Data Mining

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

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD),[1] an interdisciplinary subfield of computer science,[2][3][4] 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.[2] 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.[2] 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.[2]

The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amount of data, not the extraction 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 popular 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", or "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 are 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.

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[11] 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.

2.1 Research and evolution

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).[12][13] Since 1989 this ACM SIG has hosted an annual international conference and published its proceedings,[14] and since 1999 it has published a biannual academic journal titled "SIGKDD Explorations".[15]

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
- 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
- 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

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.[1]

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, and 2007 show that the CRISP-DM methodology is the leading methodology used by data miners.[16][17][18] 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,[19][20] and Azevedo and Santos conducted a comparison of CRISP-DM and SEMMA in 2008.[21]

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:[1]

- 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 do not actually predict future behavior 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 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.[22]

5 Notable uses

5.1 Games

Since the early 1960s, with the availability of oracles for certain combinatorial games, also called tablebases (e.g. for 3x3-chess) with any beginning configuration, small-board dots-and-boxes, small-board-hex, and certain endgames in chess, dots-and-boxes, and hex; a new area for data mining has been opened. This is the extraction of human-usable strategies from these oracles. Current pattern recognition approaches do not seem to fully acquire the high level of abstraction required to be applied successfully. Instead, extensive experimentation with the tablebases – combined with an intensive study of tablebase-answers to well designed problems, and with knowledge of prior art (i.e., pre-tablebase knowledge) – is used to yield insightful patterns. Berlekamp (in dots-and-boxes, etc.) and John Nunn (in chess endgames) are notable examples of researchers doing this work, though they were not – and are not – involved in tablebase generation.

5.2 Business

In business, data mining is the analysis of historical business activities, stored as static data in data warehouse databases. The goal is to reveal hidden patterns and trends. Data mining software uses advanced pattern recognition algorithms to sift through large amounts of data to assist in discovering previously unknown strategic business information. Examples of what businesses use data mining for include performing market analysis to identify new product bundles, finding the root cause of manufacturing problems, to prevent customer attrition and acquire new customers, cross-selling to existing customers, and profiling customers with more accuracy.[23]

- In today's world raw data is being collected by companies at an exploding rate. For example, Walmart processes over 20 million point-of-sale transactions every day. This information is stored in a centralized database, but would be useless without some type of data mining software to analyze it. If Walmart analyzed their point-of-sale data with data mining techniques they would be able to determine sales trends, develop marketing campaigns, and more accurately predict customer loyalty.[24]

- Every time a credit card or a store loyalty card is being used, or a warranty card is being filled, data is being collected about the users behavior. Many people find the amount of information stored about us from companies, such as Google, Facebook, and Amazon, disturbing and are concerned about privacy. Although there is the potential for our personal data to be used in harmful, or unwanted, ways it is also being used to make our lives better. For example, Ford and Audi hope to one day collect information about customer driving patterns so they can recommend safer routes and warn drivers about dangerous road conditions.[25]

- Data mining in customer relationship management applications can contribute significantly to the bottom line. Rather than randomly contacting a prospect or customer through a call center or sending mail, a company can concentrate its efforts on prospects that are predicted to have a high likelihood of responding to an offer. More sophisticated methods may be used to optimize resources across campaigns so that one may predict to which channel and to which offer an individual is most likely to respond (across all potential offers). Additionally, sophisticated applications could be used to automate mailing. Once the results from data mining (potential prospect/customer and channel/offer) are determined, this "sophisticated application" can either automatically send an e-mail or a regular mail. Finally, in cases where many people will take an action without an offer, "uplift modeling" can be used to determine which people have the greatest increase in response if given an offer. Uplift modeling thereby enables marketers to focus mailings and offers on persuadable people, and not to send offers to people who will buy the product without an offer. Data clustering can also be used to automatically discover the segments or groups within a customer data set.

- Businesses employing data mining may see a return on investment, but also they recognize that the number of predictive models can quickly become very large. For example, rather than using one model to predict how many customers will churn, a business may choose to build a separate model for each region and customer type. In situations where a large number of models need to be maintained, some businesses turn to more automated data mining methodologies.

- Data mining can be helpful to human resources (HR) departments in identifying the characteristics of their most successful employees. Information obtained – such as universities attended by highly successful employees – can help HR focus recruiting efforts accordingly. Additionally, Strategic Enterprise Management applications help a company translate corporate-level goals, such as profit and margin share targets, into operational decisions, such as production plans and workforce levels.[26]

- Market basket analysis, relates to data-mining use in retail sales. If a clothing store records the purchases of customers, a data mining system could identify those customers who favor silk shirts over cotton ones. Although some explanations of relationships may be difficult, taking advantage of it is easier. The example deals with association rules within transaction-based data. Not all data are transaction based and logical, or inexact rules may also be present within a database.

- Market basket analysis has been used to identify the purchase patterns of the Alpha Consumer. Analyzing the data collected on this type of user has allowed companies to predict future buying trends and forecast supply demands.

- Data mining is a highly effective tool in the catalog marketing industry. Catalogers have a rich database of history of their customer transactions for millions of customers dating back a number of years. Data mining tools can identify patterns among customers and help identify the most likely customers to respond to upcoming mailing campaigns.

- Data mining for business applications can be integrated into a complex modeling and decision making process.[27] Reactive business intelligence (RBI) advocates a "holistic" approach that integrates data mining, modeling, and interactive visualization into an end-to-end discovery and continuous innovation process powered by human and automated learning.[28]

- In the area of decision making, the RBI approach has been used to mine knowledge that is progressively acquired from the decision maker, and then self-tune the decision method accordingly.[29] The relation between the quality of a data mining system and the amount of investment that the decision maker is willing to make was formalized by providing an economic perspective on the value of "extracted knowledge" in terms of its payoff to the organization[27] This decision-theoretic classification framework[27] was applied to a real-world semiconductor wafer manufacturing line, where decision rules for effectively monitoring and controlling the semiconductor wafer fabrication line were developed.[30]

- An example of data mining related to an integrated-circuit (IC) production line is described in the paper "Mining IC Test Data to Optimize VLSI Testing."[31] In this paper, the application of data mining and decision analysis to the problem of die-level functional testing is described. Experiments mentioned demonstrate the ability to apply a system of mining historical die-test data to create a probabilistic model of patterns of die failure. These patterns are then utilized to decide, in real time, which die to test next and when to stop testing. This system has been shown, based on experiments with historical test data, to have the potential to improve profits on mature IC products. Other examples[32][33] of the application of data mining methodologies in semiconductor manufacturing environments suggest that data mining methodologies may be particularly useful when data is scarce, and the various physical and chemical parameters that affect the process exhibit highly complex interactions. Another implication is that on-line monitoring of the semiconductor manufacturing process using data mining may be highly effective.

5.3 Science and engineering

In recent years, data mining has been used widely in the areas of science and engineering, such as bioinformatics, genetics, medicine, education and electrical power engineering.

- In the study of human genetics, sequence mining helps address the important goal of understanding the mapping relationship between the inter-individual variations in human DNA sequence and the variability in disease susceptibility. In simple terms, it aims to find out how the changes in an individual's DNA sequence affects the risks of developing common diseases such as cancer, which is of great importance to improving methods of diagnosing, preventing, and treating these diseases. One data mining method that is used to perform this task is known as multifactor dimensionality reduction.[34]

- In the area of electrical power engineering, data mining methods have been widely used for condition monitoring of high voltage electrical equipment. The purpose of condition monitoring is to obtain valuable information on, for example, the status of the insulation (or other important safety-related parameters). Data clustering techniques – such as the self-organizing map (SOM), have been applied to vibration monitoring and analysis of transformer on-load tap-changers (OLTCS). Using vibration monitoring, it can be observed that each tap change operation generates a signal that contains information about the condition of the tap changer contacts and the drive mechanisms. Obviously, different tap positions will generate different signals. However, there was considerable variability amongst normal condition signals for exactly the same tap position. SOM has been applied to detect abnormal conditions and to hypothesize about the nature of the abnormalities.[35]

- Data mining methods have been applied to dissolved gas analysis (DGA) in power transformers. DGA, as a diagnostics for power transformers, has been available for many years. Methods such as SOM has been applied to analyze generated data and to determine trends which are not obvious to the standard DGA ratio methods (such as Duval Triangle).[35]

- In educational research, where data mining has been used to study the factors leading students to choose to engage in behaviors which reduce their learning,[36] and to understand factors influencing university student retention.[37] A similar example of social application of data mining is its use in expertise finding systems, whereby descriptors of human expertise are extracted, normalized, and classified so as to facilitate the finding of experts, particularly in scientific and technical fields. In this way, data mining can facilitate institutional memory.

- Data mining methods of biomedical data facilitated by domain ontologies,[38] mining clinical trial data,[39] and traffic analysis using SOM.[40]

- In adverse drug reaction surveillance, the Uppsala Monitoring Centre has, since 1998, used data mining methods to routinely screen for reporting patterns indicative of emerging drug safety issues in the WHO global database of 4.6 million suspected adverse drug reaction incidents.[41] Recently, similar methodology has been developed to mine large collections of electronic health records for temporal patterns associating drug prescriptions to medical diagnoses.[42]

- Data mining has been applied to software artifacts within the realm of software engineering: Mining Software Repositories.

5.4 Human rights

Data mining of government records – particularly records of the justice system (i.e., courts, prisons) – enables the discovery of systemic human rights violations in connection to generation and publication of invalid or fraudulent legal records by various government agencies.[43][44]

5.5 Medical data mining

In 2011, the case of Sorrell v. IMS Health, Inc., decided by the Supreme Court of the United States, ruled that pharmacies may share information with outside companies. This practice was authorized under the 1st Amendment of the Constitution, protecting the “freedom of speech.”[45] However, the passage of the Health Information Technology for Economic and Clinical Health Act (HITECH Act) helped to initiate the adoption of the electronic health record (EHR) and supporting technology in the United States.[46] The HITECH Act was signed into law on February 17, 2009 as part of the American Recovery and Reinvestment Act (ARRA) and helped to open the door to medical data mining.[47] Prior to the signing of this law, estimates of only 20% of United States based physicians were utilizing electronic patient records.[46] Søren Brunak notes that “the patient record becomes as information-rich as possible” and thereby “maximizes the data mining opportunities.”[46] Hence, electronic patient records further expands the possibilities regarding medical data mining thereby opening the door to a vast source of medical data analysis.

5.6 Spatial data mining

Spatial data mining is the application of data mining methods to spatial data. The end objective of spatial data mining is to find patterns in data with respect to geography. So far, data mining and Geographic Information Systems (GIS) have existed as two separate technologies, each with its own methods, traditions, and approaches to visualization and data analysis. Particularly, most contemporary GIS have only very basic spatial analysis functionality. The immense explosion in geographically referenced data occasioned by developments in IT, digital mapping, remote sensing, and the global diffusion of GIS emphasizes the importance of developing data-driven inductive approaches to geographical analysis and modeling.

Data mining offers great potential benefits for GIS-based applied decision-making. Recently, the task of integrating these two technologies has become of critical importance, especially as various public and private sector organizations possessing huge databases with thematic and geographically referenced data begin to realize the huge potential of the information contained therein. Among those organizations are:

- offices requiring analysis or dissemination of geo-referenced statistical data
- public health services searching for explanations of disease clustering
- environmental agencies assessing the impact of changing land-use patterns on climate change
- geo-marketing companies doing customer segmentation based on spatial location.

Challenges in Spatial mining: Geospatial data repositories tend to be very large. Moreover, existing GIS datasets are often splintered into feature and attribute components that are conventionally archived in hybrid data management systems. Algorithmic requirements differ substantially for relational (attribute) data management and for topological (feature) data management.[48] Related to this is the range and diversity of geographic data formats, which present unique challenges. The digital geographic data revolution is creating new types of data formats beyond the traditional “vector” and “raster” formats. Geographic data repositories increasingly include ill-structured data, such as imagery and geo-referenced multi-media.[49]

There are several critical research challenges in geographic knowledge discovery and data mining. Miller and Han[50] offer the following list of emerging research topics in the field:

- **Developing and supporting geographic data warehouses (GDW’s)**: Spatial properties are often reduced to simple aspatial attributes in mainstream data warehouses. Creating an integrated GDW requires solving issues of spatial and temporal data interoperability – including differences in semantics, referencing systems, geometry, accuracy, and position.

- **Better spatio-temporal representations in geographic knowledge discovery**: Current geographic knowledge discovery (GKD) methods generally use very simple representations of geographic objects and spatial relationships. Geographic data mining methods should recognize more complex geographic objects (i.e., lines and polygons) and relationships (i.e., non-Euclidean distances, direction, connectivity, and interaction through attributed geographic space such as terrain). Furthermore, the time dimension needs to be more fully integrated into these geographic representations and relationships.

- **Geographic knowledge discovery using diverse data types**: GKD methods should be developed that can handle diverse data types beyond the traditional raster and vector models, including imagery and geo-referenced multimedia, as well as dynamic data types (video streams, animation).

5.7 Temporal data mining

Data may contain attributes generated and recorded at different times. In this case finding meaningful relationships in the data may require considering the temporal order of the attributes. A temporal relationship may indicate a causal relationship, or simply an association.

5.8 Sensor data mining

Wireless sensor networks can be used for facilitating the collection of data for spatial data mining for a variety of applications such as air pollution monitoring.[51] A characteristic of such networks is that nearby sensor nodes monitoring an environmental feature typically register similar values. This kind of data redundancy due to the spatial correlation between sensor observations inspires the techniques for in-network data aggregation and mining. By measuring the spatial correlation between data sampled by different sensors, a wide class of specialized algorithms can be developed to develop more efficient spatial data mining algorithms.[52]

5.9 Visual data mining

In the process of turning from analogical into digital, large data sets have been generated, collected, and stored discovering statistical patterns, trends and information which is hidden in data, in order to build predictive patterns. Studies suggest visual data mining is faster and much more intuitive than is traditional data mining.[53][54][55] See also Computer vision.

5.10 Music data mining

Data mining techniques, and in particular co-occurrence analysis, has been used to discover relevant similarities among music corpora (radio lists, CD databases) for purposes including classifying music into genres in a more objective manner.[56]

5.11 Surveillance

Data mining has been used by the U.S. government. Programs include the Total Information Awareness (TIA) program, Secure Flight (formerly known as Computer-Assisted Passenger Prescreening System (CAPPS II)), Analysis, Dissemination, Visualization, Insight, Semantic Enhancement (ADVISE),[57] and the Multi-state Anti-Terrorism Information Exchange (MATRIX).[58] These programs have been discontinued due to controversy over whether they violate the 4th Amendment to the United States Constitution, although many programs that were formed under them continue to be funded by different organizations or under different names.[59]

In the context of combating terrorism, two particularly plausible methods of data mining are “pattern mining” and “subject-based data mining”.

5.12 Pattern mining

“Pattern mining” is a data mining method that involves finding existing patterns in data. In this context *patterns* often means association rules. The original motivation for searching association rules came from the desire to analyze supermarket transaction data, that is, to examine customer behavior in terms of the purchased products. For example, an association rule “beer ⇒ potato chips (80%)" states that four out of five customers that bought beer also bought potato chips.

In the context of pattern mining as a tool to identify terrorist activity, the National Research Council provides the following definition: “Pattern-based data mining looks for patterns (including anomalous data patterns) that might be associated with terrorist activity — these patterns might be regarded as small signals in a large ocean of noise.”[60][61][62] Pattern Mining includes new areas such a Music Information Retrieval (MIR) where patterns seen both in the temporal and non temporal domains are imported to classical knowledge discovery search methods.

5.13 Subject-based data mining

"Subject-based data mining" is a data mining method involving the search for associations between individuals in data. In the context of combating terrorism, the National Research Council provides the following definition: "Subject-based data mining uses an initiating individual or other datum that is considered, based on other information, to be of high interest, and the goal is to determine what other persons or financial transactions or movements, etc., are related to that initiating datum."[61]

5.14 Knowledge grid

Knowledge discovery "On the Grid" generally refers to conducting knowledge discovery in an open environment using grid computing concepts, allowing users to integrate data from various online data sources, as well make use of remote resources, for executing their data mining tasks. The earliest example was the Discovery Net,[63][64] developed at Imperial College London, which won the "Most Innovative Data-Intensive Application Award" at the ACM SC02 (Supercomputing 2002) conference and exhibition, based on a demonstration of a fully interactive distributed knowledge discovery application for a bioinformatics application. Other examples include work conducted by researchers at the University of Calabria, who developed a Knowledge Grid architecture for distributed knowledge discovery, based on grid computing.[65][66]

6 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).[67]

The ways in which data mining can be used can in some cases and contexts raise questions regarding privacy, legality, and ethics.[68] 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.[69][70]

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).[71] 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.[72][73][74]

It is recommended that an individual is made aware of the following **before** data are collected:[71]

- 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.[71] 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.[75]

6.1 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.

6.2 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.'*[76] *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.

7 Copyright Law

7.1 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[77] 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.[78] 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.[79]

7.2 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.[80]

8 Software

8.1 Free open-source data mining software and applications

- 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.
- 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.
- Massive Online Analysis (MOA): a real-time big data stream mining with concept drift tool in the Java programming 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.
- RapidMiner: An environment for machine learning and data mining experiments.
- SCaViS: Java cross-platform data analysis framework developed at Argonne National Laboratory.
- SenticNet API: A semantic and affective resource for opinion mining and sentiment analysis.
- Tanagra: A visualisation-oriented data mining software, also for teaching.
- 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.

8.2 Commercial data-mining software and applications

- 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.
- Grapheme: data mining and visualization software provided by iChrome.
- LIONsolver: an integrated software application for data mining, business intelligence, and modeling that implements the Learning and Intelligent OptimizatioN (LION) approach.
- Microsoft Analysis Services: data mining software provided by Microsoft.
- NetOwl: suite of multilingual text and entity analytics products that enable data mining.
- Oracle Data Mining: data mining software by Oracle.
- SAS Enterprise Miner: data mining software provided by the SAS Institute.
- STATISTICA Data Miner: data mining software provided by StatSoft.
- Qlucore Omics Explorer: data mining software provided by Qlucore.

8.3 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:

- 2011 Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery[81]
- Rexer Analytics Data Miner Surveys (2007–2013)[82]
- Forrester Research 2010 Predictive Analytics and Data Mining Solutions report[83]
- Gartner 2008 “Magic Quadrant” report[84]
- Robert A. Nisbet’s 2006 Three Part Series of articles “Data Mining Tools: Which One is Best For CRM?"[85]
- Haughton et al.'s 2003 Review of Data Mining Software Packages in *The American Statistician*[86]
- Goebel & Gruenwald 1999 “A Survey of Data Mining a Knowledge Discovery Software Tools” in SIGKDD Explorations[87]

9 See also

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
- Online analytical processing (OLAP)

Application domains

- Analytics
- Bioinformatics
- Business intelligence
- Data analysis
- Data warehouse
- Decision support system

- 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
- Information extraction
- Information integration
- Named-entity recognition
- Profiling (information science)
- Web scraping

10 References

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

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

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

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

[5] Han, Jiawei; Kamber, Micheline (2001). *Data mining: concepts and techniques*. Morgan Kaufmann. p. 5. ISBN 9781558604896. Thus, data mining should habe 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] Kantardzic, Mehmed (2003). *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons. ISBN 0-471-22852-4. OCLC 50055336.

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

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

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

[15] SIGKDD Explorations, ACM, New York.

[16] Gregory Piatetsky-Shapiro (2002) *KDnuggets Methodology Poll*

[17] Gregory Piatetsky-Shapiro (2004) *KDnuggets Methodology Poll*

[18] Gregory Piatetsky-Shapiro (2007) *KDnuggets Methodology Poll*

[19] 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.

[20] 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

[21] 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.

[22] 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 9781450308373.

[23] O'Brien, J. A., & Marakas, G. M. (2011). Management Information Systems. New York, NY: McGraw-Hill/Irwin.

[24] Alexander, D. (n.d.). Data Mining. Retrieved from The University of Texas at Austin: College of Liberal Arts: http://www.laits.utexas.edu/~{}anorman/BUS.FOR/course.mat/Alex/

[25] Goss, S. (2013, April 10). Data-mining and our personal privacy. Retrieved from The Telegraph: http://www.macon.com/2013/04/10/2429775/data-mining-and-our-personal-privacy.html

[26] Monk, Ellen; Wagner, Bret (2006). *Concepts in Enterprise Resource Planning, Second Edition*. Boston, MA: Thomson Course Technology. ISBN 0-619-21663-8. OCLC 224465825.

[27] Elovici, Yuval; Braha, Dan (2003). "A Decision-Theoretic Approach to Data Mining" (PDF). *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans* **33** (1).

[28] Battiti, Roberto; and Brunato, Mauro; *Reactive Business Intelligence. From Data to Models to Insight*, Reactive Search Srl, Italy, February 2011. ISBN 978-88-905795-0-9.

[29] Battiti, Roberto; Passerini, Andrea (2010). "Brain-Computer Evolutionary Multi-Objective Optimization (BC-EMO): a genetic algorithm adapting to the decision maker". *IEEE Transactions on Evolutionary Computation* **14** (15): 671–687. doi:10.1109/TEVC.2010.2058118.

[30] Braha, Dan; Elovici, Yuval; Last, Mark (2007). "Theory of actionable data mining with application to semiconductor manufacturing control" (PDF). *International Journal of Production Research* **45** (13).

[31] Fountain, Tony; Dietterich, Thomas; and Sudyka, Bill (2000); *Mining IC Test Data to Optimize VLSI Testing*, in Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM Press, pp. 18–25

[32] Braha, Dan; Shmilovici, Armin (2002). "Data Mining for Improving a Cleaning Process in the Semiconductor Industry" (PDF). *IEEE Transactions on Semiconductor Manufacturing* **15** (1).

[33] Braha, Dan; Shmilovici, Armin (2003). "On the Use of Decision Tree Induction for Discovery of Interactions in a Photolithographic Process" (PDF). *IEEE Transactions on Semiconductor Manufacturing* **16** (4).

[34] Zhu, Xingquan; Davidson, Ian (2007). *Knowledge Discovery and Data Mining: Challenges and Realities*. New York, NY: Hershey. p. 18. ISBN 978-1-59904-252-7.

[35] McGrail, Anthony J.; Gulski, Edward; Allan, David; Birtwhistle, David; Blackburn, Trevor R.; Groot, Edwin R. S. "Data Mining Techniques to Assess the Condition of High Voltage Electrical Plant". *CIGRÉ WG 15.11 of Study Committee 15*.

[36] Baker, Ryan S. J. d. "Is Gaming the System State-or-Trait? Educational Data Mining Through the Multi-Contextual Application of a Validated Behavioral Model". *Workshop on Data Mining for User Modeling 2007*.

[37] Superby Aguirre, Juan Francisco; Vandamme, Jean-Philippe; Meskens, Nadine. "Determination of factors influencing the achievement of the first-year university students using data mining methods". *Workshop on Educational Data Mining 2006*.

[38] Zhu, Xingquan; Davidson, Ian (2007). *Knowledge Discovery and Data Mining: Challenges and Realities*. New York, NY: Hershey. pp. 163–189. ISBN 978-1-59904-252-7.

[39] Zhu, Xingquan; Davidson, Ian (2007). *Knowledge Discovery and Data Mining: Challenges and Realities*. New York, NY: Hershey. pp. 31–48. ISBN 978-1-59904-252-7.

[40] Chen, Yudong; Zhang, Yi; Hu, Jianming; Li, Xiang (2006). "Traffic Data Analysis Using Kernel PCA and Self-Organizing Map". *IEEE Intelligent Vehicles Symposium*.

[41] Bate, Andrew; Lindquist, Marie; Edwards, I. Ralph; Olsson, Sten; Orre, Roland; Lansner, Anders; de Freitas, Rogelio Melhado (Jun 1998). "A Bayesian neural network method for adverse drug reaction signal generation" (PDF). *European Journal of Clinical Pharmacology* **54** (4): 315–21. doi:10.1007/s002280050466. PMID 9696956.

[42] Norén, G. Niklas; Bate, Andrew; Hopstadius, Johan; Star, Kristina; and Edwards, I. Ralph (2008); Temporal Pattern Discovery for Trends and Transient Effects: Its Application to Patient Records. *Proceedings of the Fourteenth International Conference on Knowledge Discovery and Data Mining (SIGKDD 2008), Las Vegas, NV*, pp. 963–971.

[43] Zernik, Joseph; Data Mining as a Civic Duty – Online Public Prisoners' Registration Systems, *International Journal on Social Media: Monitoring, Measurement, Mining*, 1: 84–96 (2010)

[44] Zernik, Joseph; Data Mining of Online Judicial Records of the Networked US Federal Courts, *International Journal on Social Media: Monitoring, Measurement, Mining*, 1:69–83 (2010)

[45] David G. Savage (2011-06-24). "Pharmaceutical industry: Supreme Court sides with pharmaceutical industry in two decisions". *Los Angeles Times*. Retrieved 2012-11-07.

[46] Analyzing Medical Data. (2012). *Communications of the ACM* 55(6), 13-15. doi:10.1145/2184319.2184324

[47] http://searchhealthit.techtarget.com/definition/HITECH-Act

[48] Healey, Richard G. (1991); *Database Management Systems*, in Maguire, David J.; Goodchild, Michael F.; and Rhind, David W., (eds.), *Geographic Information Systems: Principles and Applications*, London, GB: Longman

[49] Camara, Antonio S.; and Raper, Jonathan (eds.) (1999); *Spatial Multimedia and Virtual Reality*, London, GB: Taylor and Francis

[50] Miller, Harvey J.; and Han, Jiawei (eds.) (2001); *Geographic Data Mining and Knowledge Discovery*, London, GB: Taylor & Francis

[51] Ma, Y.; Richards, M.; Ghanem, M.; Guo, Y.; Hassard, J. (2008). "Air Pollution Monitoring and Mining Based on Sensor Grid in London". *Sensors* **8** (6): 3601. doi:10.3390/s8063601.

[52] Ma, Y.; Guo, Y.; Tian, X.; Ghanem, M. (2011). "Distributed Clustering-Based Aggregation Algorithm for Spatial Correlated Sensor Networks". *IEEE Sensors Journal* **11** (3): 641. doi:10.1109/JSEN.2010.2056916.

[53] Zhao, Kaidi; and Liu, Bing; Tirpark, Thomas M.; and Weimin, Xiao; *A Visual Data Mining Framework for Convenient Identification of Useful Knowledge*

[54] Keim, Daniel A.; *Information Visualization and Visual Data Mining*

[55] Burch, Michael; Diehl, Stephan; Weißgerber, Peter; *Visual Data Mining in Software Archives*

[56] Pachet, François; Westermann, Gert; and Laigre, Damien; *Musical Data Mining for Electronic Music Distribution*, Proceedings of the 1st WedelMusic Conference,Firenze, Italy, 2001, pp. 101–106.

[57] Government Accountability Office, *Data Mining: Early Attention to Privacy in Developing a Key DHS Program Could Reduce Risks*, GAO-07-293 (February 2007), Washington, DC

[58] Secure Flight Program report, MSNBC

[59] "Total/Terrorism Information Awareness (TIA): Is It Truly Dead?". *Electronic Frontier Foundation (official website)*. 2003. Retrieved 2009-03-15.

[60] Agrawal, Rakesh; Mannila, Heikki; Srikant, Ramakrishnan; Toivonen, Hannu; and Verkamo, A. Inkeri; *Fast discovery of association rules*, in *Advances in knowledge discovery and data mining*, MIT Press, 1996, pp. 307–328

[61] National Research Council, *Protecting Individual Privacy in the Struggle Against Terrorists: A Framework for Program Assessment*, Washington, DC: National Academies Press, 2008

[62] Haag, Stephen; Cummings, Maeve; Phillips, Amy (2006). *Management Information Systems for the information age*. Toronto: McGraw-Hill Ryerson. p. 28. ISBN 0-07-095569-7. OCLC 63194770.

[63] Ghanem, Moustafa; Guo, Yike; Rowe, Anthony; Wendel, Patrick (2002). "Grid-based knowledge discovery services for high throughput informatics". *Proceedings 11th IEEE International Symposium on High Performance Distributed Computing*. p. 416. doi:10.1109/HPDC.2002.1029946. ISBN 0-7695-1686-6.

[64] Ghanem, Moustafa; Curcin, Vasa; Wendel, Patrick; Guo, Yike (2009). "Building and Using Analytical Workflows in Discovery Net". *Data Mining Techniques in Grid Computing Environments*. p. 119. doi:10.1002/9780470699904.ch8. ISBN 9780470699904.

[65] Cannataro, Mario; Talia, Domenico (January 2003). "The Knowledge Grid: An Architecture for Distributed Knowledge Discovery". *Communications of the ACM* **46** (1): 89–93. doi:10.1145/602421.602425. Retrieved 17 October 2011.

[66] Talia, Domenico; Trunfio, Paolo (July 2010). "How distributed data mining tasks can thrive as knowledge services". *Communications of the ACM* **53** (7): 132–137. doi:10.1145/1785414.1785451. Retrieved 17 October 2011.

[67] Seltzer, William. "The Promise and Pitfalls of Data Mining: Ethical Issues".

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

[69] 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.

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

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

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

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

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

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

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

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

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

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

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

[81] 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.

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

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

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

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

[86] 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

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

11 Further reading

- Cabena, Peter; Hadjnian, Pablo; Stadler, Rolf; Verhees, Jaap; and 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; and Sanger, James; *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

12 External links

13 Text and image sources, contributors, and licenses

13.1 Text

- **Data mining** *Source:* http://en.wikipedia.org/wiki/Data%20mining?oldid=649387557 *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, Dhart, Ixfd64, Lament, Alfio, CesarB, Ahoerstemeier, Haakon, Ronz, Angela, Den fjättrade ankan, Netsnipe, Jfitzg, Tristanb, Hike395, Mydogategodshat, Dcoetzee, Andrevan, Jay, Fuzheado, WhisperToMe, Epic, Tpbradbury, Furrykef, Traroth, Nickshanks, Joy, Shantavira, Pakcw, Robbot, ZimZalaBim, Altenmann, Henrygb, Ojigiri, Sunray, Aetheling, Apogr, Wile E. Heresiarch, Tobias Bergemann, Filemon, Adam78, Alan Liefting, Giftlite, ShaunMacPherson, Sepreece, Philwelch, Tom harrison, Jks, 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, H005, Phantomsteve, AVM, Hede2000, Splash, SpuriousQ, Ansell, RadioFan, Hydrargyrum, Gaius Cornelius, 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, Unyoyega, Cutter, KocjoBot, Bhikubhadwa, Thunderboltz, CommodiCast, Comp8956, Delldot, Eskimbot, Slhumph, Onebravemonkey, Ohnoitsjamie, Skizzik, Somewherepurple, Leo505, MK8, Thumperward, DHN-bot, Tdelamater, Antonrojo, Differentview, Janvo, Can't sleep, clown will eat me, Sergio.ballestrero, Frap, Nixeagle, Serenity-Fr, Thefriedone, JonHarder, Propheci, Joinarnold, Bennose, Mackseem, Radagast83, Nibuod, Daqu, DueSouth, Blake-, Krexer, Weregerbil, Vina-iwbot, 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, Connormah, Bongwarrior, VoABot II, Tedickey, Giggy, JJ Harrison, David Eppstein, Chivista, 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, Jeff G., Jimmaths, DataExp, Philip Trueman, Adamminstead, TXiKiBoT, Deleet, Udufruduhu, Deanabb, Valerie928, TyrantX, OlavN, Arpabr, Vlad.gerchikov, Don4of4, Raymondwinn, Mannafredo, 1yesfan, Bearian, Jkosik1, Wykypydya, Billinghurst, Atannir, Hadleywickham, Hherbert, Falcon8765, Sebastjanmm, Pjoef, Mattelsen, AlleborgoBot, Burkeangirl, NHRHS2010, Rknasc, Pdfpdf, Equilibrioception, Calliopejen1, VerySmartNiceGuy, Euryalus, Dawn Bard, Estard, Srp33, Jerryobject, Kexpert, Mark Klamberg, Curuxz, Flyer22, Eikoku, JCLately, Powtroll, Jpcedenog, Strife911, Pyromaniaman, Oxymoron83, Gpswiki, Dodabe, Gargvikram07, Mátyás, Fratrep, Chrisguyot, Odo Benus, Stfg, 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, Kkarimi, Blanchardb, Edayapattiarun, Lbertolotti, Shaw76, Verticalsearch, Sebleouf, Hanifbbz, Abrech, Sterdeus, DrCroco, Nano5656, Aseld, Amossin, Dekisugi, SchreiberBike, 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, 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, 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, Elseviereditormath, 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, 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, PhnomPencil, Lisasolomonsalford, Uksas, Naeemmalik036, 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, Epicgenius, THill182, Delafé, Melonkelon, Herpderp1235689999, Revengetechy, Amykam32, The hello doctor, Mimarios1, Huang cynthia, DavidLeighEllis, Gnust, Rbrandon87, Astigitana, Alihaghi, Philip Habing, Wccsnow, Jianhui67, Tahmina.tithi, Yeda123, Skr15081997, Charlotth, Jfrench7, Zjl9191, Davidhart007, Routerdecomposer, Augt.pelle, Justincahoon, Gstoel, Wiki-jonne, MatthewP42, LiberumConsilium, Ran0512, Daniel Bachar, Galaktikasoft, Prof PD Hoy, Gary2015 and Anonymous: 969

13.2 Images

- **File:Commons-logo.svg** *Source:* http://upload.wikimedia.org/wikipedia/en/4/4a/Commons-logo.svg *License:* ? *Contributors:* ? *Original artist:* ?
- **File:Fisher_iris_versicolor_sepalwidth.svg** *Source:* http://upload.wikimedia.org/wikipedia/commons/4/40/Fisher_iris_versicolor_sepalwidth.svg *License:* CC BY-SA 3.0 *Contributors:* en:Image:Fisher iris versicolor sepalwidth.png *Original artist:* en:User:Qwfp (original); Pbroks13 (talk) (redraw)
- **File:Internet_map_1024.jpg** *Source:* http://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:* http://upload.wikimedia.org/wikipedia/en/f/fd/Portal-puzzle.svg *License:* Public domain *Contributors:* ? *Original artist:* ?
- **File:Splitsection.svg** *Source:* http://upload.wikimedia.org/wikipedia/commons/e/ea/Splitsection.svg *License:* Public domain *Contributors:* Tracing of File:Splitsection.gif, performed by Anomie *Original artist:* Original GIF: David Levy
- **File:Wiki_letter_w.svg** *Source:* http://upload.wikimedia.org/wikipedia/en/6/6c/Wiki_letter_w.svg *License:* Cc-by-sa-3.0 *Contributors:* ? *Original artist:* ?

13.3 Content license

- Creative Commons Attribution-Share Alike 3.0