Data Mining

Data mining is to process of analyzing data from different perspectives and summarizing it into useful information, including discovery of previously unknown interesting patterns, unusual records or dependencies.

1. Potential Business Benefits from effective data mining

* Identify previously unseen relationships between business data sets
* Better predict future trends & behaviors
* Extract value(e.g. performance insights) from big data sets
* Generate business actions built on data insight

1. Some examples of how data mining can help a business improve competitiveness

* Sales forecasting analyzing when customers bought to predict when they will buy again
* Database marketing examining customer purchasing patterns and looking at the demographics and psychographics of customers to build predictive profiles.
* Market segmentation a classic use of data mining, using data to break down a market into meaningful segments like age, income, occupation or gender
* E-commerce basket analysis using mined data to predict future customer behavior by past performance, including purchases and preferences.

1. The classic example of data mining

Dunnhumby pioneered data mining to help Tesco better understand its customers.

Dunnhumby launched the Tesco Club card loyalty program. Using data about past customer purchase habits, Tesco was able to stock its stores based on predictions about what customers might want in the future. It was revolutionary for the UK retail market.

1. Example of Data mining

As a result of its data mining, US supermarket giant Walmart discovered that sales of Strawberry Pop-tarts increase by seven times prior to a hurricane.

Since this discovery, Walmart places the Strawberry Pop-tarts at the checkouts prior to a hurricane.

**Data mining**

From Wikipedia, the free encyclopedia

[Jump to navigation](https://en.wikipedia.org/wiki/Data_mining#mw-head) [Jump to search](https://en.wikipedia.org/wiki/Data_mining#p-search)

**Data mining** is the process of discovering patterns in large [data sets](https://en.wikipedia.org/wiki/Data_set) involving methods at the intersection of [machine learning](https://en.wikipedia.org/wiki/Machine_learning), [statistics](https://en.wikipedia.org/wiki/Statistics), and [database systems](https://en.wikipedia.org/wiki/Database_system).[[1]](https://en.wikipedia.org/wiki/Data_mining#cite_note-acm-1) Data mining is an [interdisciplinary](https://en.wikipedia.org/wiki/Interdisciplinary) subfield of [computer science](https://en.wikipedia.org/wiki/Computer_science) and [statistics](https://en.wikipedia.org/wiki/Statistics) with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use.[[1]](https://en.wikipedia.org/wiki/Data_mining#cite_note-acm-1)[[2]](https://en.wikipedia.org/wiki/Data_mining#cite_note-brittanica-2)[[3]](https://en.wikipedia.org/wiki/Data_mining#cite_note-elements-3)[[4]](https://en.wikipedia.org/wiki/Data_mining#cite_note-4) Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.[[5]](https://en.wikipedia.org/wiki/Data_mining#cite_note-Fayyad-5) Aside from the raw analysis step, it also involves database and [data management](https://en.wikipedia.org/wiki/Data_management) aspects, [data pre-processing](https://en.wikipedia.org/wiki/Data_pre-processing), [model](https://en.wikipedia.org/wiki/Statistical_model) and [inference](https://en.wikipedia.org/wiki/Statistical_inference) considerations, interestingness metrics, [complexity](https://en.wikipedia.org/wiki/Computational_complexity_theory) considerations, post-processing of discovered structures, [visualization](https://en.wikipedia.org/wiki/Data_visualization), and [online updating](https://en.wikipedia.org/wiki/Online_algorithm).[[1]](https://en.wikipedia.org/wiki/Data_mining#cite_note-acm-1) The difference between [data analysis](https://en.wikipedia.org/wiki/Data_analysis) and data mining is that data analysis is to summarize the history such as analyzing the effectiveness of a marketing campaign, in contrast, data mining focuses on using specific machine learning and statistical models to predict the future and discover the patterns among data.[[6]](https://en.wikipedia.org/wiki/Data_mining#cite_note-6)

The term "data mining" is in fact a [misnomer](https://en.wikipedia.org/wiki/Misnomer), because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (*mining*) of data itself.[[7]](https://en.wikipedia.org/wiki/Data_mining#cite_note-han-kamber-7) It also is a [buzzword](https://en.wikipedia.org/wiki/Buzzword)[[8]](https://en.wikipedia.org/wiki/Data_mining#cite_note-8) and is frequently applied to any form of large-scale data or [information processing](https://en.wikipedia.org/wiki/Information_processing) ([collection](https://en.wikipedia.org/wiki/Data_collection), [extraction](https://en.wikipedia.org/wiki/Information_extraction), [warehousing](https://en.wikipedia.org/wiki/Data_warehouse), analysis, and statistics) as well as any application of [computer decision support system](https://en.wikipedia.org/wiki/Decision_support_system), including [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence) (e.g., machine learning) and [business intelligence](https://en.wikipedia.org/wiki/Business_intelligence). The book *Data mining: Practical machine learning tools and techniques with Java*[[9]](https://en.wikipedia.org/wiki/Data_mining#cite_note-witten-9) (which covers mostly machine learning material) was originally to be named just *Practical machine learning*, and the term *data mining* was only added for marketing reasons.[[10]](https://en.wikipedia.org/wiki/Data_mining#cite_note-10) Often the more general terms (*large scale*) [*data analysis*](https://en.wikipedia.org/wiki/Data_analysis) and [*analytics*](https://en.wikipedia.org/wiki/Analytics) – or, when referring to actual methods, *artificial intelligence* and *machine learning* – are more appropriate.

The actual data mining task is the semi-automatic or automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records ([cluster analysis](https://en.wikipedia.org/wiki/Cluster_analysis)), unusual records ([anomaly detection](https://en.wikipedia.org/wiki/Anomaly_detection)), and dependencies ([association rule mining](https://en.wikipedia.org/wiki/Association_rule_mining), [sequential pattern mining](https://en.wikipedia.org/wiki/Sequential_pattern_mining)). This usually involves using database techniques such as [spatial indices](https://en.wikipedia.org/wiki/Spatial_index). These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and [predictive analytics](https://en.wikipedia.org/wiki/Predictive_analytics). For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a [decision support system](https://en.wikipedia.org/wiki/Decision_support_system). Neither the data collection, data preparation, nor result interpretation and reporting is part of the data mining step, but do belong to the overall KDD process as additional steps.

The related terms [*data dredging*](https://en.wikipedia.org/wiki/Data_dredging), *data fishing*, and *data snooping* refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.



**Contents**

* [1 Etymology](https://en.wikipedia.org/wiki/Data_mining#Etymology)
* [2 Background](https://en.wikipedia.org/wiki/Data_mining#Background)
* [3 Process](https://en.wikipedia.org/wiki/Data_mining#Process)
  + [3.1 Pre-processing](https://en.wikipedia.org/wiki/Data_mining#Pre-processing)
  + [3.2 Data mining](https://en.wikipedia.org/wiki/Data_mining#Data_mining)
  + [3.3 Results validation](https://en.wikipedia.org/wiki/Data_mining#Results_validation)
* [4 Research](https://en.wikipedia.org/wiki/Data_mining#Research)
* [5 Standards](https://en.wikipedia.org/wiki/Data_mining#Standards)
* [6 Notable uses](https://en.wikipedia.org/wiki/Data_mining#Notable_uses)
* [7 Privacy concerns and ethics](https://en.wikipedia.org/wiki/Data_mining#Privacy_concerns_and_ethics)
  + [7.1 Situation in Europe](https://en.wikipedia.org/wiki/Data_mining#Situation_in_Europe)
  + [7.2 Situation in the United States](https://en.wikipedia.org/wiki/Data_mining#Situation_in_the_United_States)
* [8 Copyright law](https://en.wikipedia.org/wiki/Data_mining#Copyright_law)
  + [8.1 Situation in Europe](https://en.wikipedia.org/wiki/Data_mining#Situation_in_Europe_2)
  + [8.2 Situation in the United States](https://en.wikipedia.org/wiki/Data_mining#Situation_in_the_United_States_2)
* [9 Software](https://en.wikipedia.org/wiki/Data_mining#Software)
  + [9.1 Free open-source data mining software and applications](https://en.wikipedia.org/wiki/Data_mining#Free_open-source_data_mining_software_and_applications)
  + [9.2 Proprietary data-mining software and applications](https://en.wikipedia.org/wiki/Data_mining#Proprietary_data-mining_software_and_applications)
  + [9.3 Marketplace surveys](https://en.wikipedia.org/wiki/Data_mining#Marketplace_surveys)
* [10 See also](https://en.wikipedia.org/wiki/Data_mining#See_also)
* [11 References](https://en.wikipedia.org/wiki/Data_mining#References)
* [12 Further reading](https://en.wikipedia.org/wiki/Data_mining#Further_reading)
* [13 External links](https://en.wikipedia.org/wiki/Data_mining#External_links)

**Etymology**

In the 1960s, statisticians and economists used terms like *data fishing* or *data dredging* to refer to what they considered the bad practice of analyzing data without an a-priori hypothesis. The term "data mining" was used in a similarly critical way by economist [Michael Lovell](https://en.wikipedia.org/wiki/Michael_Lovell) in an article published in the [Review of Economic Studies](https://en.wikipedia.org/wiki/Review_of_Economic_Studies) 1983. Lovell indicates that the practice "masquerades under a variety of aliases, ranging from "experimentation" (positive) to "fishing" or "snooping" (negative).[[11]](https://en.wikipedia.org/wiki/Data_mining#cite_note-11)

The term *data mining* appeared around 1990 in the database community, generally with positive connotations. For a short time in 1980s, a phrase "database mining"™, was used, but since it was trademarked by HNC, a San Diego-based company, to pitch their Database Mining Workstation;[[12]](https://en.wikipedia.org/wiki/Data_mining#cite_note-Mena-12) researchers consequently turned to *data mining*. Other terms used include *data archaeology*, *information harvesting*, *information discovery*, *knowledge extraction*, etc. [Gregory Piatetsky-Shapiro](https://en.wikipedia.org/wiki/Gregory_I._Piatetsky-Shapiro) coined the term "knowledge discovery in databases" for the first workshop on the same topic [(KDD-1989)](http://www.kdnuggets.com/meetings/kdd89/) and this term became more popular in [AI](https://en.wikipedia.org/wiki/Artificial_intelligence) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning) community. However, the term data mining became more popular in the business and press communities.[[13]](https://en.wikipedia.org/wiki/Data_mining#cite_note-13) Currently, the terms *data mining* and *knowledge discovery* are used interchangeably.

In the academic community, the major forums for research started in 1995 when the First International Conference on Data Mining and Knowledge Discovery ([KDD-95](https://en.wikipedia.org/w/index.php?title=KDD-95&action=edit&redlink=1)) was started in Montreal under [AAAI](https://en.wikipedia.org/wiki/AAAI) sponsorship. It was co-chaired by [Usama Fayyad](https://en.wikipedia.org/wiki/Usama_Fayyad) and Ramasamy Uthurusamy. A year later, in 1996, Usama Fayyad launched the journal by Kluwer called [Data Mining and Knowledge Discovery](https://en.wikipedia.org/wiki/Data_Mining_and_Knowledge_Discovery) as its founding editor-in-chief. Later he started the [SIGKDD](https://en.wikipedia.org/wiki/SIGKDD) Newsletter SIGKDD Explorations.[[14]](https://en.wikipedia.org/wiki/Data_mining#cite_note-SIGKDD-explorations-14) The KDD International conference became the primary highest quality conference in data mining with an acceptance rate of research paper submissions below 18%. The journal *Data Mining and Knowledge Discovery* is the primary research journal of the field.

**Background**

The manual extraction of patterns from [data](https://en.wikipedia.org/wiki/Data) has occurred for centuries. Early methods of identifying patterns in data include [Bayes' theorem](https://en.wikipedia.org/wiki/Bayes%27_theorem) (1700s) and [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) (1800s). The proliferation, ubiquity and increasing power of computer technology has dramatically increased data collection, storage, and manipulation ability. As [data sets](https://en.wikipedia.org/wiki/Data_set) have grown in size and complexity, direct "hands-on" data analysis has increasingly been augmented with indirect, automated data processing, aided by other discoveries in computer science, such as [neural networks](https://en.wikipedia.org/wiki/Neural_networks), [cluster analysis](https://en.wikipedia.org/wiki/Cluster_analysis), [genetic algorithms](https://en.wikipedia.org/wiki/Genetic_algorithms) (1950s), [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) and [decision rules](https://en.wikipedia.org/wiki/Decision_rules) (1960s), and [support vector machines](https://en.wikipedia.org/wiki/Support_vector_machines) (1990s). Data mining is the process of applying these methods with the intention of uncovering hidden patterns[[15]](https://en.wikipedia.org/wiki/Data_mining" \l "cite_note-Kantardzic-15) in large data sets. It bridges the gap from [applied statistics](https://en.wikipedia.org/wiki/Applied_statistics) and artificial intelligence (which usually provide the mathematical background) to [database management](https://en.wikipedia.org/wiki/Database_management) by exploiting the way data is stored and indexed in databases to execute the actual learning and discovery algorithms more efficiently, allowing such methods to be applied to ever larger data sets.

**Process**

The *knowledge discovery in databases (KDD) process* is commonly defined with the stages:

1. Selection
2. Pre-processing
3. Transformation
4. *Data mining*
5. Interpretation/evaluation.[[5]](https://en.wikipedia.org/wiki/Data_mining#cite_note-Fayyad-5)

It exists, however, in many variations on this theme, such as the [Cross Industry Standard Process for Data Mining](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining) (CRISP-DM) which defines six phases:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

or a simplified process such as (1) Pre-processing, (2) Data Mining, and (3) Results Validation.

Polls conducted in 2002, 2004, 2007 and 2014 show that the CRISP-DM methodology is the leading methodology used by data miners.[[16]](https://en.wikipedia.org/wiki/Data_mining#cite_note-16) The only other data mining standard named in these polls was [SEMMA](https://en.wikipedia.org/wiki/SEMMA). However, 3–4 times as many people reported using CRISP-DM. Several teams of researchers have published reviews of data mining process models,[[17]](https://en.wikipedia.org/wiki/Data_mining" \l "cite_note-Marban-17)[[18]](https://en.wikipedia.org/wiki/Data_mining#cite_note-kurgan-18) and Azevedo and Santos conducted a comparison of CRISP-DM and SEMMA in 2008.[[19]](https://en.wikipedia.org/wiki/Data_mining#cite_note-AzevedoSantos-19)

**Pre-processing**

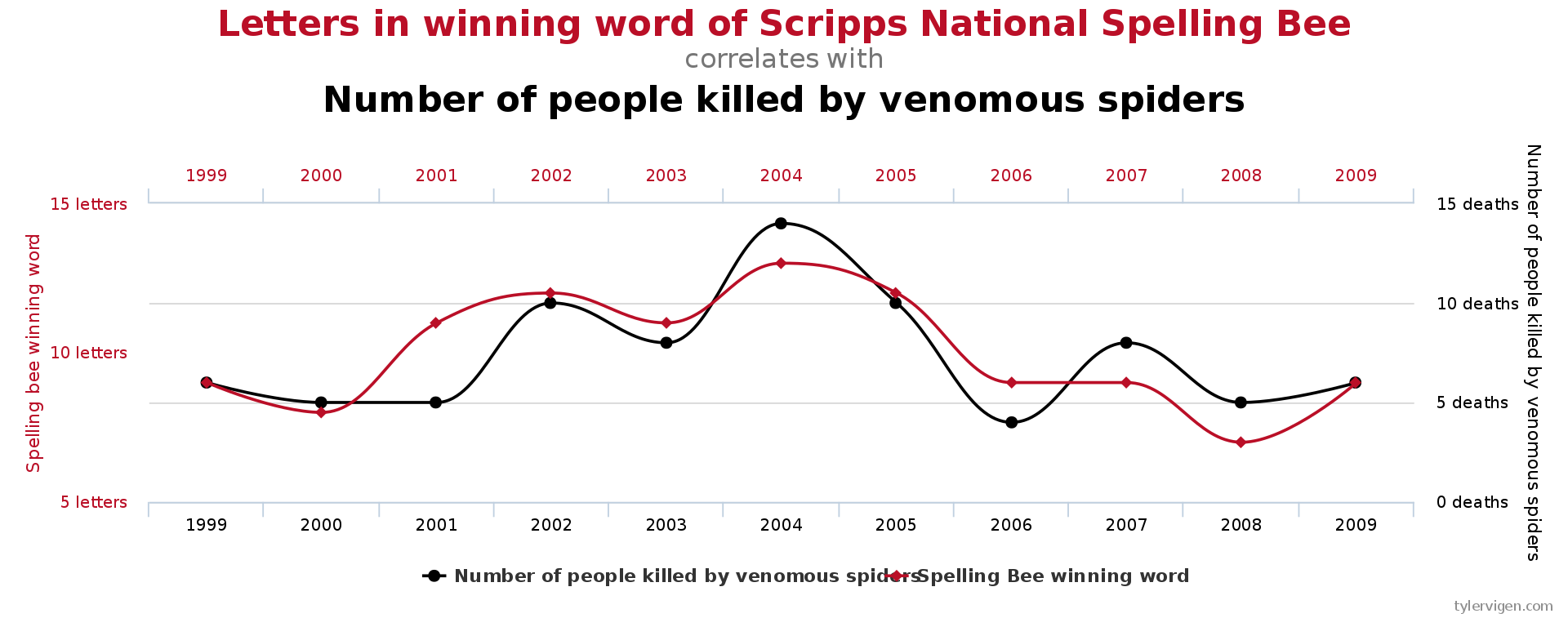
Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns actually present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. A common source for data is a [data mart](https://en.wikipedia.org/wiki/Data_mart) or [data warehouse](https://en.wikipedia.org/wiki/Data_warehouse). Pre-processing is essential to analyze the [multivariate](https://en.wikipedia.org/wiki/Multivariate_statistics) data sets before data mining. The target set is then cleaned. Data cleaning removes the observations containing [noise](https://en.wikipedia.org/wiki/Statistical_noise) and those with [missing data](https://en.wikipedia.org/wiki/Missing_data).

**Data mining**

Data mining involves six common classes of tasks:[[5]](https://en.wikipedia.org/wiki/Data_mining" \l "cite_note-Fayyad-5)

* [Anomaly detection](https://en.wikipedia.org/wiki/Anomaly_detection) (outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.
* [Association rule learning](https://en.wikipedia.org/wiki/Association_rule_learning) (dependency modelling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
* [Clustering](https://en.wikipedia.org/wiki/Cluster_analysis) – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
* [Classification](https://en.wikipedia.org/wiki/Statistical_classification) – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
* [Regression](https://en.wikipedia.org/wiki/Regression_analysis) – attempts to find a function which models the data with the least error that is, for estimating the relationships among data or datasets.
* [Summarization](https://en.wikipedia.org/wiki/Automatic_summarization) – providing a more compact representation of the data set, including visualization and report generation.

**Results validation**



An example of data produced by [data dredging](https://en.wikipedia.org/wiki/Data_dredging) through a bot operated by statistician Tyler Vigen, apparently showing a close link between the best word winning a spelling bee competition and the number of people in the United States killed by venomous spiders. The similarity in trends is obviously a coincidence.

Data mining can unintentionally be misused, and can then produce results which appear to be significant; but which do not actually predict future behaviour and cannot be [reproduced](https://en.wikipedia.org/wiki/Reproducibility) on a new sample of data and bear little use. Often this results from investigating too many hypotheses and not performing proper [statistical hypothesis testing](https://en.wikipedia.org/wiki/Statistical_hypothesis_testing). A simple version of this problem in [machine learning](https://en.wikipedia.org/wiki/Machine_learning) is known as [overfitting](https://en.wikipedia.org/wiki/Overfitting), but the same problem can arise at different phases of the process and thus a train/test split - when applicable at all - may not be sufficient to prevent this from happening.[[20]](https://en.wikipedia.org/wiki/Data_mining#cite_note-hawkins-20)

|  |  |
| --- | --- |
| [Wiki letter w.svg](https://en.wikipedia.org/wiki/File:Wiki_letter_w.svg) | This section **is missing information about non-classification tasks in data mining. It only covers** [**machine learning**](https://en.wikipedia.org/wiki/Machine_learning). Please expand the section to include this information. Further details may exist on the [talk page](https://en.wikipedia.org/wiki/Talk:Data_mining). *(September 2011)* |

The final step of knowledge discovery from data is to verify that the patterns produced by the data mining algorithms occur in the wider data set. Not all patterns found by the data mining algorithms are necessarily valid. It is common for the data mining algorithms to find patterns in the training set which are not present in the general data set. This is called [overfitting](https://en.wikipedia.org/wiki/Overfitting). To overcome this, the evaluation uses a [test set](https://en.wikipedia.org/wiki/Test_set) of data on which the data mining algorithm was not trained. The learned patterns are applied to this test set, and the resulting output is compared to the desired output. For example, a data mining algorithm trying to distinguish "spam" from "legitimate" emails would be trained on a [training set](https://en.wikipedia.org/wiki/Training_set) of sample e-mails. Once trained, the learned patterns would be applied to the test set of e-mails on which it had *not* been trained. The accuracy of the patterns can then be measured from how many e-mails they correctly classify. A number of statistical methods may be used to evaluate the algorithm, such as [ROC curves](https://en.wikipedia.org/wiki/Receiver_operating_characteristic).

If the learned patterns do not meet the desired standards, subsequently it is necessary to re-evaluate and change the pre-processing and data mining steps. If the learned patterns do meet the desired standards, then the final step is to interpret the learned patterns and turn them into knowledge.

**Research**

The premier professional body in the field is the [Association for Computing Machinery](https://en.wikipedia.org/wiki/Association_for_Computing_Machinery)'s (ACM) Special Interest Group (SIG) on Knowledge Discovery and Data Mining ([SIGKDD](https://en.wikipedia.org/wiki/SIGKDD)).[[21]](https://en.wikipedia.org/wiki/Data_mining#cite_note-21)[[22]](https://en.wikipedia.org/wiki/Data_mining#cite_note-22) Since 1989, this ACM SIG has hosted an annual international conference and published its proceedings,[[23]](https://en.wikipedia.org/wiki/Data_mining#cite_note-23) and since 1999 it has published a biannual [academic journal](https://en.wikipedia.org/wiki/Academic_journal) titled "SIGKDD Explorations".[[24]](https://en.wikipedia.org/wiki/Data_mining#cite_note-24)

Computer science conferences on data mining include:

* [CIKM Conference](https://en.wikipedia.org/wiki/CIKM_Conference) – ACM [Conference on Information and Knowledge Management](https://en.wikipedia.org/wiki/Conference_on_Information_and_Knowledge_Management)
* [European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases](https://en.wikipedia.org/wiki/European_Conference_on_Machine_Learning_and_Principles_and_Practice_of_Knowledge_Discovery_in_Databases)
* [KDD Conference](https://en.wikipedia.org/wiki/KDD_Conference) – ACM SIGKDD [Conference on Knowledge Discovery and Data Mining](https://en.wikipedia.org/wiki/Conference_on_Knowledge_Discovery_and_Data_Mining)

Data mining topics are also present on many [data management/database conferences](https://en.wikipedia.org/wiki/List_of_computer_science_conferences#Data_Management) such as the ICDE Conference, [SIGMOD Conference](https://en.wikipedia.org/wiki/SIGMOD) and [International Conference on Very Large Data Bases](https://en.wikipedia.org/wiki/International_Conference_on_Very_Large_Data_Bases)

**Standards**

There have been some efforts to define standards for the data mining process, for example the 1999 European [Cross Industry Standard Process for Data Mining](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining) (CRISP-DM 1.0) and the 2004 [Java Data Mining](https://en.wikipedia.org/wiki/Java_Data_Mining) standard (JDM 1.0). Development on successors to these processes (CRISP-DM 2.0 and JDM 2.0) was active in 2006, but has stalled since. JDM 2.0 was withdrawn without reaching a final draft.

For exchanging the extracted models – in particular for use in [predictive analytics](https://en.wikipedia.org/wiki/Predictive_analytics) – the key standard is the [Predictive Model Markup Language](https://en.wikipedia.org/wiki/Predictive_Model_Markup_Language) (PMML), which is an [XML](https://en.wikipedia.org/wiki/XML)-based language developed by the Data Mining Group (DMG) and supported as exchange format by many data mining applications. As the name suggests, it only covers prediction models, a particular data mining task of high importance to business applications. However, extensions to cover (for example) [subspace clustering](https://en.wikipedia.org/wiki/Subspace_clustering) have been proposed independently of the DMG.[[25]](https://en.wikipedia.org/wiki/Data_mining#cite_note-25)

**Notable uses**

Main article: [Examples of data mining](https://en.wikipedia.org/wiki/Examples_of_data_mining)

See also: [Category:Applied data mining](https://en.wikipedia.org/wiki/Category:Applied_data_mining).

Data mining is used wherever there is digital data available today. Notable [examples of data mining](https://en.wikipedia.org/wiki/Examples_of_data_mining) can be found throughout business, medicine, science, and surveillance.

**Privacy concerns and ethics**

While the term "data mining" itself may have no ethical implications, it is often associated with the mining of information in relation to peoples' behavior (ethical and otherwise).[[26]](https://en.wikipedia.org/wiki/Data_mining#cite_note-26)

The ways in which data mining can be used can in some cases and contexts raise questions regarding privacy, legality, and ethics.[[27]](https://en.wikipedia.org/wiki/Data_mining#cite_note-27) In particular, data mining government or commercial data sets for national security or law enforcement purposes, such as in the [Total Information Awareness](https://en.wikipedia.org/wiki/Total_Information_Awareness) Program or in [ADVISE](https://en.wikipedia.org/wiki/ADVISE), has raised privacy concerns.[[28]](https://en.wikipedia.org/wiki/Data_mining#cite_note-28)[[29]](https://en.wikipedia.org/wiki/Data_mining#cite_note-29)

Data mining requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations. A common way for this to occur is through [data aggregation](https://en.wikipedia.org/wiki/Aggregate_function). Data aggregation involves combining data together (possibly from various sources) in a way that facilitates analysis (but that also might make identification of private, individual-level data deducible or otherwise apparent).[[30]](https://en.wikipedia.org/wiki/Data_mining#cite_note-NASCIO-30) This is not data mining *per se*, but a result of the preparation of data before – and for the purposes of – the analysis. The threat to an individual's privacy comes into play when the data, once compiled, cause the data miner, or anyone who has access to the newly compiled data set, to be able to identify specific individuals, especially when the data were originally anonymous.[[31]](https://en.wikipedia.org/wiki/Data_mining#cite_note-31)[[32]](https://en.wikipedia.org/wiki/Data_mining#cite_note-32)[[33]](https://en.wikipedia.org/wiki/Data_mining#cite_note-33)

It is recommended that an individual is made aware of the following **before** data are collected:[[30]](https://en.wikipedia.org/wiki/Data_mining" \l "cite_note-NASCIO-30)

* the purpose of the data collection and any (known) data mining projects;
* how the data will be used;
* who will be able to mine the data and use the data and their derivatives;
* the status of security surrounding access to the data;
* how collected data can be updated.

Data may also be modified so as to *become* anonymous, so that individuals may not readily be identified.[[30]](https://en.wikipedia.org/wiki/Data_mining#cite_note-NASCIO-30) However, even "de-identified"/"anonymized" data sets can potentially contain enough information to allow identification of individuals, as occurred when journalists were able to find several individuals based on a set of search histories that were inadvertently released by AOL.[[34]](https://en.wikipedia.org/wiki/Data_mining#cite_note-34)

The inadvertent revelation of [personally identifiable information](https://en.wikipedia.org/wiki/Personally_identifiable_information) leading to the provider violates Fair Information Practices. This indiscretion can cause financial, emotional, or bodily harm to the indicated individual. In one instance of privacy violation, the patrons of Walgreens filed a lawsuit against the company in 2011 for selling prescription information to data mining companies who in turn provided the data to pharmaceutical companies.[[35]](https://en.wikipedia.org/wiki/Data_mining#cite_note-35)

**Situation in Europe**

Europe has rather strong privacy laws, and efforts are underway to further strengthen the rights of the consumers. However, the [U.S.-E.U. Safe Harbor Principles](https://en.wikipedia.org/wiki/International_Safe_Harbor_Privacy_Principles) currently effectively expose European users to privacy exploitation by U.S. companies. As a consequence of [Edward Snowden](https://en.wikipedia.org/wiki/Edward_Snowden)'s [global surveillance disclosure](https://en.wikipedia.org/wiki/Global_surveillance_disclosure), there has been increased discussion to revoke this agreement, as in particular the data will be fully exposed to the [National Security Agency](https://en.wikipedia.org/wiki/National_Security_Agency), and attempts to reach an agreement have failed.[[*citation needed*](https://en.wikipedia.org/wiki/Wikipedia:Citation_needed)]

**Situation in the United States**

In the United States, privacy concerns have been addressed by the [US Congress](https://en.wikipedia.org/wiki/US_Congress) via the passage of regulatory controls such as the [Health Insurance Portability and Accountability Act](https://en.wikipedia.org/wiki/Health_Insurance_Portability_and_Accountability_Act) (HIPAA). The HIPAA requires individuals to give their "informed consent" regarding information they provide and its intended present and future uses. According to an article in *Biotech Business Week*, "'[i]n practice, HIPAA may not offer any greater protection than the longstanding regulations in the research arena,' says the AAHC. More importantly, the rule's goal of protection through informed consent is approach a level of incomprehensibility to average individuals."[[36]](https://en.wikipedia.org/wiki/Data_mining#cite_note-36) This underscores the necessity for data anonymity in data aggregation and mining practices.

U.S. information privacy legislation such as HIPAA and the [Family Educational Rights and Privacy Act](https://en.wikipedia.org/wiki/Family_Educational_Rights_and_Privacy_Act) (FERPA) applies only to the specific areas that each such law addresses. Use of data mining by the majority of businesses in the U.S. is not controlled by any legislation.

**Copyright law**

**Situation in Europe**

Due to a lack of flexibilities in European copyright and [database law](https://en.wikipedia.org/wiki/Database_Directive), the mining of in-copyright works such as [web mining](https://en.wikipedia.org/wiki/Web_mining) without the permission of the copyright owner is not legal. Where a database is pure data in Europe there is likely to be no copyright, but database rights may exist so data mining becomes subject to regulations by the [Database Directive](https://en.wikipedia.org/wiki/Database_Directive). On the recommendation of the [Hargreaves review](https://en.wikipedia.org/wiki/Hargreaves_review) this led to the UK government to amend its copyright law in 2014[[37]](https://en.wikipedia.org/wiki/Data_mining#cite_note-37) to allow content mining as a [limitation and exception](https://en.wikipedia.org/wiki/Limitations_and_exceptions_to_copyright). Only the second country in the world to do so after Japan, which introduced an exception in 2009 for data mining. However, due to the restriction of the [Copyright Directive](https://en.wikipedia.org/wiki/Copyright_Directive), the UK exception only allows content mining for non-commercial purposes. UK copyright law also does not allow this provision to be overridden by contractual terms and conditions. The [European Commission](https://en.wikipedia.org/wiki/European_Commission) facilitated stakeholder discussion on text and data mining in 2013, under the title of Licences for Europe.[[38]](https://en.wikipedia.org/wiki/Data_mining#cite_note-38) The focus on the solution to this legal issue being licences and not limitations and exceptions led to representatives of universities, researchers, libraries, civil society groups and [open access](https://en.wikipedia.org/wiki/Open_access) publishers to leave the stakeholder dialogue in May 2013.[[39]](https://en.wikipedia.org/wiki/Data_mining#cite_note-39)

**Situation in the United States**

By contrast to Europe, the flexible nature of US copyright law, and in particular [fair use](https://en.wikipedia.org/wiki/Fair_use) means that content mining in America, as well as other fair use countries such as Israel, Taiwan and South Korea is viewed as being legal. As content mining is transformative, that is it does not supplant the original work, it is viewed as being lawful under fair use. For example, as part of the [Google Book settlement](https://en.wikipedia.org/wiki/Google_Book_Search_Settlement_Agreement) the presiding judge on the case ruled that Google's digitisation project of in-copyright books was lawful, in part because of the transformative uses that the digitisation project displayed - one being text and data mining.[[40]](https://en.wikipedia.org/wiki/Data_mining#cite_note-40)

**Software**

See also: [Category:Data mining and machine learning software](https://en.wikipedia.org/wiki/Category:Data_mining_and_machine_learning_software).

**Free open-source data mining software and applications**

The following applications are available under free/open source licenses. Public access to application source code is also available.

* [Carrot2](https://en.wikipedia.org/wiki/Carrot2): Text and search results clustering framework.
* [Chemicalize.org](https://en.wikipedia.org/wiki/Chemicalize.org): A chemical structure miner and web search engine.
* [ELKI](https://en.wikipedia.org/wiki/ELKI): A university research project with advanced [cluster analysis](https://en.wikipedia.org/wiki/Cluster_analysis) and [outlier detection](https://en.wikipedia.org/wiki/Anomaly_detection) methods written in the [Java](https://en.wikipedia.org/wiki/Java_(programming_language)) language.
* [GATE](https://en.wikipedia.org/wiki/General_Architecture_for_Text_Engineering): a [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing) and language engineering tool.
* [KNIME](https://en.wikipedia.org/wiki/KNIME): The Konstanz Information Miner, a user friendly and comprehensive data analytics framework.
* [Massive Online Analysis (MOA)](https://en.wikipedia.org/wiki/MOA_(Massive_Online_Analysis)): a real-time big data stream mining with concept drift tool in the [Java](https://en.wikipedia.org/wiki/Java_(programming_language)) programming language.
* [MEPX](https://en.wikipedia.org/wiki/Multi_expression_programming) - cross platform tool for regression and classification problems based on a Genetic Programming variant.
* ML-Flex: A software package that enables users to integrate with third-party machine-learning packages written in any programming language, execute classification analyses in parallel across multiple computing nodes, and produce HTML reports of classification results.
* [mlpack](https://en.wikipedia.org/wiki/Mlpack): a collection of ready-to-use machine learning algorithms written in the [C++](https://en.wikipedia.org/wiki/C%2B%2B) language.
* [NLTK](https://en.wikipedia.org/wiki/NLTK) ([Natural Language Toolkit](https://en.wikipedia.org/wiki/Natural_Language_Toolkit)): A suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) language.
* [OpenNN](https://en.wikipedia.org/wiki/OpenNN): Open [neural networks](https://en.wikipedia.org/wiki/Neural_networks) library.
* [Orange](https://en.wikipedia.org/wiki/Orange_(software)): A component-based data mining and [machine learning](https://en.wikipedia.org/wiki/Machine_learning) software suite written in the [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) language.
* [R](https://en.wikipedia.org/wiki/R_(programming_language)): A [programming language](https://en.wikipedia.org/wiki/Programming_language) and software environment for [statistical](https://en.wikipedia.org/wiki/Statistics) computing, data mining, and graphics. It is part of the [GNU Project](https://en.wikipedia.org/wiki/GNU_Project).
* [scikit-learn](https://en.wikipedia.org/wiki/Scikit-learn) is an open source machine learning library for the Python programming language
* [Torch](https://en.wikipedia.org/wiki/Torch_(machine_learning)): An [open-source](https://en.wikipedia.org/wiki/Open_source_model) [deep learning](https://en.wikipedia.org/wiki/Deep_learning) library for the [Lua](https://en.wikipedia.org/wiki/Lua_(programming_language)) programming language and [scientific computing](https://en.wikipedia.org/wiki/Scientific_computing) framework with wide support for [machine learning](https://en.wikipedia.org/wiki/Machine_learning) algorithms.
* [UIMA](https://en.wikipedia.org/wiki/UIMA): The UIMA (Unstructured Information Management Architecture) is a component framework for analyzing unstructured content such as text, audio and video – originally developed by IBM.
* [Weka](https://en.wikipedia.org/wiki/Weka_(machine_learning)): A suite of machine learning software applications written in the [Java](https://en.wikipedia.org/wiki/Java_(programming_language)) programming language.