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

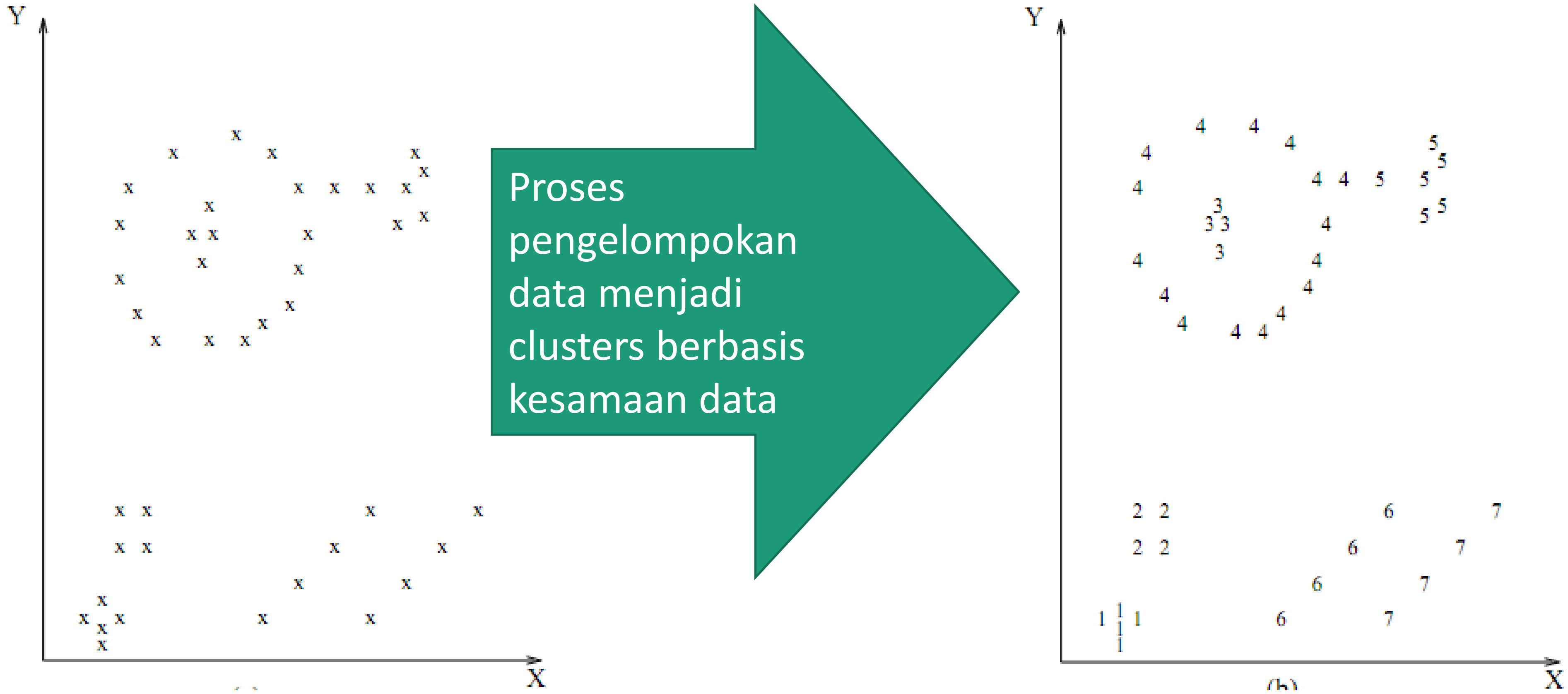
01 Clustering: What & Why?

IF3270 - Pembelajaran Mesin
(Machine Learning)

Outline

- Clustering: What?
- Tahapan Clustering
- Representasi Cluster
- Clustering: Why?
- Kategori Metode

Clustering: What ?



Proses pengelompokan data menjadi clusters berbasis kesamaan data

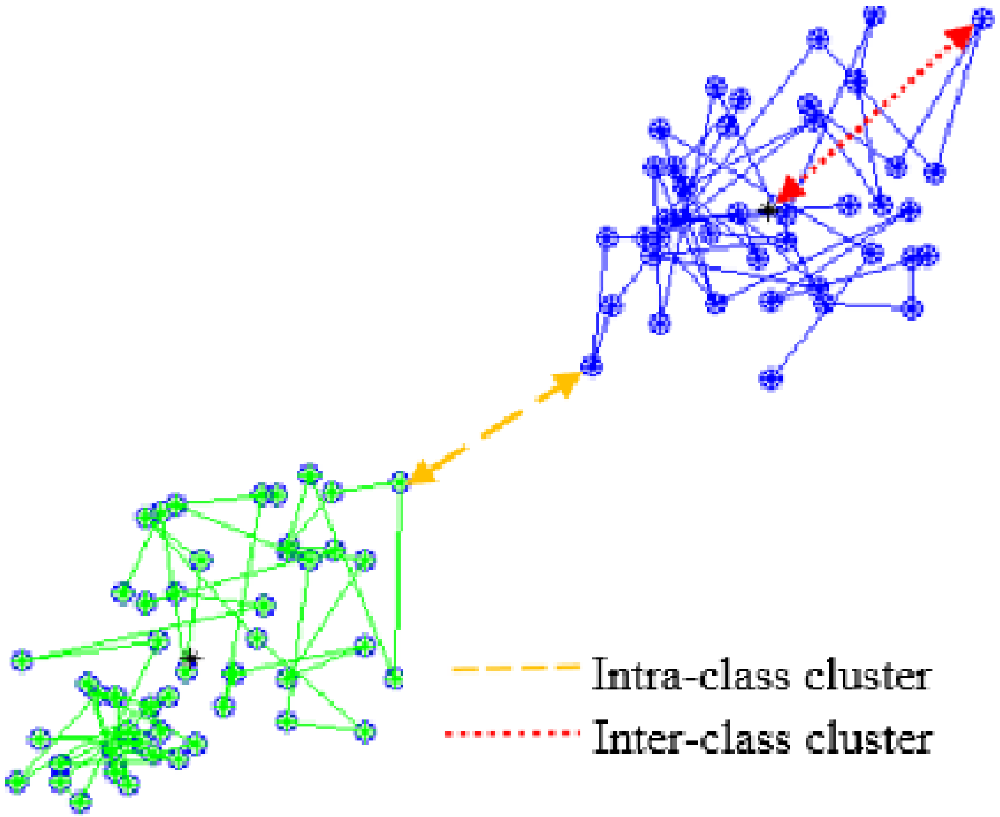
Input Data

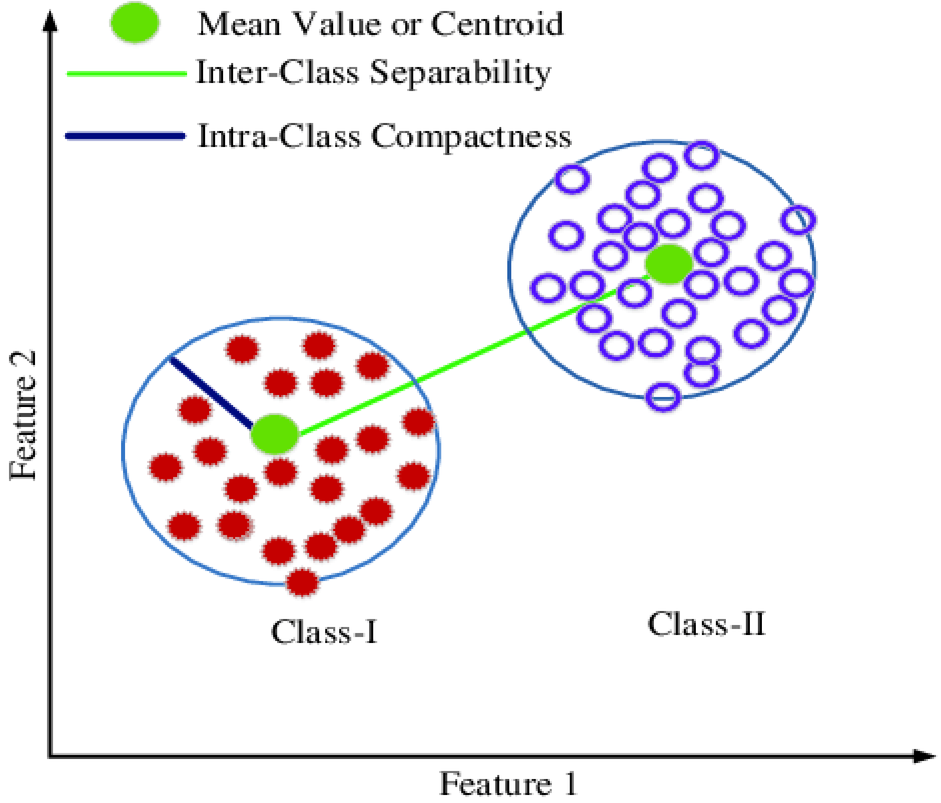
Desired cluster

Unsupervised learning = learning from raw data

Sumber: Jain dkk (1999)

Clustering:

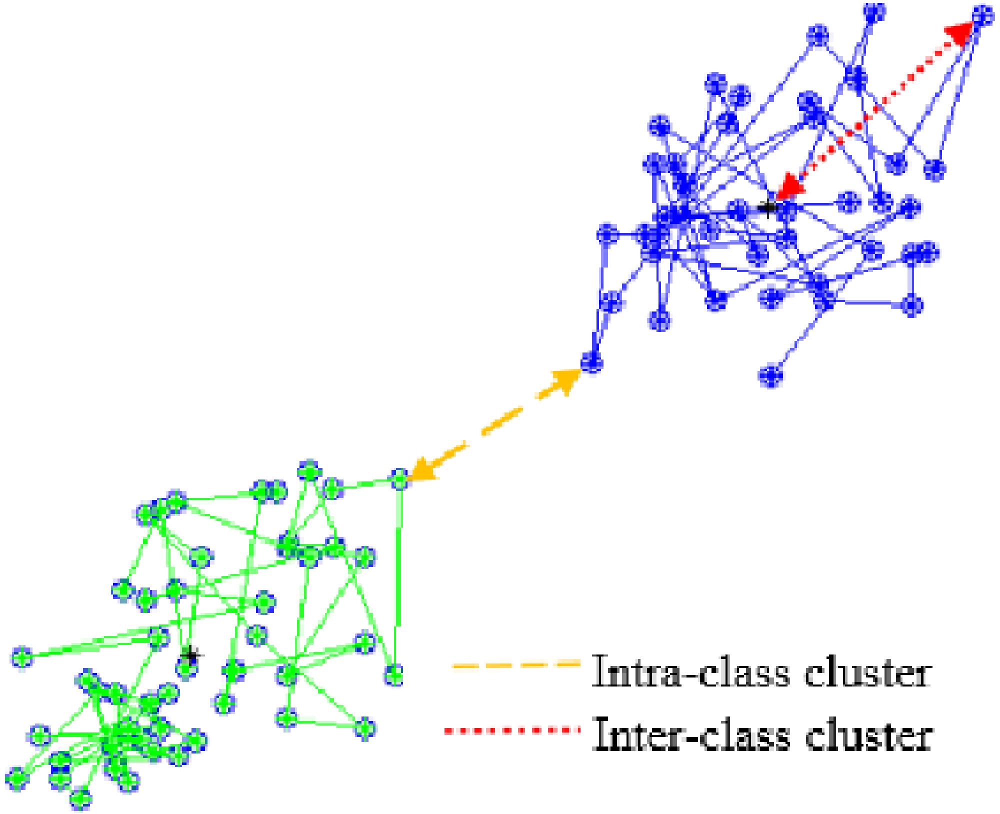




https://www.researchgate.net/figure/The-concept-of-intra-class-compactness-and-inter-class-separability-in-a-two-dimension_fig3_325095062

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)

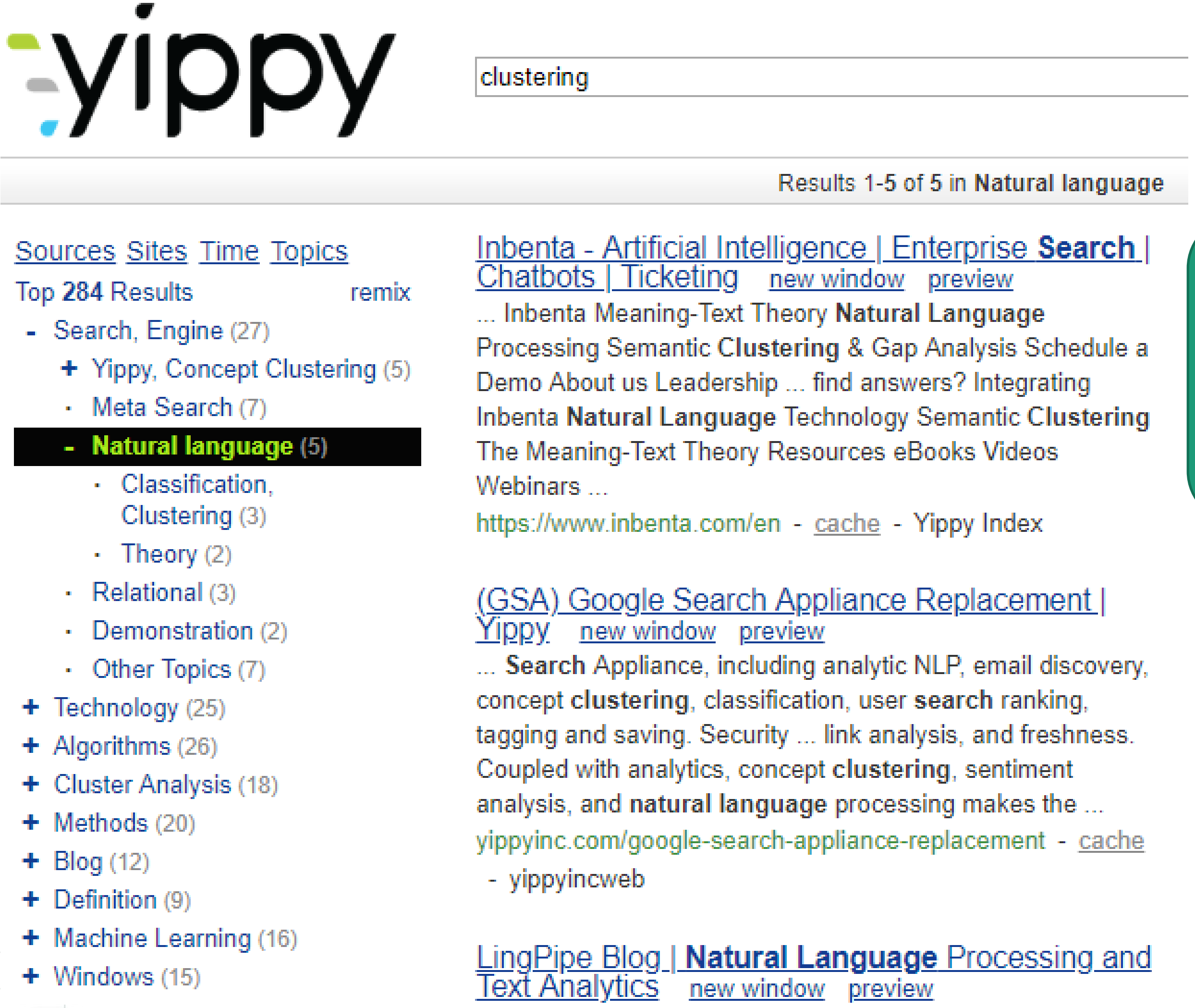


https://www.researchgate.net/profile/Sharifah_Sakinah_Syed_Ahmad/publication/280627665/

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



clustering

Results 1-5 of 5 in **Natural language**

Sources Sites Time Topics

Top 284 Results remix

- Search, Engine (27)
  - + Yippy, Concept Clustering (5)
  - Meta Search (7)
  - - **Natural language** (5)
    - Classification, Clustering (3)
    - Theory (2)
  - Relational (3)
  - Demonstration (2)
  - Other Topics (7)
- + Technology (25)
- + Algorithms (26)
- + Cluster Analysis (18)
- + Methods (20)
- + Blog (12)
- + Definition (9)
- + Machine Learning (16)
- + Windows (15)

Inbenta - Artificial Intelligence | Enterprise **Search** | Chatbots | Ticketing new window preview

... Inbenta Meaning-Text Theory **Natural Language** Processing Semantic **Clustering** & Gap Analysis Schedule a Demo About us Leadership ... find answers? Integrating Inbenta **Natural Language** Technology Semantic **Clustering** The Meaning-Text Theory Resources eBooks Videos Webinars ...

https://www.inbenta.com/en - cache - Yippy Index

(GSA) Google Search Appliance Replacement | Yippy new window preview

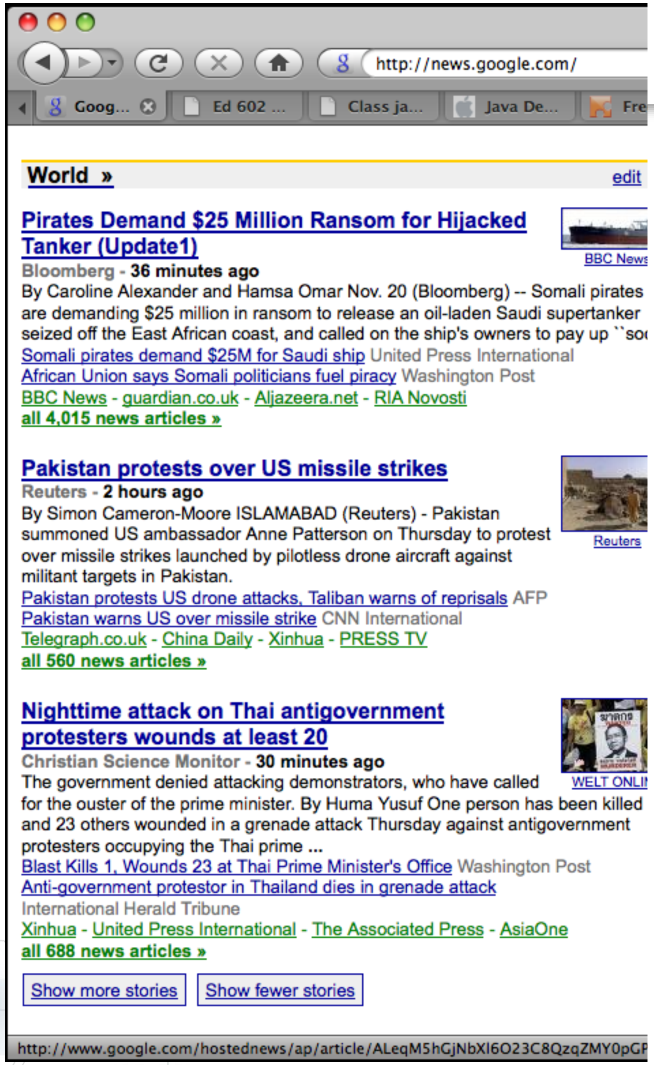
... **Search** Appliance, including analytic NLP, email discovery, concept **clustering**, classification, user **search** ranking, tagging and saving. Security ... link analysis, and freshness. Coupled with analytics, concept **clustering**, sentiment analysis, and **natural language** processing makes the ...

yippyinc.com/google-search-appliance-replacement - cache - yippyincweb

LingPipe Blog | **Natural Language** Processing and Text Analytics new window preview

Struktur internal hasil pencarian

Why: Clustering pada News Aggregator



World » edit

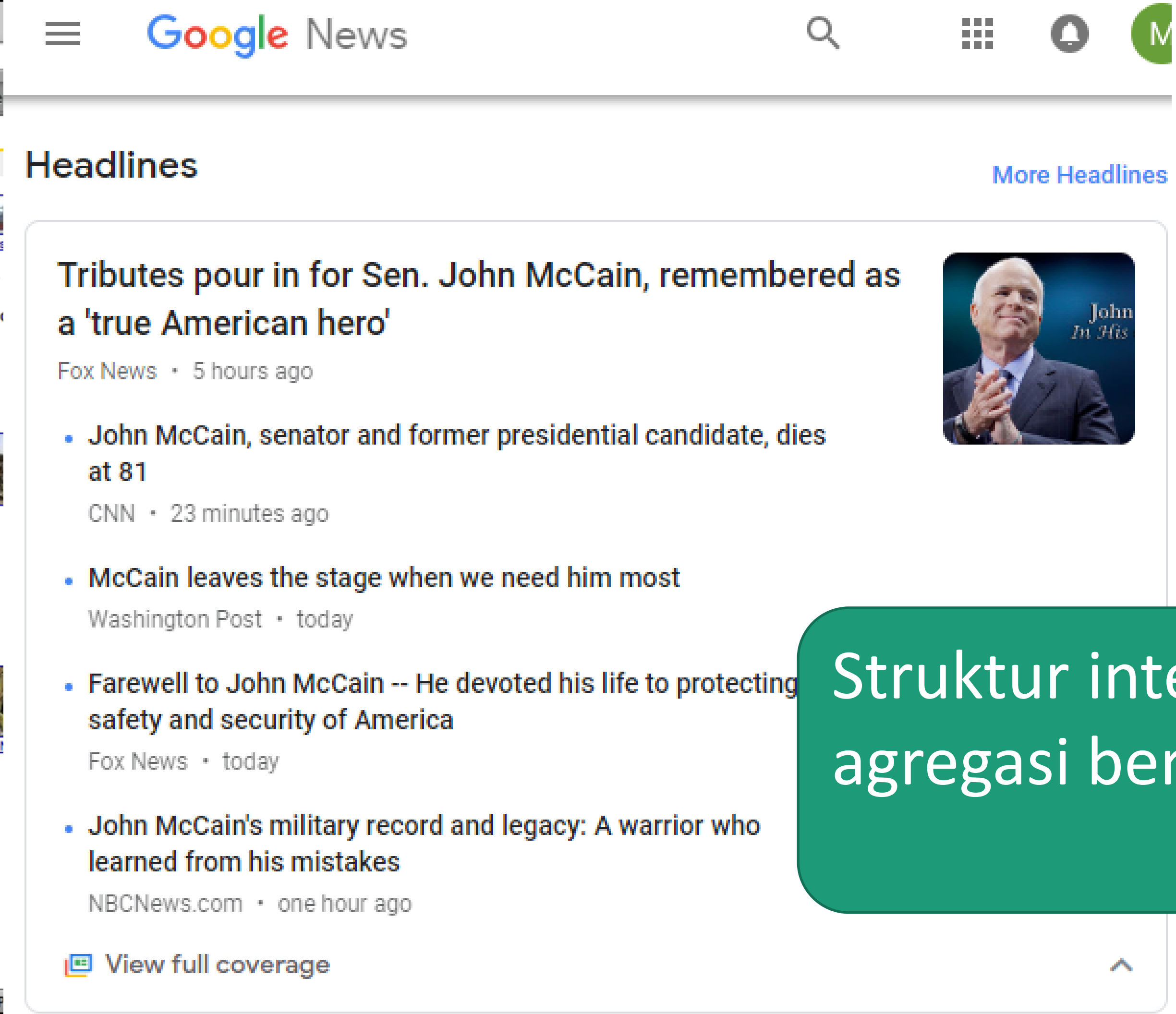
Pirates Demand $25 Million Ransom for Hijacked Tanker (Update1)
BBC News
Bloomberg - 36 minutes ago
By Caroline Alexander and Hamsa Omar Nov. 20 (Bloomberg) -- Somali pirates are demanding $25 million in ransom to release an oil-laden Saudi supertanker seized off the East African coast, and called on the ship's owners to pay up ``so
Somali pirates demand $25M for Saudi ship United Press International
African Union says Somali politicians fuel piracy Washington Post
BBC News - guardian.co.uk - Aljazeera.net - RIA Novosti
all 4,015 news articles »

Pakistan protests over US missile strikes
Reuters
Reuters - 2 hours ago
By Simon Cameron-Moore ISLAMABAD (Reuters) - Pakistan summoned US ambassador Anne Patterson on Thursday to protest over missile strikes launched by pilotless drone aircraft against militant targets in Pakistan.
Pakistan protests US drone attacks, Taliban warns of reprisals AFP
Pakistan warns US over missile strike CNN International
Telegraph.co.uk - China Daily - Xinhua - PRESS TV
all 560 news articles »

Nighttime attack on Thai antigovernment protesters wounds at least 20
WELT ONLI
Christian Science Monitor - 30 minutes ago
The government denied attacking demonstrators, who have called for the ouster of the prime minister. By Huma Yusuf One person has been killed and 23 others wounded in a grenade attack Thursday against antigovernment protesters occupying the Thai prime ...
Blast Kills 1, Wounds 23 at Thai Prime Minister's Office Washington Post
Anti-government protestor in Thailand dies in grenade attack International Herald Tribune
Xinhua - United Press International - The Associated Press - AsiaOne
all 688 news articles »

Show more stories | Show fewer stories

http://www.google.com/hostednews/ap/article/ALeqM5hGjNbXI6O23C8QzqZMY0pGF



Headlines

More Headlines

Tributes pour in for Sen. John McCain, remembered as a 'true American hero'
Fox News • 5 hours ago

- John McCain, senator and former presidential candidate, dies at 81
  CNN • 23 minutes ago
- McCain leaves the stage when we need him most
  Washington Post • today
- Farewell to John McCain -- He devoted his life to protecting safety and security of America
  Fox News • today
- John McCain's military record and legacy: A warrior who learned from his mistakes
  NBCNews.com • one hour ago

View full coverage

Struktur internal agregasi berita

Why: Clustering pada Gen

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

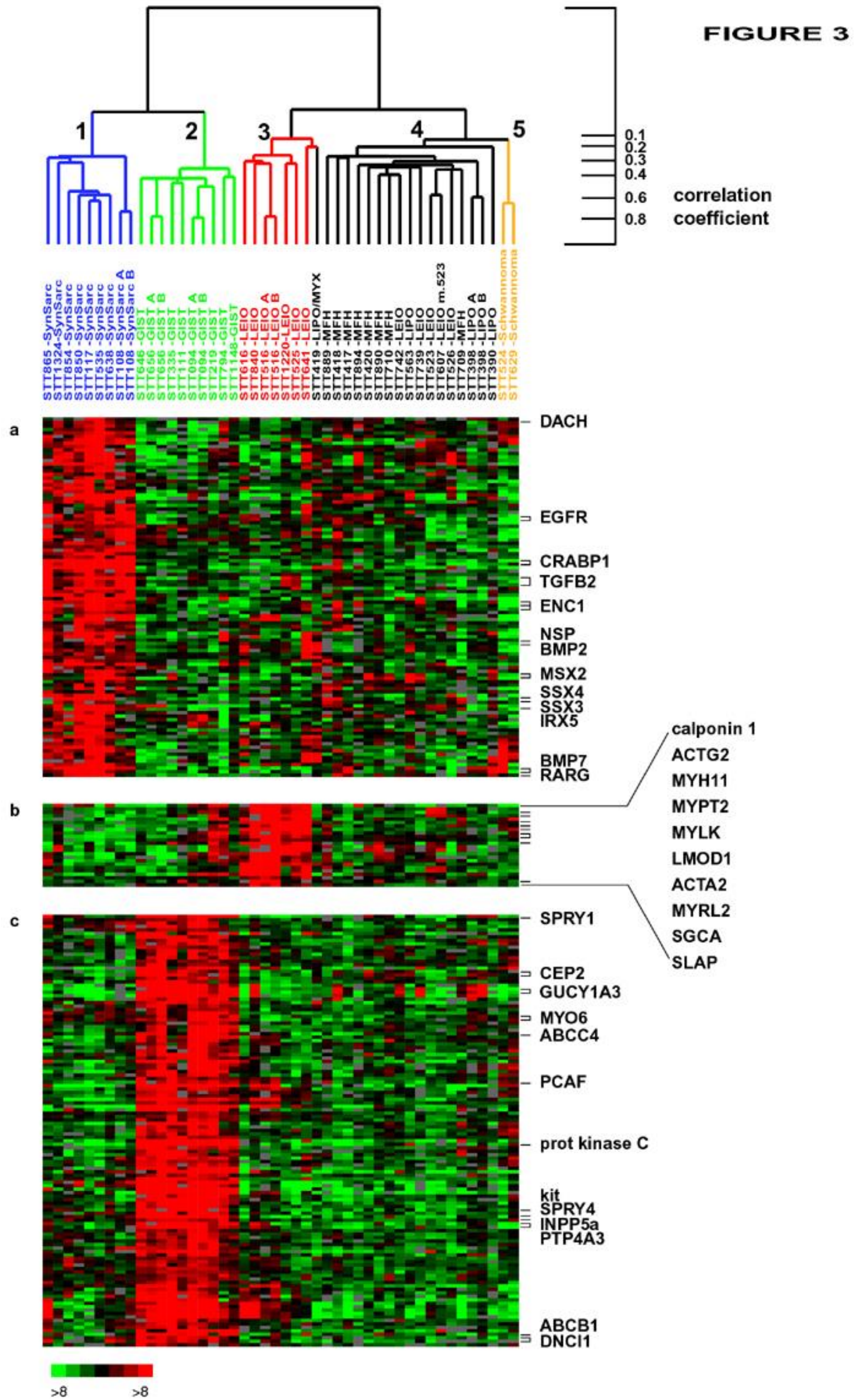
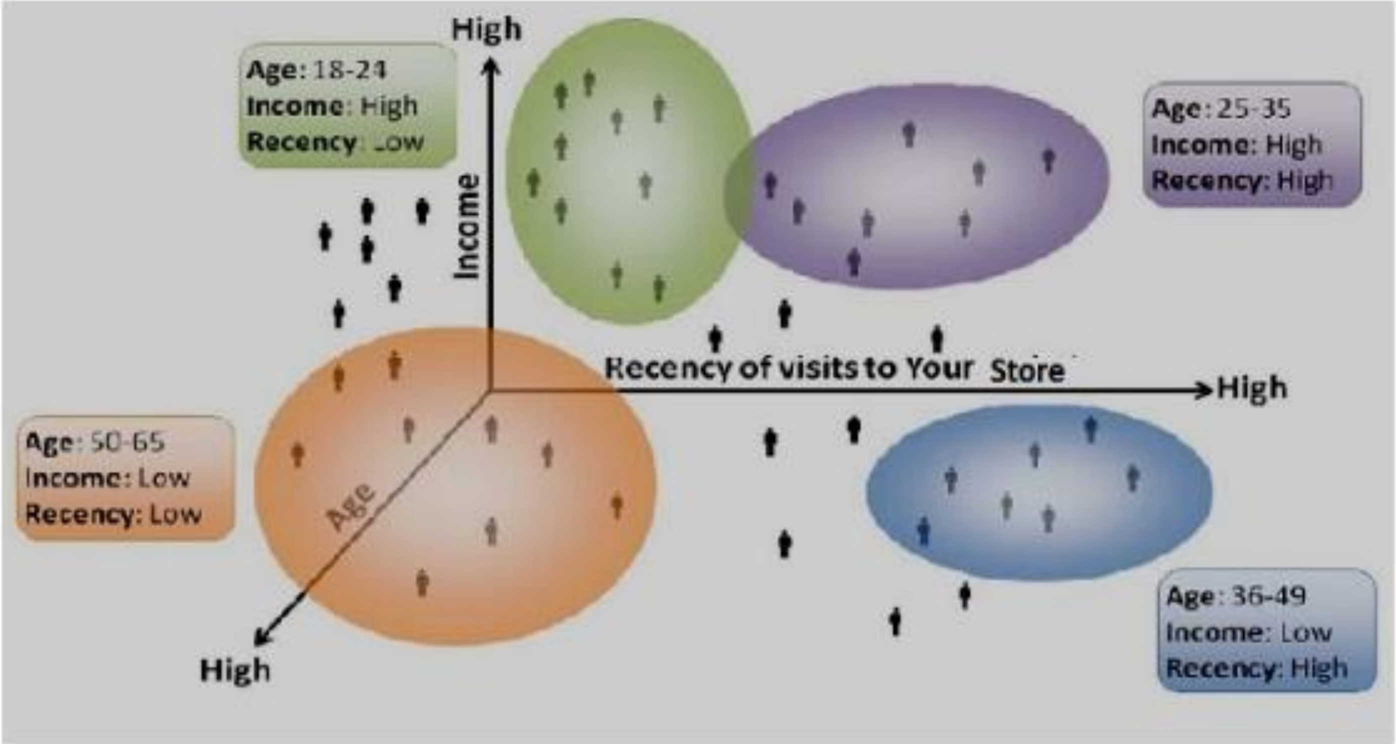


FIGURE 3

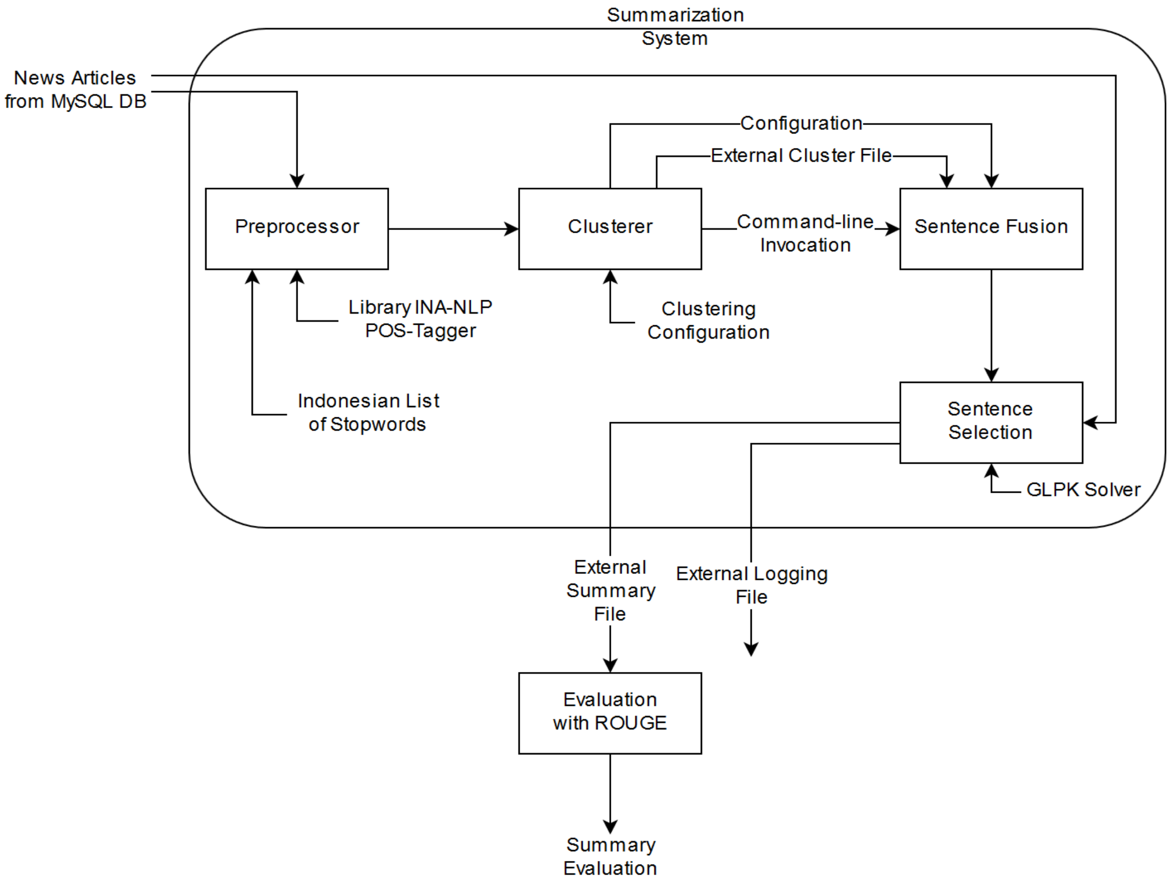
Why: Clustering untuk Segmentasi

Example - Clusters using Age, Income & Recency



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

Why: Clustering-based Approach

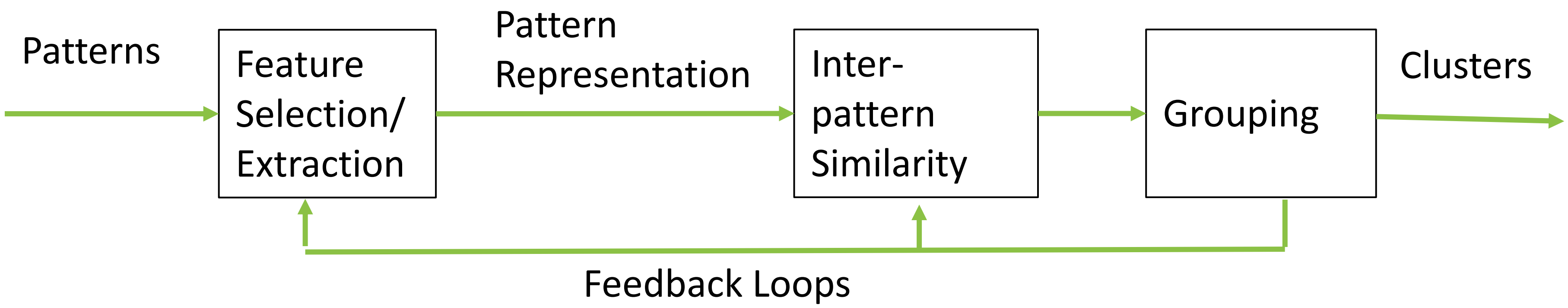


Clustering-based summarization

Clustering-based outlier detection

Clustering-based analysis

Tahapan Clustering



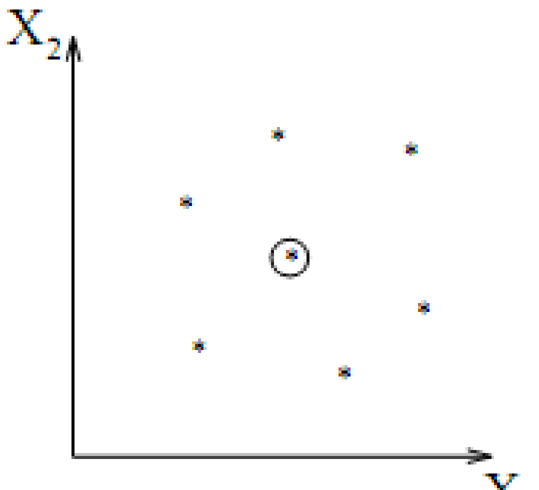
- Tahapan utama:
  1) Feature selection: original features → subset of features
     Feature extraction: transformation into new features
  2) Pattern proximity/similarity measure
  3) Grouping
- Clustering output: hard atau soft (membership degree)

Tahapan Clustering (lanjutan)

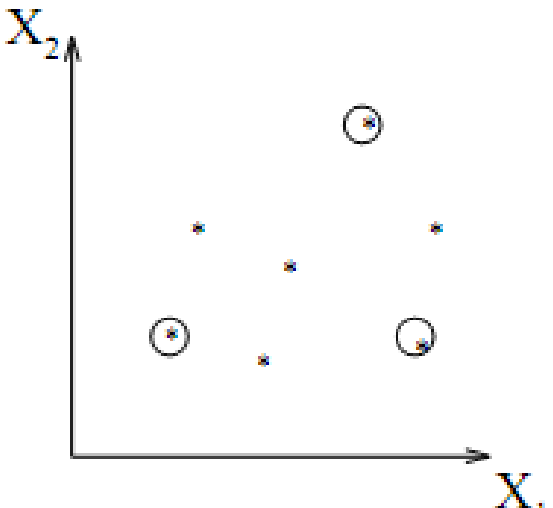
- Tahapan opsional:
  4) data abstraction
  5) assessment of output (good or poor)

Representasi Cluster (1)

- Centroid atau set of distant point



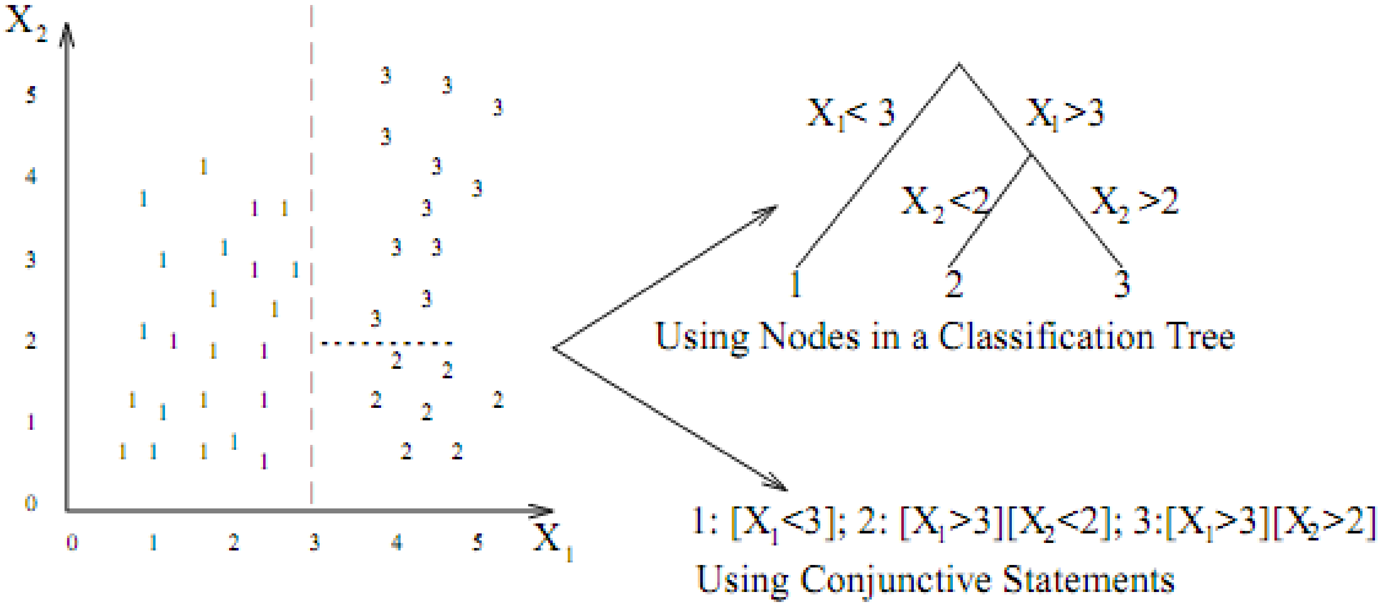
By The Centroid



By Three Distant Points

Sumber: Jain dkk (1999)

Representasi Cluster (2)



Using Nodes in a Classification Tree

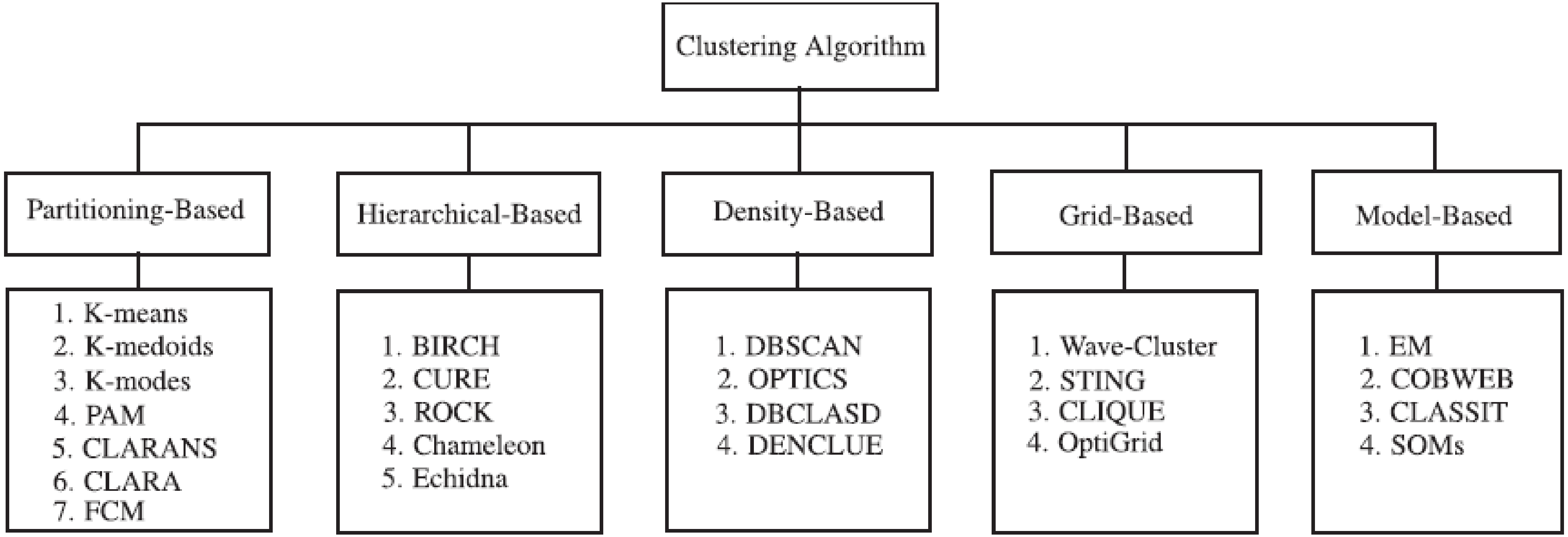
1: $[X_1<3]$; 2: $[X_1>3][X_2<2]$; 3: $[X_1>3][X_2>2]$

Using Conjunctive Statements

- Pohon klasifikasi
- Conjunctive statements

Sumber: Jain dkk (1999)

Kategori Metode Clustering (Fahad, 2014)



Clustering Algorithm

Partitioning-Based

1. K-means
2. K-medoids
3. K-modes
4. PAM
5. CLARANS
6. CLARA
7. FCM

Hierarchical-Based

1. BIRCH
2. CURE
3. ROCK
4. Chameleon
5. Echidna

Density-Based

1. DBSCAN
2. OPTICS
3. DBCLASD
4. DENCLUE

Grid-Based

1. Wave-Cluster
2. STING
3. CLIQUE
4. OptiGrid

Model-Based

1. EM
2. COBWEB
3. CLASSIT
4. SOMs

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