# Data mining

Not to be confused with analytics, information extraction, or data analysis.

**Data mining** is an interdisciplinary subfield of computer science.[1][2][3] It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.[1] The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.[1] Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.[1] Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.[4]

The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (*mining*) of data itself.[5] It also is a buzzword[6] and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence, machine learning, and business intelligence. The book *Data mining: Practical machine learning tools and techniques with Java*[7] (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.[8] Often the more general terms (*large scale*) *data analysis* and *analytics* – or, when referring to actual methods, *artificial intelligence* and *machine learning* – are more appropriate.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. 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. 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. 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*, *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.

## 1 Etymology

In the 1960s, statisticians 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" appeared around 1990 in the database community. 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;[9] researchers consequently turned to "data mining". Other terms used include Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction, etc. Gregory Piatetsky-Shapiro coined the term "Knowledge Discovery in Databases" for the first workshop on the same topic (KDD-1989) and this term became more popular in AI and Machine Learning Community. However, the term data mining became more popular in the business and press communities.[10] Currently, Data Mining and Knowledge Discovery are used interchangeably. Since about 2007, "Predictive Analytics" and since 2011, "Data Science" terms were also used to describe this field.

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) was started in Montreal under AAAI sponsorship. It was co-chaired by Usama Fayyad and Ramasamy Uthurusamy. A year later, in 1996, Usama Fayyad launched the journal by Kluwer called Data Mining and Knowledge Discovery as its founding Editor-in-Chief. Later he started the SIGKDDD Newsletter SIGKDD Explorations.[11] 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.

# 2 Background

The manual extraction of patterns from data has occurred for centuries. Early methods of identifying patterns in data include Bayes' theorem (1700s) and regression analysis (1800s). The proliferation, ubiquity and increasing power of computer technology has dramatically increased data collection, storage, and manipulation ability. As data sets 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, cluster analysis, genetic algorithms (1950s), decision trees and decision rules (1960s), and support vector machines (1990s). Data mining is the process of applying these methods with the intention of uncovering hidden patterns[12] in large data sets. It bridges the gap from applied statistics and artificial intelligence (which usually provide the mathematical background) to 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.

# 3 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.[4]

It exists, however, in many variations on this theme, such as the 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.[13] The only other data mining standard named in these polls was SEMMA. However, 3–4 times as many people reported using CRISP-DM. Several teams of researchers have published reviews of data mining process models,[14][15] and Azevedo and Santos conducted a comparison of CRISP-DM and SEMMA in 2008.[16]

## 3.1 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 or data warehouse. Pre-processing is essential to analyze the multivariate data sets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise and those with missing data.

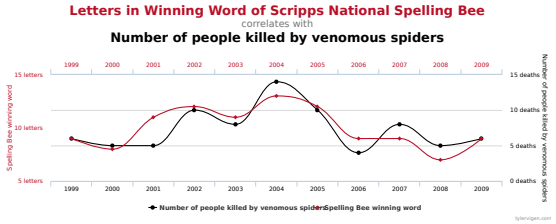
## 3.2 Data mining

Data mining involves six common classes of tasks:[4]

- 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 (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 – 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 – 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 – attempts to find a function which models the data with the least error.
- Summarization – providing a more compact representation of the data set, including visualization and report generation.

## 3.3 Results validation

Data mining can unintentionally be misused, and can then produce results which appear to be significant; but which




*An example of data produced by data dredging through a bot operated by statistician Tyler Viglen, 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.*

do not actually predict future behaviour and cannot be reproduced 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. A simple version of this problem in machine learning is known as 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.

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. To overcome this, the evaluation uses a 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 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.

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.

# 4 Research

The premier professional body in the field is the Association for Computing Machinery's (ACM) Special Interest Group (SIG) on Knowledge Discovery and Data Mining (SIGKDD).[17][18] Since 1989 this ACM SIG has hosted an annual international conference and published its proceedings,[19] and since 1999 it has published a biannual academic journal titled "SIGKDD Explorations".[20]

Computer science conferences on data mining include:

- CIKM Conference – ACM Conference on Information and Knowledge Management
- DMIN Conference – International Conference on Data Mining
- DMKD Conference – Research Issues on Data Mining and Knowledge Discovery
- DSAA Conference – IEEE International Conference on Data Science and Advanced Analytics
- ECDM Conference – European Conference on Data Mining
- ECML-PKDD Conference – European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases
- EDM Conference – International Conference on Educational Data Mining
- INFOCOM Conference – IEEE INFOCOM
- ICDM Conference – IEEE International Conference on Data Mining
- KDD Conference – ACM SIGKDD Conference on Knowledge Discovery and Data Mining
- MLDM Conference – Machine Learning and Data Mining in Pattern Recognition
- PAKDD Conference – The annual Pacific-Asia Conference on Knowledge Discovery and Data Mining
- PAW Conference – Predictive Analytics World
- SDM Conference – SIAM International Conference on Data Mining (SIAM)
- SSTD Symposium – Symposium on Spatial and Temporal Databases
- WSDM Conference – ACM Conference on Web Search and Data Mining

Data mining topics are also present on many data management/database conferences such as the ICDE Conference, SIGMOD Conference and International Conference on Very Large Data Bases

# 5 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 (CRISP-DM 1.0) and the 2004 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 – the key standard is the Predictive Model Markup Language (PMML), which is an 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 have been proposed independently of the DMG.[21]

# 6 Notable uses

Main article: Examples of data mining
See also: Category:Applied data mining.

Data mining is used wherever there is digital data available today. Notable examples of data mining can be found throughout business, medicine, science, and surveillance.

# 7 Privacy concerns and ethics

While the term “data mining” itself has no ethical implications, it is often associated with the mining of information in relation to peoples’ behavior (ethical and otherwise).[22]

The ways in which data mining can be used can in some cases and contexts raise questions regarding privacy, legality, and ethics.[23] In particular, data mining government or commercial data sets for national security or law enforcement purposes, such as in the Total Information Awareness Program or in ADVISE, has raised privacy concerns.[24][25]

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. 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).[26] 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.[27][28][29]

It is recommended that an individual is made aware of the following **before** data are collected:[26]

- 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.[26] 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.[30]

The inadvertent revelation of 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.[31]

## 7.1 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 currently effectively expose European users to privacy exploitation by U.S. companies. As a consequence of Edward Snowden's 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, and attempts to reach an agreement have failed.

## 7.2 Situation in the United States

In the United States, privacy concerns have been addressed by the US Congress via the passage of regulatory controls such as the 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 undermined by the complexity of consent forms that are required of patients and participants, which approach a level of incomprehensibility to average individuals."*[32] *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 (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.

# 8 Copyright Law

## 8.1 Situation in Europe

Due to a lack of flexibilities in European copyright and database law, the mining of in-copyright works such as 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. On the recommendation of the Hargreaves review this led to the UK government to amend its copyright law in 2014[33] to allow content mining as a limitation and exception. 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, 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 facilitated stakeholder discussion on text and data mining in 2013, under the title of Licences for Europe.[34] 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 publishers to leave the stakeholder dialogue in May 2013.[35]

## 8.2 Situation in the United States

By contrast to Europe, the flexible nature of US copyright law, and in particular 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 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.[36]

# 9 Software

See also: Category:Data mining and machine learning software.

## 9.1 Free open-source data mining software and applications

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

- Carrot2: Text and search results clustering framework.
- Chemicalize.org: A chemical structure miner and web search engine.
- ELKI: A university research project with advanced cluster analysis and outlier detection methods written in the Java language.
- GATE: a natural language processing and language engineering tool.
- KNIME: The Konstanz Information Miner, a user friendly and comprehensive data analytics framework.
- Massive Online Analysis (MOA): a real-time big data stream mining with concept drift tool in the Java programming language.
- 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 library: a collection of ready-to-use machine learning algorithms written in the C++ language.
- NLTK (Natural Language Toolkit): A suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the Python language.
- OpenNN: Open neural networks library.
- Orange: A component-based data mining and machine learning software suite written in the Python language.

- R: A programming language and software environment for statistical computing, data mining, and graphics. It is part of the GNU Project.
- SCaViS: Java cross-platform data analysis framework developed at Argonne National Laboratory.
- scikit-learn is an open source machine learning library for the Python programming language
- SenticNet API: A semantic and affective resource for opinion mining and sentiment analysis.
- Torch: An open source deep learning library for the Lua programming language and scientific computing framework with wide support for machine learning algorithms.
- 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: A suite of machine learning software applications written in the Java programming language.

## 9.2 Proprietary data-mining software and applications

The following applications are available under proprietary licenses.

- Angoss KnowledgeSTUDIO: data mining tool provided by Angoss.
- Clarabridge: enterprise class text analytics solution.
- HP Vertica Analytics Platform: data mining software provided by HP.
- IBM SPSS Modeler: data mining software provided by IBM.
- KXEN Modeler: data mining tool provided by KXEN.
- LIONsolver: an integrated software application for data mining, business intelligence, and modeling that implements the Learning and Intelligent OptimizatioN (LION) approach.
- Megaputer Intelligence: data and text mining software is called PolyAnalyst.
- Microsoft Analysis Services: data mining software provided by Microsoft.
- NetOwl: suite of multilingual text and entity analytics products that enable data mining.
- OpenText™ Big Data Analytics: Visual Data Mining & Predictive Analysis by Open Text Corporation
- Oracle Data Mining: data mining software by Oracle.
- PSeven: platform for automation of engineering simulation and analysis, multidisciplinary optimization and data mining provided by DATADVANCE.
- Qlucore Omics Explorer: data mining software provided by Qlucore.
- RapidMiner: An environment for machine learning and data mining experiments.
- SAS Enterprise Miner: data mining software provided by the SAS Institute.
- STATISTICA Data Miner: data mining software provided by StatSoft.
- Tanagra: A visualisation-oriented data mining software, also for teaching.

## 9.3 Marketplace surveys

Several researchers and organizations have conducted reviews of data mining tools and surveys of data miners. These identify some of the strengths and weaknesses of the software packages. They also provide an overview of the behaviors, preferences and views of data miners. Some of these reports include:

- Hurwitz Victory Index: Report for Advanced Analytics as a market research assessment tool, it highlights both the diverse uses for advanced analytics technology and the vendors who make those applications possible.Recent-research
- 2011 Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery[37]
- Rexer Analytics Data Miner Surveys (2007–2013)[38]
- Forrester Research 2010 Predictive Analytics and Data Mining Solutions report[39]
- Gartner 2008 “Magic Quadrant” report[40]
- Robert A. Nisbet’s 2006 Three Part Series of articles “Data Mining Tools: Which One is Best For CRM?"[41]
- Haughton et al.'s 2003 Review of Data Mining Software Packages in *The American Statistician*[42]
- Goebel & Gruenwald 1999 “A Survey of Data Mining a Knowledge Discovery Software Tools” in SIGKDD Explorations[43]

# 10 See also

See also: Category:Applied data mining.

### Methods

- Anomaly/outlier/change detection
- Association rule learning
- Classification
- Cluster analysis
- Decision tree
- Factor analysis
- Genetic algorithms
- Intention mining
- Multilinear subspace learning
- Neural networks
- Regression analysis
- Sequence mining
- Structured data analysis
- Support vector machines
- Text mining
- Agent mining

### Application domains

- Analytics
- Behavior informatics
- Big Data
- Bioinformatics
- Business intelligence
- Data analysis
- Data warehouse
- Decision support system
- Domain driven data mining
- Drug discovery
- Exploratory data analysis
- Predictive analytics
- Web mining

### Application examples

- Customer analytics
- Data mining in agriculture
- Data mining in meteorology
- Educational data mining
- National Security Agency
- Police-enforced ANPR in the UK
- Quantitative structure–activity relationship
- Surveillance / Mass surveillance (e.g., Stellar Wind)

### Related topics

Data mining is about *analyzing* data; for information about extracting information out of data, see:

- Data integration
- Data transformation
- Electronic discovery
- Information extraction
- Information integration
- Named-entity recognition
- Profiling (information science)
- Web scraping

# 11 References


[1] "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27.

[2] Clifton, Christopher (2010). "Encyclopædia Britannica: Definition of Data Mining". Retrieved 2010-12-09.

[3] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Retrieved 2012-08-07.

[4] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases" (PDF). Retrieved 17 December 2008.

[5] Han, Jiawei; Kamber, Micheline (2001). *Data mining: concepts and techniques*. Morgan Kaufmann. p. 5. ISBN 978-1-55860-489-6. Thus, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long

[6] See e.g. OKAIRP 2005 Fall Conference, Arizona State University About.com: Datamining

[7] Witten, Ian H.; Frank, Eibe; Hall, Mark A. (30 January 2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3 ed.). Elsevier. ISBN 978-0-12-374856-0.

[8] Bouckaert, Remco R.; Frank, Eibe; Hall, Mark A.; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H. (2010). "WEKA Experiences with a Java open-source project". *Journal of Machine Learning Research* **11**: 2533–2541. the original title, "Practical machine learning", was changed ... The term "data mining" was [added] primarily for marketing reasons.

[9] Mena, Jesús (2011). *Machine Learning Forensics for Law Enforcement, Security, and Intelligence*. Boca Raton, FL: CRC Press (Taylor & Francis Group). ISBN 978-1-4398-6069-4.

[10] Piatetsky-Shapiro, Gregory; Parker, Gary (2011). "Lesson: Data Mining, and Knowledge Discovery: An Introduction". *Introduction to Data Mining*. KD Nuggets. Retrieved 30 August 2012.

[11] Fayyad, Usama (15 June 1999). "First Editorial by Editor-in-Chief". *SIGKDD Explorations* **1** (1): 1. doi:10.1145/2207243.2207269. Retrieved 27 December 2010.

[12] Kantardzic, Mehmed (2003). *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons. ISBN 0-471-22852-4. OCLC 50055336.

[13] Gregory Piatetsky-Shapiro (2002) *KDnuggets Methodology Poll*, Gregory Piatetsky-Shapiro (2004) *KDnuggets Methodology Poll*, Gregory Piatetsky-Shapiro (2007) *KDnuggets Methodology Poll*, Gregory Piatetsky-Shapiro (2014) *KDnuggets Methodology Poll*

[14] Óscar Marbán, Gonzalo Mariscal and Javier Segovia (2009); *A Data Mining & Knowledge Discovery Process Model*. In Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca, ISBN 978-3-902613-53-0, pp. 438–453, February 2009, I-Tech, Vienna, Austria.

[15] Lukasz Kurgan and Petr Musilek (2006); *A survey of Knowledge Discovery and Data Mining process models*. The Knowledge Engineering Review. Volume 21 Issue 1, March 2006, pp 1–24, Cambridge University Press, New York, NY, USA doi:10.1017/S0269888906000737

[16] Azevedo, A. and Santos, M. F. KDD, SEMMA and CRISP-DM: a parallel overview. In Proceedings of the IADIS European Conference on Data Mining 2008, pp 182–185.

[17] "Microsoft Academic Search: Top conferences in data mining". Microsoft Academic Search.

[18] "Google Scholar: Top publications - Data Mining & Analysis". Google Scholar.

[19] Proceedings, International Conferences on Knowledge Discovery and Data Mining, ACM, New York.

[20] SIGKDD Explorations, ACM, New York.

[21] Günnemann, Stephan; Kremer, Hardy; Seidl, Thomas (2011). "An extension of the PMML standard to subspace clustering models". *Proceedings of the 2011 workshop on Predictive markup language modeling - PMML '11*. p. 48. doi:10.1145/2023598.2023605. ISBN 978-1-4503-0837-3.

[22] Seltzer, William. "The Promise and Pitfalls of Data Mining: Ethical Issues" (PDF).

[23] Pitts, Chip (15 March 2007). "The End of Illegal Domestic Spying? Don't Count on It". *Washington Spectator*.

[24] Taipale, Kim A. (15 December 2003). "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data". *Columbia Science and Technology Law Review* **5** (2). OCLC 45263753. SSRN 546782.

[25] Resig, John; Teredesai, Ankur (2004). "A Framework for Mining Instant Messaging Services". *Proceedings of the 2004 SIAM DM Conference*.

[26] *Think Before You Dig: Privacy Implications of Data Mining & Aggregation*, NASCIO Research Brief, September 2004

[27] Ohm, Paul. "Don't Build a Database of Ruin". Harvard Business Review.

[28] Darwin Bond-Graham, Iron Cagebook - The Logical End of Facebook's Patents, Counterpunch.org, 2013.12.03

[29] Darwin Bond-Graham, Inside the Tech industry's Startup Conference, Counterpunch.org, 2013.09.11

[30] *AOL search data identified individuals*, SecurityFocus, August 2006

[31] Kshetri, Nir (2014). "Big data´s impact on privacy, security and consumer welfare". *Telecommunications Policy* **38** (11): 1134–1145. doi:10.1016/j.telpol.2014.10.002.

[32] Biotech Business Week Editors (June 30, 2008); *BIOMEDICINE; HIPAA Privacy Rule Impedes Biomedical Research*, Biotech Business Week, retrieved 17 November 2009 from LexisNexis Academic

[33] UK Researchers Given Data Mining Right Under New UK Copyright Laws. *Out-Law.com*. Retrieved 14 November 2014

[34] "Licences for Europe - Structured Stakeholder Dialogue 2013". *European Commission*. Retrieved 14 November 2014.

[35] "Text and Data Mining:Its importance and the need for change in Europe". *Association of European Research Libraries*. Retrieved 14 November 2014.

[36] "Judge grants summary judgment in favor of Google Books — a fair use victory". *Lexology.com*. Antonelli Law Ltd. Retrieved 14 November 2014.

[37] Mikut, Ralf; Reischl, Markus (September–October 2011). "Data Mining Tools". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1** (5): 431–445. doi:10.1002/widm.24. Retrieved October 21, 2011.

[38] Karl Rexer, Heather Allen, & Paul Gearan (2011); *Understanding Data Miners*, Analytics Magazine, May/June 2011 (INFORMS: Institute for Operations Research and the Management Sciences).

[39] Kobielus, James; *The Forrester Wave: Predictive Analytics and Data Mining Solutions, Q1 2010*, Forrester Research, 1 July 2008

[40] Herschel, Gareth; *Magic Quadrant for Customer Data-Mining Applications*, Gartner Inc., 1 July 2008

[41] Nisbet, Robert A. (2006); *Data Mining Tools: Which One is Best for CRM? Part 1*, Information Management Special Reports, January 2006

[42] Haughton, Dominique; Deichmann, Joel; Eshghi, Abdolreza; Sayek, Selin; Teebagy, Nicholas; and Topi, Heikki (2003); *A Review of Software Packages for Data Mining*, The American Statistician, Vol. 57, No. 4, pp. 290–309

[43] Goebel, Michael; Gruenwald, Le (1999); *A Survey of Data Mining and Knowledge Discovery Software Tools*, SIGKDD Explorations, Vol. 1, Issue 1, pp. 20–33

## 12 Further reading

- Cabena, Peter; Hadjnian, Pablo; Stadler, Rolf; Verhees, Jaap; Zanasi, Alessandro (1997); *Discovering Data Mining: From Concept to Implementation*, Prentice Hall, ISBN 0-13-743980-6
- M.S. Chen, J. Han, P.S. Yu (1996) "Data mining: an overview from a database perspective". *Knowledge and data Engineering, IEEE Transactions* on 8 (6), 866–883
- Feldman, Ronen; Sanger, James (2007); *The Text Mining Handbook*, Cambridge University Press, ISBN 978-0-521-83657-9
- Guo, Yike; and Grossman, Robert (editors) (1999); *High Performance Data Mining: Scaling Algorithms, Applications and Systems*, Kluwer Academic Publishers
- Han, Jiawei, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Morgan kaufmann, 2006.
- Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome (2001); *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, ISBN 0-387-95284-5
- Liu, Bing (2007); *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*, Springer, ISBN 3-540-37881-2
- Murphy, Chris (16 May 2011). "Is Data Mining Free Speech?". *InformationWeek* (UMB): 12.
- Nisbet, Robert; Elder, John; Miner, Gary (2009); *Handbook of Statistical Analysis & Data Mining Applications*, Academic Press/Elsevier, ISBN 978-0-12-374765-5
- Poncelet, Pascal; Masseglia, Florent; and Teisseire, Maguelonne (editors) (October 2007); "Data Mining Patterns: New Methods and Applications", *Information Science Reference*, ISBN 978-1-59904-162-9
- Tan, Pang-Ning; Steinbach, Michael; and Kumar, Vipin (2005); *Introduction to Data Mining*, ISBN 0-321-32136-7
- Theodoridis, Sergios; and Koutroumbas, Konstantinos (2009); *Pattern Recognition*, 4th Edition, Academic Press, ISBN 978-1-59749-272-0
- Weiss, Sholom M.; and Indurkhya, Nitin (1998); *Predictive Data Mining*, Morgan Kaufmann
- Witten, Ian H.; Frank, Eibe; Hall, Mark A. (30 January 2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3 ed.). Elsevier. ISBN 978-0-12-374856-0. (See also Free Weka software)
- Ye, Nong (2003); *The Handbook of Data Mining*, Mahwah, NJ: Lawrence Erlbaum

## 13 External links

Knowledge Discovery Software at DMOZ

Data Mining Tool Vendors at DMOZ

# 14 Text and image sources, contributors, and licenses

## 14.1 Text

- **Data mining** *Source:* https://en.wikipedia.org/wiki/Data_mining?oldid=723223840 *Contributors:* Dreamyshade, WojPob, Bryan Derksen, The Anome, Ap, Verloren, Andre Engels, Fcueto, Matusz, Deb, Boleslav Bobcik, Hefaistos, Mswake, N8chz, Michael Hardy, Confusss, Fred Bauder, Isomorphic, Nixdorf, Kku, Dhart, Ixfd64, Lament, Alfio, CesarB, Ahoerstemeier, Haakon, Ronz, Angela, Den fjättrade ankan~enwiki, Netsnipe, Jfitzg, Tristanb, Hike395, Mydogategodshat, Dcoetzee, Andrevan, Jay, Fuzheado, WhisperToMe, Epic~enwiki, Tpbradbury, Furrykef, Traroth, Nickshanks, Joy, Shantavira, Pakcw, Robbot, ZimZalaBim, Altenmann, Henrygb, Ojigiri~enwiki, Sunray, Aetheling, Apogr~enwiki, Wile E. Heresiarch, Tea2min, Filemon, Adam78, Alan Liefting, Giftlite, ShaunMacPherson, Sepreece, Philwelch, Tom harrison, Jkseppan, Simon Lacoste-Julien, Ianhowlett, Varlaam, LarryGilbert, Kainaw, Siroxo, Adam McMaster, Just Another Dan, Neilc, Comatose51, Chowbok, Gadfium, Pgan002, Bolo1729, SarekOfVulcan, Raand, Antandrus, Onco p53, OverlordQ, Gscshoyru, Urhixidur, Kadambarid, Mike Rosoft, Monkeyman, KeyStroke, Rich Farmbrough, Nowozin, Stephenpace, Vitamin b, Bender235, Flyskippy1, Marner, Aaronbrick, Etz Haim, Janna Isabot, Mike Schwartz, John Vandenberg, Maurreen, Ejrrjs, Nsaa, Mdd, Alansohn, Gary, Walter Görlitz, Denoir, Rd232, Jeltz, Jet57, Jamiemac, Malo, Compo, Caesura, Axeman89, Vonaurum, Oleg Alexandrov, Jefgodesky, Nuno Tavares, OwenX, Woohookitty, Mindmatrix, Katyare, TigerShark, LOL, David Haslam, Ralf Mikut, GregorB, Hynespm, Essjay, MarcoTolo, Joerg Kurt Wegner, Simsong, Lovro, Tslocum, Graham87, Deltabeignet, BD2412, Kbdank71, DePiep, CoderGnome, Chenxlee, Sjakkalle, Rjwilmsi, Gmelli, Lavishluau, Michal.burda, Bubba73, Bensin, GeorgeBills, GregAsche, HughJorgan, Twerbrou, FlaBot, Emarsee, AlexAnglin, Ground Zero, Mathbot, Jrtayloriv, Predictor, Bmicomp, Compuneo, Vonkje, Gurubrahma, BMF81, Chobot, DVdm, Bgwhite, The Rambling Man, YurikBot, Wavelength, NTBot~enwiki, H005, Phantomsteve, AVM, Hede2000, Splash, SpuriousQ, Ansell, RadioFan, Hydrargyrum, Gaius Cornelius, Rsrikanth05, Philopedia, Bovineone, Zeno of Elea, EngineerScotty, NawlinWiki, Grafen, ONEder Boy, Mshecket, Aaron Brenneman, Jpbowen, Tony1, Dlyons493, DryaUnda, Bota47, Tlevine, Ripper234, Graciella, Deville, Zzuuzz, Lt-wiki-bot, Fang Aili, Pb30, Modify, GraemeL, Wikiant, JoanneB, LeonardoRob0t, ArielGold, Katieh5584, John Broughton, SkerHawx, Capitalist, Palapa, SmackBot, Looper5920, ThreeDee912, TestPilot, McGeddon, Unyoyega, Cutter, KocjoBot~enwiki, Bhikubhadwa, Thunderboltz, CommodiCast, Comp8956, Delldot, Eskimbot, Slhumph, Onebravemonkey, Ohnoitsjamie, Skizzik, Somewhere-purple, Leo505, MK8, Thumperward, DHN-bot~enwiki, Tdelamater, Antonrojo, Differentview, Janvo, Can't sleep, clown will eat me, Sergio.ballestrero, Frap, Nixeagle, Serenity-Fr, Thefriedone, JonHarder, Propheci, Joinarnold, Bennose, Macksseem~enwiki, Radagast83, Nibuod, Daqu, DueSouth, Blake-, Krexer, Weregerbil, Vina-iwbot~enwiki, Andrei Stroe, Deepred6502, Spiritia, Lambiam, Wikiolap, Kuru, Bmhkim, Vgy7ujm, Calum MacÙisdean, Athernar, Burakordu, Feraudyh, 16@r, Beetstra, Mr Stephen, Jimmy Pitt, Julthep, Dicklyon, Waggers, Ctacmo, RichardF, Nabeth, Beefyt, Hu12, Enggakshat, Vijay.babu.k, Ft93110, Dagoldman, Veyklevar, Ralf Klinkenberg, JHP, IvanLanin, Paul Foxworthy, Adrian.walker, Linkspamremover, CRGreathouse, CmdrObot, Filip*, Van helsing, Shorespirit, Matt1299, Kushal one, CWY2190, Ipeirotis, Nilfanion, Cydebot, Valodzka, Gogo Dodo, Ar5144-06, Akhil joey, Martin Jensen, Pingku, Oli2140, Mikeputnam, Talgalili, Malleus Fatuorum, Thijs!bot, Barticus88, Nirvanalulu, Drowne, Scientio, Kxlai, Headbomb, Ubuntu2, AntiVandalBot, Seaphoto, Ajaysathe, Gwyatt-agastle, Onasraou, Spencer, Alphachimpbot, JAnDbot, Wiki0709, Barek, Sarnholm, MER-C, The Transhumanist, Bull3t, TFinn734, Andonic, Mkch, Hut 8.5, Leiluo, Jguthaaz, EntropyAS, SiobhanHansa, Timdew, Dmmd123, Magioladitis, Connormah, Bongwarrior, VoABot II, Tedickey, Giggy, JJ Harrison, David Eppstein, Chivista~enwiki, Gomm, Pmbhagat, Fourthcourse, Kgfleischmann, RoboBaby, Quanticle, ERI employee, R'n'B, Jfroelich, Tgeairn, Pharaoh of the Wizards, Trusilver, Bongomatic, Roxy1984, Andres.santana, Shwapnil, DanDoughty, Foober, Ocarbone, RepubCarrier, Gzkn, AtholM, Salih, LordAnubisBOT, Starnestommy, Jmajeremy, A m sheldon, AntiSpamBot, LeighvsOptimvsMaximvs, Ramkumar.krishnan, Shoessss, Josephjthomas, Parikshit Basrur, Doug4, Cometstyles, DH85868993, DorganBot, Bonadea, WinterSpw, Mark.hornick, Andy Marchbanks, Yecril, BernardZ, RJASE1, Idioma-bot, RonFredericks, Black Kite, Jeff G., Jimmaths, DataExp, Philip Trueman, Adamminstead, TXiKiBoT, Oshwah, Deleet, Udufruduhu, Deanabb, Valerie928, TyrantX, OlavN, Arpabr, Vlad.gerchikov, Don4of4, Raymondwinn, Mannafredo, 1yesfan, Bearian, Jkosik1, Wykypydya, Billinghurst, Atannir, Hadleywickham, Hherbert, Falcon8765, Sebastjanmm, Monty845, Pjoef, Mattelsen, AlleborgoBot, Burkeangirl, FlyingLeopard2014, Rknasc, Pdfpdf, Equilibrioception, Calliopejen1, VerySmartNiceGuy, Euryalus, Dawn Bard, Estard, Srp33, Jerryobject, Kexpert, Mark Klamberg, Curuxz, Flyer22 Reborn, Eikoku, JCLately, Powtroll, Jpcedenog, Strife911, Pyromaniaman, Oxymoron83, Gpswiki, Dodabe~enwiki, Gargvikram07, Mátyás, Fratrep, Chrisguyot, Odo Benus, Stfg, Brylie, StaticGull, Sanya r, DixonD, Kjtobo, Melcombe, 48states, LaUs3r, Pinkadelica, Ypouliot, Denisarona, Sbacle, Kotsiantis, Loren.wilton, Sfan00 IMG, Nezza 4 eva, ClueBot, The Thing That Should Not Be, EoGuy, Supertouch, Mild Bill Hiccup, Kkarimi, Blanchardb, Edayapattiarun, Lbertolotti, Shaw76, Verticalsearch, Sebleouf, Hanifbbz, Abrech, Sterdeus, DrCroco, Nano5656, Aseld, Amossin, Dekisugi, Schreiber-Bike, DyingIce, Atallcostsky, 9Nak, Dank, Versus22, Katanada, Qwfp, DumZiBoT, Sunsetsky, XLinkBot, Articdawg, Cgfjpfg, Ecmalthouse, Little Mountain 5, WikHead, SilvonenBot, Badgernet, Foxyliah, Freestyle-69, Texterp, Addbot, DOI bot, Mabdul, Landon1980, Mhahsler, AndrewHZ, Elsendero, Matt90855, Jpoelma13, Cis411, Drkknightbatman, MrOllie, Download, RTG, M.r santosh kumar., Glane23, Delaszk, Chzz, Swift-Epic (Refectory), AtheWeatherman, Fauxstar, Jesuja, Luckas-bot, Yobot, Adelpine, Bunnyhop11, Ptbotgourou, Cflm001, Hulek, Alusayman, Ryanscraper, Carleas, Nallimbot, SOMart, Tiffany9027, AnomieBOT, Rjanag, Jim1138, JackieBot, Fahadsadah, ChristopheS, OptimisticCynic, Dudukeda, Materialscientist, Citation bot, Schul253, Cureden, Capricorn42, Gtfjbl, Lark137, Liwaste, The Evil IP address, Tomwsulcer, BluePlateSpecial, Dr Oldekop, Rosannel, Rugaaad, RibotBOT, Charvest, Tareq300, Cmccormick8, Smallman12q, Lbcao, Andrzejrauch, Davgrig04, Stekre, Whizzdumb, Thehelpfulbot, Kyleamiller, OlafvanD, FrescoBot, Mark Renier, Ph92, W Nowicki, X7q, Colewaldron, Er.piyushkp, HamburgerRadio, Atlantia, Webzie, Citation bot 1, Killian441, Manufan 11, Rustyspatula, Pinethicket, Guerrerocarlos, Toohuman1, BRUTE, Elsevierediterormath, Stpasha, MastiBot, SpaceFlight89, Jackverr, UngerJ, Juliustch, Priyank782, TobeBot, Pamparam, Btcoal, Kmettler, Jonkerz, GregKaye, Glenn Maddox, Jayrde, Angelorf, Reaper Eternal, Chenzheruc, Pmauer, DARTH SIDIOUS 2, Mean as custard, RjwilmsiBot, Mike78465, D vandyke67, Ripchip Bot, Slon02, Aaronzat, Helwr, Ericmortenson, EmausBot, Acather96, BillyPreset, Fly by Night, Primefac, WirlWhind, GoingBatty, Emilescheepers444, Stheodor, Lawrykid, Uploadvirus, Wikipelli, Dcirovic, Joanlofe, Anir1uph, Chire, Cronk28, Zedutchgandalf, Vangelis12, T789, Rick jens, Donner60, Terryholmsby, MainFrame, Phoglenix, Raomohsinkhan, ClueBot NG, Mathstat, Aiwing, Nuwanmenuka, Statethatiamin, CherryX, Candace Gillhoolley, Robiminer, Leonardo61, Twillisjr, Widr, WikiMSL, Luke145, EvaJamax, Debuntu, Helpful Pixie Bot, AlbertoBetulla, HMSSolent, Ngorman, Inoshika, Data.mining, ErinRea, BG19bot, Wanming149, Northamerica1000, PhnomPencil, Lisasolomonsalford, Uksas, Naeemmalik036, Marcocapelle, Chafe66, Onewhohelps, Netra Nahar, Aranea Mortem, Jasonem, Flaticida, Funkykeith777, Moshiurbd, Nathanashleywild, Anilkumar 0587, Mpaye, Rabarbaro70, Thundertide, BattyBot, Aacruzr, Warrenxu, IjonTichyIjonTichy, Harsh 2580, Dexbot, Webclient101, Mogism, TwoTwoHello, Frosty, Bradhill14, 7376a73b3bf0a490fa04bea6b76f4a4b, L8fortee, Dougs campbell, Mark viking, Cmartines, Phamnhatkhanh, Epicgenius, THill182, Delafé, Melonkelon, Herpderp1235689999, Revengetechy, Amykam32, The hello doctor, Blythwood, Mimarios1, Huang cynthia, NYBrook098, DavidLeighEllis, Gnust, Rbrandon87, Astigitana, Alihaghi, Philip Habing, Wccsnow, Jianhui67, Tahmina.tithi, Yeda123, Skr15081997, Charlotth, Jfrench7, Zjl9191, Davidhart007, Routerdecomposer, Waggie, Augt.pelle, Justincahoon, Gstoel, Wiki-jonne, MatthewP42, 115ash, LiberumConsilium, Ran0512, Daniel Bachar,

Galaktikasoft, Prof PD Hoy, GoldCoastPrior, Gary2015, HelpUsStopSpam, W Ansari, KasparBot, GrahamJClark, Baharsahu, Hillbilly Dragon Farmer, Howardbm, Cfeng97, Mahda133, JonFredriksen, Randomuser959, Kellyshintaku, Nishant.23j, Fmadd, Nnavarro unomaha.edu, Fatisara2016, Foram karia and Anonymous: 1013

## 14.2 Images

- **File:Commons-logo.svg** *Source:* https://upload.wikimedia.org/wikipedia/en/4/4a/Commons-logo.svg *License:* CC-BY-SA-3.0 *Contributors:* ? *Original artist:* ?
- **File:Internet_map_1024.jpg** *Source:* https://upload.wikimedia.org/wikipedia/commons/d/d2/Internet_map_1024.jpg *License:* CC BY 2.5 *Contributors:* Originally from the English Wikipedia; description page is/was here. *Original artist:* The Opte Project
- **File:Portal-puzzle.svg** *Source:* https://upload.wikimedia.org/wikipedia/en/f/fd/Portal-puzzle.svg *License:* Public domain *Contributors:* ? *Original artist:* ?
- **File:Spurious_correlations_-_spelling_bee_spiders.svg** *Source:* https://upload.wikimedia.org/wikipedia/commons/0/0c/Spurious_correlations_-_spelling_bee_spiders.svg *License:* CC BY 4.0 *Contributors:* Spurious Correlations website *Original artist:* Tyler Viglen
- **File:Wiki_letter_w.svg** *Source:* https://upload.wikimedia.org/wikipedia/en/6/6c/Wiki_letter_w.svg *License:* Cc-by-sa-3.0 *Contributors:* ? *Original artist:* ?

## 14.3 Content license

- Creative Commons Attribution-Share Alike 3.0