Clustering: Unsupervised Learning

IF-3270 Pembelajaran Mesin

Teknik Informatika ITB

Modul 7: Clustering

Fariska Z. Ruskanda, S.T., M.T.
(fariska@informatika.org)

KK IF -Teknik Informatika - STEI ITB

01 Clustering: What & Why?

IF3270 - Pembelajaran Mesin
(Machine Learning)

Outline

Clustering: What?

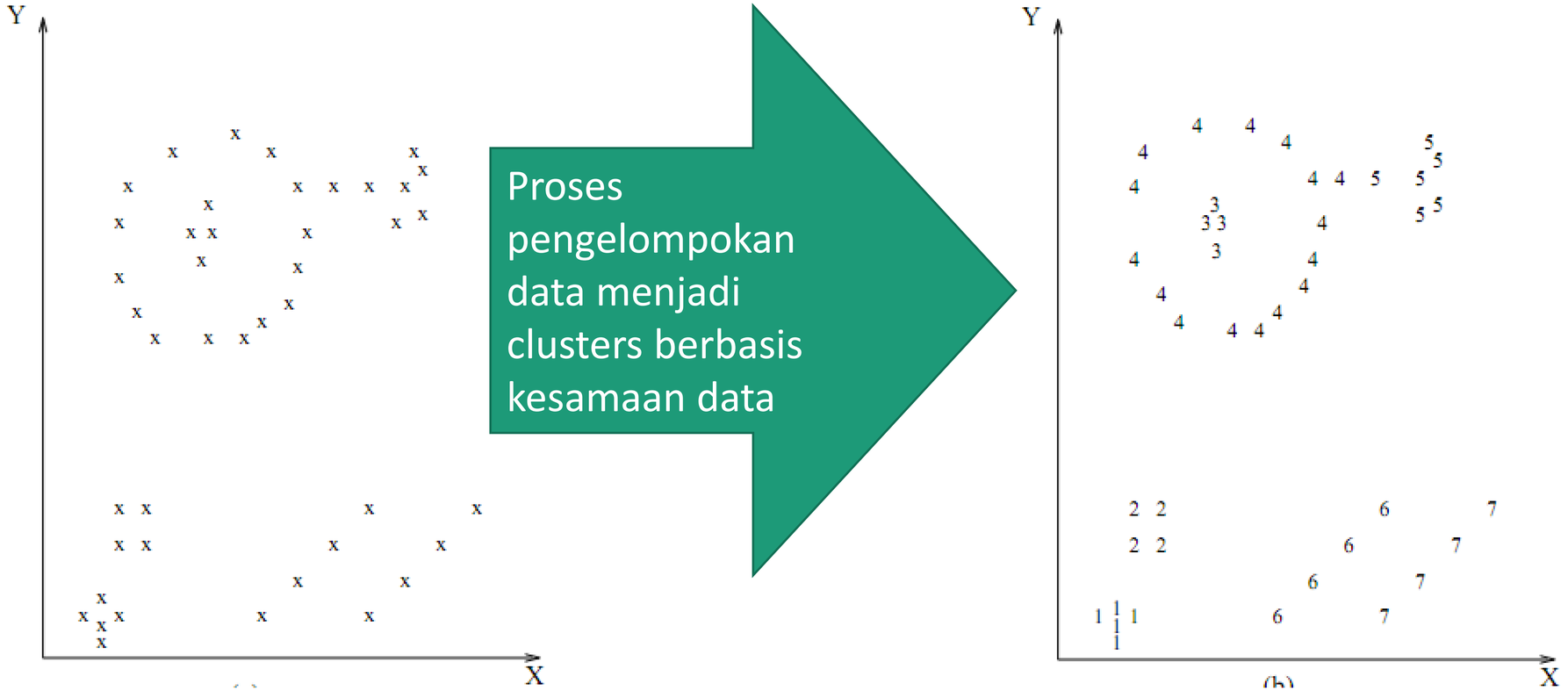
Tahapan Clustering

Representasi Cluster

Clustering: Why?

Kategori Metode

Clustering: What ?



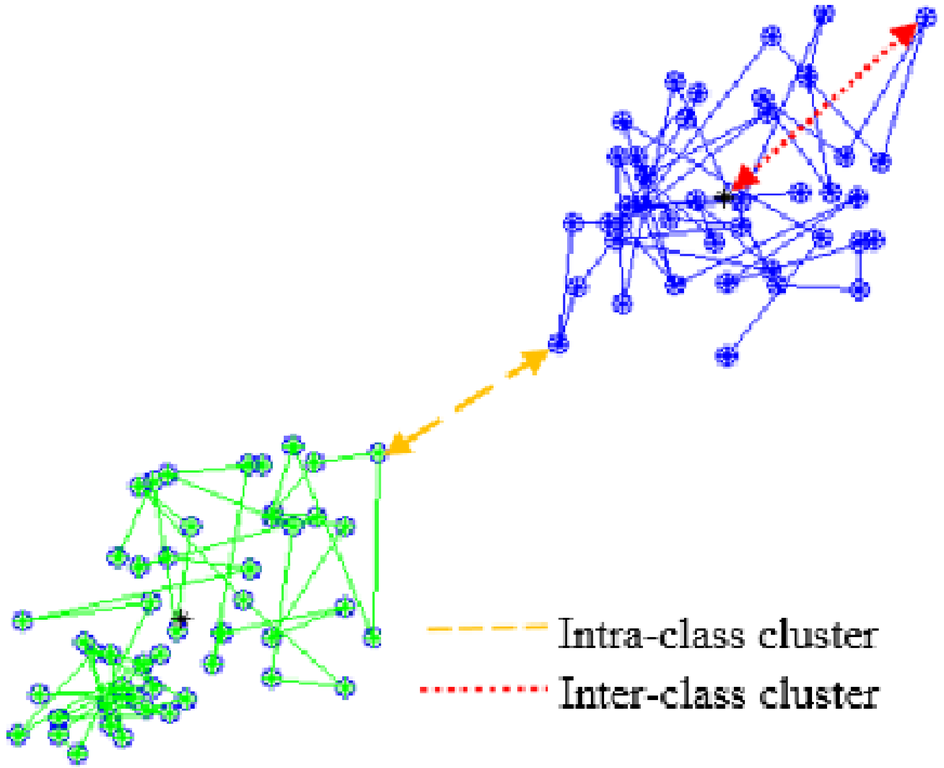
Input Data

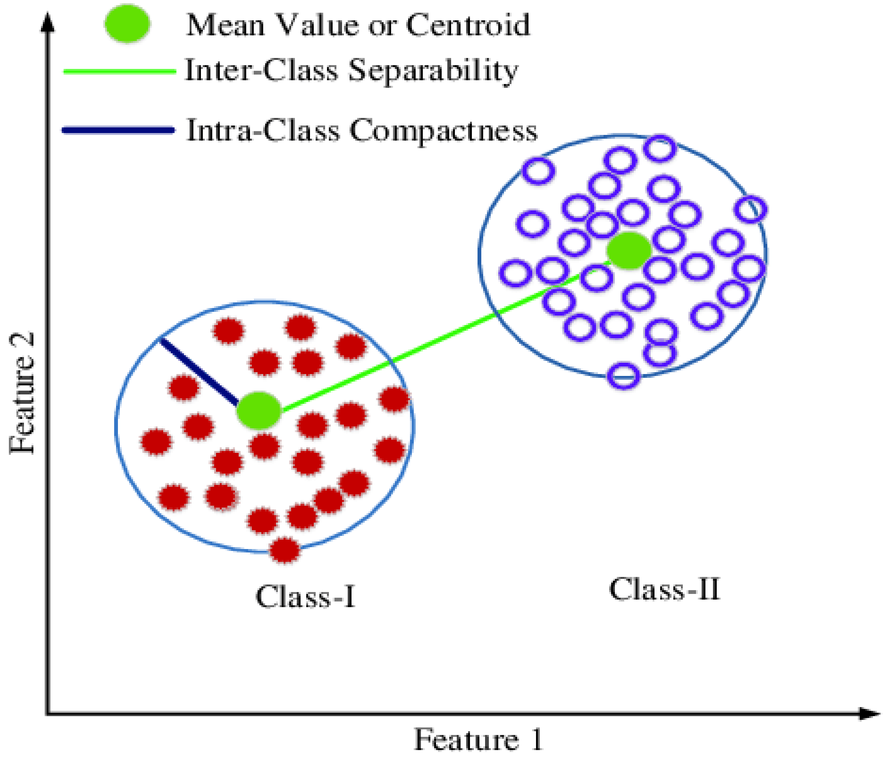
Desired cluster

Unsupervised learning = learning from raw data

Sumber: Jain dkk (1999)

Clustering:

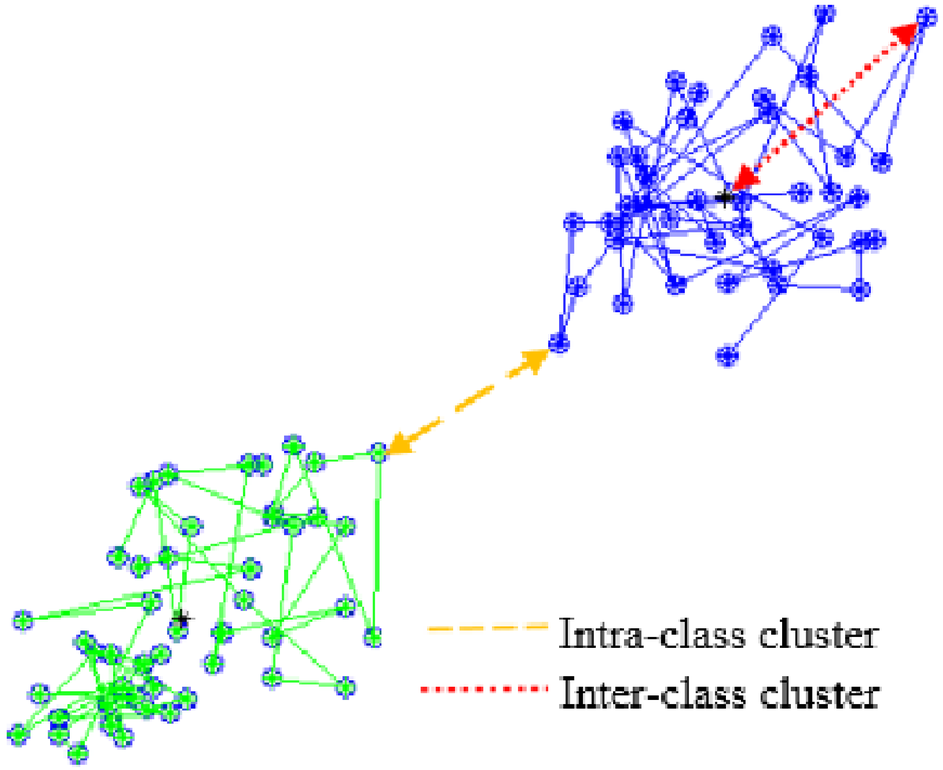




https://www.researchgate.net/figure/The-concept-of-intra-class-compactness-and-inter-class-separability-in-a-two-dimension_fig3_325095062

Clustering: Finding natural groups

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

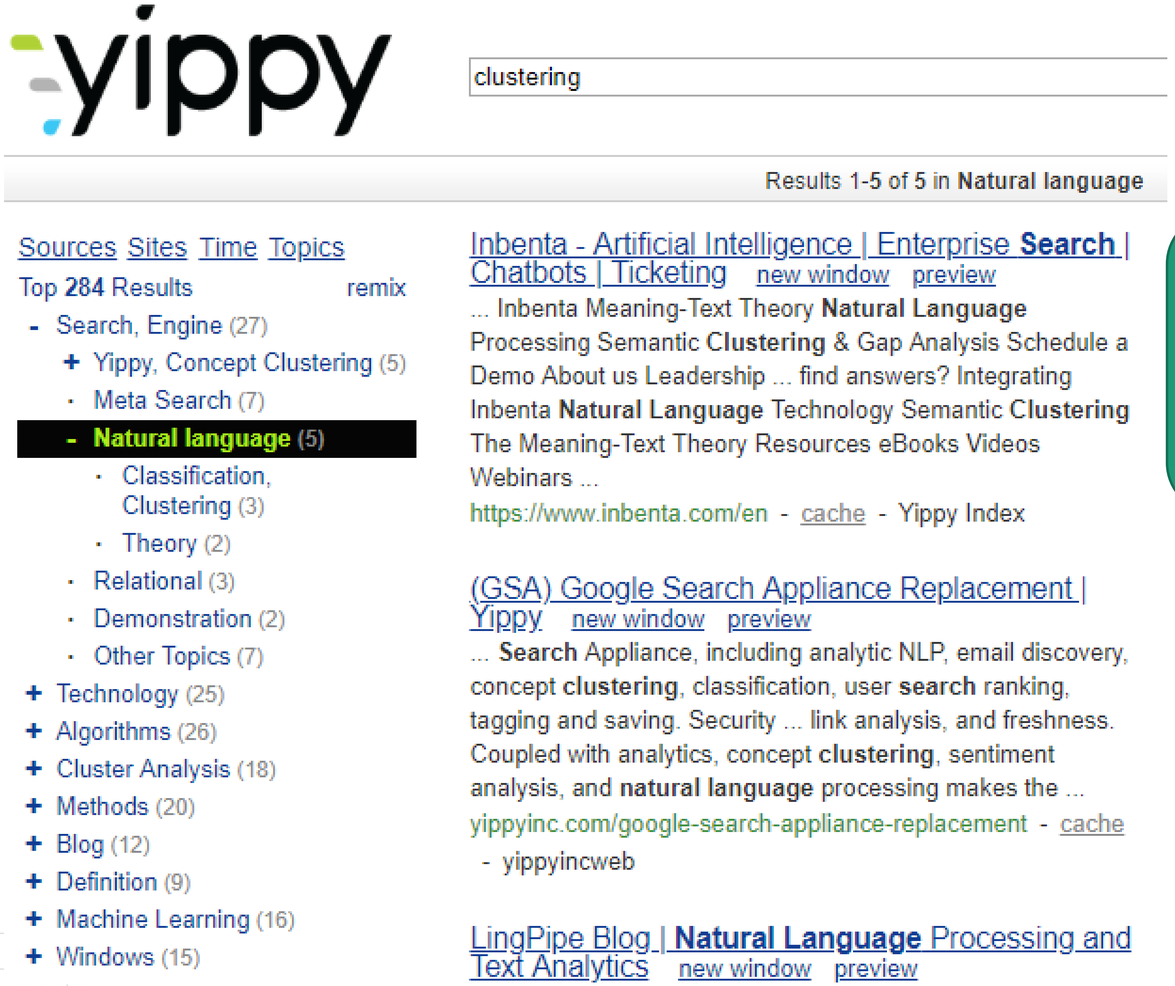


https://www.researchgate.net/profile/Sharifah_Sakinah_Syed_Ahmad/publication/280627665/

Clustering: Why ?

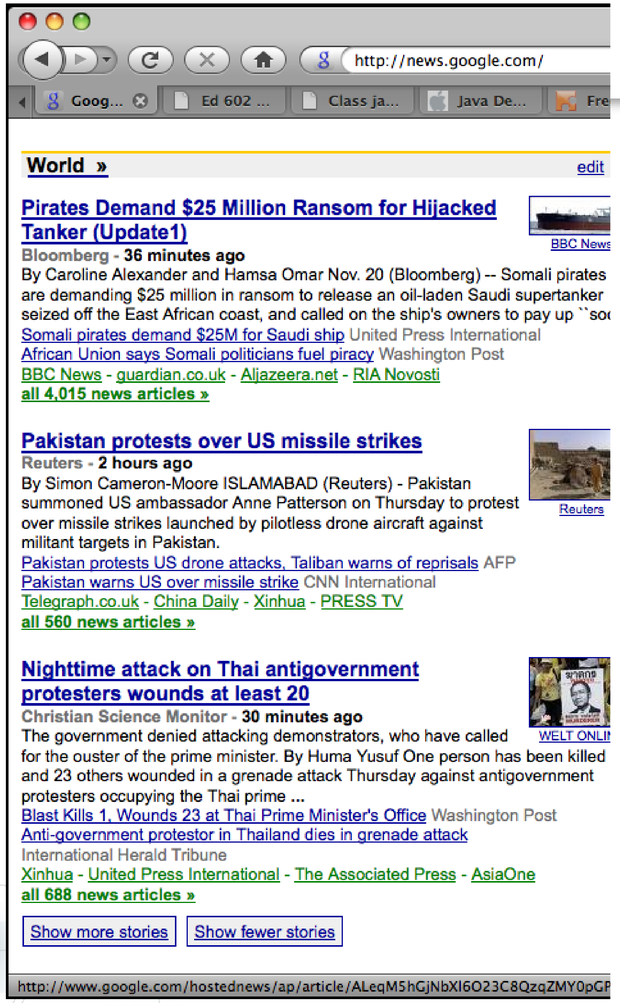
- Data discovery (cluster = struktur internal data)
  - Contoh: search engine, news aggregator, gen
- Tujuan awalnya partisi / pengelompokan
  - Contoh: segmentasi pasar, segmentasi gambar
- Bagian dari teknik lainnya
  - Contoh: peringkasan berbasis clustering

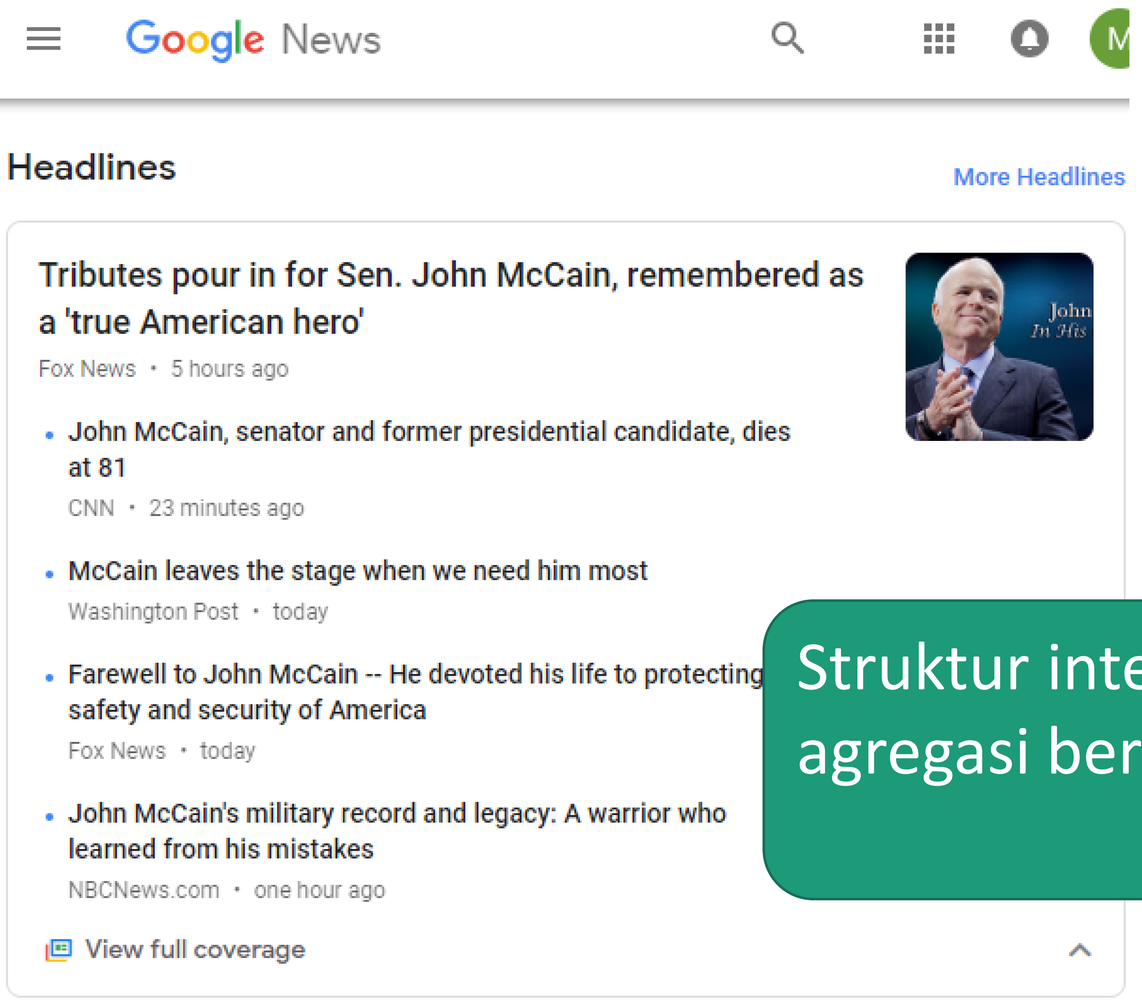
Why: Clustering pada Search Engine



Struktur internal hasil pencarian

Why: Clustering pada News Aggregator

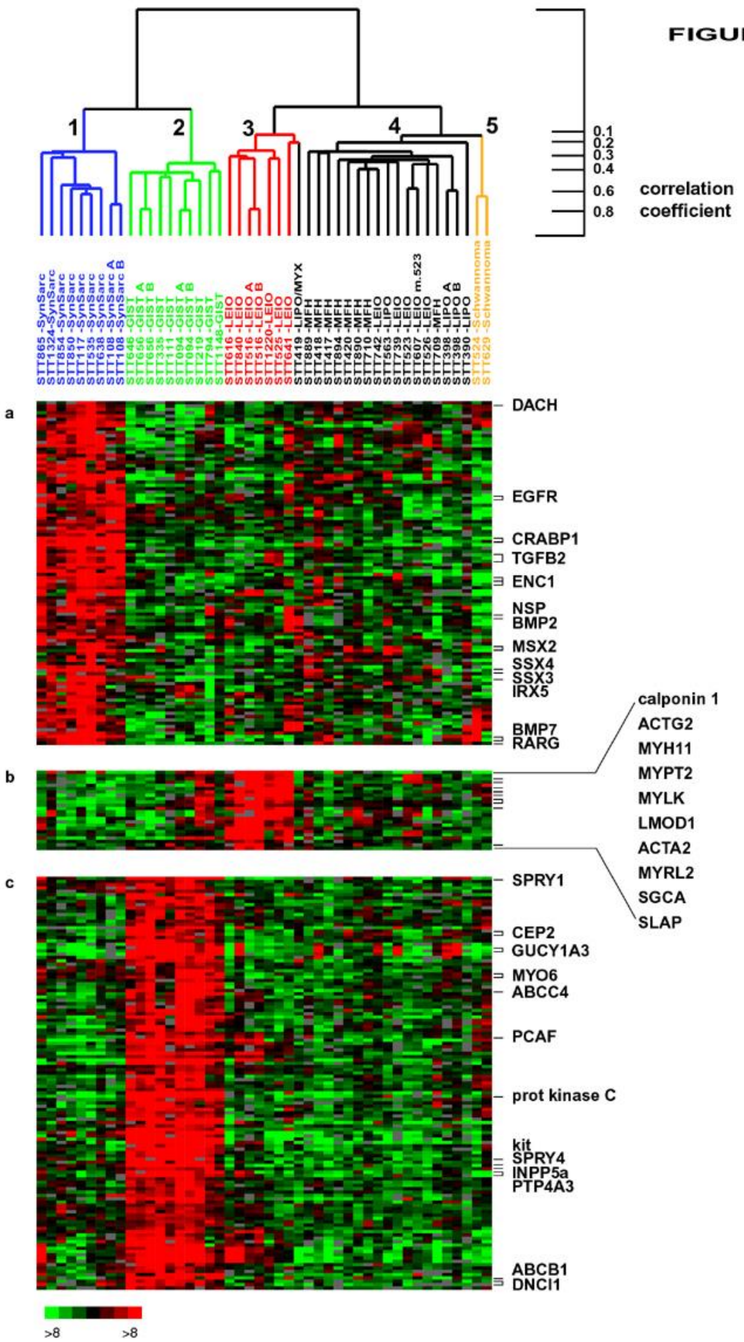




Struktur internal agregasi berita

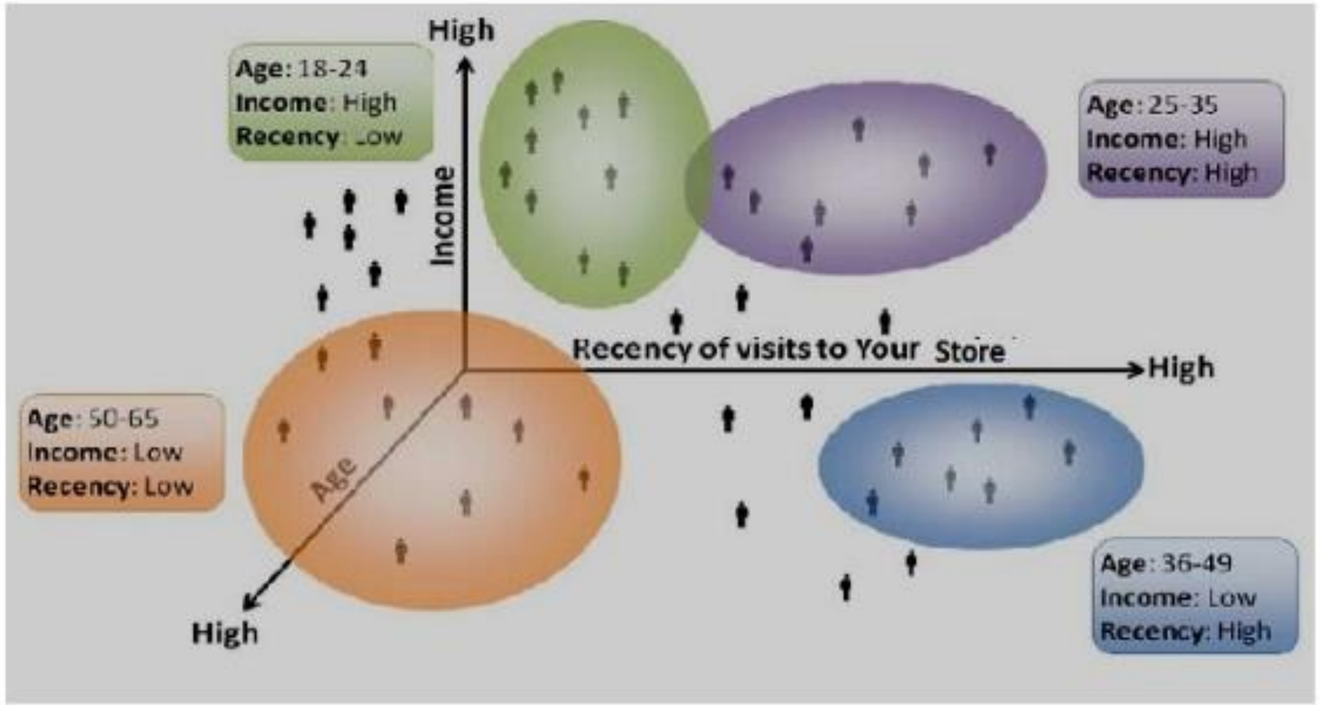
Why: Clustering pada Gen

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



Why: Clustering untuk Segmentasi

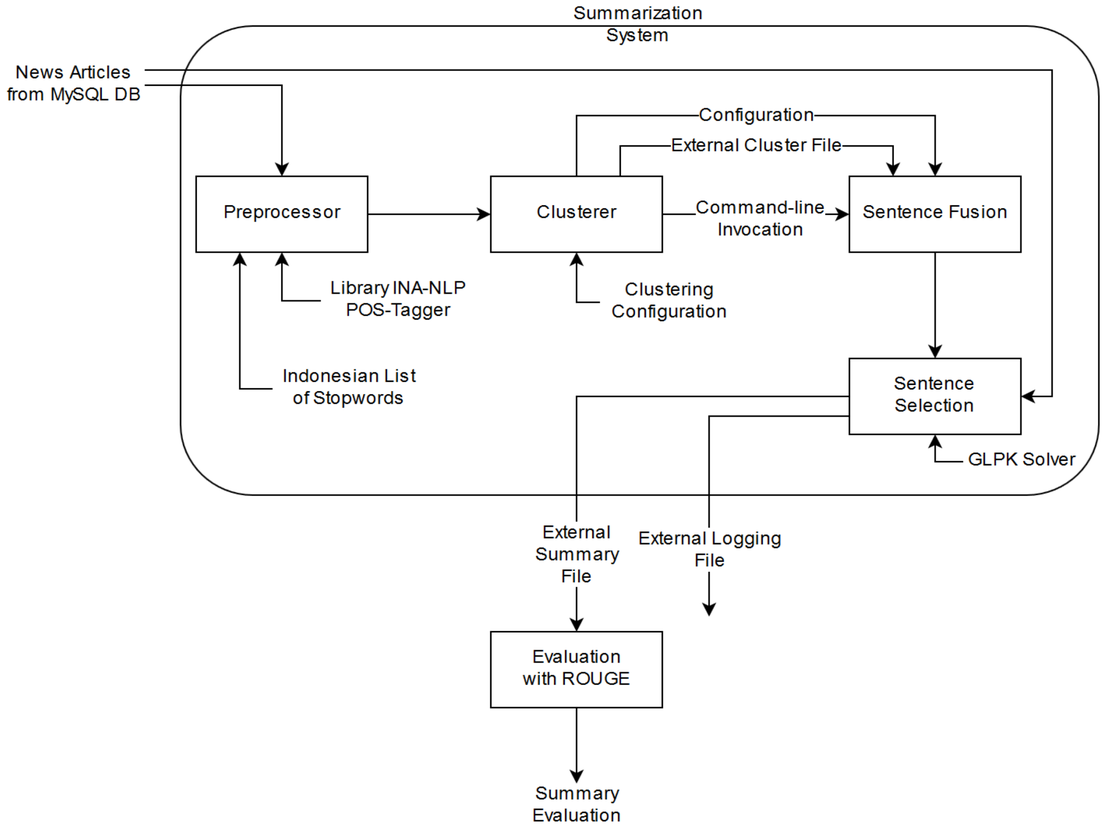
Example - Clusters using Age, Income & Recency



Canvass

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

Why: Clustering-based Approach

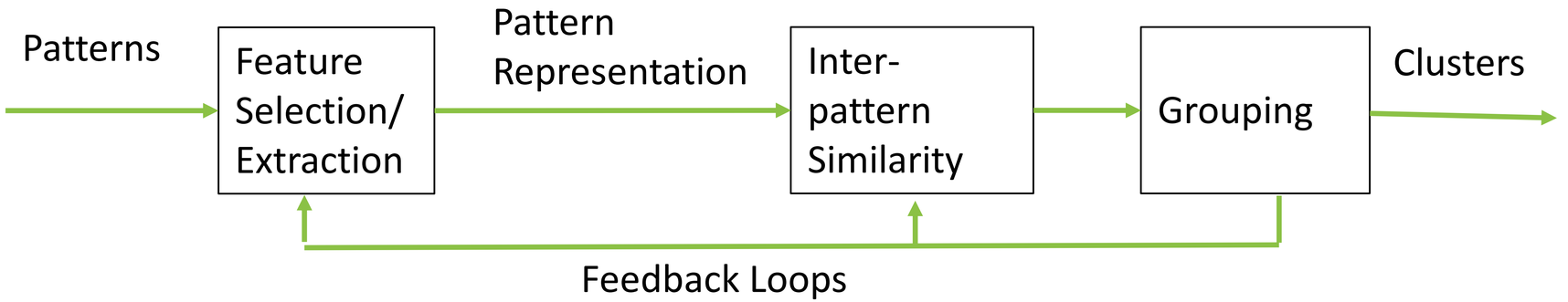


Clustering-based summarization

Clustering-based outlier detection

Clustering-based analysis

Tahapan Clustering



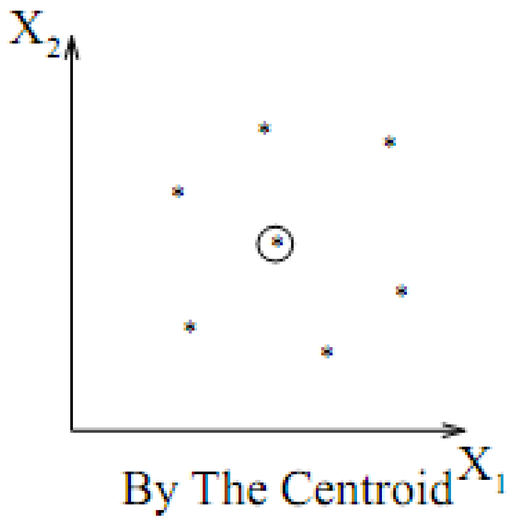
- Tahapan utama:
  1) Feature selection: original features → subset of features
     Feature extraction: transformation into new features
  2) Pattern proximity/similarity measure
  3) Grouping
- Clustering output: hard atau soft (membership degree)

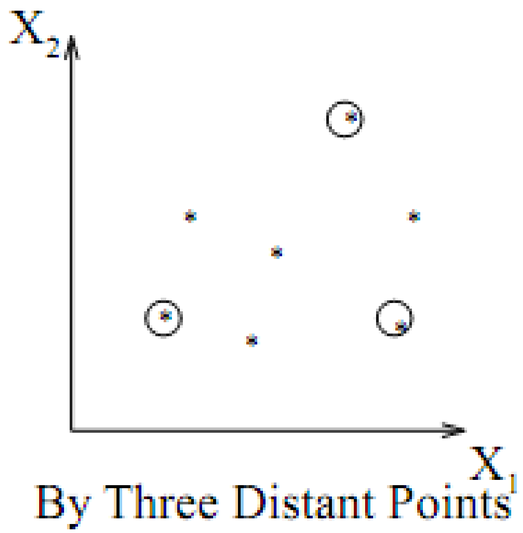
Tahapan Clustering (lanjutan)

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

Representasi Cluster (1)

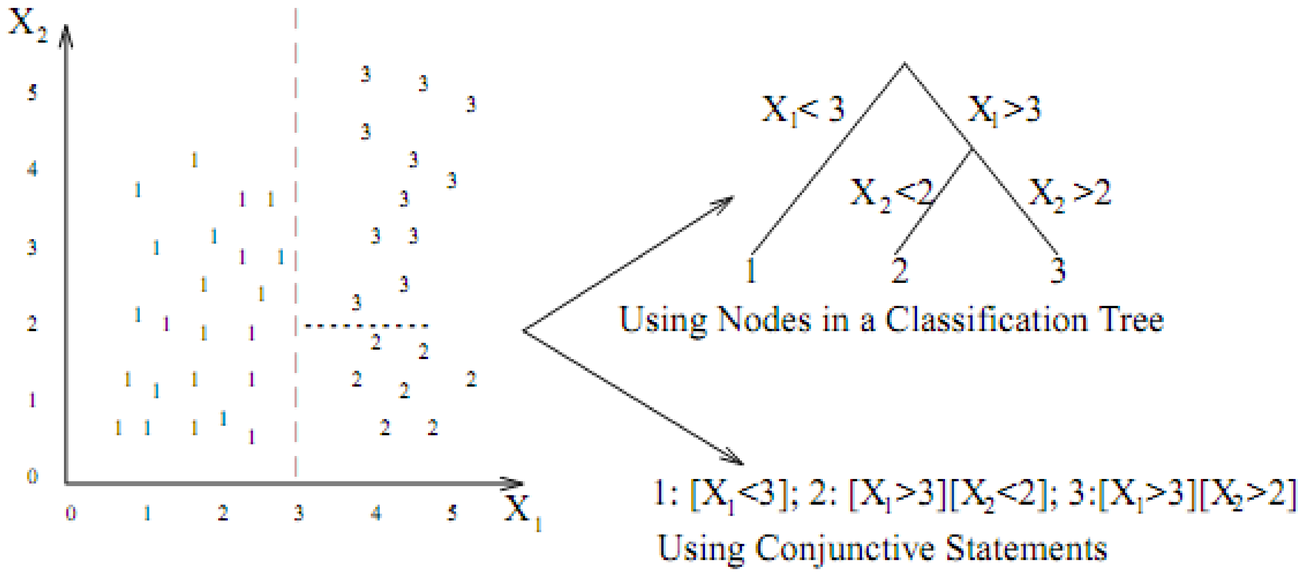
- Centroid atau set of distant point





Sumber: Jain dkk (1999)

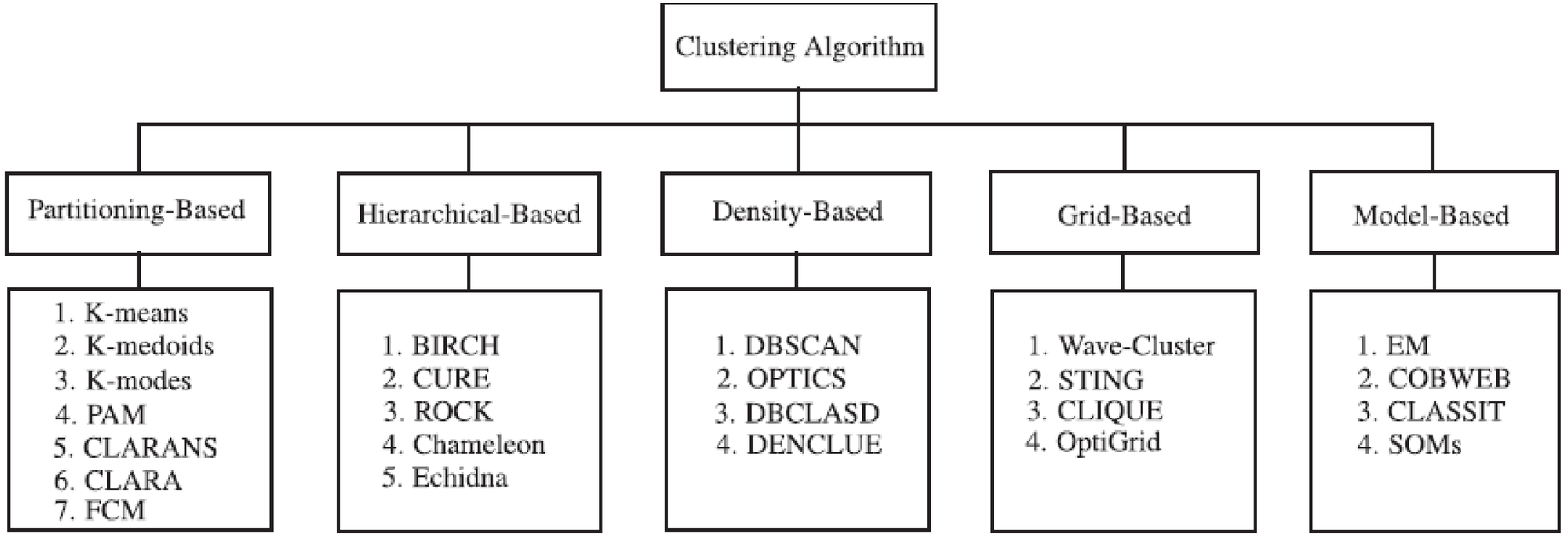
Representasi Cluster (2)



- Pohon klasifikasi
- Conjunctive statements

Sumber: Jain dkk (1999)

Kategori Metode Clustering (Fahad, 2014)



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)

02 Partitional Clustering

IF3270 Pembelajaran Mesin