

# Pengantar *Data Mining* #1: Apa itu *Data Mining*?

Isnan Mulia, S.Komp, M.Kom

# Ragam Data Di Sekitar Kita

- Banyak jenis data yang dihasilkan dari berbagai kegiatan manusia
- Perkembangan teknologi informasi menyebabkan munculnya berbagai variasi data
- Contoh:
  - Data penjualan *minimarket* M dalam 1 bulan terakhir
  - Data pengukuran daun pohon durian di kebun X
  - Data pelanggan bengkel mobil B pada bulan Januari 2022
  - Data produksi ponsel merk A dalam rentang waktu tertentu
  - Data pemesanan ojek *online*
  - Data riwayat pembelian barang di *marketplace* Z

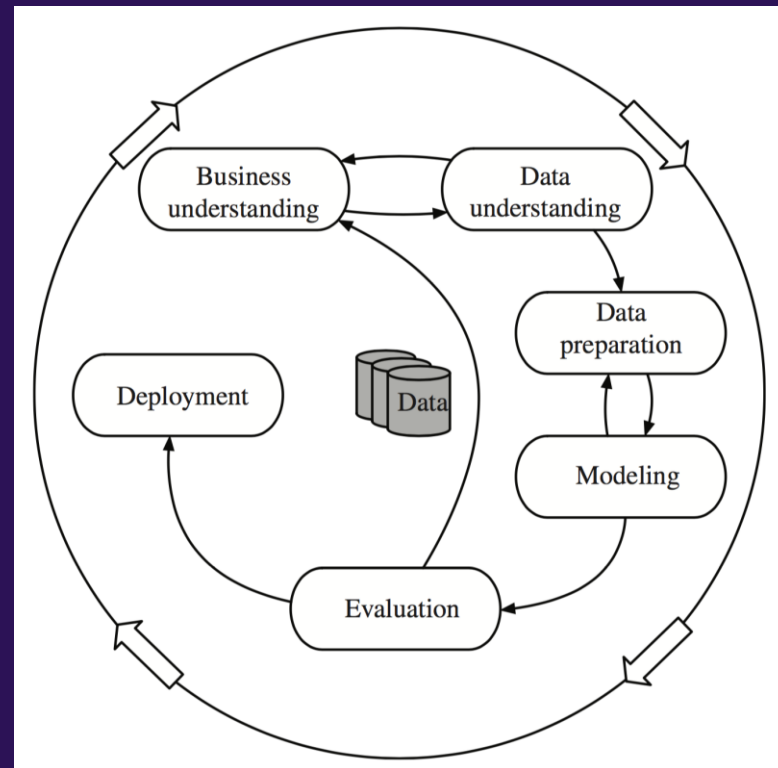
## Ragam Data Di Sekitar Kita

- Pertanyaan yang mungkin muncul:
  - Informasi apa yang dapat diambil dari data ini?
  - Apakah pelanggan yang membeli tepung terigu selalu membeli mentega secara bersamaan?
  - Apakah varietas durian dapat dibedakan dari ukuran daunnya?
  - Bagaimana proporsi profil pengguna ojek *online* di siang hari?
  - Berapa GB rata-rata penggunaan paket data oleh pelanggan "*silver*"?
  - Berapa banyak pelanggan yang memeriksakan kendaraannya tepat waktu?
  - Kelompok barang apa saja yang dapat disarankan kepada seorang pengguna di *marketplace*, berdasarkan riwayat pencariannya?
- Kita dapat menjawab pertanyaan tersebut menggunakan pendekatan *Data Mining*

# Definisi *Data Mining*

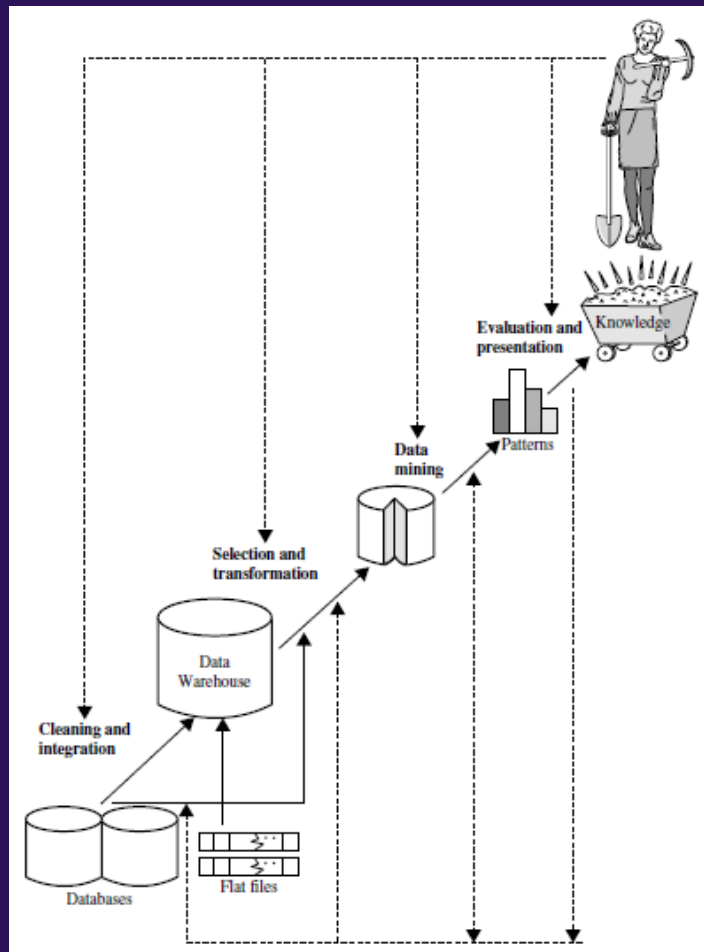
- Proses penggalian informasi dan pola yang menarik dari data yang sangat besar
- Tujuan: mengekstrak pengetahuan dari kumpulan data sehingga didapatkan struktur dan informasi yang dimengerti manusia
- Nama lain:
  - *Knowledge Discovery from Data* (KDD)
  - *Knowledge Extraction*

# Tahapan *Data Mining*



Sumber: Witten I et al. 2019. *Data Mining: Practical Machine Learning Tools and Techniques*. 4<sup>th</sup> edition. Massachusetts: Morgan Kaufmann.

## Tahapan Data Mining

























Sumber: Han J *et al.* 2012. *Data Mining: Concepts and Techniques*. 3<sup>rd</sup> edition. Massachusetts: Morgan Kaufmann.

Secara garis besar:

- Memahami tujuan yang ingin dicapai
  - ➔ Informasi apa yang ingin dicari/dihasilkan
- Memahami data yang digunakan
  - ➔ Bagaimana kelengkapan & kualitas data yang akan digunakan
- Praproses data
  - ➔ Bagaimana cara menyiapkan data agar bisa diproses lebih lanjut
- Pemodelan *data mining*
  - ➔ Bagaimana model *data mining* yang diinginkan
- Evaluasi hasil *data mining*
  - ➔ Apakah model *data mining* yang dihasilkan sudah cukup memuaskan
- *Deployment* (menerapkan model *data mining*)
  - ➔ Bagaimana menerapkan model *data mining* untuk data lain yang serupa

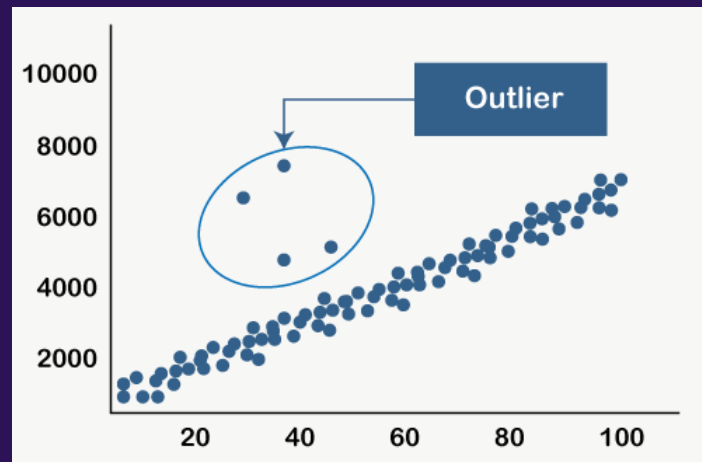
# Tugas *Data Mining*

- Pembelajaran aturan asosiasi (*association rule learning*)

|               |   |
|---------------|---|
| Transaction 1 |     |
| Transaction 2 |      |
| Transaction 3 |     |
| Transaction 4 |     |
| Transaction 5 |     |
| Transaction 6 |      |
| Transaction 7 |     |
| Transaction 8 |     |

Sumber:  
<https://annalysin.files.wordpress.com/2016/04/association-rule-support-table.png>

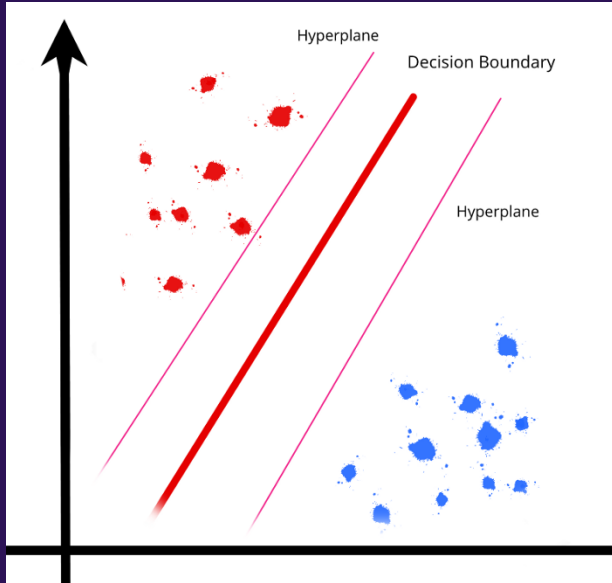
- Pendeteksian anomali/ analisis pencilan (*outlier*)



Sumber: <https://www.tutorialandexample.com/wp-content/uploads/2021/02/Outlier-Analysis-2.png>



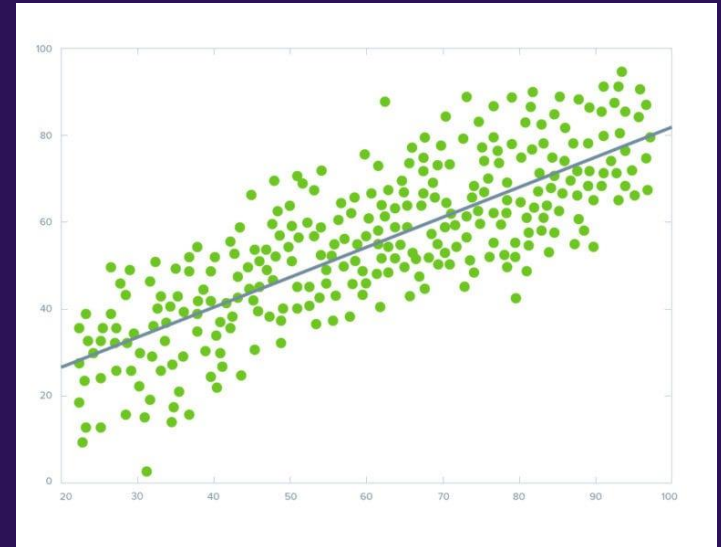
- Klasifikasi & regresi



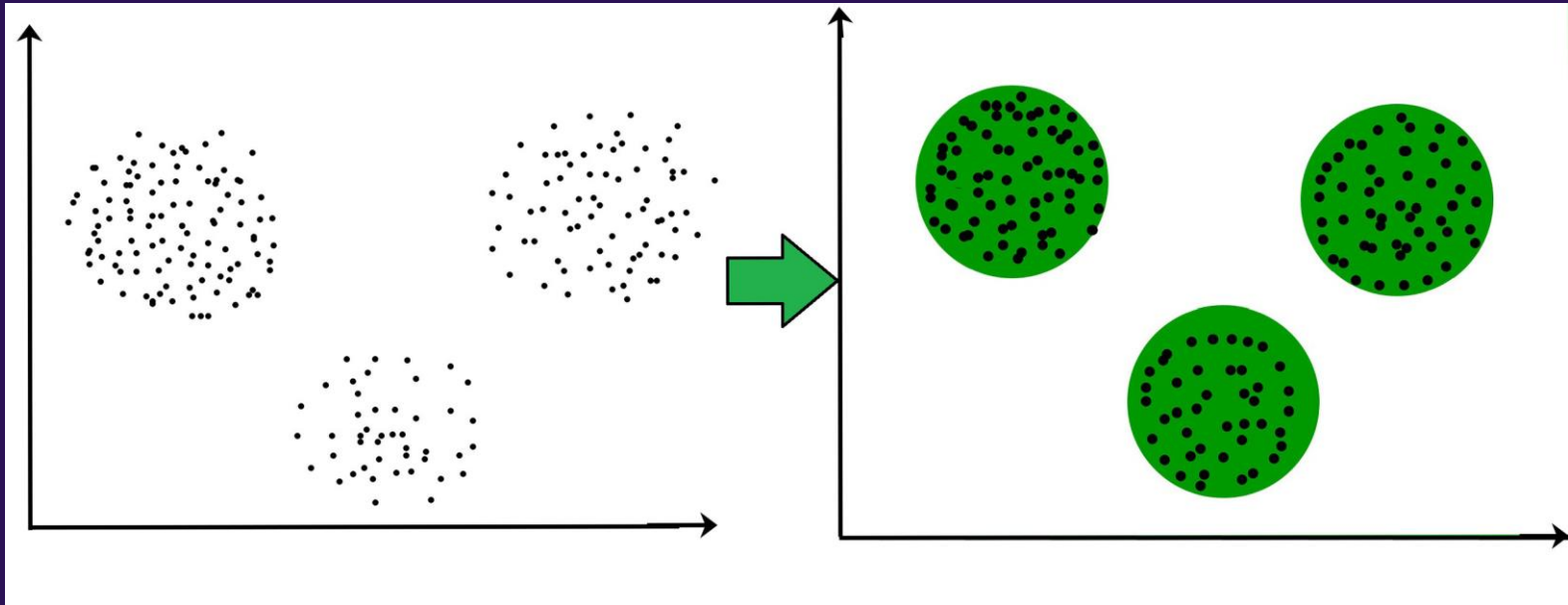
Sumber:

[https://miro.medium.com/max/700/1\\*Jsdi4EdZDDozpp6ZplzOIg.png](https://miro.medium.com/max/700/1*Jsdi4EdZDDozpp6ZplzOIg.png)

Sumber: <https://149695847.v2.pressablecdn.com/wp-content/uploads/2017/09/graphic2-1x-769x600.jpg>

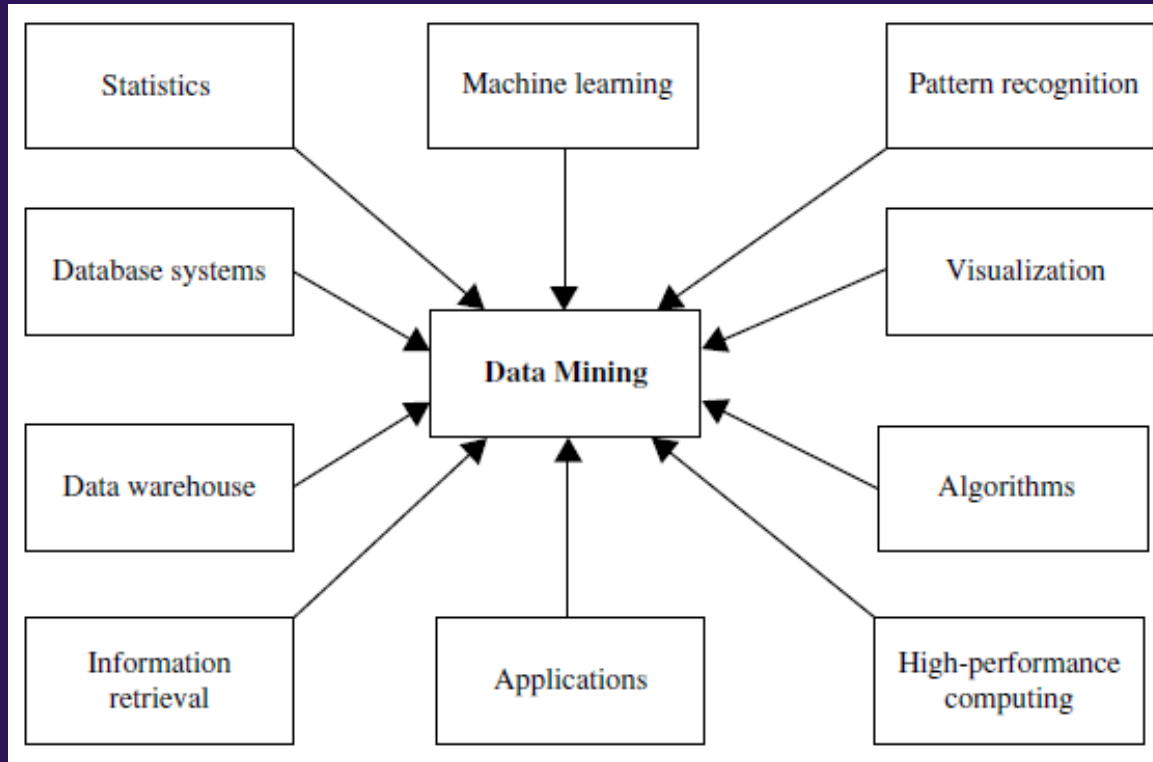


- Analisis *cluster*



Sumber: <https://media.geeksforgeeks.org/wp-content/uploads/merge3cluster.jpg>

# Teknologi yang Digunakan



Sumber: Han J *et al.* 2012. *Data Mining: Concepts and Techniques*. 3<sup>rd</sup> edition. Massachusetts: Morgan Kaufmann.

# Variasi *Data Mining*

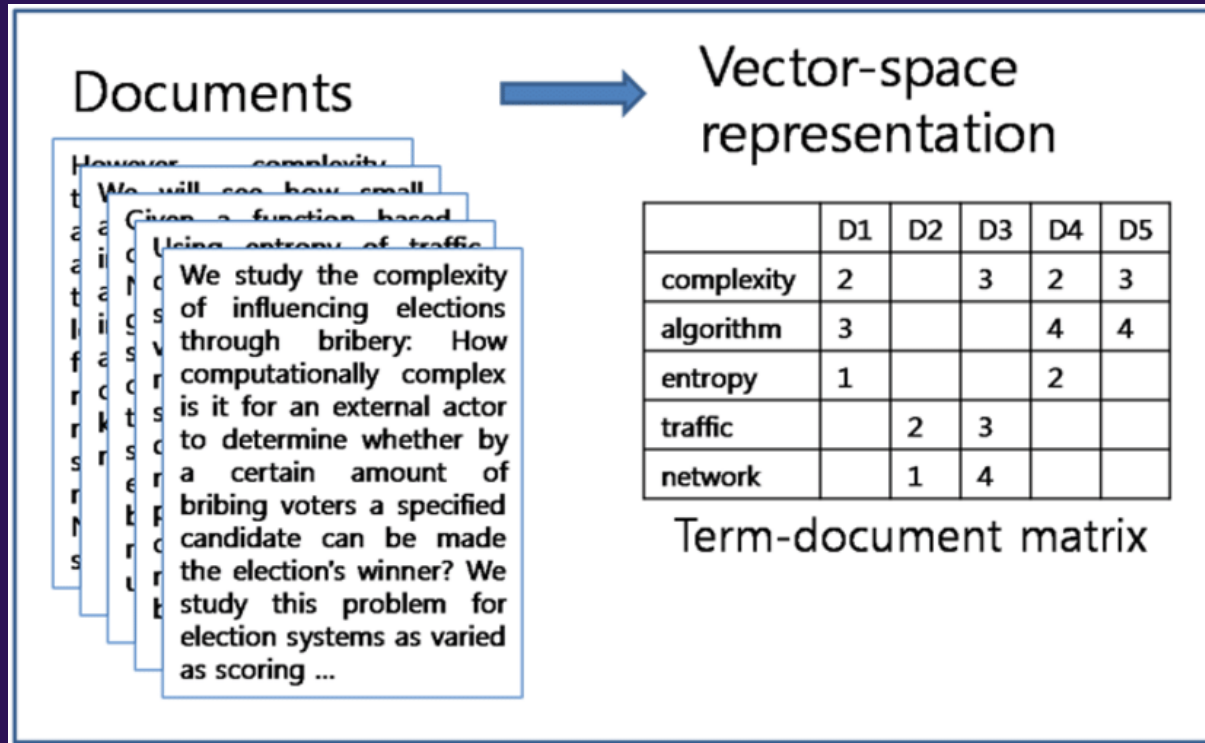
- *Data mining* (secara umum) → menggunakan sumber data numerik/kategorik yang terstruktur

| case ID |             | predictors |            |     | target        |
|---------|-------------|------------|------------|-----|---------------|
| CUST_ID | CUST_GENDER | EDUCATION  | OCCUPATION | AGE | AFFINITY_CARD |
| 101501  | F           | Masters    | Prof.      | 41  | 0             |
| 101502  | M           | Bach.      | Sales      | 27  | 0             |
| 101503  | F           | HS-grad    | Cleric.    | 20  | 0             |
| 101504  | M           | Bach.      | Exec.      | 45  | 1             |
| 101505  | M           | Masters    | Sales      | 34  | 1             |
| 101506  | M           | HS-grad    | Other      | 38  | 0             |
| 101507  | M           | < Bach.    | Sales      | 28  | 0             |
| 101508  | M           | HS-grad    | Sales      | 19  | 0             |
| 101509  | M           | Bach.      | Other      | 52  | 0             |
| 101510  | M           | Bach.      | Sales      | 27  | 1             |

Sumber:

[https://docs.oracle.com/cd/E18283\\_01/datamine.112/e16808/img/class\\_sampledata.gif](https://docs.oracle.com/cd/E18283_01/datamine.112/e16808/img/class_sampledata.gif)

- *Text mining* → menggunakan sumber data teks yang tidak terstruktur



Sumber:

<https://www.researchgate.net/profile/Ibtehal-Baazeem/publication/312471174/figure/fig1/AS:451427626688517@1484640141976/Figure4DocumentrepresentationintheVectorSpaceModel22.png>

- *Spatial data mining* → menggunakan sumber data spasial (keruangan), biasanya hasil pencitraan satelit



Sumber: <https://eos.com/wp-content/uploads/2019/04/Main.jpg.webp>

# Penerapan *Data Mining*

- Penyusunan barang-barang di rak, di pasar swalayan/minimarket
- Sistem rekomendasi di *marketplace* berdasarkan riwayat pencarian
- Pendeteksian kecurangan (*fraud detection*) yang dilakukan oleh pelanggan
- *Sentiment analysis* pada *tweet* warganet mengenai suatu topik yang viral → “drone emprit” (DE)
- Menganalisis sekuens DNA
- Menganalisis perubahan penggunaan lahan pada suatu daerah dari waktu ke waktu

# Data Mining vs Data Science

| S.No. | Data Science   | Data Mining  |
|-------|--|--|
| 1     | Data Science is an area.   | Data Mining is a technique.  |
| 2     | It is about collection, processing, analyzing and utilizing of data into various operations. It is more conceptual.                        | It is about extracting the vital and valuable information from the data.                         |
| 3     | It is a field of study just like the Computer Science, Applied Statistics or Applied Mathematics.  | It is a technique which is a part of the Knowledge Discovery in Data Base processes (KDD).       |
| 4     | The goal is to build data-dominant products for a venture.   | The goal is to make data more vital and usable i.e. by extracting only important information.    |
| 5     | It deals with the all types of data i.e. structured, unstructured or semi-structured.  | It mainly deals with the structured forms of the data.   |
| 6     | It is a super set of Data Mining as data science consists of Data scrapping, cleaning, visualization, statistics and many more techniques. | It is a sub set of Data Science as mining activities which is in a pipeline of the Data science. |
| 7     | It is mainly used for scientific purposes.   | It is mainly used for business purposes.   |
| 8     | It broadly focuses on the science of the data.   | It is more involved with the processes.  |

Sumber: <https://www.geeksforgeeks.org/difference-between-data-science-and-data-mining/>



## Recap

- Definisi *data mining*
- Tahapan *data mining*
- Variasi *data mining*
- Penerapan *data mining*
- *Data mining vs data science*

Next: apa saja jenis atribut data?

つづく