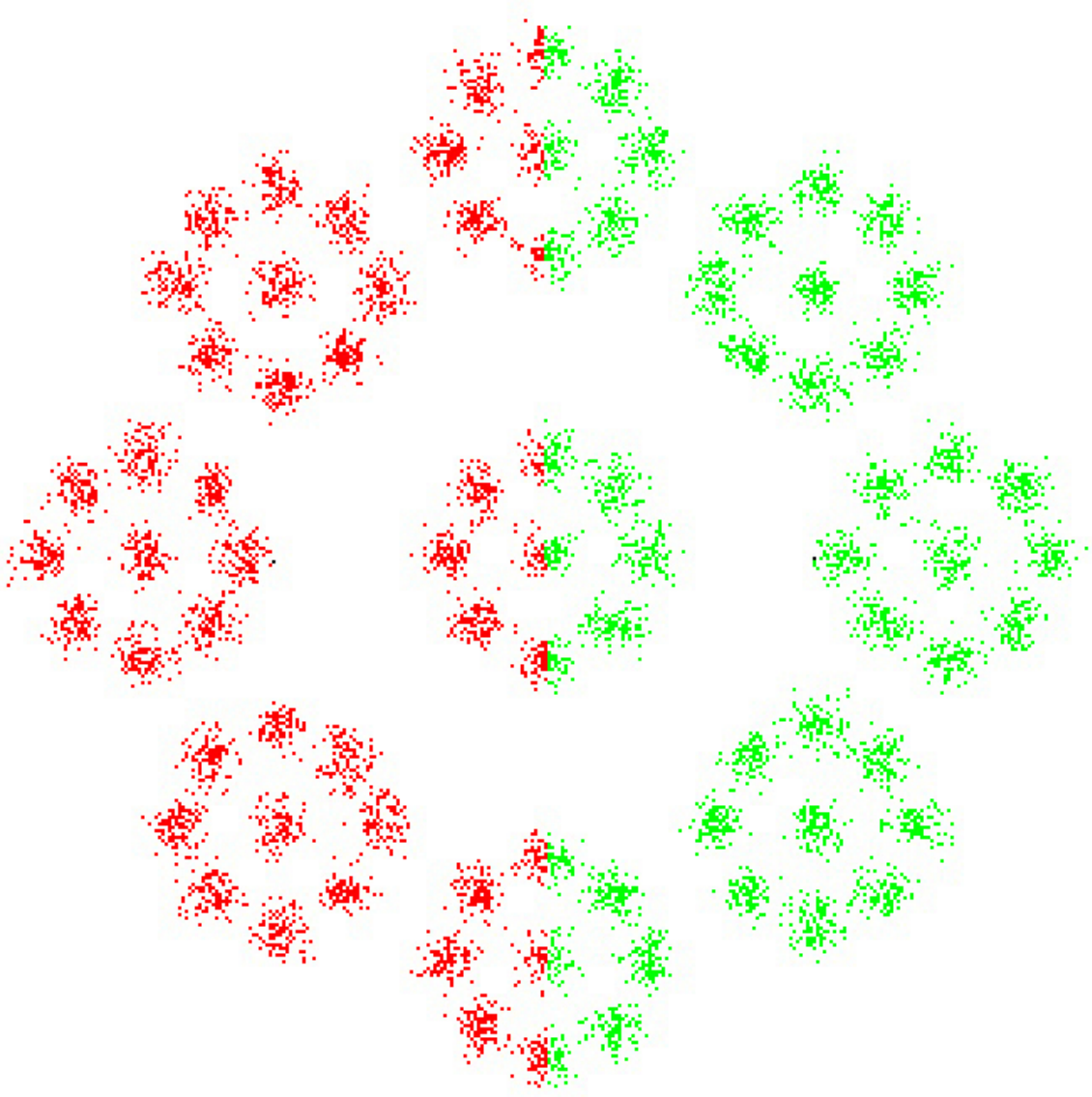
5 x 500 punktów, 5 skupisk



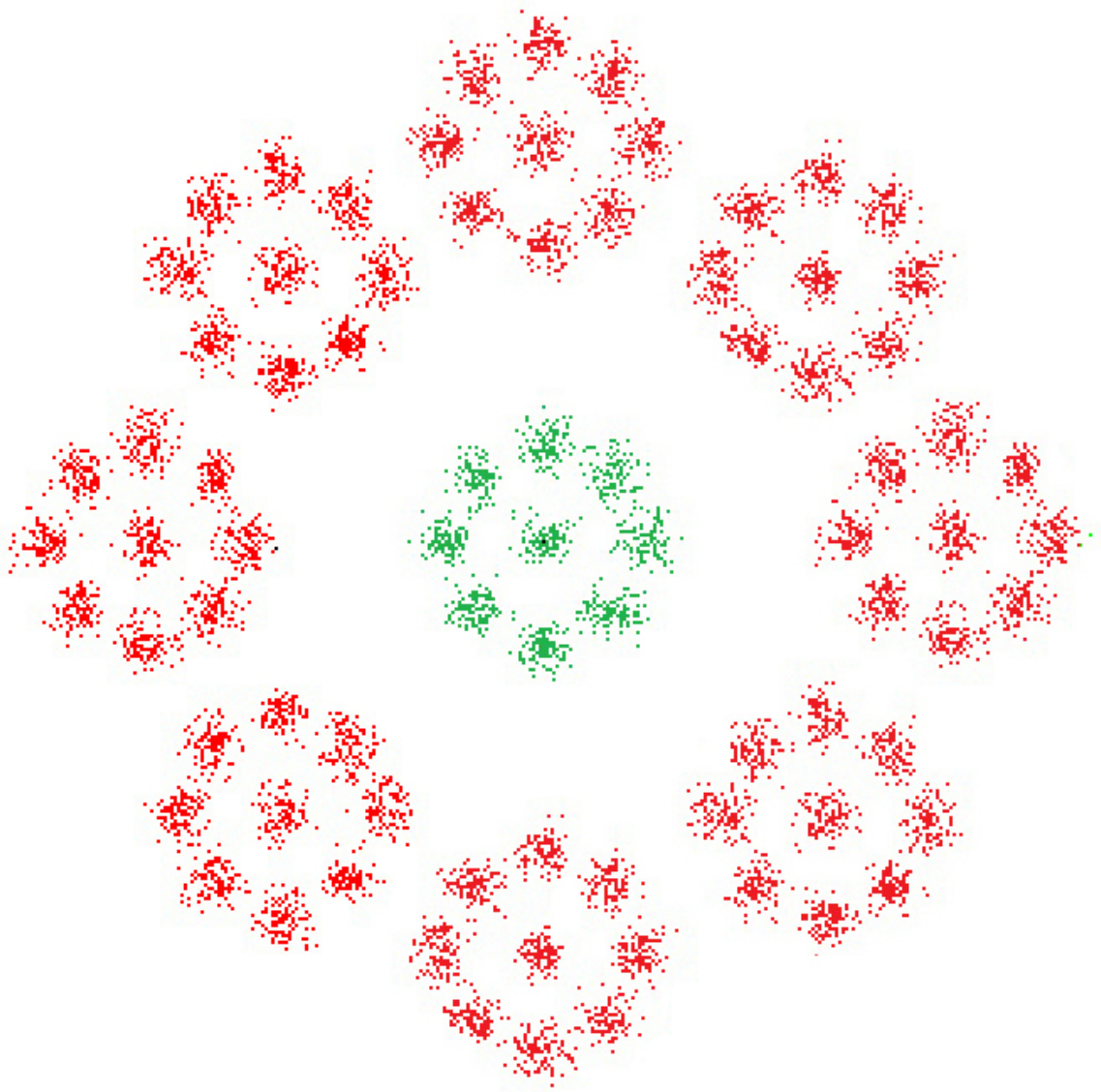
81 x 100 punktów



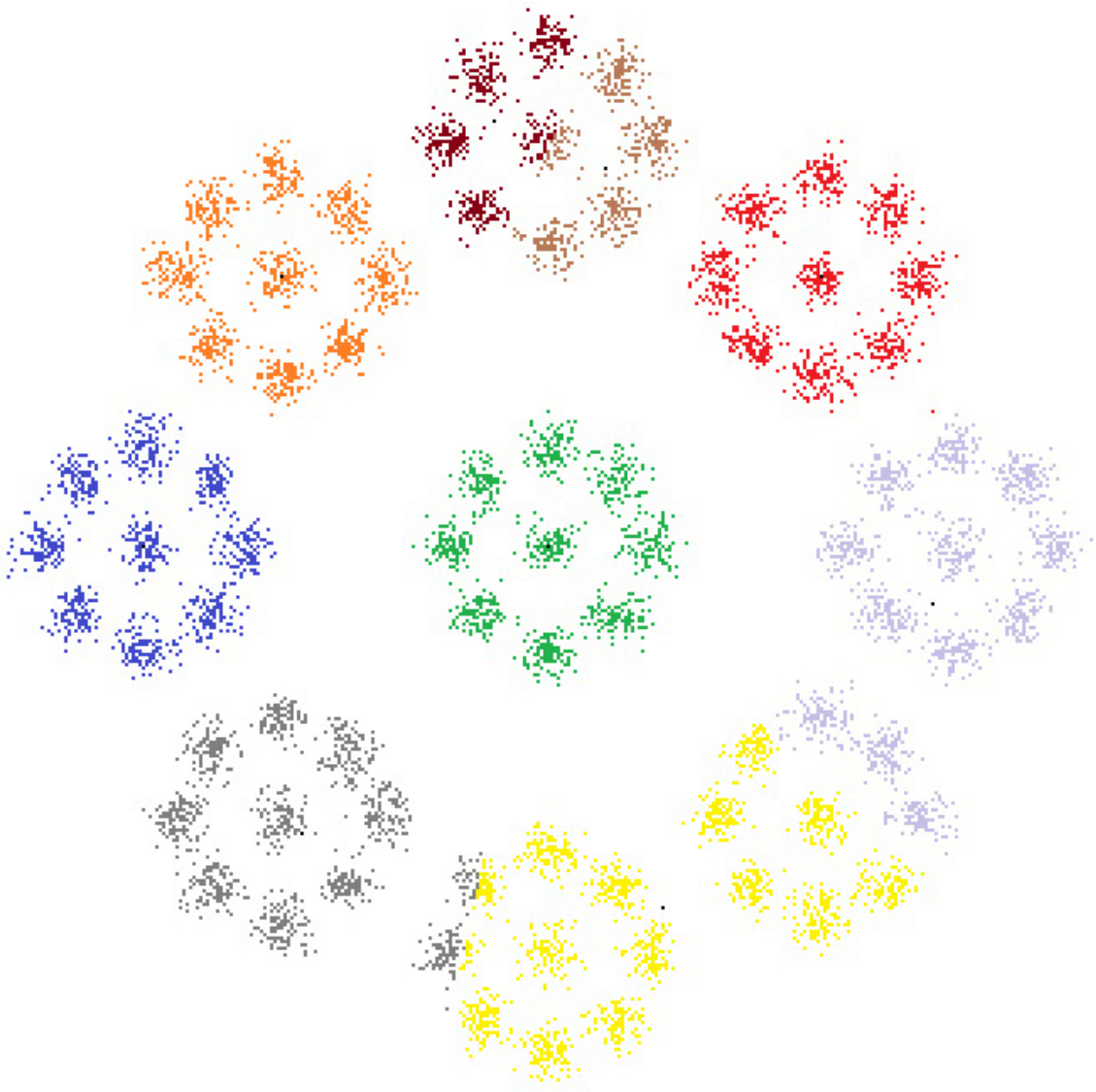
81 x 100 punktów, 2 skupiska (1)



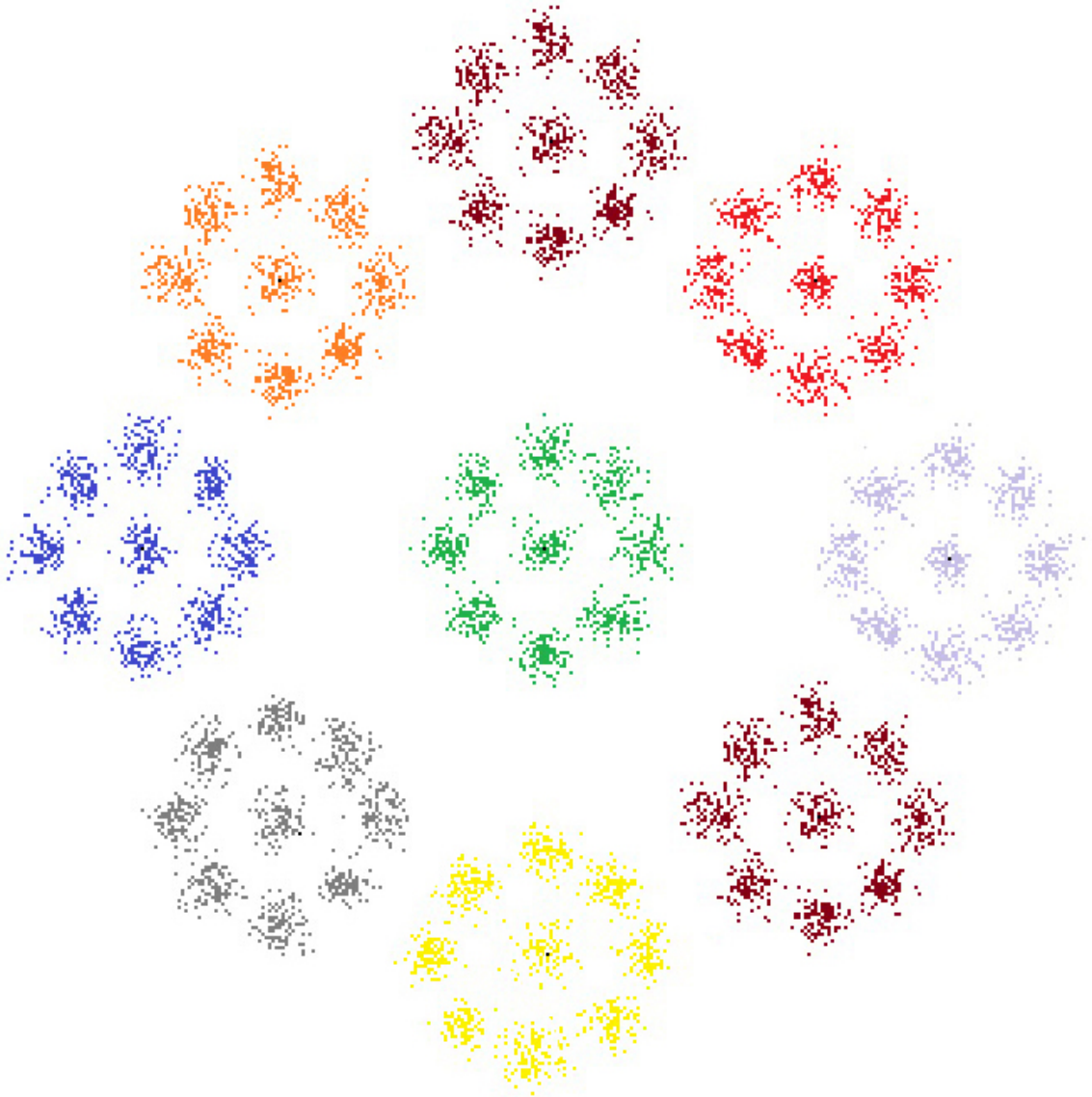
81 x 100 punktów, 2 skupiska (2)



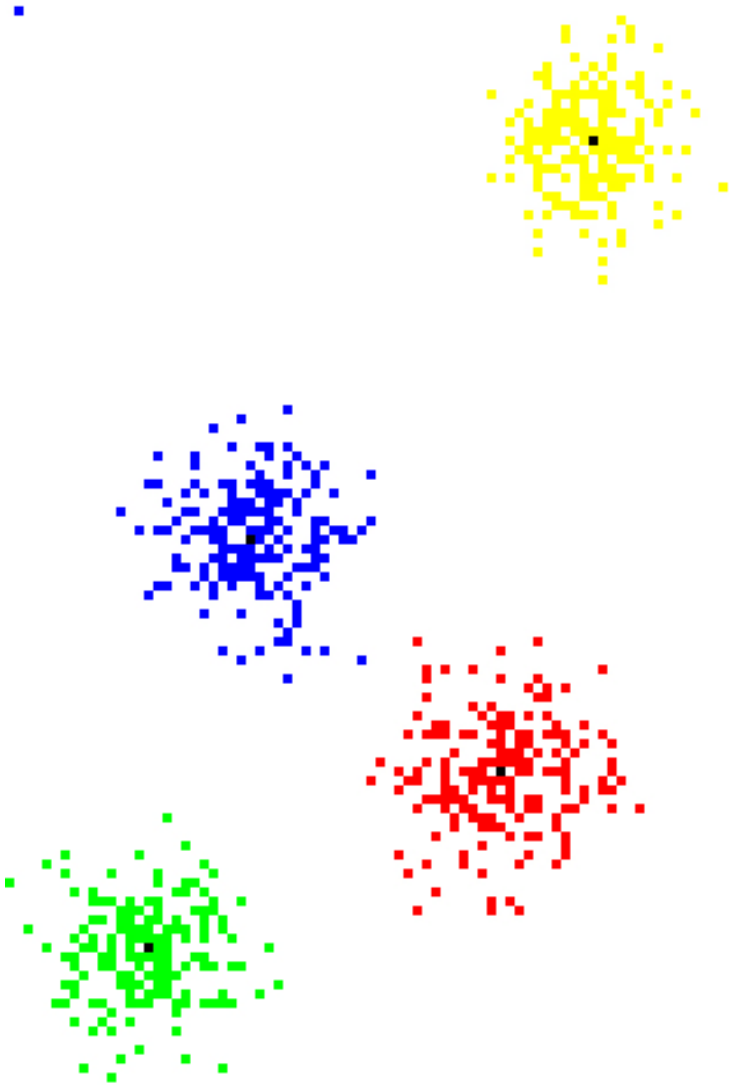
81 x 100 punktów, 9 skupisk (1)



81 x 100 punktów, 9 skupisk (2)

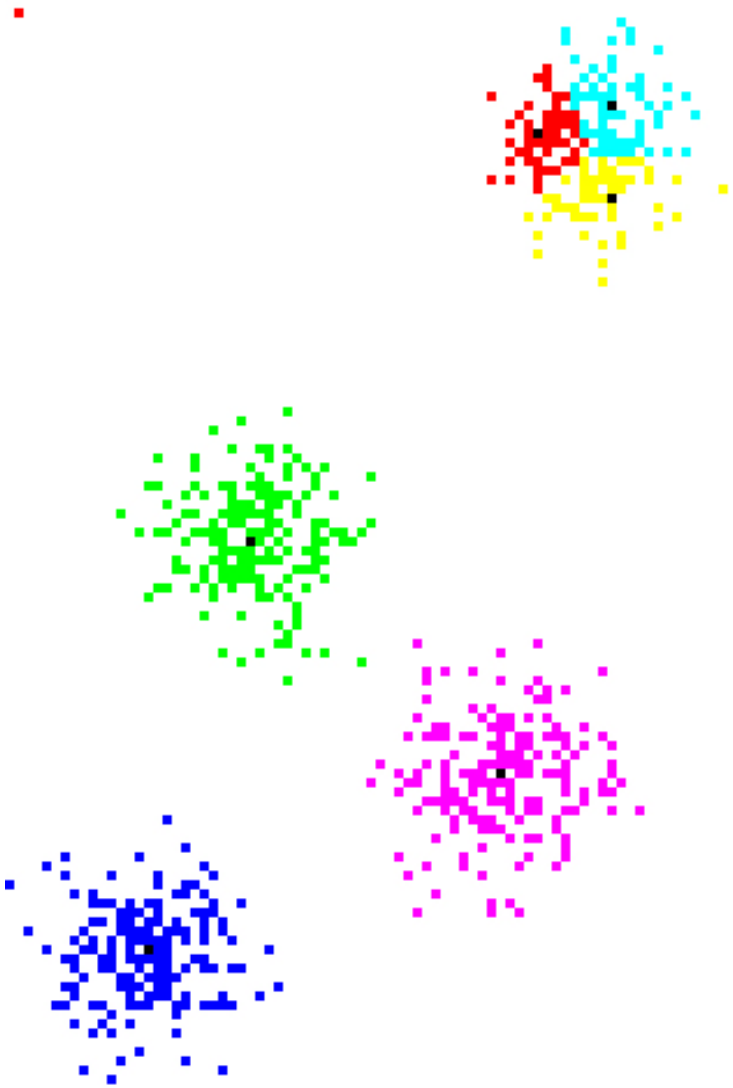


Punkty osobliwe



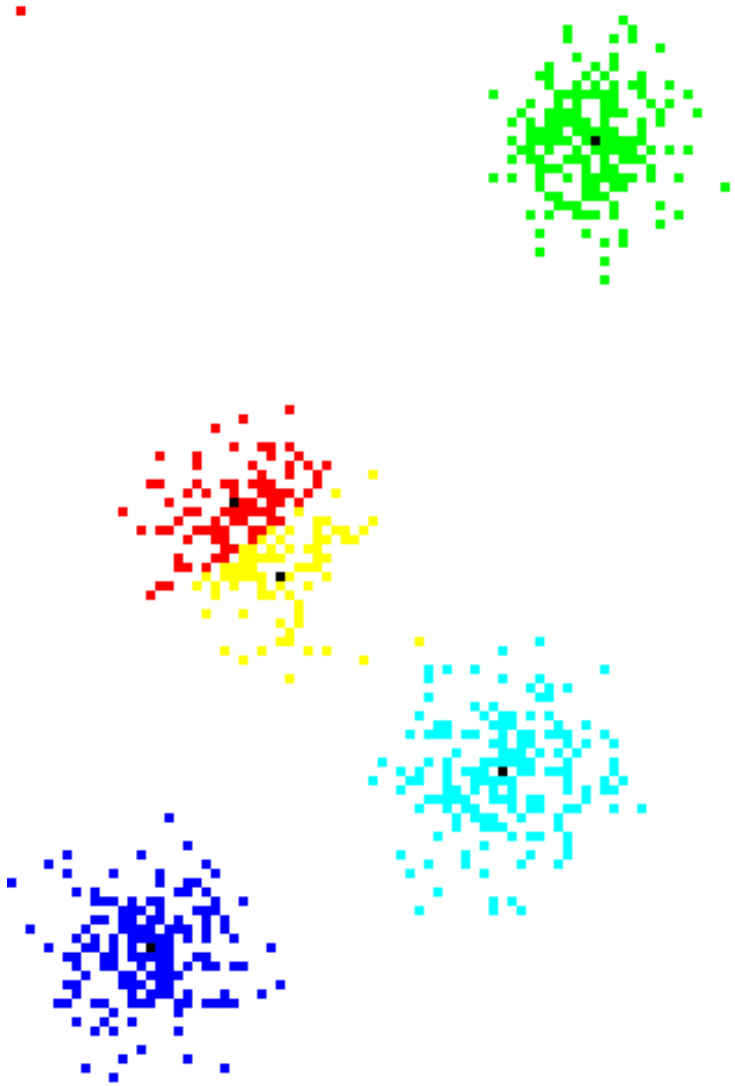
Cztery klastry

Punkty osobliwe



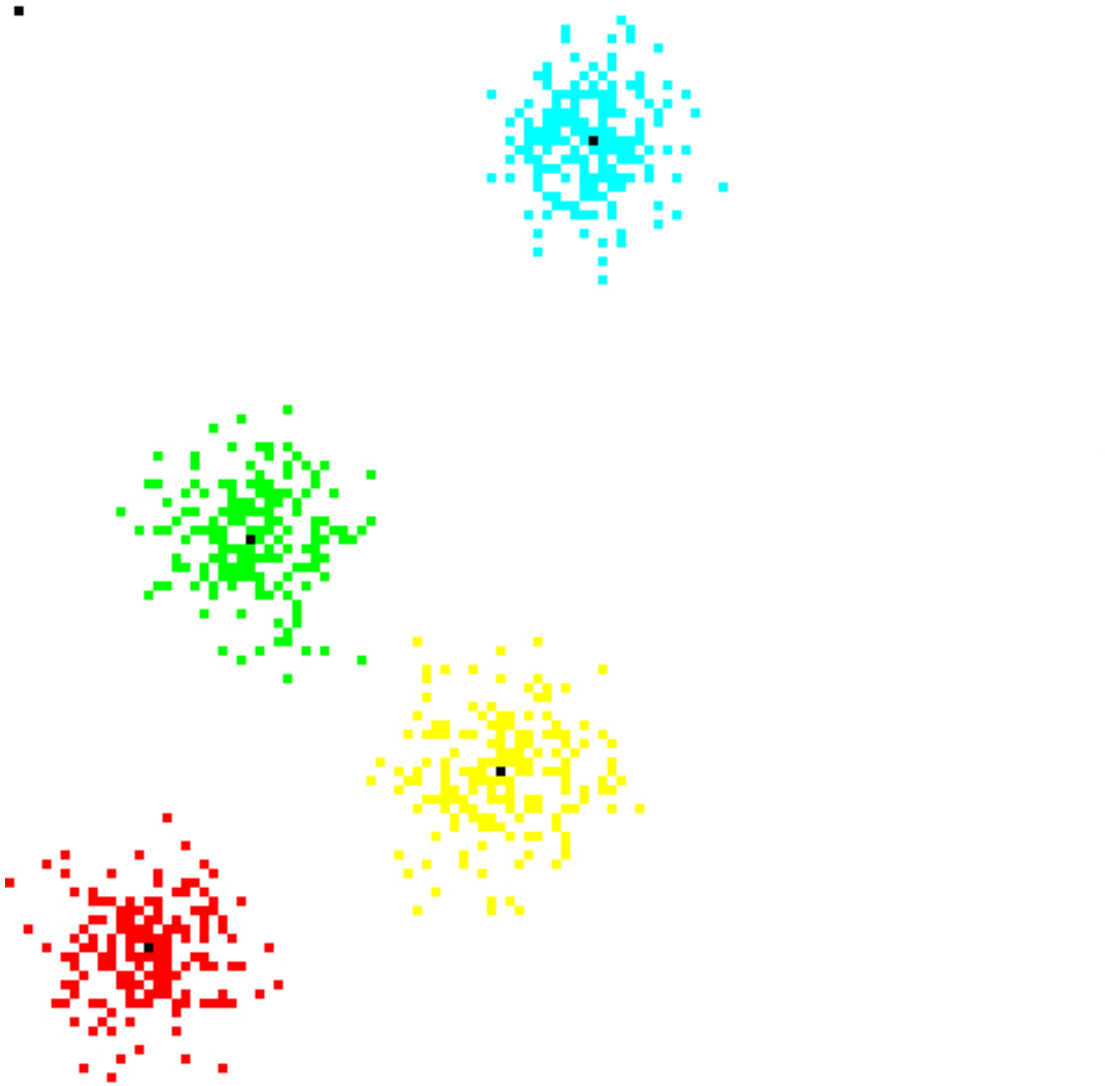
Sześć klastrów

Punkty osobliwe



Sześć klastrów — jeden punkt osobliwy zidentyfikowany

Punkty osobliwe



Sześć klastrów — dwa punkty osobliwe zidentyfikowane

Eksploracja danych zmienia świat

Według przewidywań internetowego magazynu *ZDNET News* z 8 lutego 2001 roku eksploracja danych miała być *jednym z najbardziej rewoulcyjnych osiągnięć następnej dekady*. Rzeczywiście, *MIT Technology Review* wybrało eksplorację danych jako jedną z dziesięciu nowych technologii, które zmienią świat.

Jak sobie radzić z nadmiarem informacji?

Należy torturować dane tak długo, aż zaczną zeznawać.

Jak sobie radzić z nadmiarem informacji?

Należy torturować dane tak długo, aż zaczną zeznawać.

Koniec

Dziękuję za uwagę.