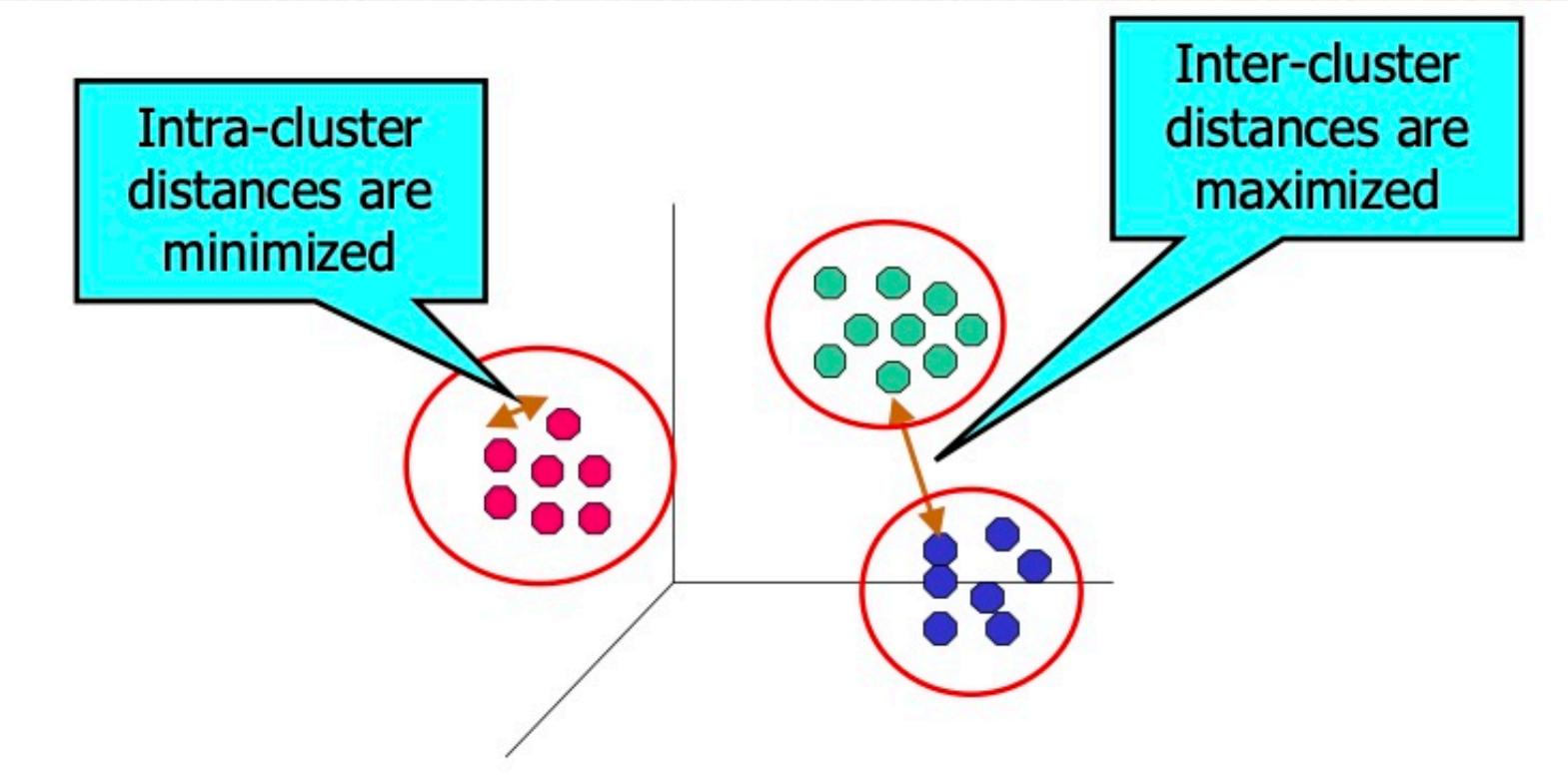
4 Grupowanie.

Czym jest grupowanie?

Znajdowanie takich grup obiektów, aby obiekty w grupie były do siebie podobne (lub pokrewne) i różniły się od (lub nie były pokrewne) obiektów w innych grupach:



Zastosowanie grupowania

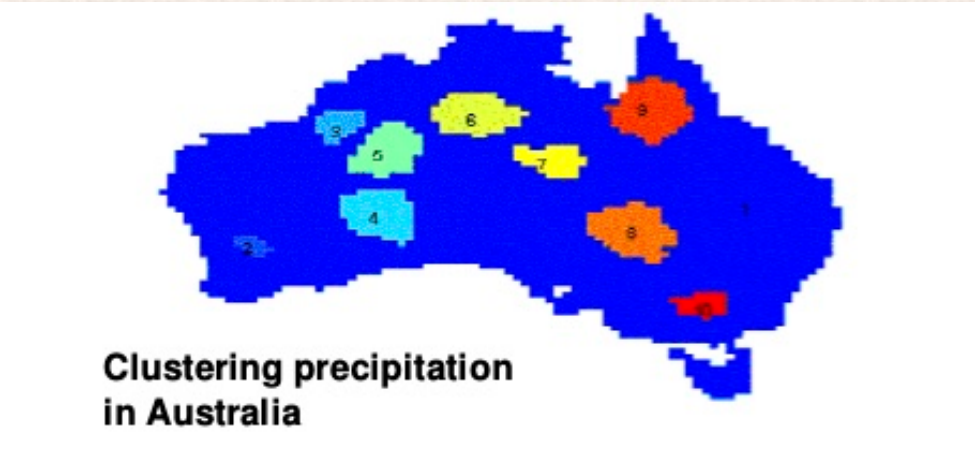
Zrozumienie:

- Grupuj dokumenty do przeglądania, grupuj geny i białka, które mają 2 podobne funkcje lub grupuj akcje z podobnymi 3 fluktuacjami cen:

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

Podsumowanie:

- Redukcja rozmiaru dużych zbiorów danych:


Clustering precipitation
in Australia

Przykłady i zastosowania grupowania:

- Segmentuj bazę klientów na podstawie podobnych wzorców zakupowych
- Grupuj domy w mieście w dzielnice na podstawie podobnych cech
- Grupuj rośliny w kategorie
- Zidentyfikuj podobne wzorce korzystania z sieci

Czym nie jest grupowanie?

• Nadzorowana klasyfikacja

- Dysponuj informacjami na etykiecie klasy

• Prosta segmentacja

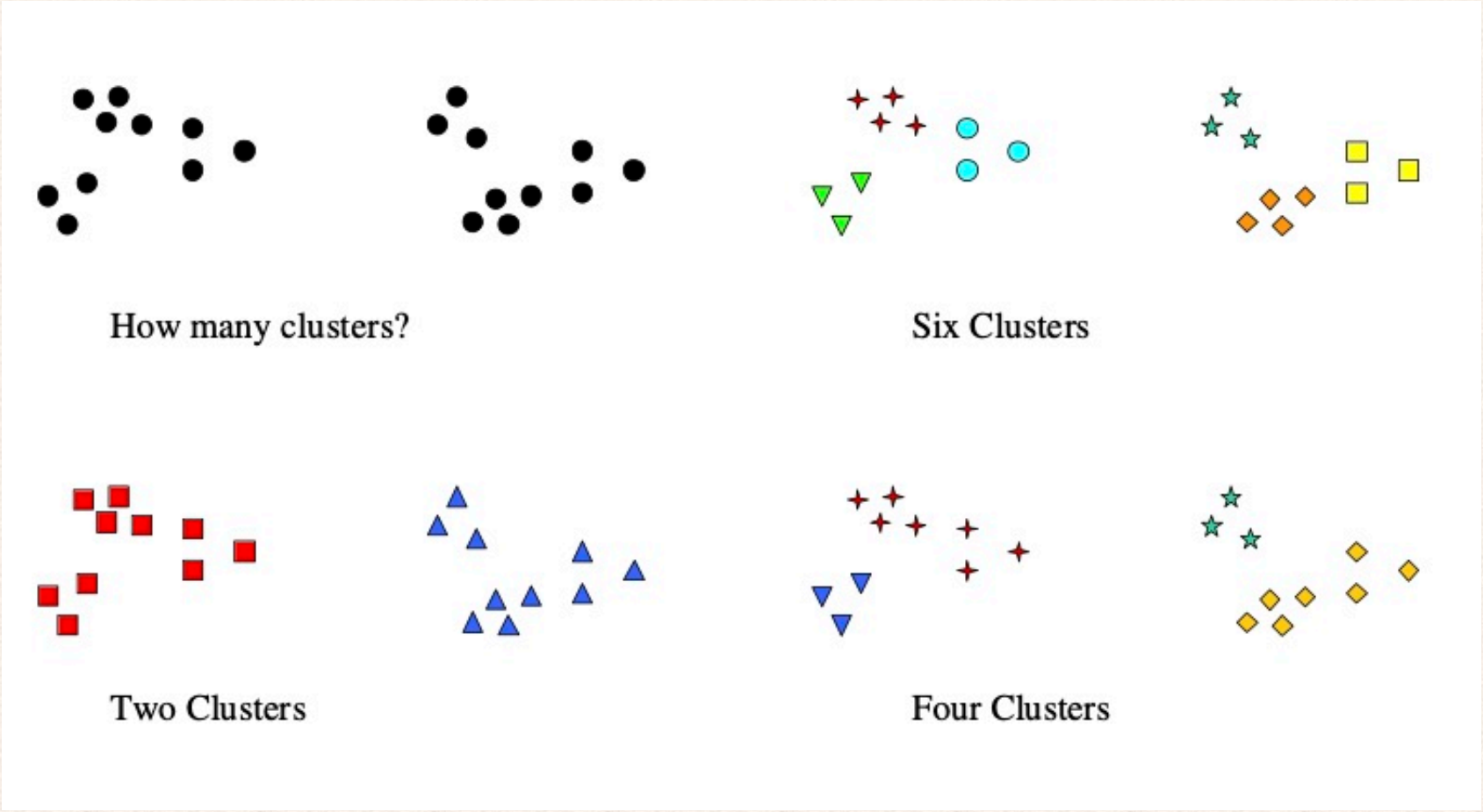
- Podział uczniów na różne grupy rejestracyjne alfabetycznie według nazwiska

• Grupuj wyniki zapytania

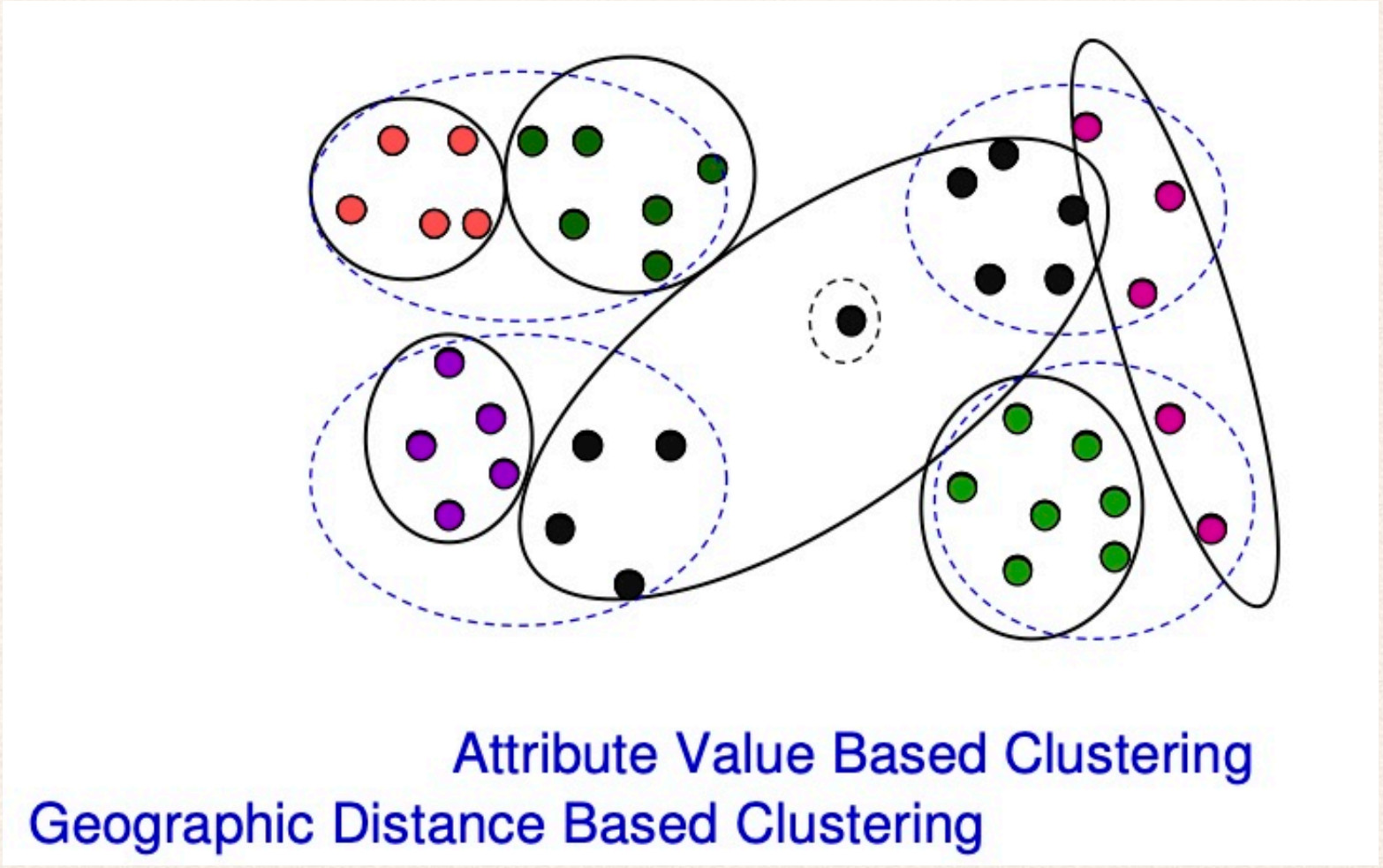
- Gdy grupowanie jest wynikiem zewnętrznej specyfikacji

• Partycjonowanie wykresów

Zdefiniowanie grupy może być niejednoznaczne...



Grupowanie może być przeprowadzane na wiele sposobów



Problem grupowania

• Przy danej bazie danych $D = \{t_1, t_2, ..., t_n\}$ krotek i być może również wartości całkowitej k, problem klastrowania polega na zdefiniowaniu odwzorowania $f: D \to \{1, .., k\}$ takiego, że każda t_i jest przypisana do jednego klastra K_j, $1 \leq j \leq k$.

• Klaster Kj zawiera dokładnie te krotki, które zostały na niego odwzorowane.

Klasyfikacja vs. grupowanie

• W klasyfikacji podawany jest zbiór grup, do których należą obiekty, a zadaniem jest rozróżnienie grup na podstawie wartości atrybutów obiektów

• W klastrach nie ma wcześniejszej wiedzy: znaczenie (i często liczba) klastrów jest nieznane, a celem jest ich odkrycie

• Tworzenie klastrów jest czasami nazywane uczeniem się bez nadzoru

Typy grupowania

• Klastrowanie to zbiór klastrów

• Ważne rozróżnienie między hierarchicznymi i partycjonalnymi zbiorami klastrów

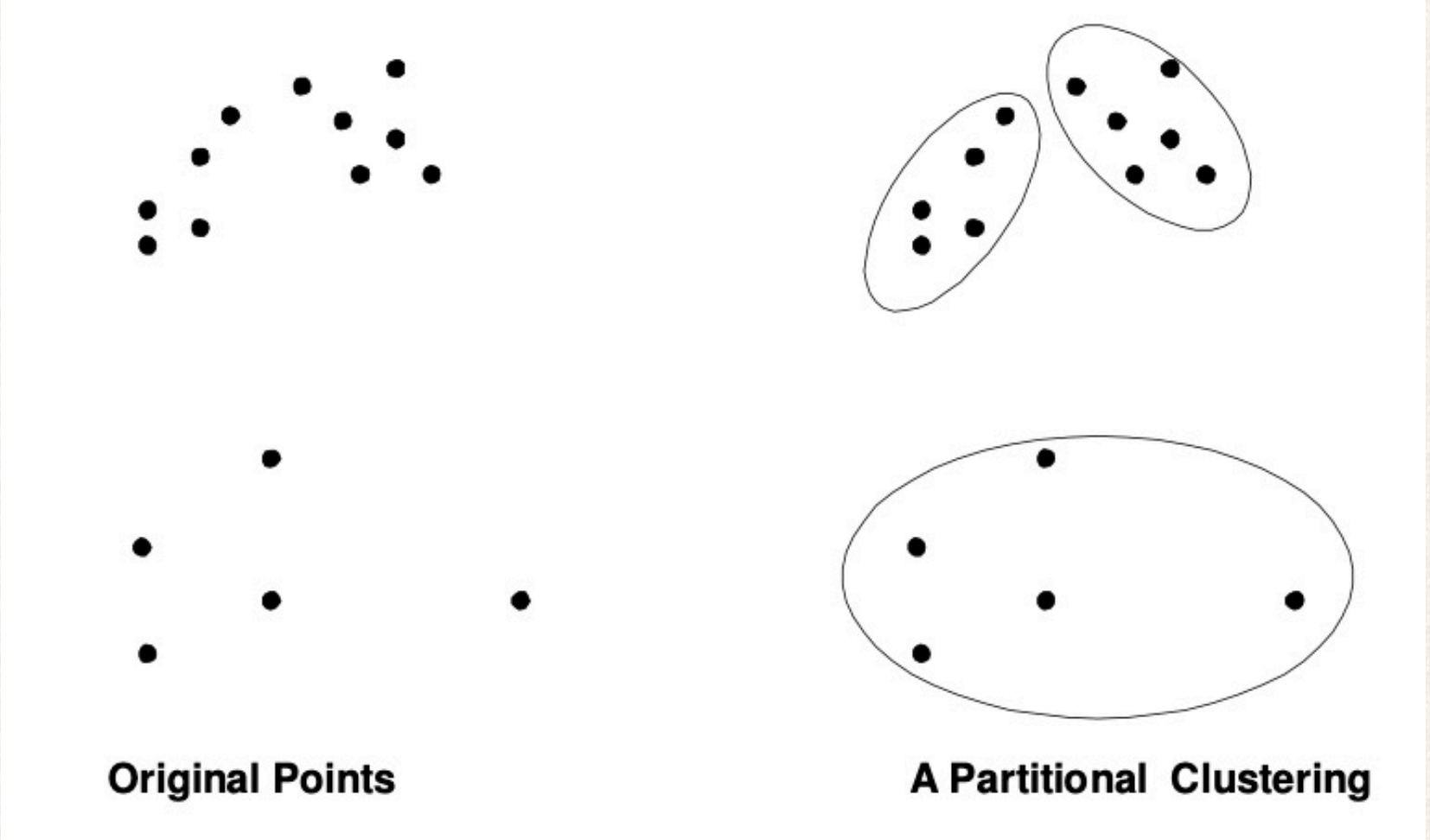
• Grupowanie partycjonalne

- Podział obiektów danych na nienakładające się podzbiory (klastry) w taki sposób, że każdy obiekt danych znajduje się dokładnie w jednym podzbiorze

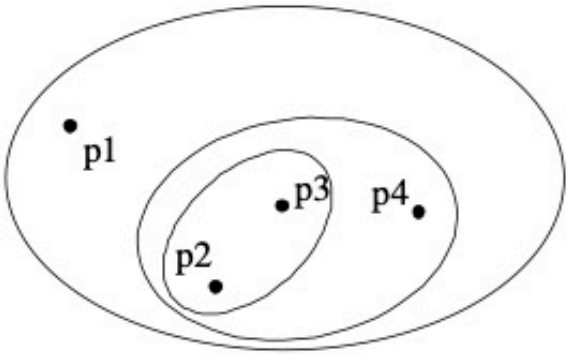
• Grupowanie hierarchiczne

- Zestaw zagnieżdżonych klastrów zorganizowanych w postaci hierarchicznego drzewa

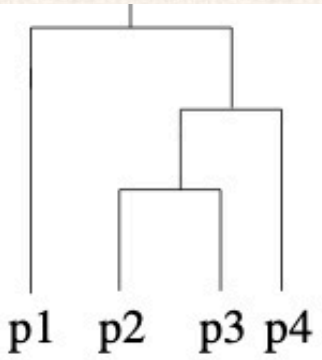
Grupowanie partycjonalne



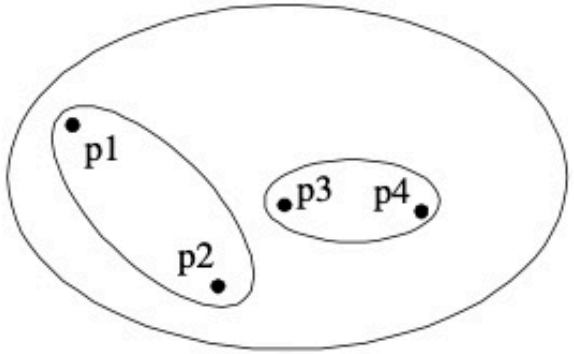
Grupowanie hierarchiczne



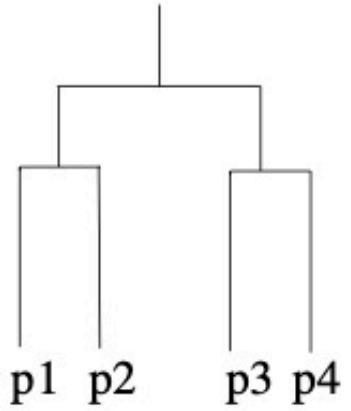
Traditional Hierarchical Clustering



Traditional Dendrogram

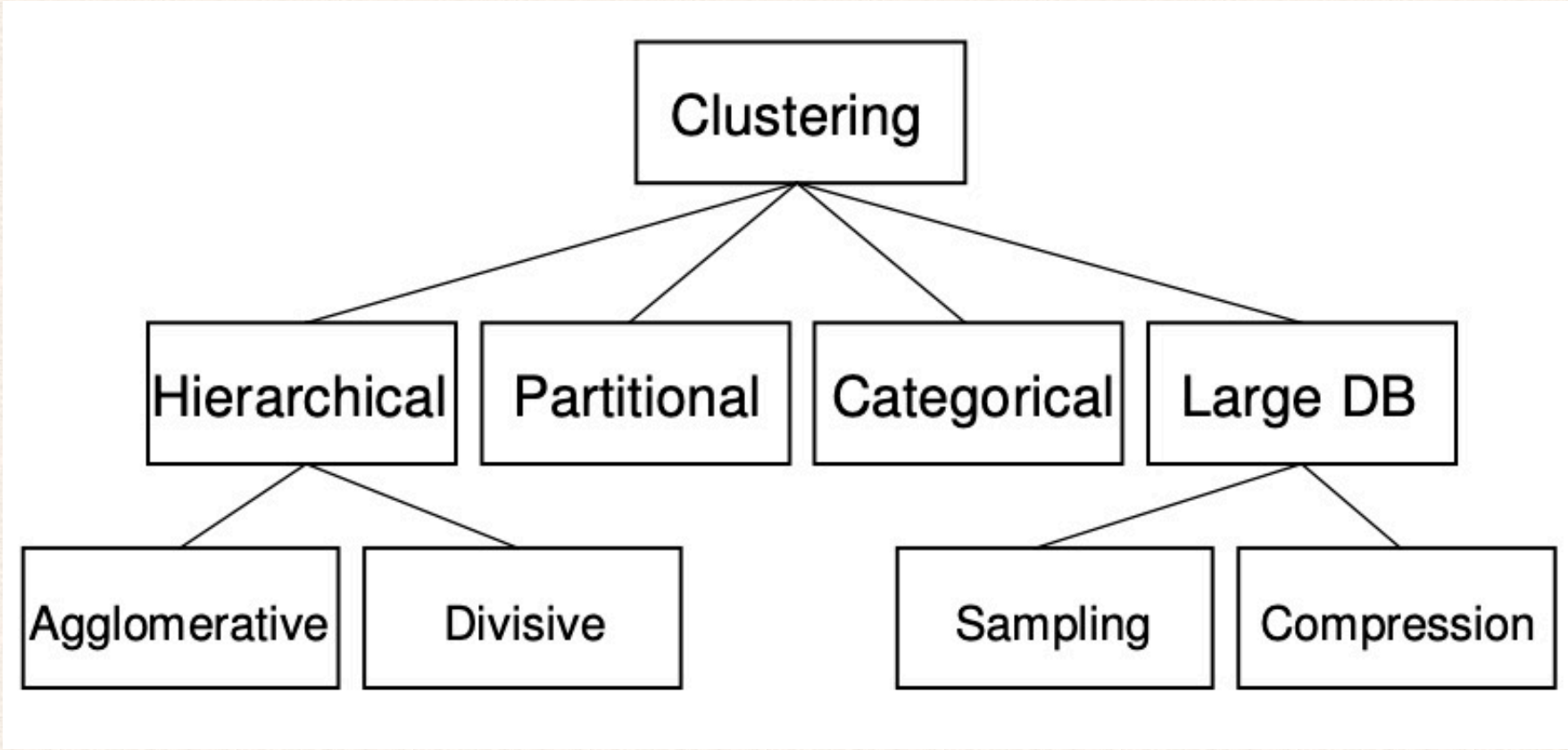


Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

Podejścia do grupowania



Inne klasyfikacje grupowania:

• Wyłączne vs. niewyłączne

- Klastry w grupowaniu niewyłącznym mogą zawierać punkty należące do wielu klastrów.

- Może reprezentować wiele klas lub punktów granicznych

• Rozmyte vs. nierozmyte

- W grupwaniu rozmytym do każdego klastra przypisujemy wagę od 0 do 1

- Wagi muszą sumować się do 1

- Klastrowanie probabilistyczne ma podobne cechy

• Częściowe vs. całkowite

- W pewnych przypadkach możemy chcieć grupować tylko część

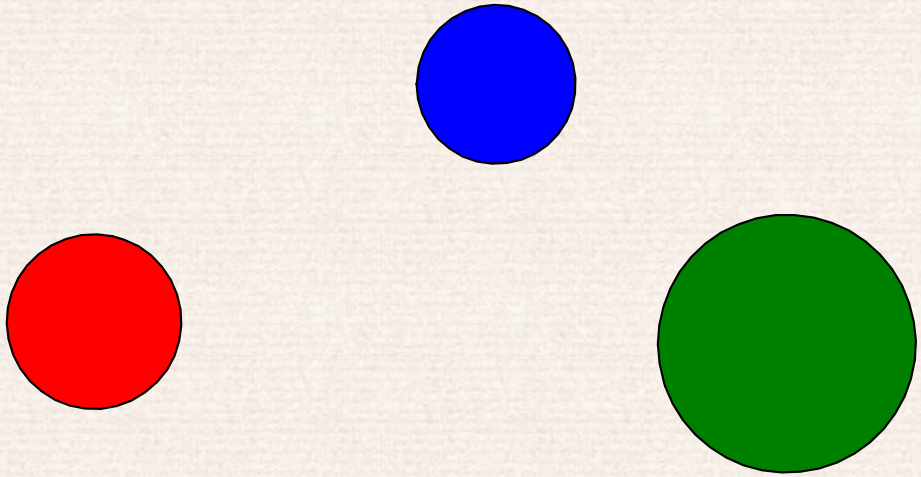
• Heterogeniczne kontra jednorodne

- Klaster szeroko róťni siŚ rozmiarami, kształtami i gŚstoąciĄ

Rodzaje klastrów: dobrze rozdzielone.

• Dobrze rozdzielone klastry:

- Klaster to taki zbiór punktów, że każdy punkt w klastrze jest bliżej do każdego innego punktu w klastrze niż do dowolnego punktu spoza klastra:



Rodzaje klastrów: oparte na centrum

• Oparte na centrum

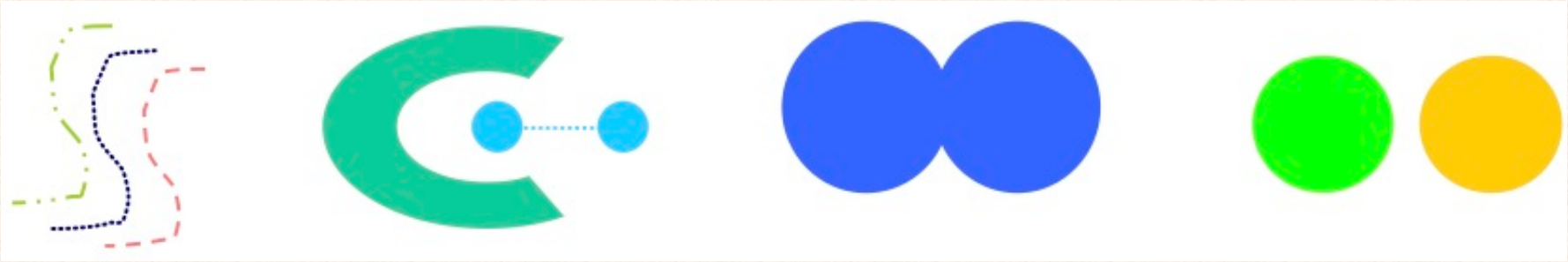
- Klaster to taki zbiór obiektów, że obiekt w klastrze jest bliżej do „środka" klastra niż do środka jakiegokolwiek innego skupienia

- Środek klastra jest często centroidem, średnią wszystkich punktów w klastrze lub medoidą, najbardziej „reprezentatywnym" punktem klastra:

Rodzaje klastrów: ciągłe

• Ciągły klaster (najbliższy sąsiad lub przechodni)

- Klaster to taki zbiór punktów, że punkt w klastrze jest bliższy do jednego lub więcej innych punktów w klastrze niż do dowolnego punktu spoza klastra:

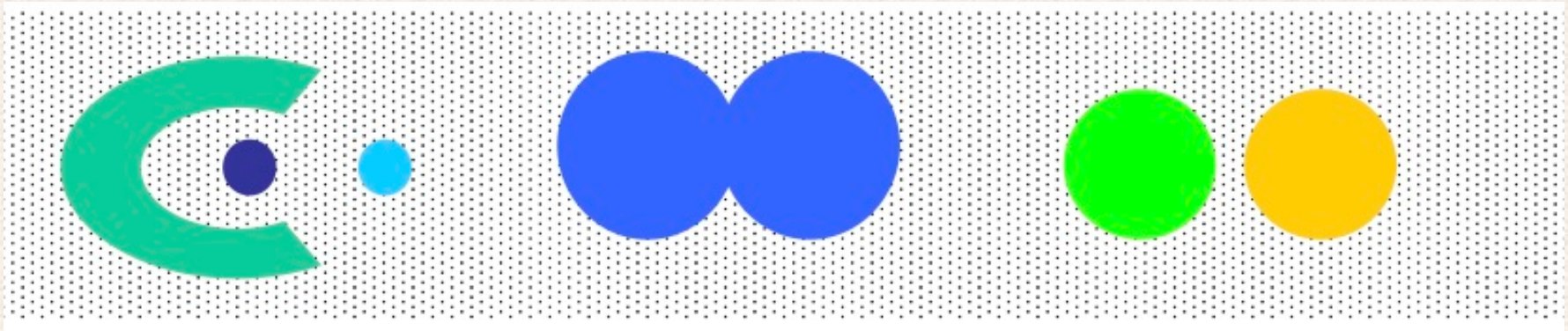


Rodzaje klastrów: oparte na gęstości

• Oparte na gęstości

- Klaster to gęsty region punktów, który jest oddzielony regionami o małej gęstości od innych regionów o dużej gęstości.

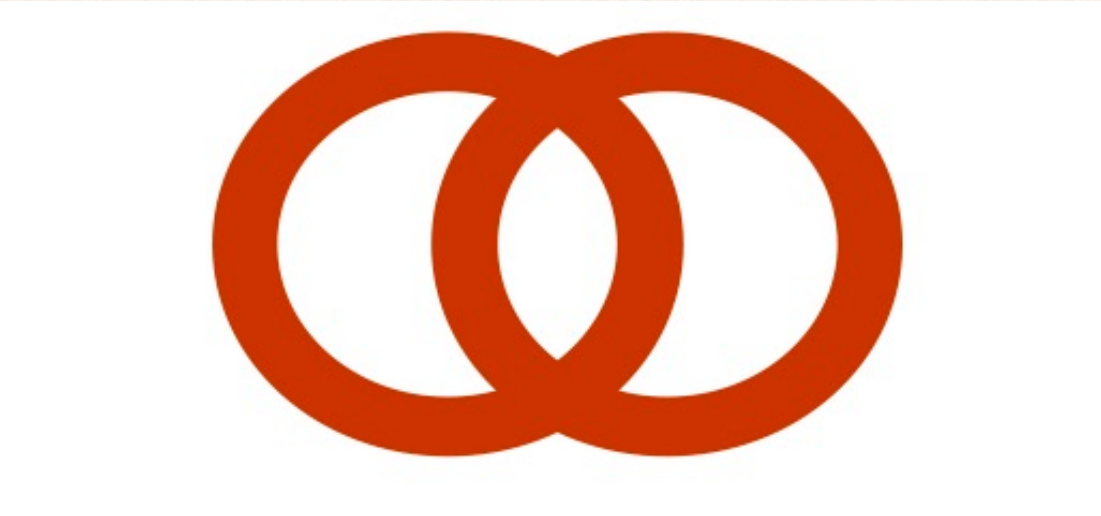
- Używane, gdy klastry są nieregularne lub splecione, oraz gdy występuje szum i wartości odstające:



Rodzaje klastrów: koncepcyjne

• Wspólna własność lub klastry koncepcyjne

- Znajduje klastry, które mają wspólną właściwość lub reprezentują okreąloną koncepcję:



Rodzaje danych są ważne!

• Rodzaj miernika bliskości lub gęstości

- Ten środek jest nietypowy, ale skupiony centralnie

• Rzadkość

- Podobieństwo typu dyktatu

- Dodatkowa efektywność

• Typ atrybutu

- Dyktowanie typu podobieństwa

• Rodzaj danych

- Dyktowanie typu podobieństwa

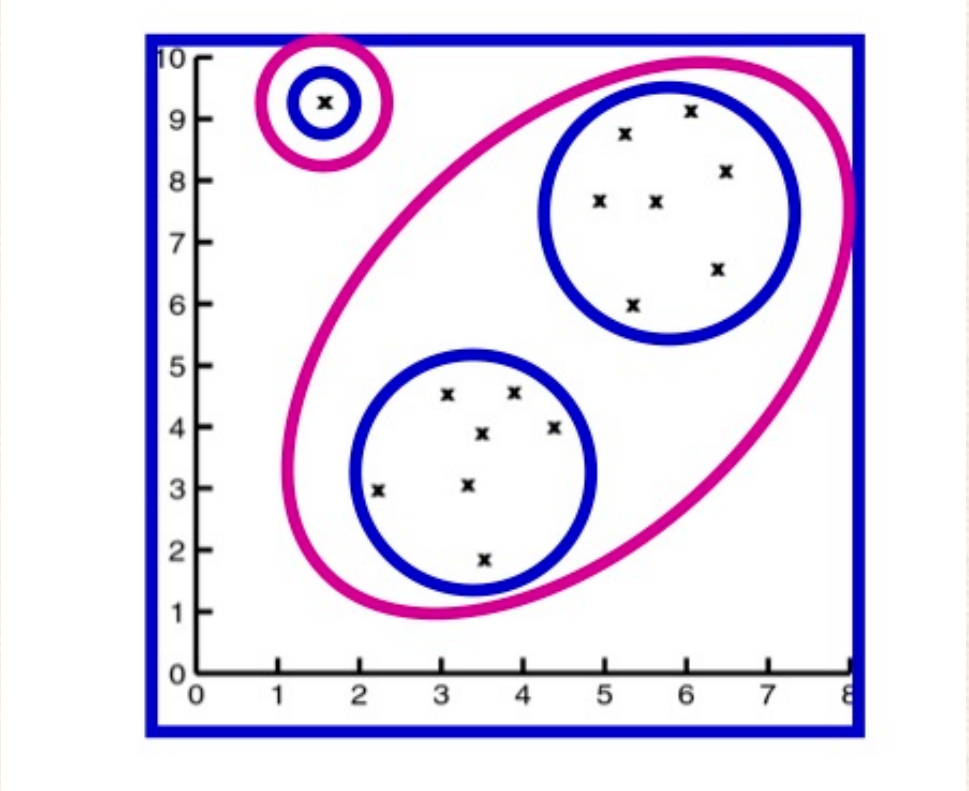
- Inne cechy, np. autokorelacja

• Wymiarowość

• Szum i wartości odstające

• Rodzaj dystrybucji

Wpływ wartości odstających na grupowanie



Parametry klastra

$$centroid = C_m = \frac{\sum_{i=1}^{N}(t_{mi})}{N}$$

$$radius = R_m = \sqrt{\frac{\sum_{i=1}^{N}(t_{mi} - C_m)^2}{N}}$$

$$diameter = D_m = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(t_{mi} - t_{mj})^2}{(N)(N-1)}}$$

Algorytmy grupujące

- Algorytmy partycjonujące

- K-means i jego warianty

- Hierarchiczne algorytmy grupowania

- Algorytmy grupowania oparte na gęstości

Grupowanie K-means

- Partycjonujące podejście do grupowania
- Każdy klaster jest powiązany z centroidą
- Każdy punkt jest przypisany do klastra z najbliższą centroidą
- Należy określić liczbę klastrów K
- Podstawowy algorytm jest bardzo prosty:

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

Grupowanie K-means: przykład

- Dany jest klaster $K_i = \{t_{i1}, t_{i2}, ..., t_{im}\}$, niech centroidą klastra będzie

$$m_i = (1/m)(t_{i1} + ... + t_{im})$$

Niech dane będzie: $\{2, 4, 10, 12, 3, 20, 30, 11, 25\}, k = 2$

- Losowo wybierz kilka początkowych średnich: $m_1 = 3, \quad m_2 = 4$
- $K_1 = \{2, 3\}, \quad K_2 = \{4, 10, 12, 20, 30, 11, 25\}, \quad m_1 = 2,5, \quad m_2 = 16$
- $K_1 = \{2, 3, 4\}, \quad K_2 = \{10, 12, 20, 30, 11, 25\}, \quad m_1 = 3, \quad m_2 = 18$
- $K_1 = \{2, 3, 4, 10\}, \quad K_2 = \{12, 20, 30, 11, 25\}, \quad m_1 = 4,75, \quad m_2 = 19,6$
- $K_1 = \{2, 3, 4, 10, 11, 12\}, \quad K_2 = \{20, 30, 25\}, \quad m_1 = 7, \quad m_2 = 25$

Zatrzymaj się, ponieważ klastry z tymi centroidami są takie same.

Grupowanie K-means: szczegóły

• Początkowe centroidy są często wybierane losowo.

- Klastry wyprodukowane na różne sposoby.

• Centroida jest (zazwyczaj) średnią punktów w klastrze.

• „Bliskość" jest mierzona odległością euklidesową, podobieństwem cosinusowym, korelacją itp.

• K-means będą zbieżne dla wspólnych miar podobieństwa wymienionych powyżej.

• Zbieżność zachodzi w pierwszych kilku iteracjach.

- Często warunek zatrzymujący algorytm jest zmieniany na „dopóki stosunkowo niewiele punktów zienia klastry"

• Złożoność jest rzędu $O(n\,K\,I\,d)$, gdzie: n = liczba punktów, K = liczba klastrów, I = liczba iteracji, d = liczba atrybutów

Dwa różne grupowania K-means