Institut Informatika & Bisnis DARMAJAYA
Yayasan Alfian Husin

SAINS DATA
Institut Informatika & Bisnis Darmajaya

Kampus Merdeka INDONESIA JAYA

MERDEKA BELAJAR

DATA SCIENCE DARMAJAYA
"YOUR BEST FUTURE IN DATA"

PERTEMUAN KE: 9

# *Unsupervised Learning*
# Konsep *Clustering*

## KULIAH

OLEH: Nurjoko

[ 12-09-2023 ] | [KODE MK: SD230411 | MATAKULIAH: MATEMATIKA | SKS: 4 | VERSI: 01

# Learning Objectives

- **Mahasiswa mampu menjelaskan algoritma clustering berbasis partisi, serta penggunaannya pada suatu aplikasi.**

- **Mahasiswa mampu menjelaskan algoritma analisa asosiasi, serta penggunaannya pada suatu aplikasi**
- **Learning Objective 3**
- **Learning Objective 4**
- **Learning Objective 5**

# Konsep *clustering*

## APA ITU *CLUSTERING*?

- ***Clustering*** adalah proses membagi kumpulan data (objek) ke dalam kelompok-kelompok yang disebut ***cluster***.
- Cluster adalah kumpulan objek yang dikelompokkan berdasarkan **kemiripan** antar objek.
  - Suatu cluster berisi kumpulan objek yang mirip
  - Objek pada satu cluster berbeda (tidak mirip) dengan objek pada cluster yang lain
- Kemiripan objek ditentukan menggunakan **jarak** (distance measure).
  - Dua objek yang mirip, memiliki jarak yang dekat (kecil)
- Clustering : **unsupervised learning, no predefined classes**.
- Membantu user memahami struktur dalam kumpulan objek

# APA ITU *CLUSTERING*?

## (ilustrasi)

**Kriteria kemiripan tabung**

- Warna
- Tinggi
- Jari-jari
- Tinggi dan jari-jari

# APA ITU *CLUSTERING*?

## (ilustrasi)

**Kriteria kemiripan tabung**

- **Warna**
- **Tinggi**
- **Jari-jari**
- **Tinggi dan jari-jari**

# APA ITU *CLUSTERING*?

## (ilustrasi)

**Kriteria kemiripan tabung**

- **Warna**
- **Tinggi**
- **Jari-jari**
- **Tinggi dan jari-jari**

# APA ITU *CLUSTERING*?

## (ilustrasi)

**Kriteria kemiripan tabung**

- **Warna**
- **Tinggi**
- **Jari-jari**
- **Tinggi dan jari-jari**

# APA ITU *CLUSTERING*?

## (ilustrasi)

**Kriteria kemiripan tabung**

- **Warna**
- **Tinggi**
- **Jari-jari**
- **Tinggi dan jari-jari**

# APA ITU *CLUSTERING*?

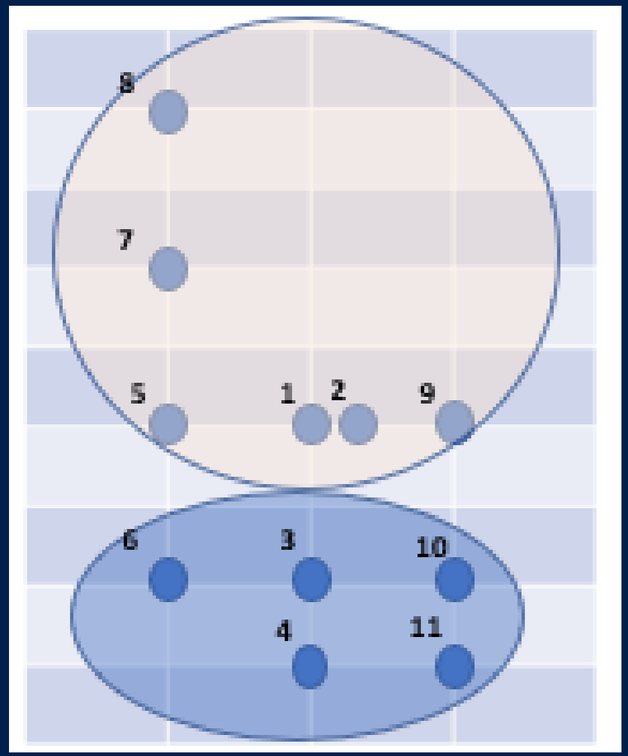
## (ilustrasi)

| Jari-jari | Tinggi |
|---|---|
| 1 | 2 |
| 1.1 | 2 |
| 1 | 1 |
| 1 | 0.5 |
| 0.5 | 2 |
| 0.5 | 1 |
| 0.5 | 3 |
| 0.5 | 4 |
| 1.5 | 2 |
| 1.5 | 1 |
| 1.5 | 0.5 |

# APA ITU *CLUSTERING*?
## (ilustrasi)

| Jari-jari | Tinggi |
|---|---|
| 1 | 2 |
| 1.1 | 2 |
| 1 | 1 |
| 1 | 0.5 |
| 0.5 | 2 |
| 0.5 | 1 |
| 0.5 | 3 |
| 0.5 | 4 |
| 1.5 | 2 |
| 1.5 | 1 |
| 1.5 | 0.5 |

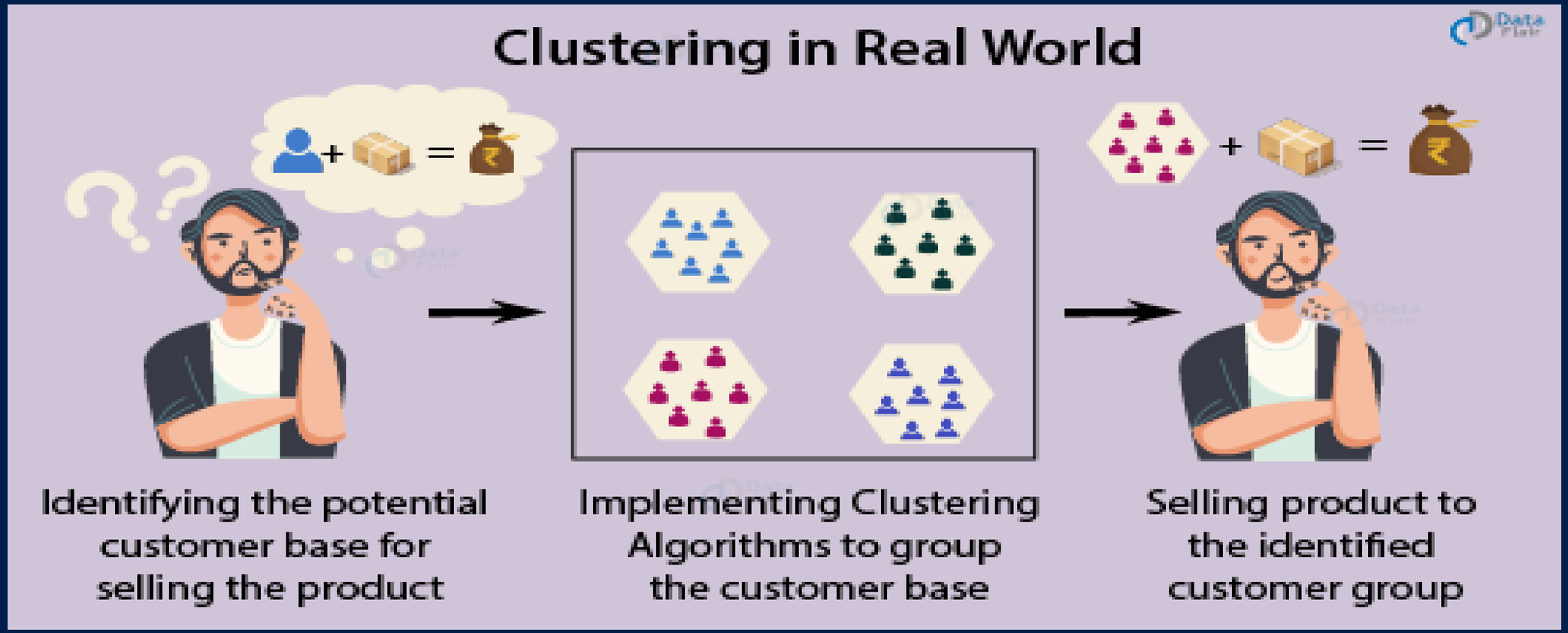
Tinggi




Jari-jari

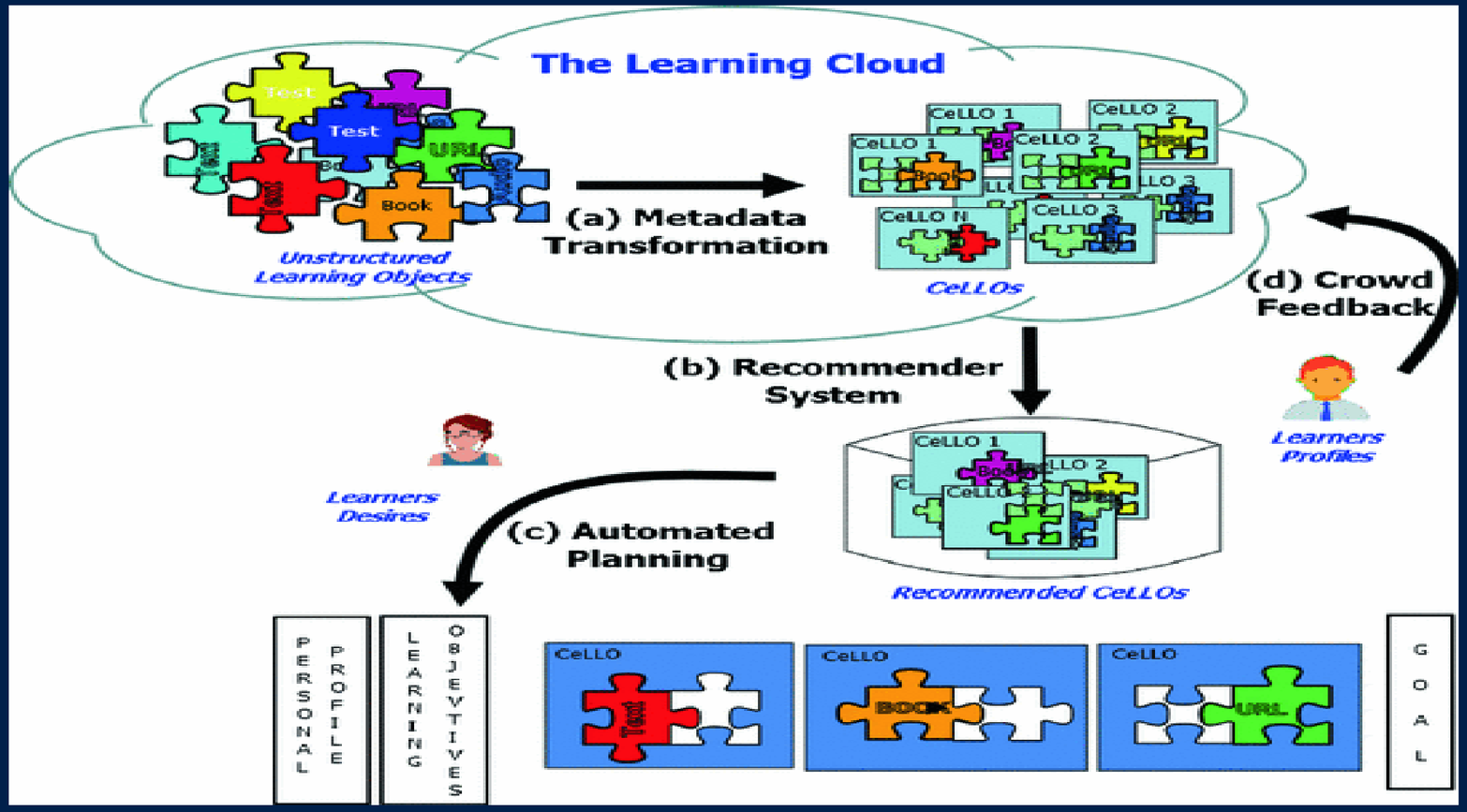
- Mirip → jarak yang dekat
- Rumus jarak
  - Minkowsky
    - Euclidean
    - Manhattan
  - Cosine
- Teknik clustering berbasis jarak
  - Algoritma Partisi
    - K-Means
    - K-Medians
    - K-Medoids
  - Algoritma Hirarki
    - Agglomerative vs divisive methods

# Contoh penggunaan clustering



# Contoh penggunaan clustering

Pireva K., Kefalas P. (2018) **A Recommender System Based on Hierarchical Clustering for Cloud e-Learning**. In: Ivanović M., Bădică C., Dix J., Jovanović Z., Malgeri M., Savić M. (eds) Intelligent Distributed Computing XI. IDC 2017. Studies in Computational Intelligence, vol 737. Springer, Cham. https://doi.org/10.1007/978-3-319-66379-1_21

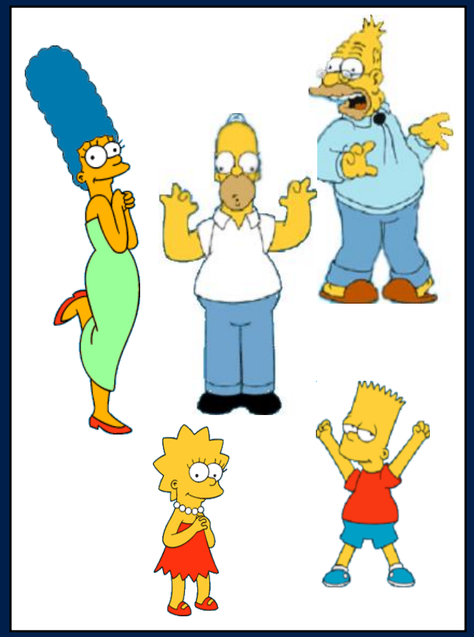


# Dua tipe *clustering* berbasis jarak

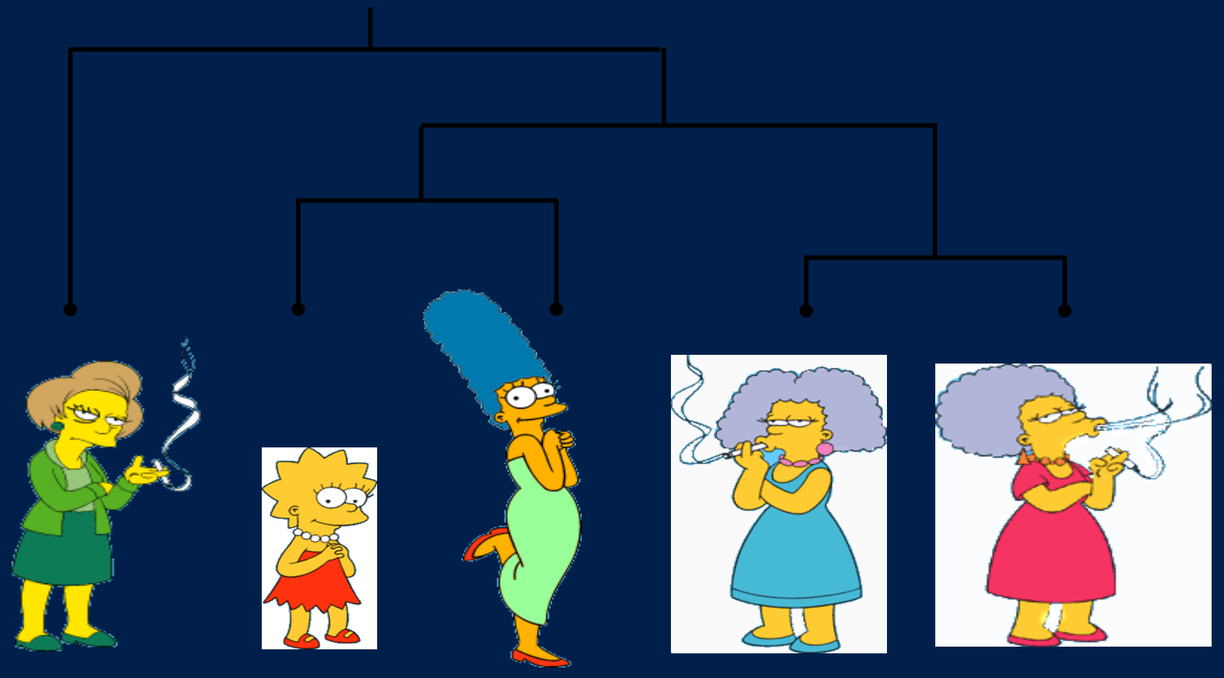
- Algoritma partisi: Menyusun beberapa partisi dan mengevaluasinya menggunakan kriteria tertentu.
  - *k*-means, *k*-medoids, *k*-median
  - Fuzzy *c*-means
- Algoritma hirarki: Membuat dekomposisi hirarki dari kumpulan objek menggunakan kriteria tertentu.
  - Agglomerative ("bottom-up"): Dimulai dengan menjadikan setiap objek sebagai satu cluster dan selanjutnya menggabungkannya menjadi cluster yang lebih besar.
  - Divisive ("top-down"): Dimulai dari satu cluster yang besar dan selanjutnya membagi menjadi cluster-cluster yang lebih kecil.

# Dua tipe *clustering* berbasis jarak

## Partisi



## Hirarki



# Rumus jarak untuk data dengan atribut numerik (1)

- Rumus jarak menentukan perhitungan kemiripan antara dua objek.
- Rumus jarak mempengaruhi bentuk cluster.
- Rumus jarak Minkowski:

$$d(i,j) = \sqrt[p]{\left|x_{i1} - x_{j1}\right|^p + \left|x_{i2} - x_{j2}\right|^p + \cdots + \left|x_{il} - x_{jl}\right|^p}$$

Dengan $i = (x_{i1}, x_{i2}, \ldots, x_{il})$ dan $j = \left(x_{j1}, x_{j2}, \ldots, x_{jl}\right)$ adalah data dengan ***l*** dimensi dan ***p*** adalah order (jarak disebut L-p norm).

# Rumus jarak untuk data dengan atribut numerik (2)

- p=1: (L1 norm) → Jarak Manhattan (city block)
  - Jarak Hamming: jumlah bit yang berbeda di antara dua vektor biner

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{il} - x_{jl}|$$

- p=2: (L2 norm) → Jarak Euclidean

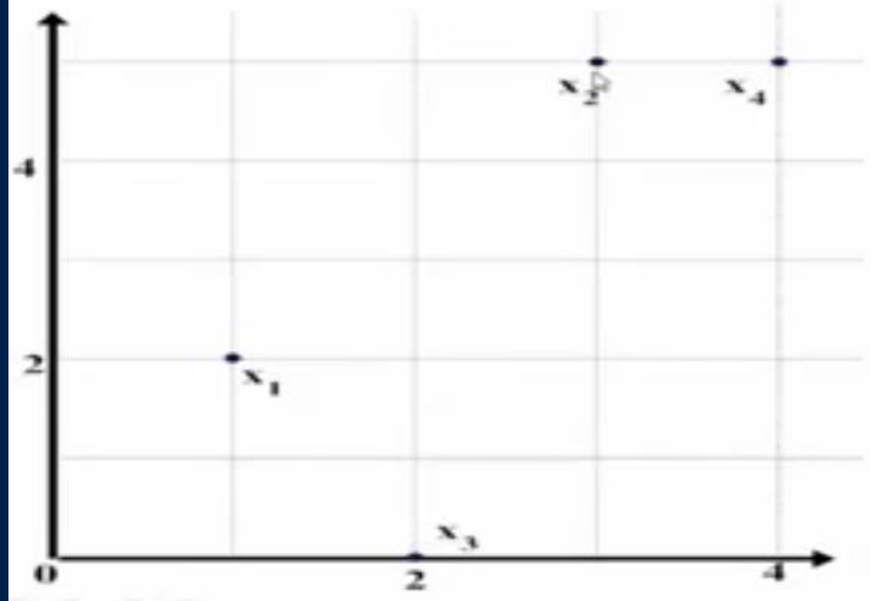
$$d(i.j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{il} - x_{jl}|^2}$$

- P→∞: (Lmax norm, L∞ norm) → Jarak supremum
  - Jarak maksimum antara semua atribut pada vektor

$$d(i.j) = \lim_{p \to \infty} \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p} = \max_{1 \le f \le l} |x_{if} - x_{jf}|$$

# Rumus jarak untuk data dengan atribut numerik (3)

| point | attribute 1 | attribute 2 |
|---|---|---|
| x1 | 1 | 2 |
| x2 | 3 | 5 |
| x3 | 2 | 0 |
| x4 | 4 | 5 |




## Manhattan ($L_1$)

| L | x1 | x2 | x3 | x4 |
|---|---|---|---|---|
| x1 | 0 | | | |
| x2 | 5 | 0 | | |
| x3 | 3 | 6 | 0 | |
| x4 | 6 | 1 | 7 | 0 |

## Euclidean ($L_2$)

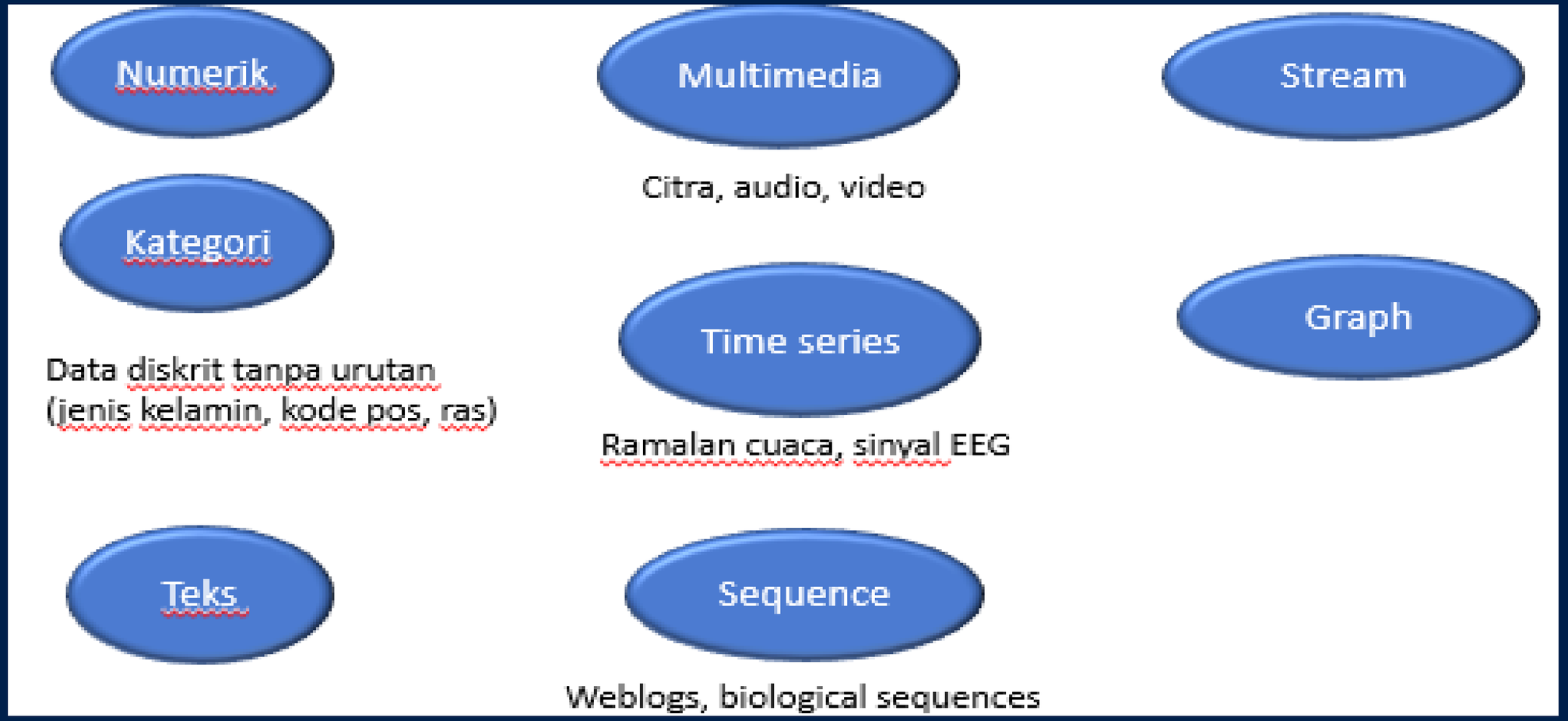
| L2 | x1 | x2 | x3 | x4 |
|---|---|---|---|---|
| x1 | 0 | | | |
| x2 | 3.61 | 0 | | |
| x3 | 2.24 | 5.1 | 0 | |
| x4 | 4.24 | 1 | 5.39 | 0 |

## Supremum ($L_\infty$)

| $L_\infty$ | x1 | x2 | x3 | x4 |
|---|---|---|---|---|
| x1 | 0 | | | |
| x2 | 3 | 0 | | |
| x3 | 2 | 5 | 0 | |
| x4 | 3 | 1 | 5 | 0 |

# Beberapa tipe metodologi *clustering*

# Clustering pada berbagi tipe data



# Association Rules
# (*Kaidah Asosiasi*)

# Association Rules

- **Association Rule Mining disebut juga Frequent Itemset Mining**
- **Adalah proses mendeteksi kumpulan atribut-atribut yang muncul bersamaan (co-occur) dalam frekuensi yang sering (itemset), dan membentuk sejumlah kaidah(rules) dari kumpulan-kumpulan tersebut.**
- **Karena aplikasi association rules yang sangat luas untuk menganalisa keranjang belanja di pasar swalayan, association rules sering juga disebut sebagai *market basket analysis.***
- **Tujuan dari algoritma association rules adalah untuk menghasilkan algoritma yang efisien untuk analisa pola frekuensi tinggi (frequent pattern mining)**
- **Contoh : 90% orang yang berbelanja di suatu supermarket yang membeli roti juga membeli selai, dan 60% dari semua orang yang berbelanja membeli keduanya.**
- **Contoh 2 *: "70% dari orang-orang yang membeli mie, juice dan saus akan membeli* juga roti tawar".**

# Definisi Association Rules (dari bbrp pakar)

- *Association rule mining  adalah analisa* dari kebiasaan belanja konsumen dengan mencari asosiasi dan korelasi antara item-item berbeda yang diletakkan konsumen dalam keranjang belanjaannya(Yang, 2003)
- Dengan kemajuan teknologi, data penjualan dapat disimpan dalam jumlah besar yang disebut dengan "*basket data.*"
- *Aturan asosiasi yang didefinisikan* pada basket data tersebut, dapat digunakan untuk menganalisa data dalam rangka :
  - keperluan desain katalog promosi,
  - proses pembuatan keputusan bisnis,
    - Kampanye pemasaran dengan diskon atau potongan harga
    - Pengaturan tata letak display barang, (barang A didekatkan dengan barang B)
  - segmentasi konsumen dan
  - target pemasaran.

# Contoh aplikasi kaidah asosiasi

- **Marketing and Sales Promotion**
  - **Misal :**
    - **Ketergantungan {bagels, … } → {Potato Chips}**
    - **Potato Chips sebagai consequent → dapat digunakan untuk menentukan apa yang dilakukan untuk meningkatkan penjualan**
    - **Bagels in the antecedent → dapat digunakan untuk melihat produk mana yang akan terkena dampak jika toko tersebut tidak lagi menjual bagels.**
    - **Bagels in antecedent and Potato chips in consequent → Dapat digunakan untuk melihat produk apa yang harus dijual dengan bagels untuk mempromosikan penjualan potato chips.**

# Contoh aplikasi kaidah asosiasi

- **Supermarket Shelf Management**
  - Tujuan untuk mengenali item2 yang dibeli bersama-sama(dalam sekali transaksi) oleh beberapa pelanggan.
  - Pendekatan : memproses data point of sale dengan pemindai barcode untuk dicari ketergantungan antar item.
  - Implementasi real pada promosi di supermarket atau swalayan, akan jamak kita jumpai pembelian 6 pack keju cheedar yang dibundling dengan 1 pack roti tawar.
  - Atau kita jumpai, penataan pampers yang berdekatan dengan tissue,

- **Inventory Management**
  - Tujuan : seorang pelanggan perusahaan perbaikan peralatan mengharapkan keaslian dari perbaikan produk konsumen dan menjaga pelayanan dengan menggunakan suku cadang yang baik untuk mengurangi jumlah kunjungan ke rumah pelanggan.
  - Pendekatan : memproses data peralatan dan suku cadang yang dibutuhkan pada perbaikan sebelumnya di tempat pelanggan yang berbeda dan menemukan pola kejadian yang berulang.

# Association Rules Pattern

- **Bentuk dari aturan asosiasi umumnya dinyatakan dalam bentuk :**

**X => Y , dimana X dan Y adalah itemset**
**Contoh : {roti, mentega} → {susu} (support = 40%, confidence = 50%)**

- **Aturan tersebut berarti "50%" dari transaksi di database yang memuat item roti dan mentega juga memuat item susu. Sedangkan 40% dari seluruh transaksi yang ada di database memuat ketiga item itu.**
- **Dapat juga diartikan : seorang konsumen yang membeli roti dan mentega punya kemungkinan 50% untuk juga membeli susu. Aturan ini cukup signifikan karena mewakili 40% dari catatan transasksi selama ini.**

# Association Rule Mining

- **Jika terdapat sebuah himpunan transaksi T, maka tujuan dari association rules mining adalah untuk menemukan semua aturan yang mempunyai support >= *minsup* dan confidence >= *minconf***

# SUPPORT

$$Support\,(A) = \frac{Jumlah\ transaksi\ mengandung\ A}{Total\ Transaksi}$$

$$Support\,(A,B) = P(A \cap B) = \frac{Jumlah\ transaksi\ mengandung\ A\ dan\ B}{Total\ Transaksi}$$

# CONFIDENCE ATURAN A→B

$$Confidence = P(B \mid A) = \frac{Jumlah\ transaksi\ mengandung\ A\ dan\ B}{Jumlah\ Transaksi\ mengandung\ A}$$

# Tabel Representasi Biner

**Tabel pembelian**

| Id_Trans | Id_Cust | Tanggal | Item | Jumlah |
|---|---|---|---|---|
| 111 | 201 | 5/1/2007 | Pena | 2 |
| 111 | 201 | 5/1/2007 | Tinta | 1 |
| 111 | 201 | 5/1/2007 | Susu | 3 |
| 111 | 201 | 5/1/2007 | Jus | 6 |
| 112 | 105 | 6/3/2007 | Pena | 1 |
| 112 | 105 | 6/3/2007 | Tinta | 1 |
| 112 | 105 | 6/3/2007 | Susu | 1 |
| 113 | 106 | 5/10/2007 | Pena | 1 |
| 113 | 106 | 5/10/2007 | Susu | 1 |
| 114 | 201 | 6/1/2007 | Pena | 2 |
| 114 | 201 | 6/1/2007 | Tinta | 2 |
| 114 | 201 | 6/1/2007 | Jus | 4 |
| 114 | 201 | 6/1/2007 | Air | 1 |

**Tabel representasi biner**

| Id_Trans | Pena | Tinta | Susu | Jus | Air |
|---|---|---|---|---|---|
| 111 | 1 | 1 | 1 | 1 | 0 |
| 112 | 1 | 1 | 1 | 0 | 0 |
| 113 | 1 | 0 | 1 | 0 | 0 |
| 114 | 1 | 1 | 0 | 1 | 1 |

Jika diamati ada redundancy pada table pembelian.

Pembuatan table ‘denormalized’ untuk mempermudah data mining dilakukan pada tahap data preparation pada CRISP-DM.

# Support - Minimum Support - Frequent itemset

**Catatan : Itemset dapat berisi hanya satu item.**

| Id_Trans | Pena | Tinta | Susu | Jus | Air |
|---|---|---|---|---|---|
| 111 | 1 | 1 | 1 | 1 | 0 |
| 112 | 1 | 1 | 1 | 0 | 0 |
| 113 | 1 | 0 | 1 | 0 | 0 |
| 114 | 1 | 1 | 0 | 1 | 1 |

- **Pada table diatas, dapat kita hitung :**
- **Support count (·), merupakan jumlah transaksi yang berisi suatu itemset tertentu atau merupakan frekuensi kejadian dari suatu itemset.**
- **Support dari suatu item adalah perbandingan dari transaksi dalam basisdata yang berisi semua item dalam itemset.**
- **Dalam contoh diatas, itemset {pena,tinta} memiliki support 75% dalam table pembelian. Itemset {susu,jus} supportnya hanya 25%**
- **Frequent itemset menunjukkan itemset yang memiliki frekuensi kemunculan lebih dari nilai minimum yang telah ditentukan (·)**
- **Jika missal, ditentukan minimum support adalah 70%, maka *frequent-itemset* pada contoh diatas adalah {pena}, {tinta}, {susu}, {pena,tinta}, dan {pena,susu}**

# Support and Confidence

- **Support (*s*) dan Confidence (*c*) merupakan metrik yang digunakan pada Association Rule.**
- **Support menunjukkan persentasi jumlah transaksi yang berisi X dan Y.**
- **Sedangkan Confidence menunjukkan persentasi banyaknya Y pada transaksi yang mengandung X.**
- **Bentuk persamaan matematisnya dapat dituliskan seperti ini:**
- **Contoh : {Milk, Diaper} => {Beer}**

$$support(X \Rightarrow Y) = P(X \cup Y)$$

$$confidence(X \Rightarrow Y) = P(Y \mid X)$$

$$support(\{\text{Milk,Diaper}\} \Rightarrow \{\text{Beer}\}) = \frac{2}{5} = 0.4 = 40\%$$

$$confidence(\{\text{Milk,Diaper}\} \Rightarrow \{\text{Beer}\}) = \frac{2}{3} = 0.667 = 66.7\%$$

# Support and Confidence

- Confidence menyatakan seberapa sering item-item dalam Y muncul dalam transaksi yang berisi X
- S (X→Y) = 𝜎 (X U Y) / N , Dimana S = Support, N = total transaksi
- C (X→Y) = 𝜎 (X U Y), / 𝜎(X) Dimana c = confidence

| Id_Trans | Pena | Tinta | Susu | Jus | Air |
|---|---|---|---|---|---|
| 111 | 1 | 1 | 1 | 1 | 0 |
| 112 | 1 | 1 | 1 | 0 | 0 |
| 113 | 1 | 0 | 1 | 0 | 0 |
| 114 | 1 | 1 | 0 | 1 | 1 |

{Pena,Tinta}→Jus

Contoh : S (X→Y) = 𝜎 ({Pena, Tinta, Jus})/4 = 2/4 = 0.5

C (X→Y) = 𝜎 ({Pena, Tinta, Jus})/𝜎({Pena, Tinta}) = 2/3 = 0.67

# {milk,cheese}→pocari , tentukan s dan c nya

| Id_Trans | Pena | Tinta | Susu | Jus | Air |
|---|---|---|---|---|---|
| 111 | 1 | 1 | 1 | 1 | 0 |
| 112 | 1 | 1 | 1 | 0 | 0 |
| 113 | 1 | 0 | 1 | 0 | 0 |
| 114 | 1 | 1 | 0 | 1 | 1 |

| Id trans | Items |
|---|---|
| 1 | Bread, milk |
| 2 | Bread, cheese, pocari, eggs |
| 3 | Milk,cheese,pocari,coke |
| 4 | Bread,milk,cheese,pocari |
| 5 | Bread,milk,cheese,coke |

**{Pena,Tinta}→Jus**
**Contoh :**
**S (X→Y) = · ({Pena, Tinta, Jus})/4 = 2/4 = 0.5**
**C (X→Y) = · ({Pena, Tinta, Jus})/·({Pena, Tinta}) = 2/3 = 0.67**

# Algorithma Association Rules

- Algoritma A priori termasuk dalam association rules.
- Algoritma lainnya yang termasuk kedalam association rules diantaranya :
  - FP-Growth
  - Generalized Rule Induction
  - Hash Based algorithm

# Learning Objective 2

Fill in .....

# CONCLUSION

Fill in .....

# REFERENCES

Fill in IEEE Style

Institut Informatika & Bisnis DARMAJAYA
Yayasan Alfian Husin

SAINS DATA
Institut Informatika & Bisnis Darmajaya

Kampus Merdeka INDONESIA JAYA

MERDEKA BELAJAR

# THANK YOU!!

DATA SCIENCE DARMAJAYA “YOUR BEST FUTURE IN DATA”