TAB
OFFICE OF TECHNOLOGY ASSESSMENT
AT THE GERMAN BUNDESTAG

Data mining – sociopolitical and legal challenges

Source: nicoelnino/123RF

TAB-Fokus no. 40 regarding report no. 203 July 2022

Summary

- Data mining refers to the recognition of patterns and structures in large data sets. Information is generated, e. g., on similarities, deviations, or anomalies, as well as mathematical-statistical models and algorithms that can be used in new situations regarding the same circumstances in order to at least support decisions.
- Data mining is considered to have an enormous innovation potential in almost all areas of life.
- However, there are concerns about non-transparent procedures as well as unequal possibilities of making use of data mining. Involved fears refer to, e. g., an end of privacy or the uncontrollability of algorithmic systems.
- Among others, there are challenges with regard to providing data, specifying the possibilities and limits of the analysis, and dealing with the results. Extensive expertise is required for implementation, verification, and monitoring.
- Impact assessments and evaluations should be based on the respective application. The report discusses examples for data mining in the fields of medicine and healthcare.

What is involved

Data mining refers to the application of mathematical-statistical methods in order to identify structures and patterns in large data sets. If we want to analyse the societal opportunities and challenges involved, we should not only consider the direct use of such methods (data mining in the narrower sense), but also the entire process of extracting information from data sets (data mining in the broader sense). This includes the following:

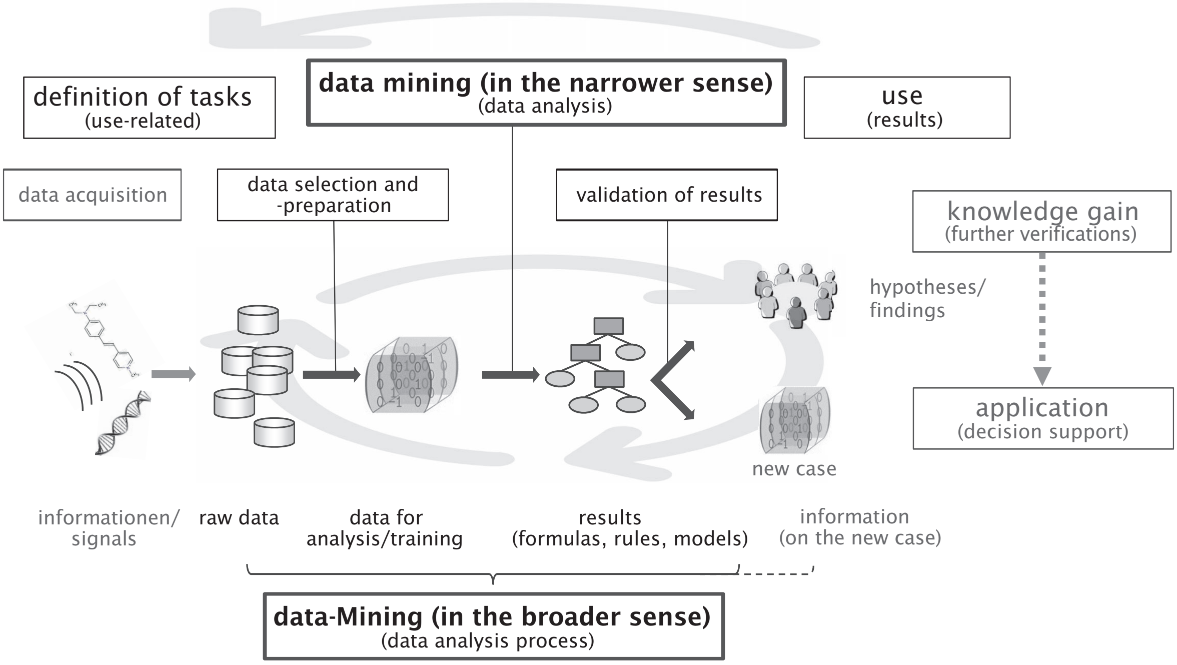
- **Definition of tasks:** search for similarities, differences or anomalies in data sets, classification/grouping of objects, derivation of rules, modelling
- **Data selection and preparation:** data suitability analysis, error correction, conversions, creation of data sets for analysis/training
- **Data analysis:** Depending on the task and the data form, different methods can be considered. Some of them have been established for decades already (e. g., cluster/regression analyses), others are currently becoming ready for application (e. g., training of artificial neural networks).
- **Validation of results:** verifying the determined parameters, formulas, or rules

Data mining is most commonly associated with analyses of existing data sets that are relinked or reused for new purposes. A high degree of automation is possible when directly applying the respective structure recognition methods to a standardised (training) data set, because algorithms define all analysis steps that are executed by a machine. A variety of human activities are required for task definition, data preparation, and validation of results. Usually, (interim) results are assessed in several loops, the analytical procedure is adjusted, and data are added or excluded until the results are considered to be sufficiently valid. Validity requirements vary depending on the context of application. Considerable expertise is required to be able to assess the suitability of data and the suitability of the mathematical approach for the respective analysis task, to distinguish anomalies from errors, or to interpret results.

The particular potential of data mining lies in the generalisation and reuse of sufficiently valid results, i.e., derived decision rules or trained mathematical-statistical models can be used in new situations on the same facts in order to evaluate the situation and support decisions. Such data analysis procedures have always been a core element of scientific work. Information services or algorithmic decision (support) systems are increasingly being developed and applied in other contexts as well.

Client
Committee on Education, Research and
Technology Assessment
+49 30 227-32861
bildungundforschung@bundestag.de

Data mining: schematic representation of the process steps



Data mining: key technology of digitisation and societal challenges involved

The term »data mining« overlaps considerably with the buzzwords »big data«, »machine learning«, or »artificial intelligence«. For several years, numerous committees have been dealing with the potentials and challenges of digitisation in general as well as with growing data sets, the possibilities and limits of data reuse and analysis, and how to deal with the results in particular. They unanimously recommend accelerating digitisation activities, driving forward the standardisation of data, improving the provision of data via data infrastructures, focusing more strongly on data use, strengthening the relevant expertise, promoting the development of data analysis applications, regulating high-risk applications more strongly, and striving for greater national or European digital sovereignty. The latter is especially important in order to ensure high standards of protection and the safeguarding of fundamental rights when dealing with data. These recommendations can also be derived from the statements given in the TAB report.

After presenting the technical and legal aspects of complex data analyses in general, the TAB report takes a closer look at some public applications and identifies specific particularities and key aspects. This allows the derivation of further application-related strengths and courses of action.

Excellent prerequisites for data mining in the field of geodata

In the case of spatial information, standards have already become widely accepted for data collection, but also for data storage, provision, and some analytical functions. For years, the national spatial data infrastructure (SDI) has been expanded, enabling an increasingly open provision of standardised official geodata. Germany is one of the European pioneers regarding the expansion of the

Application example for data mining in medizine

Mammography has long been mentioned as a potential application area for data-trained algorithms for object classification, i. a., due to the large quantity of mammography images that have been generated for years by national screening programs and that have been evaluated and reviewed by medical specialists, might be used as training data. For years, there has been a wide range of data-analytical activities in this field. In order to be able to use such algorithm-based diagnostic systems regularly, their safety and performance

Diagnosis of mammographies

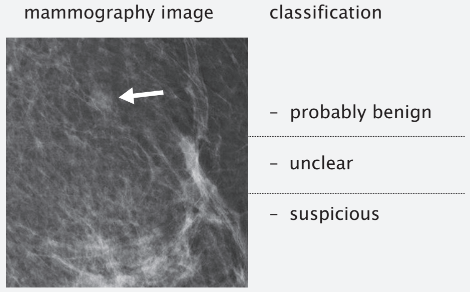


Fig. in box: according to Heindel et al. (2021): Systematische und qualitätsgesicherte Früherkennung des sporadischen Mammakarzinoms, p. 132 and according to Sickles et al. (2013): ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. 5th Edition. American College of Radiology, Reston

SDI. Other public sectors, that are just beginning to standardise data and build data infrastructures, might benefit from this experience.

So far, the focus of SDI activities has been on the provision of data, not on its further use. It is still difficult to estimate to what extent the provided geodata can/will be used in complex data analyses. In the long term, more emphasis should be placed on data demand. The data analysis potential of geodata increases as more specialised data are provided with a georeference. For this, significant efforts will be required in other disciplines.

Spatial analyses – especially when using high-resolution georeferenced data – can provide important information (e. g., to localise health hazards), but can also involve risks of violating fundamental rights (e. g., tracking of individuals) or be classified as a security concern (e. g., situational analyses of critical infrastructures). So far, there has been some legal uncertainty about how to proceed with spatial analyses involving elevated risks. In this regard, the European Artificial Intelligence Act – which is currently being coordinated – should bring more legal certainty.

Medicine: experience with data trust models, risk assessment, and quality assurance for data-trained algorithms

Medical data are collected with an increasing level of detail within the framework of medical treatment and various research and development activities. The low interoperability of these medical primary data – which are personal, stored in a decentralised way, and highly protected – and the IT systems required limit the possibilities regarding their further use. This is why improving the interoperability of medical primary data and the IT is a crucial task. So far, complex data preparation and transfers to special medical registers and data centers have been necessary in order to be able to use them further. Only a fraction of primary medical data is currently transferred to repositories that permanently hold data sets in trust, continually expand them, and review protection and use interests for analysis requests on a case-by-case basis before making necessary data available to a limited analysis team. Their audit procedures should be harmonised, simplified and accelerated. The networking of repositories and the establishment of a national or even European health data infrastructure are political goals that can only be realised by taking a considerable effort.

If medical decision rules are derived or models are trained on the basis of these data and then further developed into software supporting treatment-relevant decisions, this software will be subject to the law on medical devices. This law defines risk-adjusted procedures for product quality assurance, ranging from certifications based on safety and performance records to continuous risk monitoring during use.

Both the established trust structures and the procedures for certification and quality assurance of medical devices can be exemplary for other high-risk areas in which data analysis procedures and decision support systems are also being increasingly used (e. g., security, fin-tech, or legal-tech areas).

Even with extensive quality assurance measures, risks to affected individuals can never be ruled out completely when using data-trained algorithmic systems as part of medical treatment. Such systems reach their limits when dealing with very rare events, can reproduce actually existing discriminations and deliver results that are sometimes difficult to understand even for experts. This is why the clarification of permanent product responsibilities and of liability issues is of particular relevance for the acceptance and use of decision support systems – also, but not only, in medicine.

Data mining projects in the healthcare system must be evaluated individually

A highly relevant data source for analyses in the healthcare sector are the standardised data sets used for billing and reimbursement for medical services. In total, they depict the entire provision of services of the primary healthcare market at individual case level and also contain health-related information on all patients. Therefore, they require particular protection. The further use of these data is subject to complex regulation. On the one hand, various institutions are allowed to use these data within the scope of their legally

must be evaluated (medical device certification). Some of these systems have already received this certification. In addition, relevant medical societies and health insurances must recognise an (additional) benefit to the national status quo and recommend the use resp. bear the costs. This hurdle is particularly high in European countries, among others, where screening programs have low error rates due to the established personnel-intensive multiple diagnosis process. So far, the German medical societies are still skeptical with regard to algorithm-based systems for mammography diagnosis. They consider the conversion from 2D to 3D imaging technology to be more promising. If the imaging technology changes, new training data must be generated and algorithms must be trained, evaluated and certified again. In countries where such personnel-intensive diagnosis processes are not established, it should be easier to achieve benefits with such algorithm-based diagnostic systems.

defined tasks. On the other hand, the data are transmitted to central research data centers. However, it often takes years to integrate into the centers' data inventories and for third parties to be able to use them on a very limited basis upon request. There are frequent calls for a timely data provision and broader usability.

Data mining approaches in the healthcare sector are often embedded in complex processes, e. g., for healthcare service reimbursement and its regular updating, for quality assurance of medical care, for monitoring the health situation, or for searching for adverse drug effects. The extent to which the data used in each case are suitable and the analytical procedure is target-oriented and meaningful, whether medical care processes are improved and/or health-related added value is generated as a result, and what undesirable consequences are associated with this (e. g., disincentives in the healthcare sector), can only be assessed for each case individually.

Discussing the scope of research privilege for data mining

Data analyses as well as data mining for research purposes are legally privileged in different ways. However, some of the wording is open to interpretation. The European General Data Protection Regulation (GDPR) recommends linking the concept of research to the compliance with recognised ethical research standards, providing for an appropriate consent management, examining research intentions on a case-by-case basis, and permitting both publicly and privately funded research up to and including technological developments and application demonstrations. However, national specifications are admitted via opening clauses.

Ethical research standards and reviews of analysis requests have long been established in medicine and the healthcare system. However, consent management is a weak point. This is one of the reasons why the research privilege has so far been interpreted more narrowly in the national healthcare system and only public (research) institutions – whose analysis purposes are in the public interest – have a right of application or use. In parallel, companies that fund clinical trials to demonstrate the safety and performance of medical products are not obliged to make their trial data available for other research purposes.

Health-related risks arising from not using existing health-related data have been pointed out for years, for example, if disease risks, infection foci, or adverse side effects of treatment procedures are not identified, or algorithms are trained using data that may not accurately represent the national situation. These risks arising from the non-use of data should be focused on more strongly in the debate regarding data protection and data use in the future.

TAB report no. 203

Data-Mining – gesellschaftspolitische und rechtliche Herausforderungen

Katrin Gerlinger

TAB

BÜRO FÜR TECHNIKFOLGEN-ABSCHÄTZUNG BEIM DEUTSCHEN BUNDESTAG

Katrin Gerlinger

Data-Mining – gesellschaftspolitische und rechtliche Herausforderungen

Endbericht zum TA-Projekt

Juni 2022
Arbeitsbericht Nr. 203

Website of the project

www.tab-beim-bundestag.de/en/data-mining

Project manager and contact

Dr. Katrin Gerlinger
+49 30 28491-108
gerlinger@tab-beim-bundestag.de

The Office of Technology Assessment at the German Bundestag (TAB) advises the German Bundestag and its committees on questions of scientific and technological change. TAB has been operated by the Institute for Technology Assessment and Systems Analysis (ITAS) of the Karlsruhe Institute of Technology (KIT) since 1990. It has been cooperating with the IZT – Institute for Futures Studies and Technology Assessment and VDI/VDE Innovation + Technik GmbH since September 2013. The Committee for Education, Research and Technology Assessment decides on TAB's work programme, which also includes subjects proposed by other parliamentary committees. The standing »TA Rapporteur Group« consists of the Chairman of the Committee Kai Gehring (Bündnis 90/Die Grünen), and one member from each of the parliamentary parties: Dr. Holger Becker (SPD), Lars Rohwer (CDU/CSU), Laura Kraft (Bündnis 90/Die Grünen), Prof. Dr. Stephan Seiter (FDP), Prof. Dr. Michael Kaufmann (AfD), Ralph Lenkert (Die Linke).