Jurusan **Teknik Informatika**
Fakultas Teknologi Industri, Universitas Islam Indonesia

**Kampus UII Terpadu, Jl.Kaliurang Km.14,5 Yogyakarta,**
☎ 0274-895287 ext.122 | http://informatics.uii.ac.id | informatika@fti.uii.ac.id

Materi Kuliah – [11]:
Data Mining

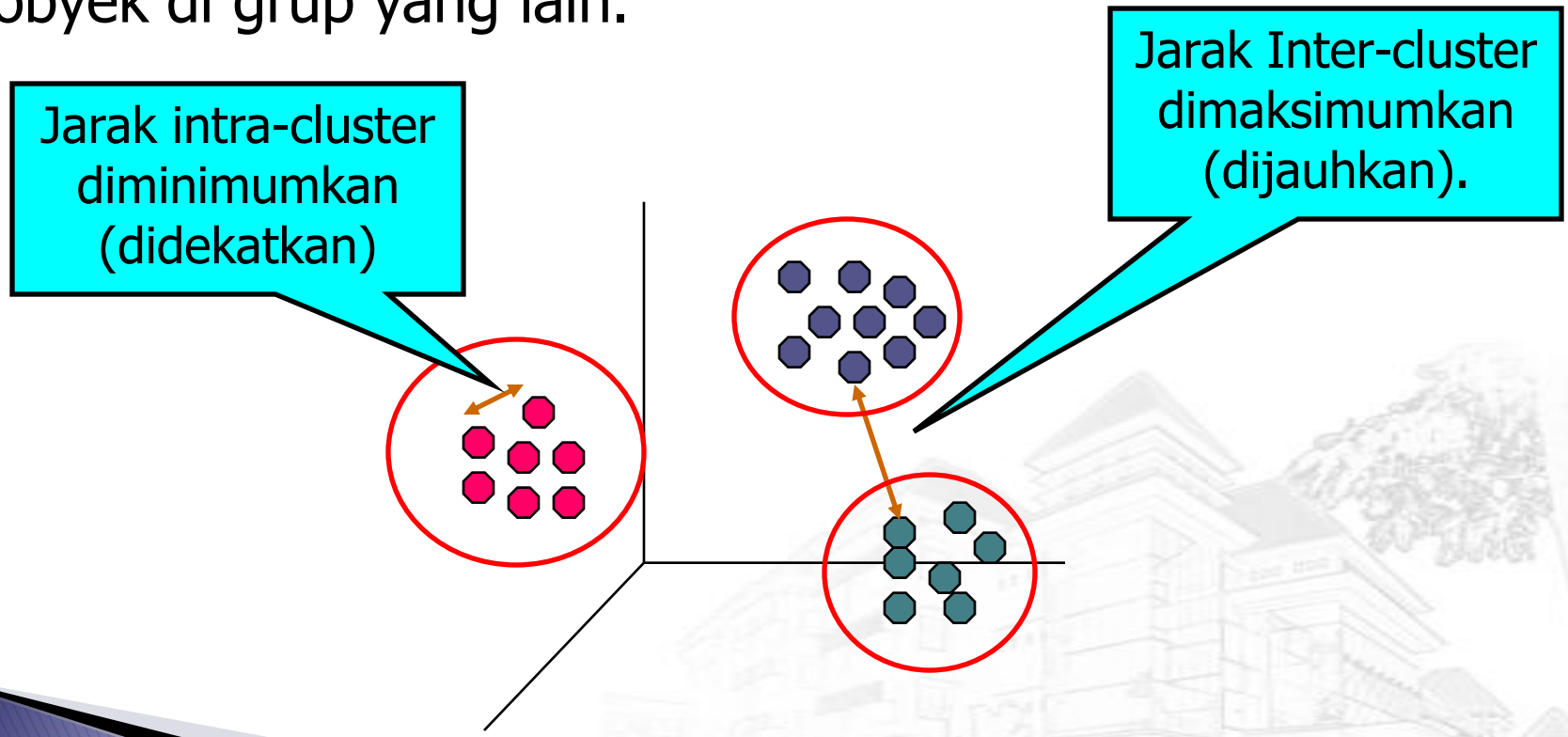
# Algoritma K-Means

## Sumber: Introduction to Data Mining by Tan, dkk.
## Discovery Knowledge in Data by Daniel T. Larose

lizda.iswari@uii.ac.id
Juni 2012

# Apakah arti "Cluster Analysis"?

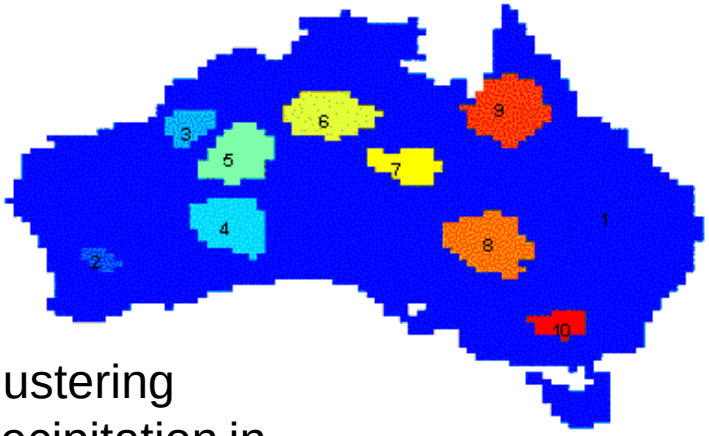
- Usaha mengelompokkan objek data sehingga obyek-obyek yang berada dalam satu grup "similar" (saling berhubungan) dan memiliki perbedaan dengan obyek-obyek di grup yang lain.



# Aplikasi Cluster Analysis

- **Understanding (pemahaman data)**
  - Mengelompokkan dokumen yang saling terkait hasil browsing , mengelompokkan gen atau protein yang memiliki fungsi yang sama, atau mengelompokkan data stok dengar fluktuasi harga yang sama.
- **Summarisasi (ringkasan data)**
  - Tujuan: mengurangi ukuran data set yang besar.

| | *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 |




Clustering precipitation in Australia

# Tidak termasuk "Cluster Analysis:?

- Klasifikasi Terawasi (Supervised classification)
  - Berlaku untuk data yang telah memiliki informasi "class label"
- Segmentasi Sederhana
  - Memisahkan data registrasi mahasiswa dalam sejumlah kelompok berdasarkan urut abjad nama belakang mereka.
- Hasil Query
  - Pengelompokkan terjadi sebagai hasil dari spesifikasi external
- Graph partitioning
  - Some mutual relevance and synergy, but areas are not identical

# Pengertian sebuah "Cluster" bisa Ambigu

Tebak ada berapa clusters?

Enam Clusters

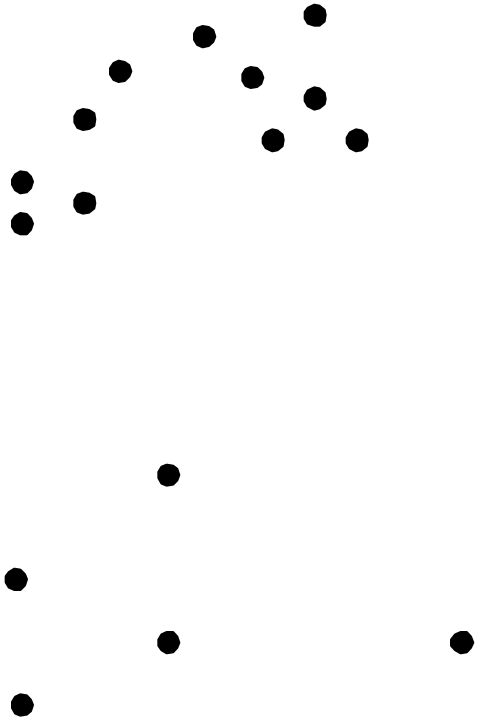
Dua Clusters

Empat Clusters

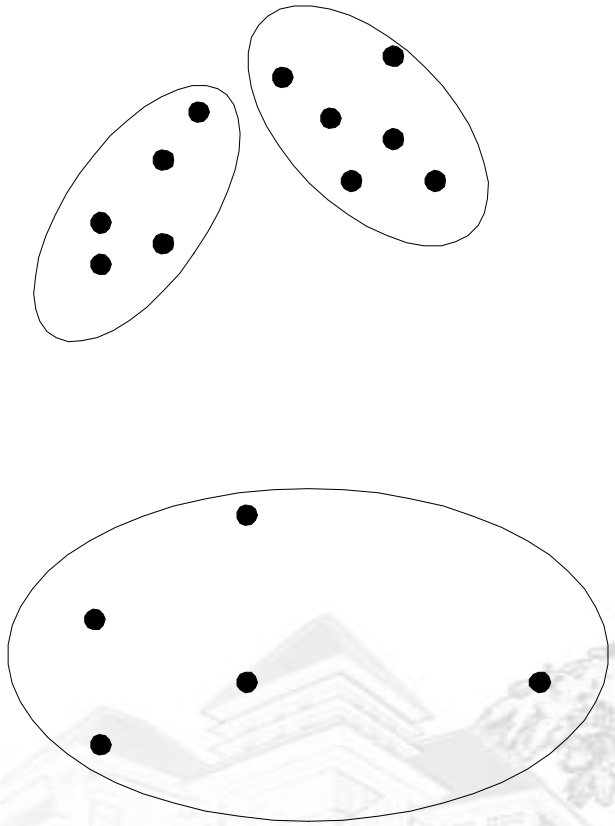
# Tipe Clusterings

- Sebuah clustering adalah kumpulan dari beberapa clusters
- Terdapat dua jenis clustering yang sangat berbeda:
  - hierarchical
  - partitional
- Partitional Clustering
  - Pembagian data obyek sebagai non-overlapping subsets (clusters) sehingga setiap data pasti berada tepat dalam satu cluster.
- Hierarchical clustering
  - Kumpulan clusters bersarang yang disusun sebagai pohon berbentuk hirarki.

# Partitional Clustering

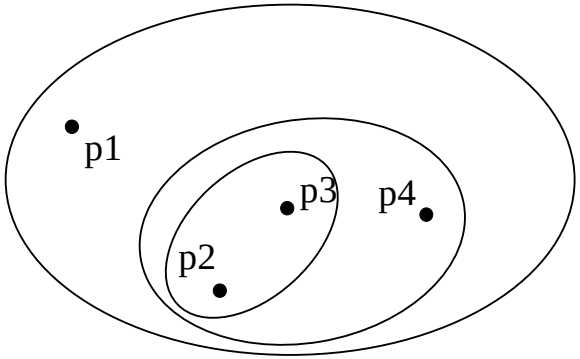


Original Points

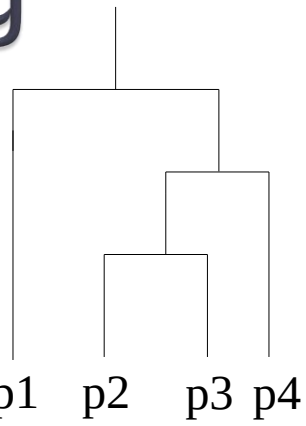


A Partitional Clustering

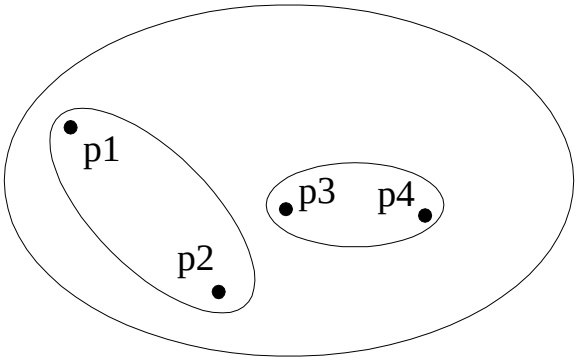
# Hierarchical Clustering



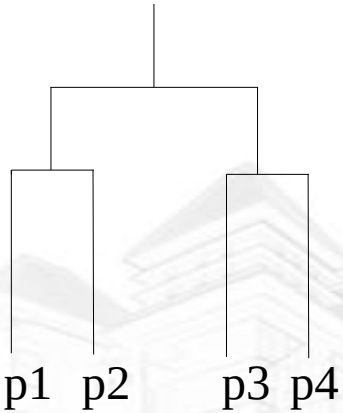

Traditional Hierarchical Clustering




Traditional Dendrogram




Non-traditional Hierarchical Clustering




Non-traditional Dendrogram

# Perbedaan Antar Sets of Clusters Lainnya

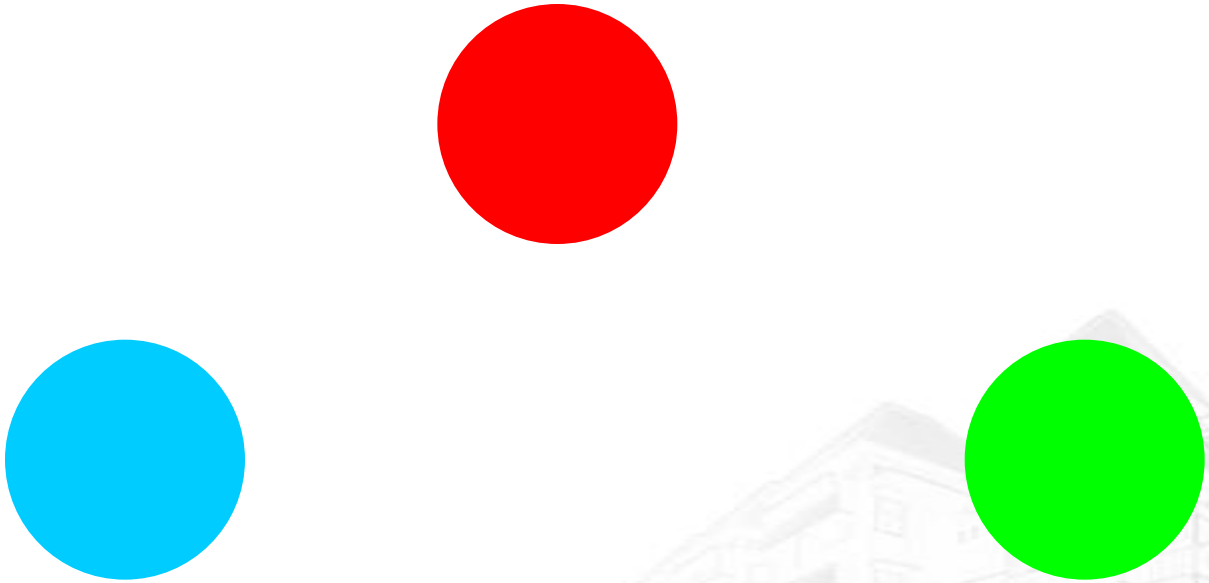
- Exclusive versus non-exclusive
  - Untuk non-exclusive clusterings, sebuah data dapat masuk ke sejumlah clusters.
  - Dapat digunakan sebagai 'border' points.
- Fuzzy versus non-fuzzy
  - Untuk fuzzy clustering, sebuah data dapat menjadi anggota setiap cluster dengan nilai bobot antara 0 dan 1
  - Bobot-bobot tersebut ketika dijumlah harus = 1.
  - Probabilistic clustering memiliki konsep yang sama.
- Partial versus complete
  - Bisa jadi dalam beberapa kasus, kita hanya ingin meng-cluster beberapa data saja.
- Heterogeneous versus homogeneous
  - Cluster yang memiliki perbedaan ukuran, bentuk, dan kepadatan data.

# Tipe-tipe Clusters

- Well-separated clusters
- Center-based clusters
- Contiguous clusters
- Density-based clusters
- Property or Conceptual
- Described by an Objective Function

# Tipe Clusters: Well-Separated

- Well-Separated Clusters:
  - Cluster dinyatakan sebagai kumpulan point sehingga satu point dalam sebuah cluster lebih dekat (atau lebih mirip) dengan setiap point di cluster yang sama dibandingkan dengan point-point yang bukan di clusternya.



3 well-separated clusters

# Tipe Clusters: Center-Based

- Center-based
  - Cluster adalah kumpulan obyek sehingga obyek dalam sebuah cluster lebih dekat (atau mirip) dengan "center" dari suatu cluster, dibandingkan dengan center cluster lainnya.
  - Pusat atau center sebuah cluster biasanya berupa titik pusat (centroid), nilai rata-rata semua point dalam sebuh cluster (medoid) dapat dijadikan sebagai point pusat cluster yang paling "representatif".

4 center-based clusters

# K-means Clustering

- Termasuk sebagai Partitional clustering
- Setiap cluster diasosiasikan dengan sebuah centroid (center point)
- Setiap point dikelompokkan ke dalam suatu cluster yang memiliki jarak ke centroid terdekat.
- Jumlah cluster, K, harus dinyatakan terlebih dahulu.
- Algoritma dasarnya sangat sederhana:

---

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

---

# K-means Clustering

- Centroid awal seringkali dipilih secara random.
  - Cluster yang dihasilkan bisa jadi berbeda antara satu "running" dengan "running" lainnya.
- Centroid (umumnya) berupa nilai rata-rata point dalam suatu cluster.
- Nilai "kedekatan" diukur menggunakan Euclidean distance, cosine similarity, correlation, etc.

# Algoritma K-Means

1. Tentukan jumlah cluster yang ingin diperoleh.
2. Tentukan pusat cluster secara random.
3. Hitung jarak setiap obyek data dengan pusat tiap cluster (gunakan Euclidean Distance).
4. Tentukan cluster tiap obyek data berdasarkan jarak terdekat dengan pusat cluster.
5. Hitung besaran rasio Between Cluster Variation (BCV) dan Within Cluster Variation (WCV).
6. Bandingkan rasio sekarang dengan yang sebelumnya.
7. Jika nilai rasio sekarang > rasio sebelumnya, perbaharui pusat cluster dengan menghitung nilai rata-rata.
8. Ulangi langkah 3 – 7.

# Contoh

- Terdapat data nasabah bank sebagai berikut:

| Nasabah | Jumlah Rumah | Jumlah Mobil |
|---|---|---|
| A | 1 | 3 |
| B | 3 | 3 |
| C | 4 | 3 |
| D | 5 | 3 |
| E | 1 | 2 |
| F | 4 | 2 |
| G | 1 | 1 |
| H | 2 | 1 |

Sumber: Larose, 2005

# Tahapan K-Means

- Jumlah cluster (k = 3)
- Pusat cluster dipilih secara random, misal:
  - C1 = B → posisi pusat data (3,3)
  - C2 = E → posisi pusat data (1,2)
  - C3 = F → posisi pusat data (4,2)

# Tahapan K-Means

- Hitung jarak tiap obyek ke pusat cluster dan tentukan clusternya!

| Nasabah | Jarak ke C1 | Jarak ke C2 | Jarak ke C3 | Cluster |
|---|---|---|---|---|
| A | 2 | 1 | 3,162 | C2 |
| B | 0 | 2,236 | 1,414 | C1 |
| C | 1 | 3,162 | 1 | C3 |
| D | 2 | 4,123 | 1,414 | C3 |
| E | 2,236 | 0 | 3 | C2 |
| F | 1,414 | 3 | 0 | C3 |
| G | 2,828 | 1 | 3,162 | C2 |
| H | 2,236 | 1,414 | 2,236 | C2 |

# Tahapan K-Means

- Diperoleh keanggotaan cluster sbb:
  - C1 = (B)
  - C2 = (A, E, G, H)
  - C3 = (C, D, F)
- Hitung BCV dan WCV
  - BCV = d(m1, m2) + d(m1, m3) + d(m2, m3)
  - Dimana ($m_i$,$m_j$) adalah jarak Euclides dari pusat cluster i ke pusat cluster j.
  - Diperoleh BCV = 6,650
  - WCV = $1^2 + 0^2 + 1^2 + 1{,}414^2 + 0^2 + 0^2 + 1^2 + 1{,}414^2 = 7$
- Rasio = BCV/ WCV = 0,950

# Tahapan K-Means

- Update pusat cluster yang baru.
  - m1 = rata-rata ($m_B$) = (3,3)
  - m2 = rata-rata($m_A$, $m_E$, $m_G$, $m_H$) = (1,25; 1,75)
  - m3 = rata-rata ($m_C$, $m_D$, $m_F$) = (4,333; 2,667)

# Tahapan K-Means

- Hitung jarak tiap obyek ke pusat cluster dan tentukan clusternya!

| Nasabah | Jarak ke C1 | Jarak ke C2 | Jarak ke C3 | Cluster |
|---|---|---|---|---|
| A | 2 | 1,275 | 3,350 | C2 |
| B | 0 | 1,768 | 1,374 | C1 |
| C | 1 | 3,021 | 0,471 | C3 |
| D | 2 | 3,953 | 0,745 | C3 |
| E | 2,236 | 0,354 | 3,399 | C2 |
| F | 1,414 | 2,813 | 0,745 | C3 |
| G | 2,828 | 0,791 | 3,727 | C2 |
| H | 2,236 | 1,061 | 2,867 | C2 |

# Tahapan K-Means

- Diperoleh keanggotaan cluster sbb:
  - C1 = {B}
  - C2 = {A, E, G, H}
  - C3 = {C, D, F}
- Hitung rasio BCV dan WCV:
  - BCV = 6,741
  - WCV = 4,833
  - Rasio (BCV/WCV) = 1,394
- Bandingkan rasio sekarang dengan sebelumnya:
  - 1,394 > 0,950

# Tahapan K-Means

- Update pusat cluster baru:
  - m1 = (3; 3)
  - m2 = (1,25; 1,75)
  - m3 = (4,333; 2,667)
- Diperoleh keanggotaan cluster sbb:
  - C1 = {B}
  - C2 = {A, E, G, H}
  - C3 = {C, D, F}
- Hitung rasio (BCV/WCV) dan diperoleh rasio = 1,394
- Bandingkan rasio sekarang dengan sebelumnya, diperoleh 1,394 = 1,394.

# Tahapan K-Means

- Diperoleh rasio tidak membesar, sehingga algoritma dapat dihentikan.
- Anggota cluster adalah:
  - C1 = {B}
  - C2 = {A, E, G, H}
  - C3 = {C, D, F}
- Analisis hasil clustering tersebut!

# Analisis Hasil K-Means Clustering

- Dari hasil clustering tersebut, dapat diketahui bahwa:
  - Kelompok nasabah I: nasabah dengan jumlah rumah sedang (3 buah) dan jumlah mobil banyak (3 buah).
  - Kelompok nasabah II: nasabah dengan jumlah rumah sedikit (sekitar 1-2 buah) dan jumlah mobil juga sedikit (sekitar 1-2 buah).
  - Kelompok nasabah III: nasabah dengan jumlah rumah banyak (sekitar 4-5 buah) dan jumlah mobil yang banyak juga (sekitar 2-3 buah).