

# 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

IF3270 - Pembelajaran Mesin  
(Machine Learning)

# Outline

Clustering: What?

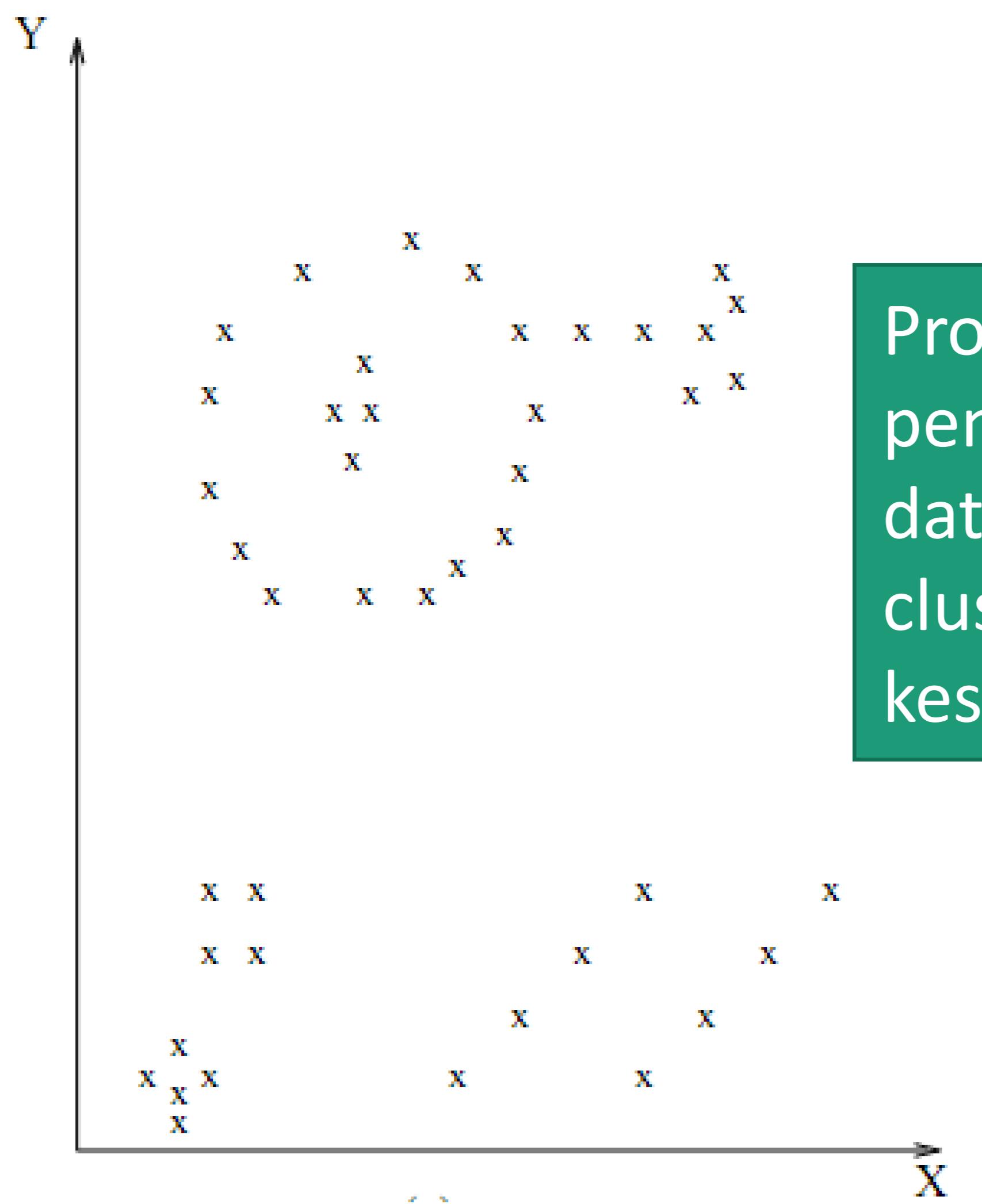
Tahapan Clustering

Representasi  
Cluster

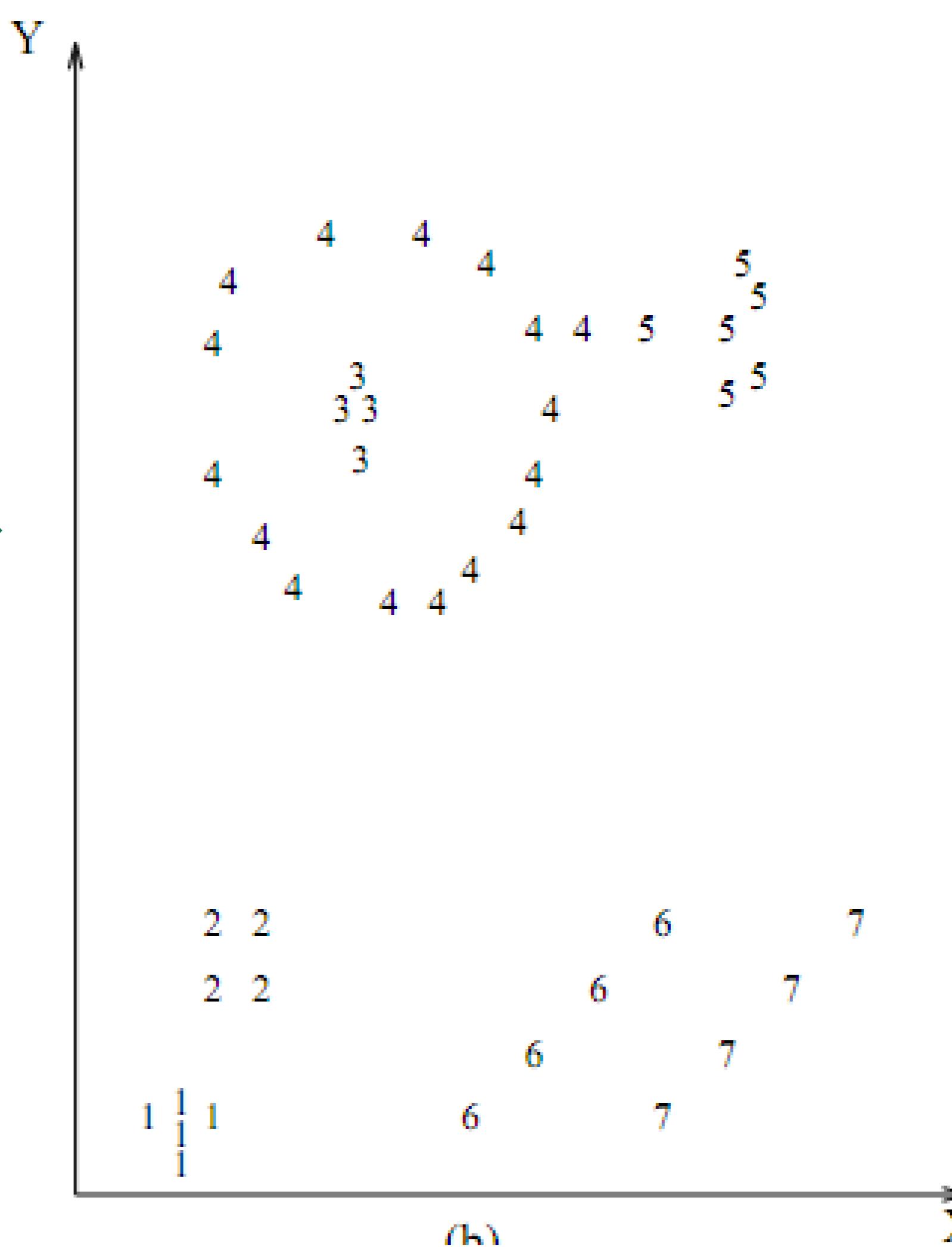
Clustering: Why?

Kategori Metode

# Clustering: What ?



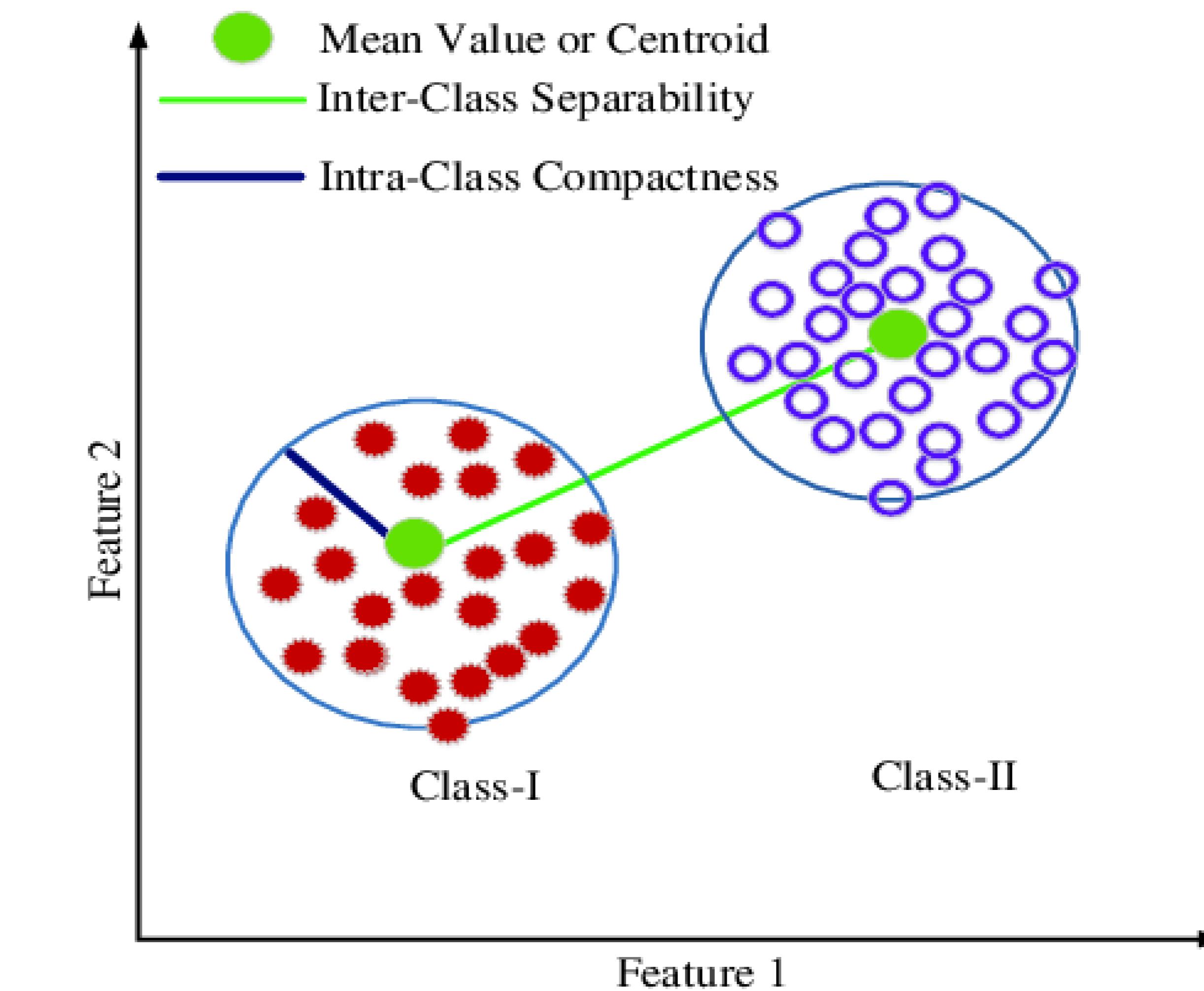
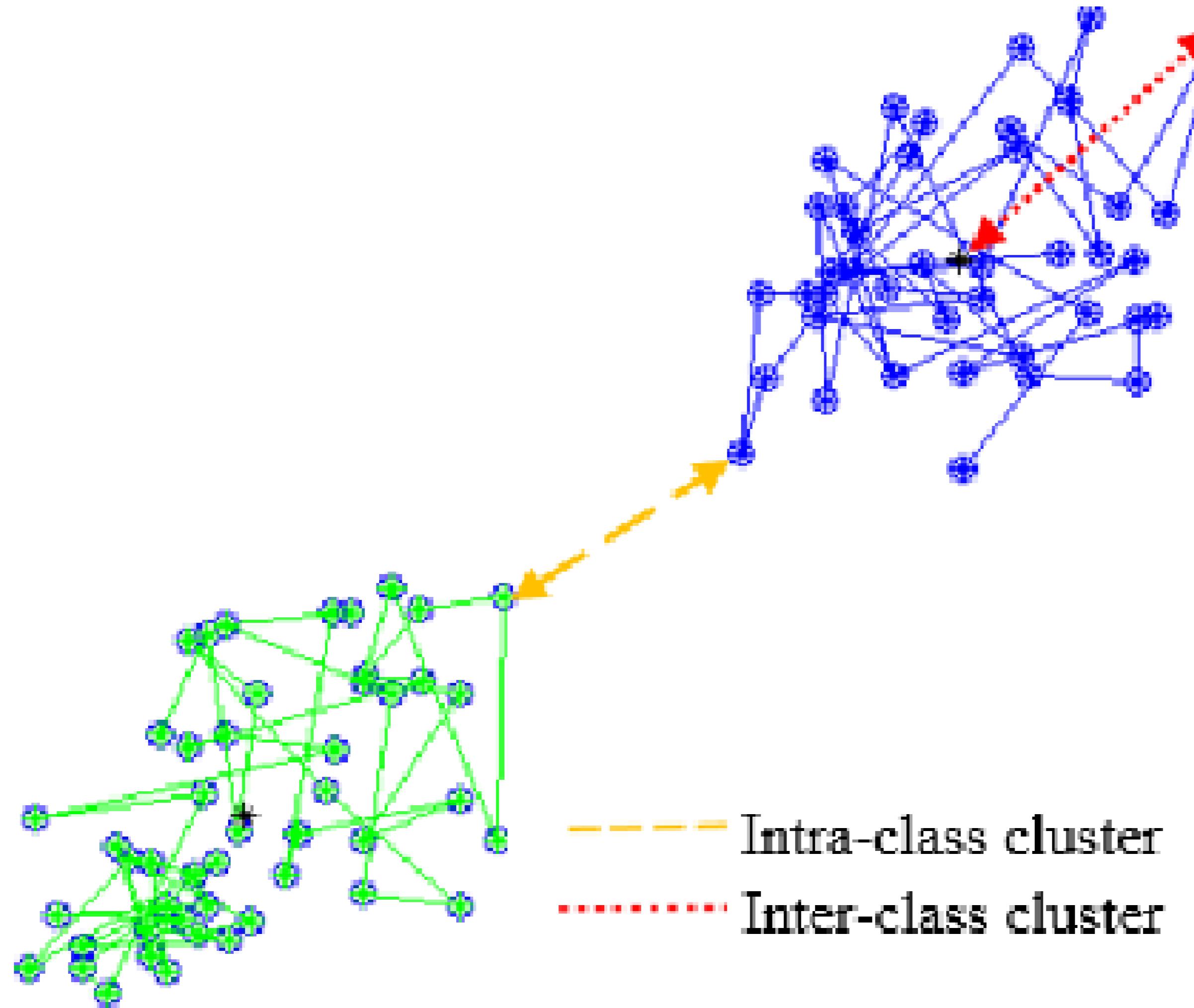
Proses pengelompokan data menjadi clusters berbasis kesamaan data



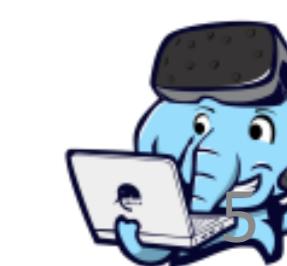
Unsupervised learning = learning from raw data

Sumber: Jain dkk (1999)

# Clustering:



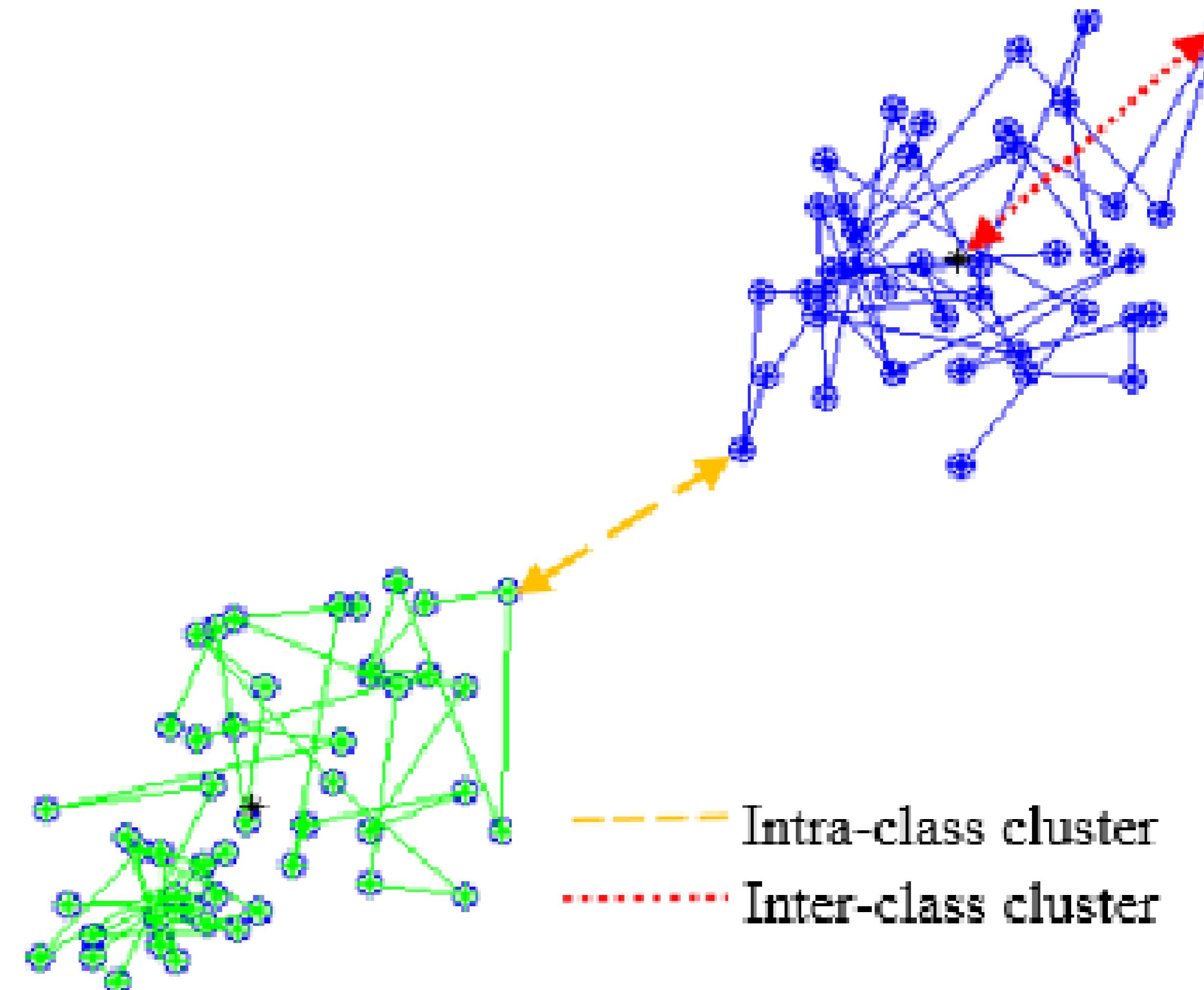
[https://www.researchgate.net/figure/The-concept-of-intra-class-compactness-and-inter-class-separability-in-a-two-dimension\\_fig3\\_325095062](https://www.researchgate.net/figure/The-concept-of-intra-class-compactness-and-inter-class-separability-in-a-two-dimension_fig3_325095062)





# 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/](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 ...

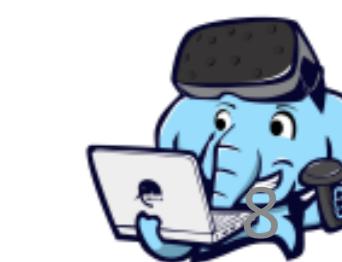
[yippyinc.com/google-search-appliance-replacement](http://yippyinc.com/google-search-appliance-replacement) - [cache](#)

- yippyincweb

[LingPipe Blog | Natural Language Processing and](#)

[Text Analytics](#) [new window](#) [preview](#)

Struktur internal  
hasil pencarian



# Why: Clustering pada News Aggregator

**Google News**

**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

McCain leaves the stage when we need him most

Farewell to John McCain -- He devoted his life to protecting safety and security of America

John McCain's military record and legacy: A warrior who learned from his mistakes

View full coverage

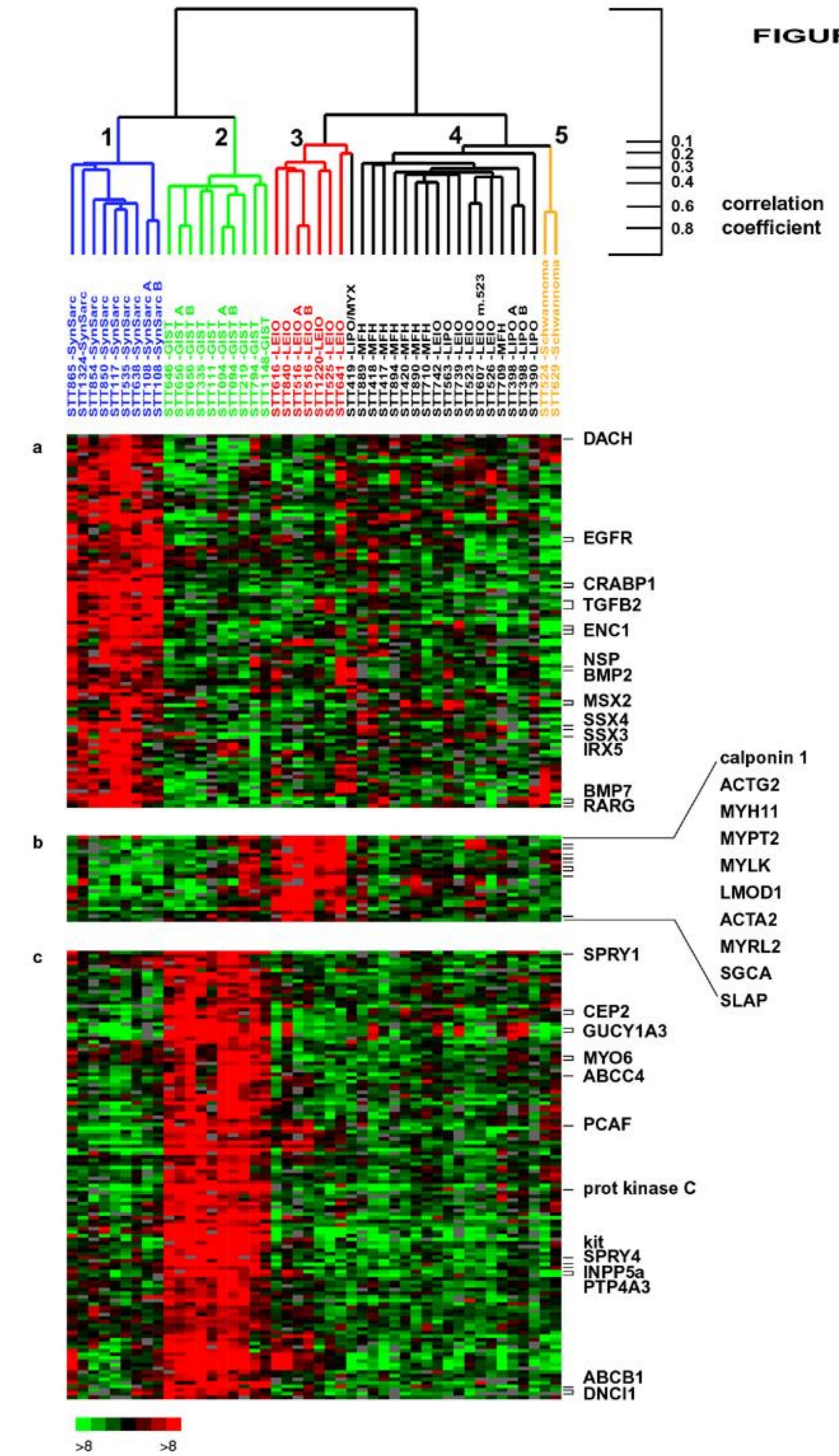
The screenshot shows the Google News homepage. On the left, there are three news stories listed under 'World': 'Pirates Demand \$25 Million Ransom for Hijacked Tanker (Update1)', 'Pakistan protests over US missile strikes', and 'Nighttime attack on Thai antigovernment protesters wounds at least 20'. Each story includes a thumbnail image, the source, and a link to 'all [number] news articles'. On the right, a large sidebar is titled 'Headlines' and features a prominent image of Senator John McCain clapping. Below the image, several news items about McCain's death are listed, each with a source and a timestamp. At the bottom of the sidebar, there is a link to 'View full coverage'.

Struktur internal  
agregasi berita



# Why: Clustering pada Gen

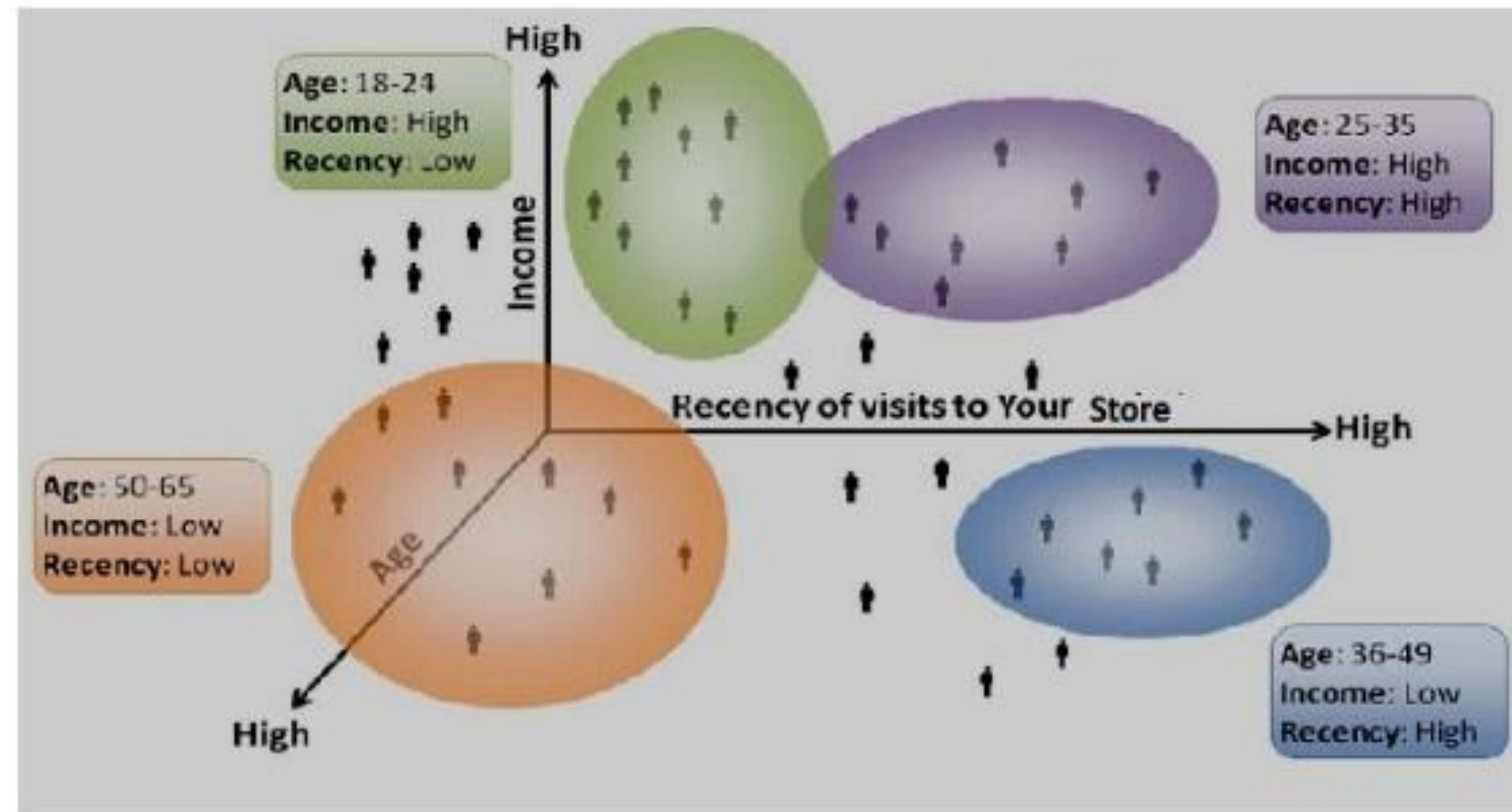
[http://genome-www.stanford.edu/sarcoma/supplemental\\_data.html](http://genome-www.stanford.edu/sarcoma/supplemental_data.html)





# Why: Clustering untuk Segmentasi

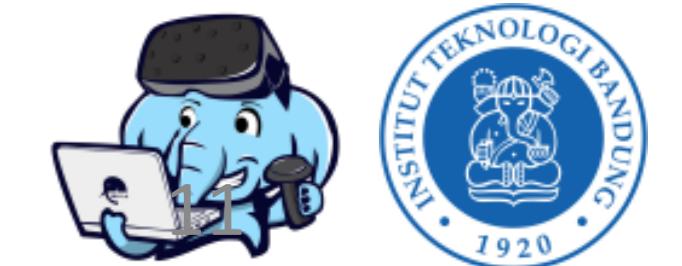
Example - Clusters using Age, Income & Recency



Canvass

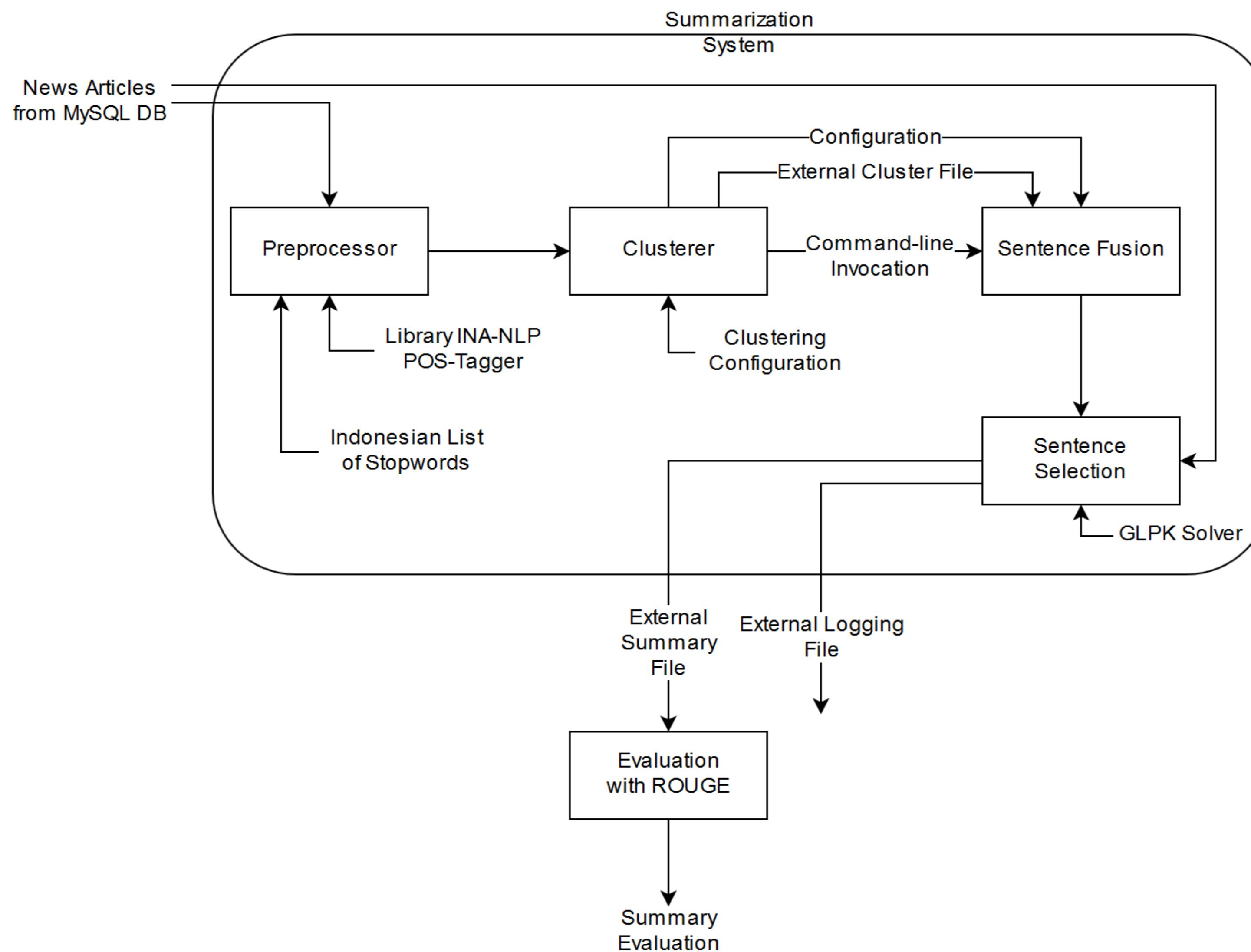
<https://www.slideshare.net/Canvass/marketing-data-analytics>

Copyright: Canvass 2013-2016





# Why: Clustering-based Approach



Clustering-based summarization

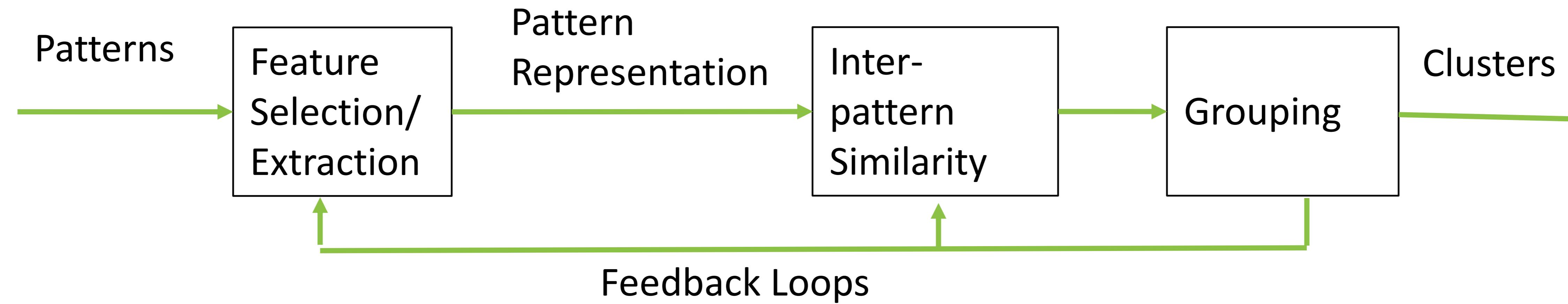
Clustering-based outlier detection

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)

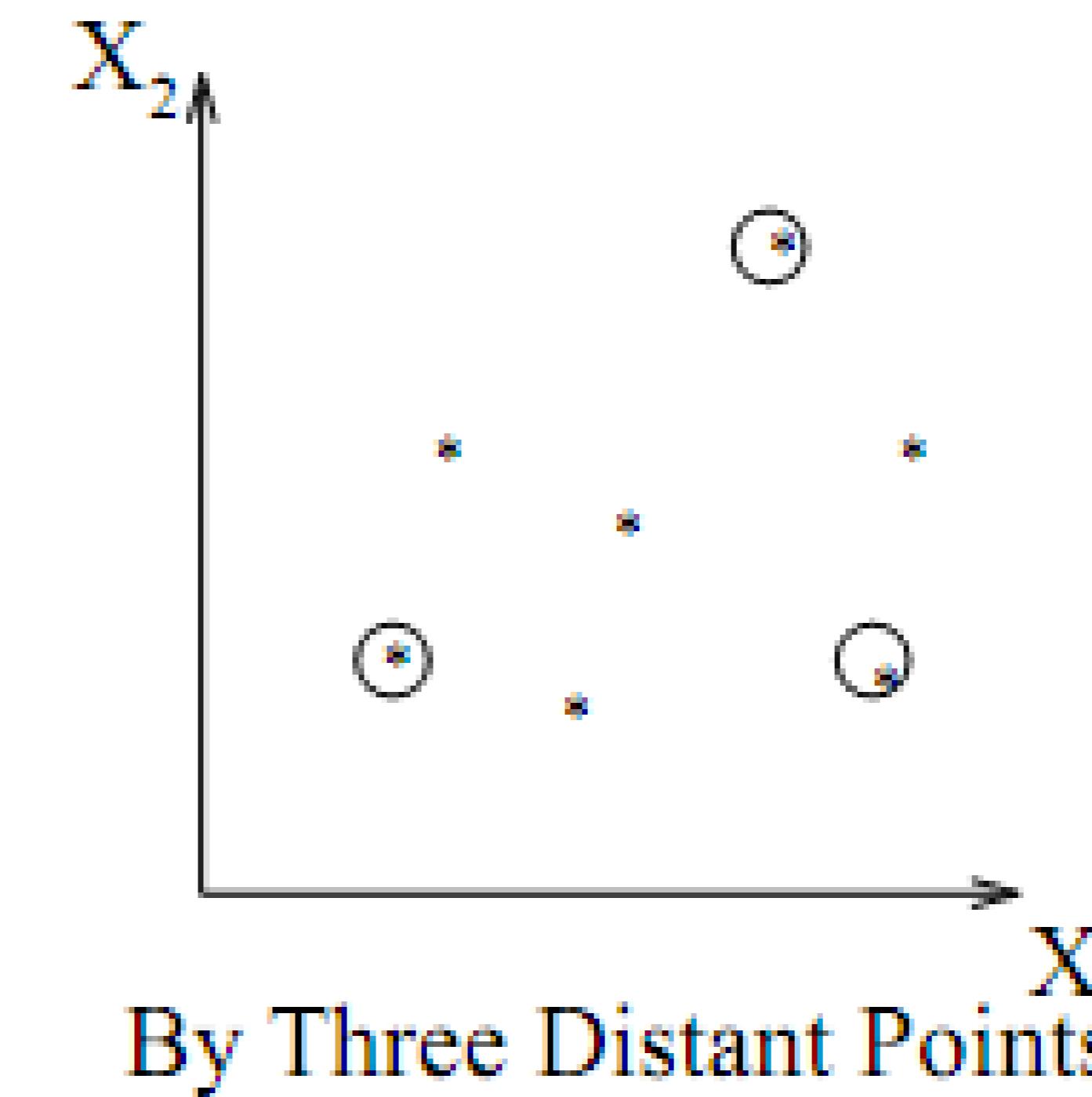
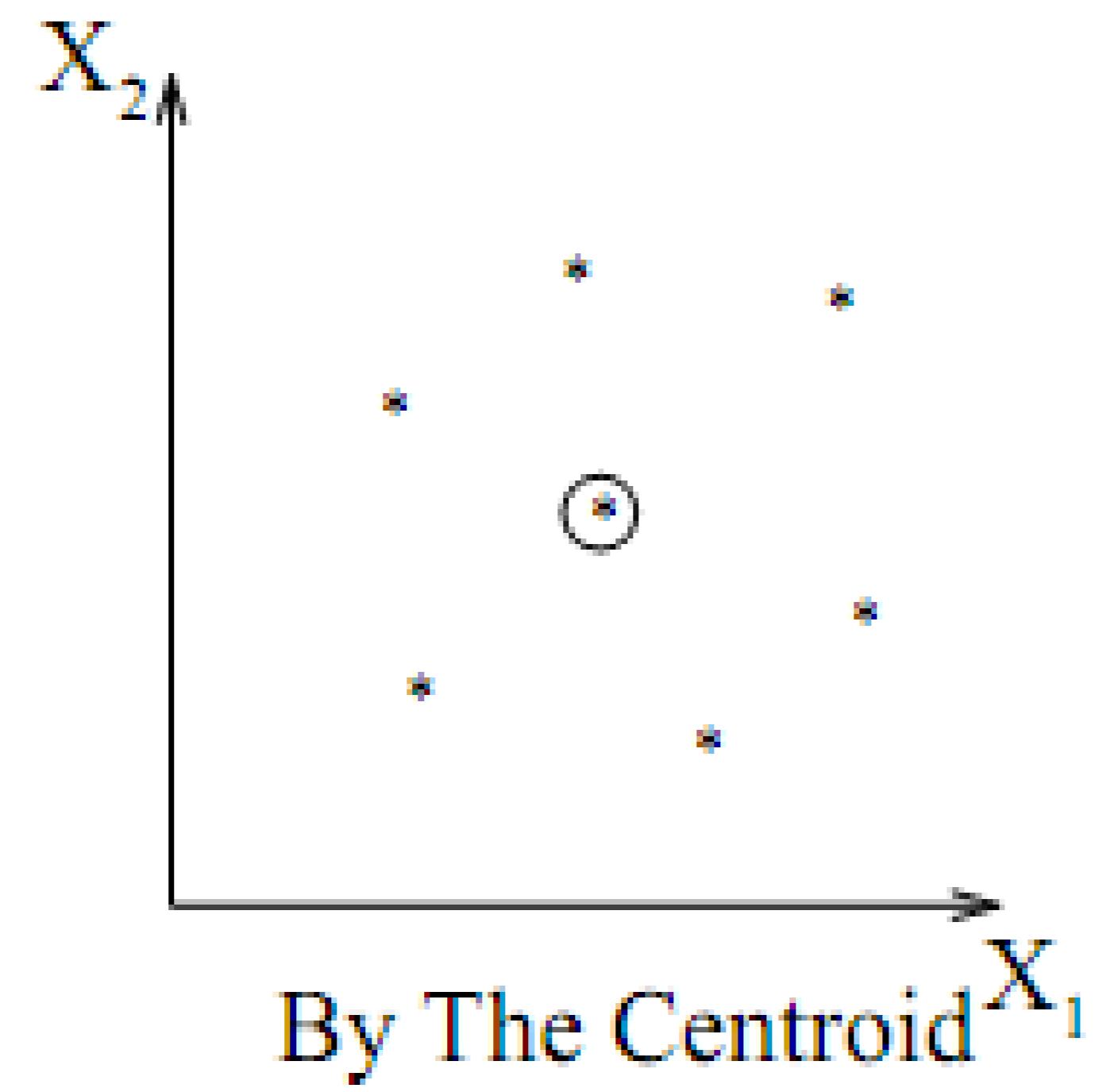
- Tahapan opsional:
  - 4) data abstraction
  - 5) assessment of output (good or poor)





# Representasi Cluster (1)

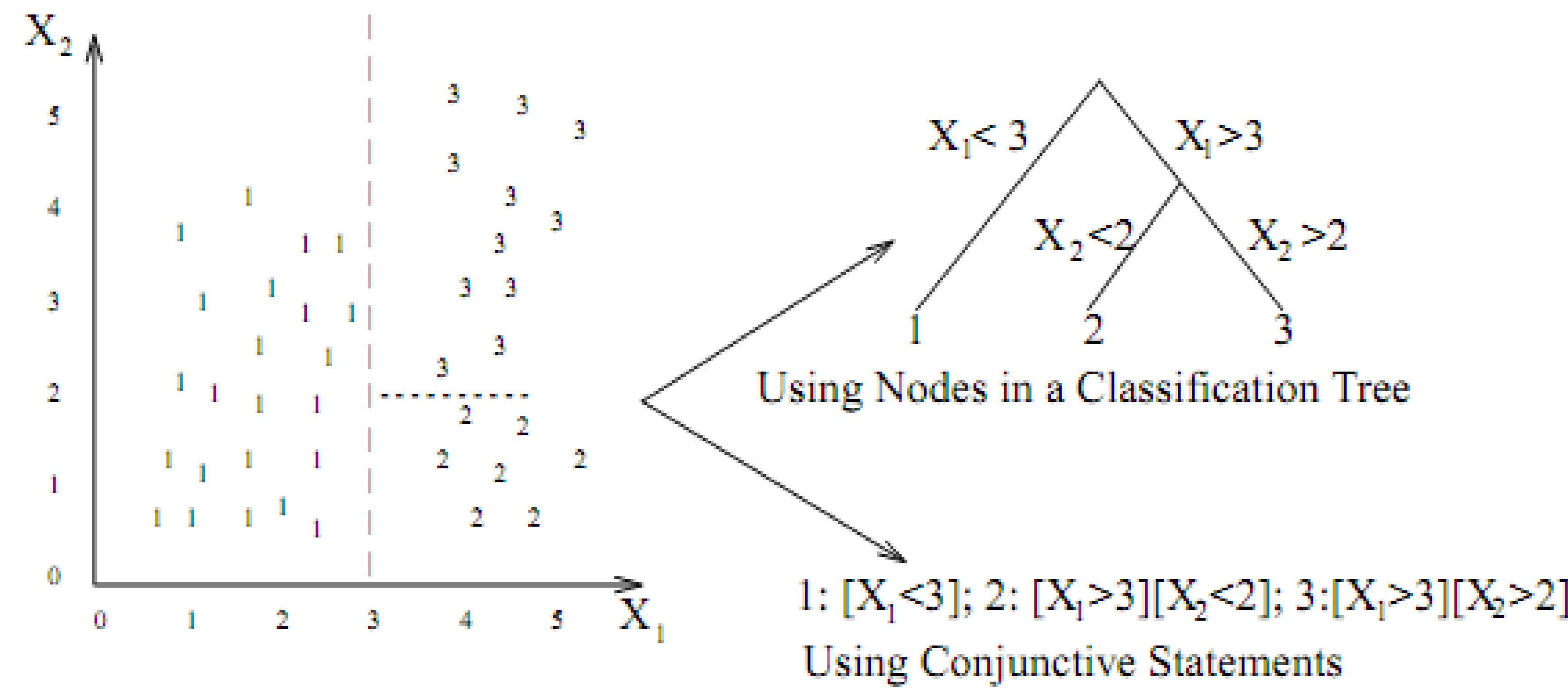
- Centroid atau set of distant point



Sumber: Jain dkk (1999)

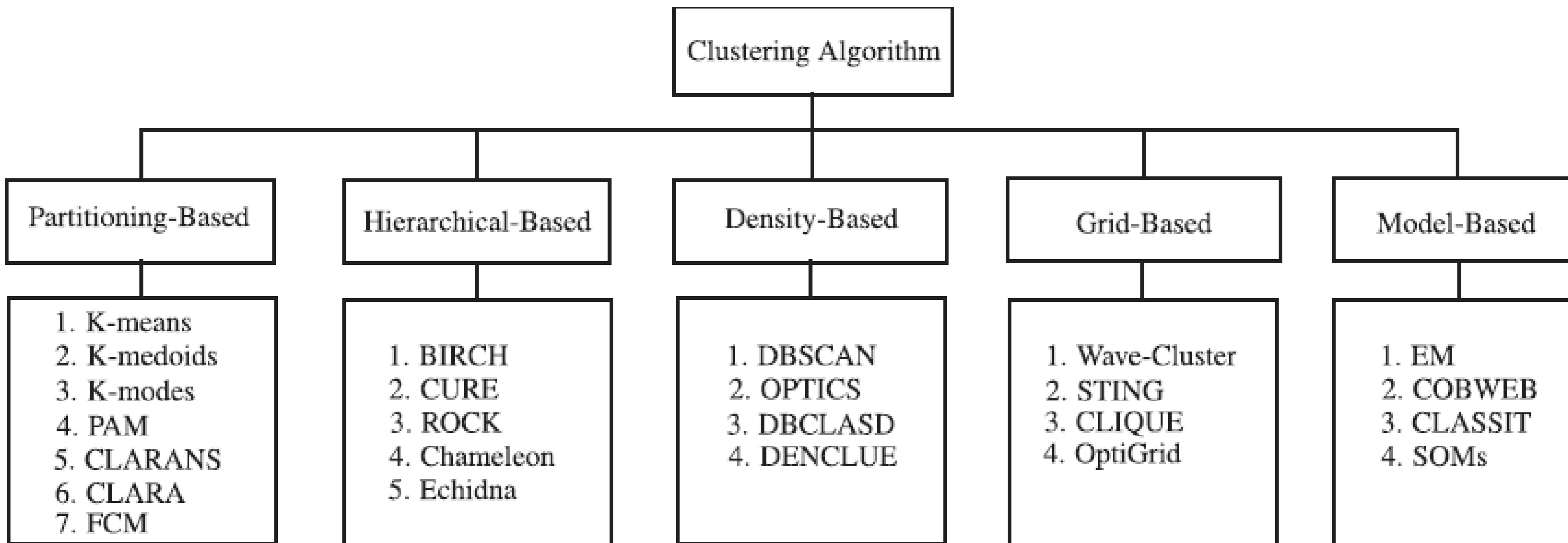


# Representasi Cluster (2)



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);  $k \leq$  jumlah data
- Contoh: K-means, k-medoids

## 2. Metode *hierarchical*

menghasilkan rangkaian partisi bersarang

- Agglomerative (bottom-up, merge):  
1 object ~ 1 cluster  $\rightarrow$  1 cluster n-object
- Divisive (top-down, split):  
1 cluster n-object  $\rightarrow$  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

