IF5181 Pengenalan Pola

Clustering

Masayu Leylia Khodra

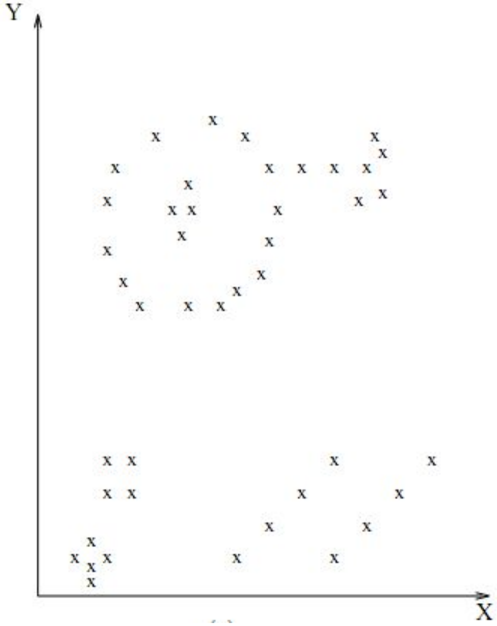
Referensi

- Bab 10 & 11 dari Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.
- Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*(2), 165-193.
- Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., ... & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing*, *2*(3), 267-279.
- Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., ... & Lin, C. T. (2017). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664-681.
- A.K. Jain, M.N. Murty, P.J. FLYNN (1999), Data Clustering: A Review. ACM computing surveys
- Pengyu Hong (2005), Introduction to Hierarchical Clustering Analysis
- Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data* (pp. 25-71). Springer, Berlin, Heidelberg.
- Rui Xu, Donald Wunsch (2005), Survey of Clustering Algorithm
- DBSCAN http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=72F8D8F33A502FAB448D4C13809D83C3?doi=10.1.1.71.1980&rep=rep1&type=pdf
- http://www.cse.buffalo.edu/~jing/cse601/fa12/materials/clustering_density.pdf

Outline

- Clustering: what, why

Clustering: What ?



Input Data

Proses pengelompokan data menjadi clusters berbasis kesamaan data

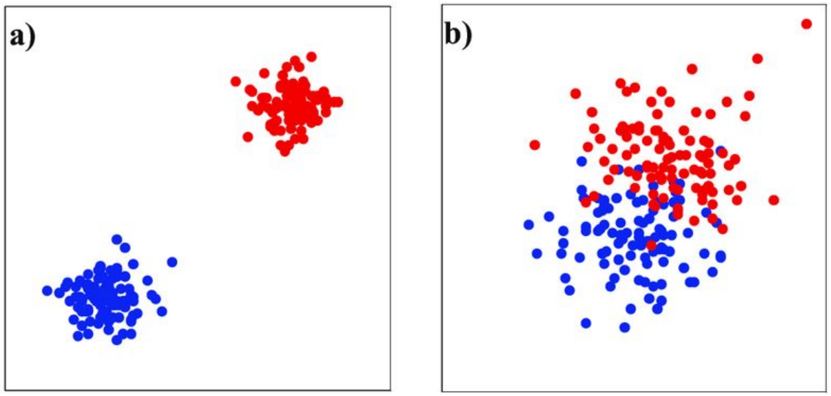
Desired cluster

Unsupervised learning = learning from raw data

Sumber: Jain dkk (1999)

Clustering: Finding natural groups

- High intra-cluster similarity/ low intra-cluster variance
  - Data pd cluster yang sama harus semirip mungkin
- Low inter-cluster similarity / high inter-cluster variance
  - Data pd cluster yang berbeda harus sejauh mungkin
- Pengukuran kemiripan dan jarak harus jelas dan punya semantik praktikal (sesuai domain)



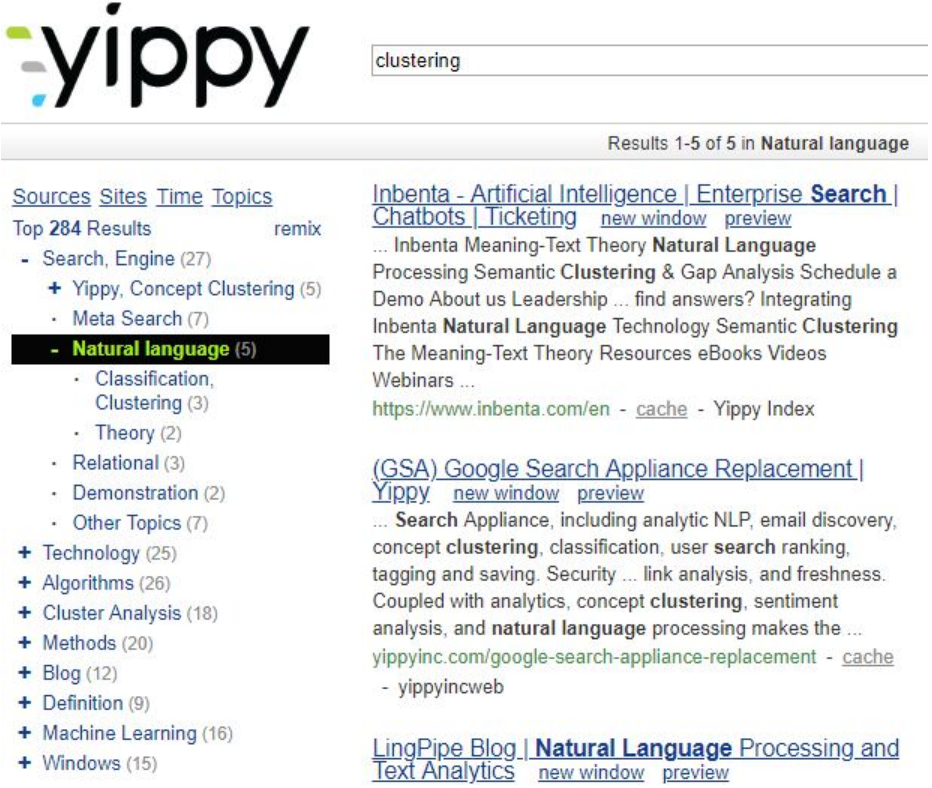
a) Low intra-class variance and high inter-class variance: compact well separated clusters. b) High intra-class variance and low inter-class variance: wide clusters without a clear frontier.

https://www.researchgate.net/figure/Inter-class-and-Intra-class-variances-concept-a-Low-intra-class-variance-and-high_fig2_278382762

Clustering: Why ?

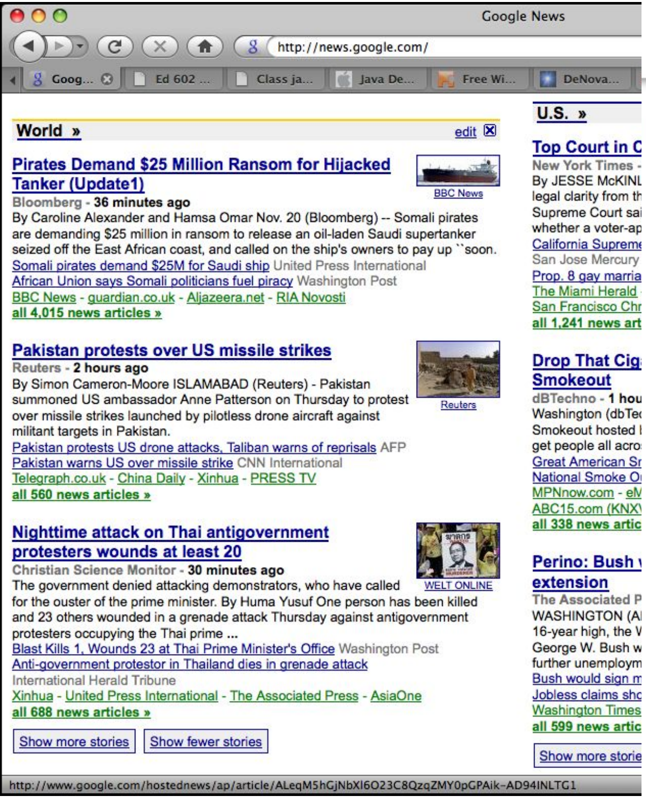
- Data discovery (cluster = struktur internal data)
  - Contoh: search engine, news aggregator, gen
- Tujuan awalnya partisi / pengelompokan
  - Contoh: segmentasi pasar
- Bagian dari teknik lainnya
  - Contoh: peringkasan berbasis clustering

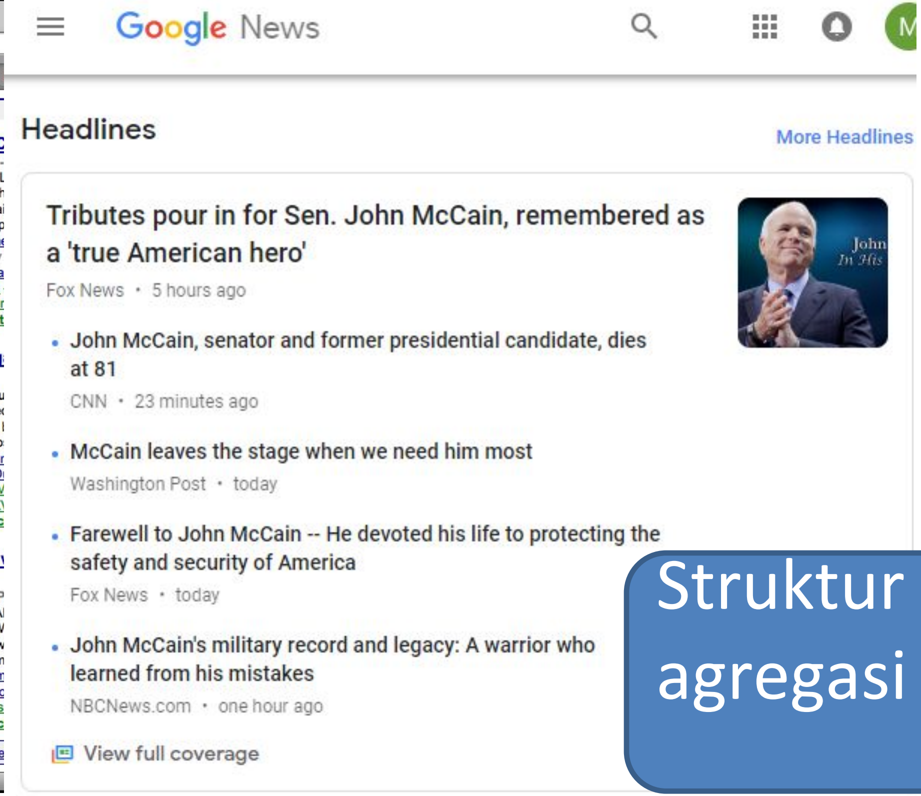
Why: Clustering pada Search Engine



Struktur internal hasil pencarian

Why: Clustering pada News Aggregator

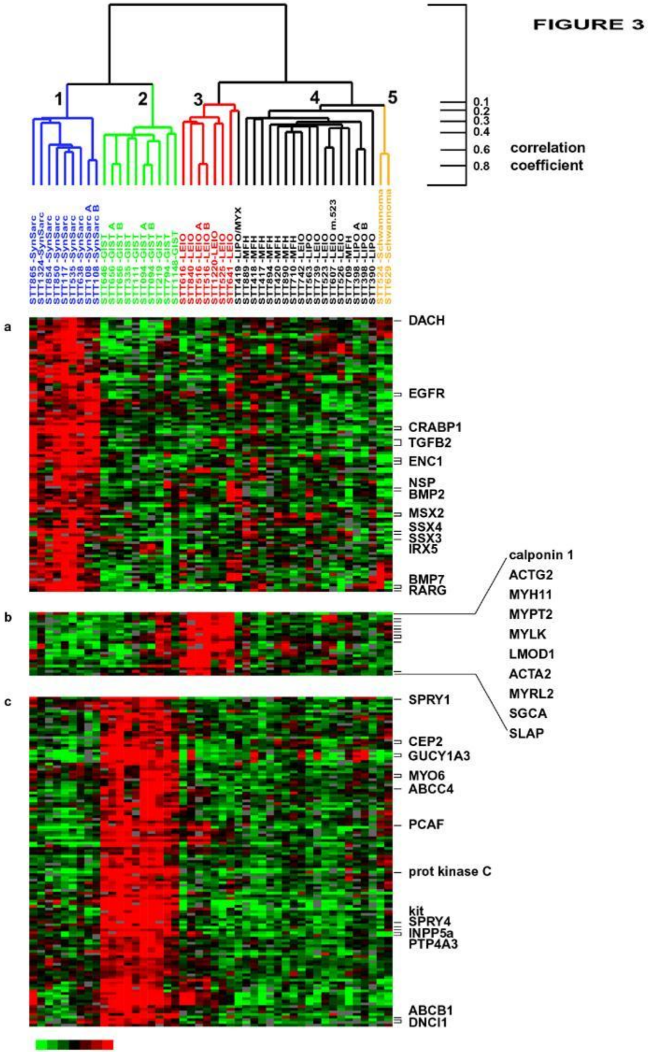




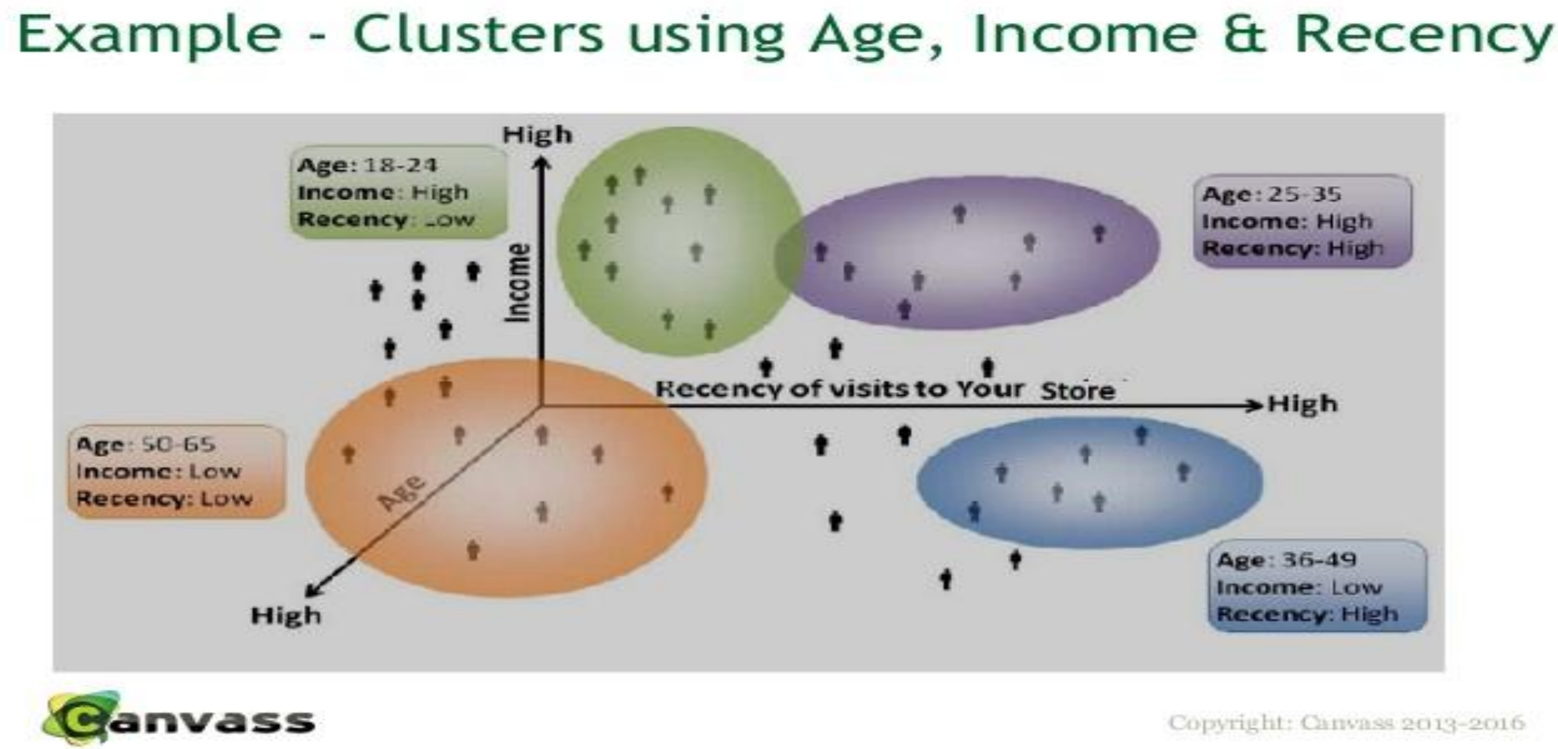
Struktur internal agregasi berita

Why: Clustering pada Gen

http://genome-www.stanford.edu/sarcoma/supplemental_data.html



Why: Clustering untuk Segmentasi



https://www.slideshare.net/Canvass/marketing-data-analytics

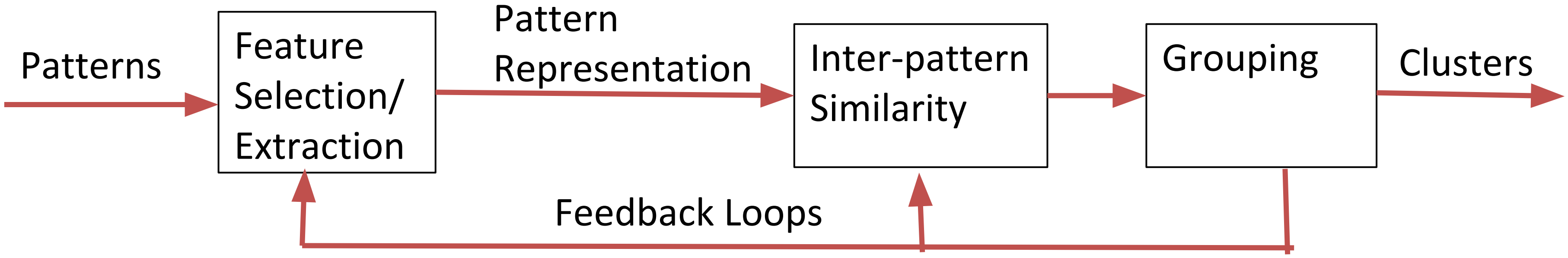
Why: Clustering-based Approach

Clustering-based summarization

Clustering-based outlier detection

Clustering-based analysis

Tahapan Clustering



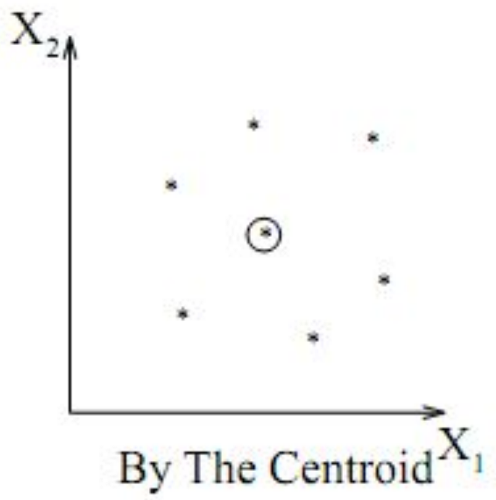
- Tahapan utama:
  1) Feature selection: original features → subset of features
     Feature extraction: transformation into new features
  2) pattern proximity/similarity measure
  3) Grouping
- Clustering output: hard atau soft (membership degree)

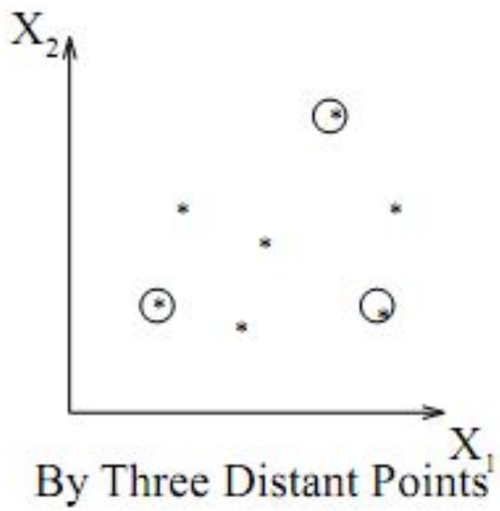
Tahapan Clustering (lanjutan)

- Tahapan opsional:
  4) data abstraction
  5) assessment of output (good or poor)

Representasi Cluster (1)

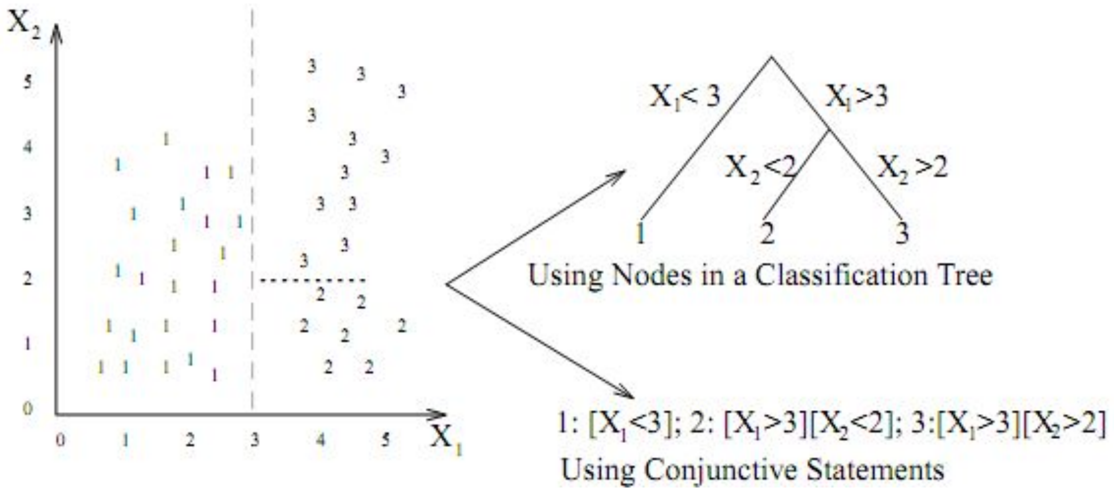
- Centroid atau set of distant point





Sumber: Jain dkk (1999)

Representasi Cluster (2)

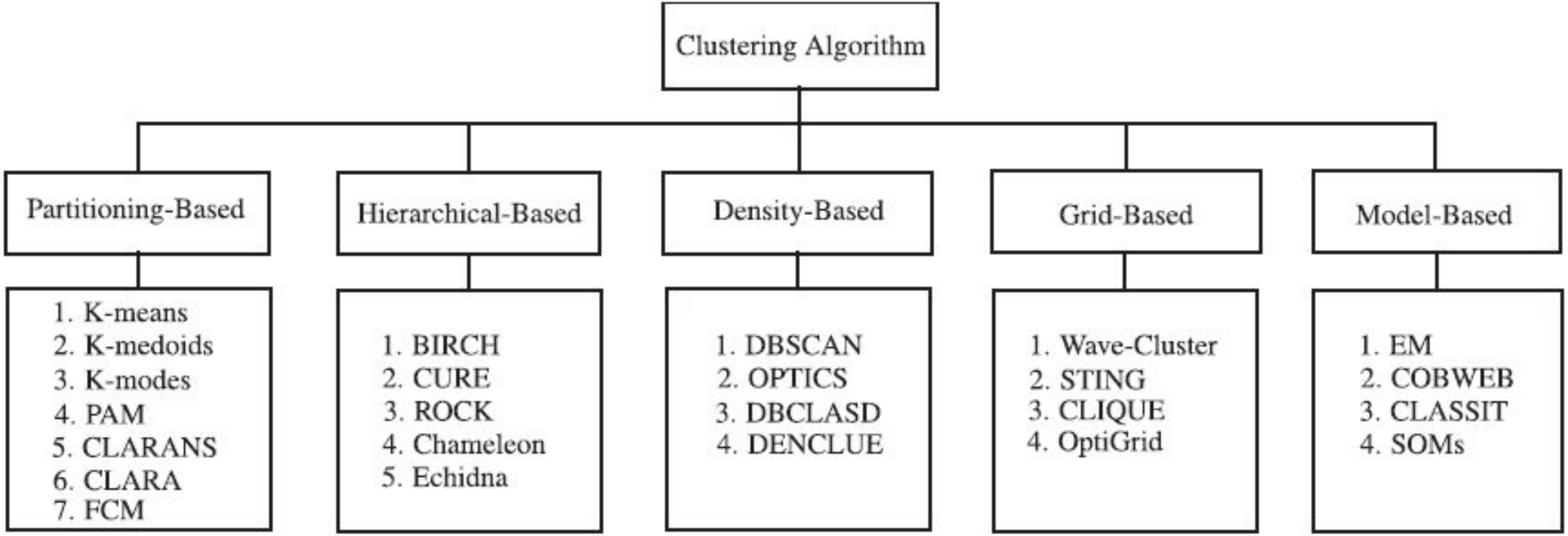


- Pohon klasifikasi
- Conjunctive statements

Sumber: Jain dkk (1999)

Kategori Metode Clustering

(Fahad, 2014; Han & Kamber, 2006)



Kategori Metode Clustering (Han & Kamber, 2006)

1. Metode *partitioning*
   - mengidentifikasi partisi yang mengoptimalkan kriteria pengelompokan (squared error, absolute error)
   - Konstruksi k-partisi data (partisi ~ cluster); k ≤ jumlah data
   - Contoh: K-means, k-medoids
2. Metode *hierarchical*
   - menghasilkan rangkaian partisi bersarang
   - Agglomerative (bottom-up, merge):
     1 object ~ 1 cluster → 1 cluster n-object
   - Divisive (top-down, split):
     1 cluster n-object → 1 object ~ 1 cluster

Kategori Metode Clustering (lanj) (Han & Kamber, 2006)

3. Metode berbasis density
    - Densitas: jumlah objek
    - Contoh: DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
4. Metode berbasis grid
    - Struktur grid, cepat, bergantung jumlah sel, tidak dipengaruhi jumlah objek, perhitungan bisa dilakukan secara paralel
    - Contoh: STING (STatistical INformation Grid)
5. Metode berbasis model
    - Contoh: EM (Expectation-Maximization), SOM (self-organizing map)

Evaluasi Cluster: Purity

Semakin besar nilai purity, semakin baik solusi clustering yang dihasilkan.

- Purity untuk setiap cluster C_r dengan ukuran n_r

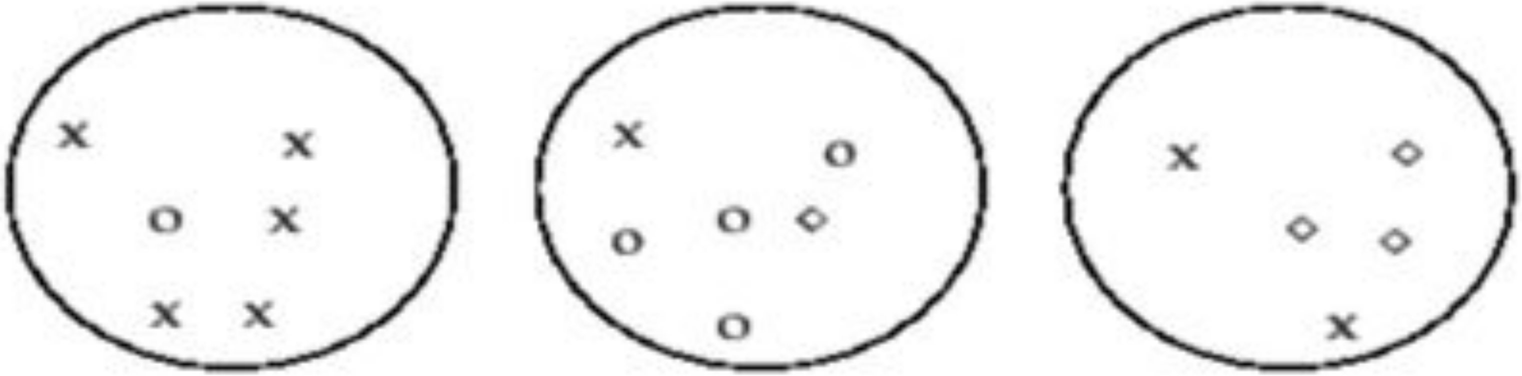
$$P(C_r) = \frac{1}{n_r} \max_i \; n_r^i$$

- Purity dari keseluruhan clustering

$$Purity\,(C) = \sum_{r=1}^{k} \frac{n_r}{n} P(C_r)$$

$$\text{or } = \frac{1}{n} \sum_{r=1}^{k} \max_i \,(n_r^i)$$

Purity: Contoh



- Purity= 1/17 * (5 + 4 + 3) ≈ 0.71

Tugas 2

- Gunakanlah dataset yang sama dengan Tugas1
- Tools: Jupyter Notebook dengan python. Gunakan library untuk konstruksi model clustering.
- Skenario:
  - Gunakanlah hasil split dataset (train:test) dari tugas1
  - Lakukanlah clustering training data , hitunglah purity dengan menggunakan. Pilihlah satu algoritma dari setiap kategori algoritma clustering.
  - Gunakanlah test data untuk mendapatkan akurasi dari setiap algoritma clustering.
- Lakukanlah analisis hasil testing.
- Tugas dikumpulkan Rabu 25 September 2019 jam 23.55