Penggalian data

Dari Wikipedia bahasa Indonesia, ensiklopedia bebas

[Loncat ke navigasi](https://id.wikipedia.org/wiki/Penggalian_data#mw-head)[Loncat ke pencarian](https://id.wikipedia.org/wiki/Penggalian_data#searchInput)

**Penggalian data** (bahasa Inggris: *data mining*) adalah ekstraksi [pola](https://id.wikipedia.org/wiki/Pola) yang menarik dari [data](https://id.wikipedia.org/wiki/Data) dalam jumlah besar [[1]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-1). Suatu pola dikatakan menarik apabila pola tersebut tidak sepele, implisit, tidak diketahui sebelumnya, dan berguna. Pola yang disajikan haruslah mudah dipahami, berlaku untuk data yang akan diprediksi dengan derajat kepastian tertentu, berguna, dan baru. Penggalian data memiliki beberapa nama alternatif, meskipun definisi eksaknya berbeda, seperti KDD (*knowledge discovery in database*), analisis pola, arkeologi data, pemanenan informasi, dan intelegensia bisnis. Penggalian data diperlukan saat data yang tersedia terlalu banyak (misalnya data yang diperoleh dari [sistem basis data](https://id.wikipedia.org/w/index.php?title=Sistem_basis_data&action=edit&redlink=1) perusahaan, [e-commerce](https://id.wikipedia.org/wiki/E-commerce), data [saham](https://id.wikipedia.org/wiki/Saham), data [sensus](https://id.wikipedia.org/wiki/Sensus) dan data [bioinformatika](https://id.wikipedia.org/wiki/Bioinformatika)), tetapi tidak tahu pola apa yang bisa didapatkan.

## **Proses Pencarian Pola**

Penggalian data adalah salah satu bagian dari proses pencarian pola. Berikut ini urutan proses pencarian pola:

1. [Pembersihan Data](https://id.wikipedia.org/w/index.php?title=Pembersihan_Data&action=edit&redlink=1): yaitu menghapus data pengganggu (*noise*) dan mengisi data yang hilang.
2. [Integrasi Data](https://id.wikipedia.org/w/index.php?title=Integrasi_Data&action=edit&redlink=1): yaitu menggabungkan berbagai sumber data.
3. [Pemilihan Data](https://id.wikipedia.org/w/index.php?title=Pemilihan_Data&action=edit&redlink=1): yaitu memilih data yang relevan.
4. [Transformasi Data](https://id.wikipedia.org/w/index.php?title=Transformasi_Data&action=edit&redlink=1): yaitu mentransformasi data ke dalam format untuk diproses dalam penggalian data.
5. Penggalian Data: yaitu menerapkan metode cerdas untuk ekstraksi pola.
6. [Evaluasi pola](https://id.wikipedia.org/w/index.php?title=Evaluasi_pola&action=edit&redlink=1): yaitu mengenali pola-pola yang menarik saja.
7. [Penyajian pola](https://id.wikipedia.org/w/index.php?title=Penyajian_pola&action=edit&redlink=1): yaitu memvisualisasi pola ke pengguna.

## **Latar belakang**

Perkembangan yang pesat di bidang pengumpulan data dan teknologi penyimpanan di berbagai bidang, menghasilkan [basis data yang terlampau besar](https://id.wikipedia.org/w/index.php?title=VLDB&action=edit&redlink=1). Namun, data yang dikumpulkan jarang dilihat lagi, karena terlalu panjang, membosankan, dan tidak menarik. Seringkali, keputusan -yang katanya berdasarkan data- dibuat tidak lagi berdasarkan data, melainkan dari intuisi para pembuat keputusan. Sehingga, lahirlah cabang ilmu penggalian data ini.

Analisis data tanpa menggunakan otomasi dari penggalian data adalah tidak memungkinkan lagi, kalau 1) data terlalu banyak, 2) [dimensionalitas data](https://id.wikipedia.org/w/index.php?title=Dimensionalitas_data&action=edit&redlink=1) terlalu besar, 3) data terlalu kompleks untuk dianalisis manual (misalnya: [data time series](https://id.wikipedia.org/w/index.php?title=Time_series&action=edit&redlink=1), [data spatiotemporal](https://id.wikipedia.org/w/index.php?title=Spatiotemporal&action=edit&redlink=1), [data multimedia](https://id.wikipedia.org/wiki/Multimedia), [data streams](https://id.wikipedia.org/w/index.php?title=Data_streams&action=edit&redlink=1)).

## **Teknik Penggalian Data**

Pada dasarnya penggalian data dibedakan menjadi dua fungsionalitas, yaitu deskripsi dan prediksi. Berikut ini beberapa fungsionalitas penggalian data yang sering digunakan:

* [Karakterisasi dan Diskriminasi](https://id.wikipedia.org/w/index.php?title=Karakterisasi_dan_Diskriminasi&action=edit&redlink=1): yaitu menggeneralisasi, merangkum, dan mengkontraskan karakteristik data.
* [Penggalian pola berulang](https://id.wikipedia.org/w/index.php?title=Penggalian_pola_berulang&action=edit&redlink=1): yaitu pencarian pola asosiasi (*association rule*) atau pola intra-transaksi, atau pola pembelian yang terjadi dalam satu kali transaksi.
* [Klasifikasi](https://id.wikipedia.org/wiki/Klasifikasi): yaitu membangun suatu model yang bisa mengklasifikasikan suatu objek berdasar atribut-atributnya. Kelas target sudah tersedia dalam data sebelumnya, sehingga fokusnya adalah bagaimana mempelajari data yang ada agar klasifikator bisa mengklasifikasikan sendiri.
* [Prediksi](https://id.wikipedia.org/w/index.php?title=Prediksi&action=edit&redlink=1): yaitu memprediksi nilai yang tidak diketahui atau nilai yang hilang, menggunakan model dari [klasifikasi](https://id.wikipedia.org/wiki/Klasifikasi).
* [Penggugusan/Cluster analysis](https://id.wikipedia.org/w/index.php?title=Penggugusan/Cluster_analysis&action=edit&redlink=1): yaitu mengelompokkan sekumpulan objek data berdasarkan kemiripannya. Kelas target tidak tersedia dalam data sebelumnya, sehingga fokusnya adalah memaksimalkan kemiripan intrakelas dan meminimalkan kemiripan antarkelas.
* [Analisis outlier](https://id.wikipedia.org/w/index.php?title=Analisis_outlier&action=edit&redlink=1): yaitu proses pengenalan data yang tidak sesuai dengan perilaku umum dari data lainnya. Contoh: mengenali *noise* dan pengecualian dalam data.
* [Analisis trend dan evolusi](https://id.wikipedia.org/w/index.php?title=Analisis_trend_dan_evolusi&action=edit&redlink=1): meliputi analisis regresi, penggalian pola sekuensial, analisis periodisitas, dan analisis berbasis kemiripan.

Berikut ini adalah 10 algoritme penggalian data yang paling populer berdasarkan konferensi ICDM '06, semua algoritme dinominasikan oleh para pemenang ACM KDD Innovation Award dan IEEE ICDM Research Contributions Award [[2]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-2):

1. [C4.5](https://id.wikipedia.org/w/index.php?title=C4.5&action=edit&redlink=1) (61 suara) [[3]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-3)
2. [k-Means](https://id.wikipedia.org/w/index.php?title=K-Means&action=edit&redlink=1) (60 suara):[[4]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-4)
3. [SVM](https://id.wikipedia.org/w/index.php?title=SVM&action=edit&redlink=1) atau *Support Vector Machine* (58 suara):[[5]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-5)
4. [Apriori](https://id.wikipedia.org/wiki/Apriori) (52 suara):[[6]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-6)
5. [EM](https://id.wikipedia.org/w/index.php?title=EM&action=edit&redlink=1) (48 suara):[[7]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-7)
6. [PageRank](https://id.wikipedia.org/wiki/PageRank) (46 suara):[[8]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-8)
7. [AdaBoost](https://id.wikipedia.org/w/index.php?title=AdaBoost&action=edit&redlink=1) (45 suara):[[9]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-9)
8. [kNN](https://id.wikipedia.org/wiki/KNN) (45 suara):[[10]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-10)
9. [Naive Bayes](https://id.wikipedia.org/w/index.php?title=Naive_Bayes&action=edit&redlink=1) (34 suara):[[11]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-11)

Berikut ini adalah yang hanya masuk nominasi:

* [CART](https://id.wikipedia.org/wiki/CART):[[12]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-12)
* [FP-Tree](https://id.wikipedia.org/w/index.php?title=FP-Tree&action=edit&redlink=1):[[13]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-13)
* [HITS](https://id.wikipedia.org/w/index.php?title=HITS&action=edit&redlink=1):[[14]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-14)
* [BIRCH](https://id.wikipedia.org/w/index.php?title=BIRCH&action=edit&redlink=1):[[15]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-15)
* [GSP](https://id.wikipedia.org/w/index.php?title=GSP&action=edit&redlink=1):[[16]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-16)
* [PrefixSpan](https://id.wikipedia.org/w/index.php?title=PrefixSpan&action=edit&redlink=1):[[17]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-17)
* [CBA](https://id.wikipedia.org/wiki/CBA):[[18]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-18)
* [Finding Reduct](https://id.wikipedia.org/w/index.php?title=Finding_Reduct&action=edit&redlink=1):[[19]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-19)
* [gSpan](https://id.wikipedia.org/w/index.php?title=GSpan&action=edit&redlink=1):[[20]](https://id.wikipedia.org/wiki/Penggalian_data#cite_note-20)

# Data Mining

By [ALEXANDRA TWIN](https://www.investopedia.com/contributors/54405/)

 Reviewed By [AMY DRURY](https://www.investopedia.com/amy-drury-4776200)

 Updated Sep 13, 2020

## **What Is Data Mining?**

Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs. Data mining depends on [effective data collection](https://www.investopedia.com/terms/b/big-data.asp), [warehousing](https://www.investopedia.com/terms/w/warehousing.asp), and computer processing.

Data mining processes are used to build machine learning models that power applications including search engine technology and website recommendation programs.

## **How Data Mining Works**

[Data mining](https://www.investopedia.com/articles/basics/03/053003.asp) involves exploring and analyzing large blocks of information to glean meaningful patterns and trends. It can be used in a variety of ways, such as database marketing, credit risk management, [fraud detection](https://www.investopedia.com/financial-edge/0512/the-most-common-types-of-consumer-fraud.aspx), spam Email filtering, or even to discern the sentiment or opinion of users.

The data mining process breaks down into five steps. First, organizations collect data and load it into their data warehouses. Next, they store and manage the data, either on in-house servers or the cloud. Business analysts, management teams and information technology professionals access the data and determine how they want to organize it. Then, application software sorts the data based on the user's results, and finally, the end-user presents the data in an easy-to-share format, such as a graph or table.

## **Data Warehousing and Mining Software**

Data mining programs analyze relationships and patterns in data based on what users request. For example, a company can use data mining software to create classes of information. To illustrate, imagine a restaurant wants to use data mining to determine when it should offer certain specials. It looks at the information it has collected and creates classes based on when customers visit and what they order.

In other cases, data miners find clusters of information based on logical relationships or look at associations and sequential patterns to draw conclusions about trends in consumer behavior.

Warehousing is an important aspect of data mining. Warehousing is when companies centralize their data into one database or program. With a data warehouse, an organization may spin off segments of the data for specific users to analyze and use.

However, in other cases, analysts may start with the data they want and create a [data warehouse](https://www.investopedia.com/terms/d/data-warehousing.asp) based on those specs. Regardless of how businesses and other entities organize their data, they use it to support management's decision-making processes.

## **Example of Data Mining**

Grocery stores are well-known users of data mining techniques. Many supermarkets offer free [loyalty cards](https://www.investopedia.com/terms/l/loyalty-program.asp) to customers that give them access to reduced prices not available to non-members. The cards make it easy for stores to track who is buying what, when they are buying it and at what price. After analyzing the data, stores can then use this data to offer customers coupons targeted to their buying habits and decide when to put items on sale or when to sell them at full price.

Data mining can be a cause for concern when a company uses only selected information, which is not representative of the overall sample group, to prove a certain hypothesis.

### **KEY TAKEAWAYS**

* Data mining is the process of analyzing a large batch of information to discern trends and patterns.
* Data mining can be used by corporations for everything from learning about what customers are interested in or want to buy to fraud detection and spam filtering.
* Data mining programs break down patterns and connections in data based on what information users request or provide.