THE BIG PICTURE

Martin Komenda, Jiří Marek

TERMINOLOGY

This book focuses on data analytics and different approaches to visualise the obtained results, which are directly used in the decision-making processes. Before briefly describing the methodological background that underpins each of the chapters, let us take a detailed look at the key concepts and their explanation within the context of this publication. In today's world, which is full of information, it is possible to find multiple definitions and descriptions of the same concept. For the purposes of this publication, a published and verified source of information was always chosen that fit well within the context of the editor's point of view on the subject.

LITERACY

Information literacy is the adoption of appropriate information behaviour to obtain, through whatever channel or medium, information well fitted to information needs, together with critical awareness of the importance of wise and ethical use of information in society [1].

Data literacy is the component of information literacy that enables individuals to access, interpret, critically assess, manage, handle and ethically use data [2].

Statistical literacy is envisaged as the component of data literacy involved in the critical appraisal, interpretation, processing and statistical analysis of data [2].

Health literacy refers to the ability to find, understand and correctly use information on health and healthcare. This includes, for example, information on disease or lifestyle risk factors, invitations to appointments, package leaflets, instructions from health professionals, basic orientation in the healthcare delivery system, knowledge of the symptoms of common diseases, knowledge of the basic functions of the human body, knowledge of the basic steps to take care of oneself or to ensure self-sufficiency in the event of illness [3]. Improving health literacy in populations provides the foundation on which citizens are enabled to play an active role in improving their own health, engage successfully with community action for health, and push governments to meet their responsibilities in addressing health and health equity. Meeting the health literacy needs of

the most disadvantaged and marginalised societies will particularly accelerate progress in reducing inequities in health and beyond [4].

KNOWLEDGE DISCOVERY

Knowledge discovery in databases (KDD) is often used in data mining. It is a non-trivial process of discovering novel and potentially useful information from large amounts of data and aims to identify new understandable patterns in the data convincingly [5].

Data mining is often a synonym for extracting useful information from databases. It refers to the application of algorithms for extracting patterns from data without the additional steps of the knowledge discovery process [6].

There needs to be more consistency in data mining (DM) and knowledge discovery in databases (KDD) in the literature and web resources. Some authors use these terms synonymously. Thus, both terms more or less mean the same thing; with KDD, the actual preparation of the data is also considered essential (Figure 1). This figure shows the complexity of data processing to extract valid and correct information or possibly knowledge. Below, the main steps are shortly described [5].

1. Data: Building a data domain based on a detailed understanding will then be worked with.

2. Selection: Choosing and creating a data set on which the discovery process will be performed.

3. Preprocessing: Application of basic operations such as handling the missing values, noise removal, data format unification, etc.

4. Transformation: Finding functional characteristics of data, dimensionality reduction, discretisation of numerical attributes, etc.

5. Data mining: Matching the given aims (step 1) to a particular data mining method (i.e. summarisation or classification), choosing the proper data mining algorithm and application of the selected algorithm to search the patterns (i.e. classification rules or trees).

6. Interpretation / Evaluation: Assessment and correct interpretation of the mined results, possibly return to any of the above-mentioned steps (i.e. 1 to 5) for further iteration.

7. Knowledge: Incorporating achieved knowledge for further action.

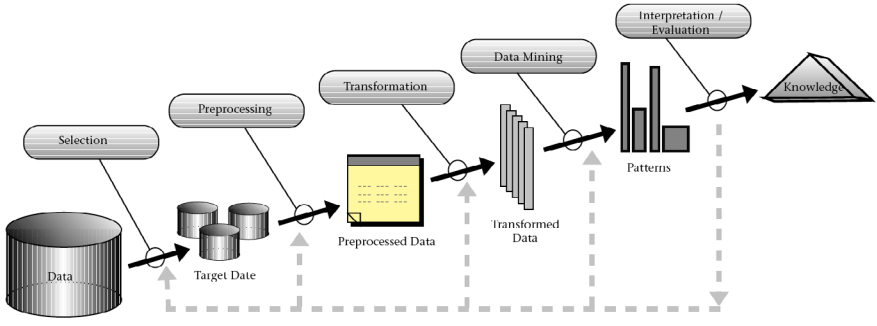

Figure 1: The overview of tile steps comprising the KDD process

Generally, the main goal of proven process methodologies (see section Methodologies for data processing) is to provide users with a unified framework for solving various knowledge discovery tasks. These methodologies allow for sharing and transferring experiences from successful projects. This book contains individual case studies based on the proven KDD methodology. The separate stages, which are appropriately linked to each other, make it easy for the user to understand each stage of the lifecycle while giving it the attention it deserves.

When used correctly, the methodology enables the correct and, above all, complete implementation of a process that leads from the successful mining of **data** (objective facts with no further information regarding patterns or context), through the acquisition of hitherto unknown **information** (contextualised, categorised, calculated, and condensed data with a specific meaning painting a bigger picture), to **knowledge**, which is closely linked to doing and implies know-how and understanding (information with added value, including a specific context) [7,8].

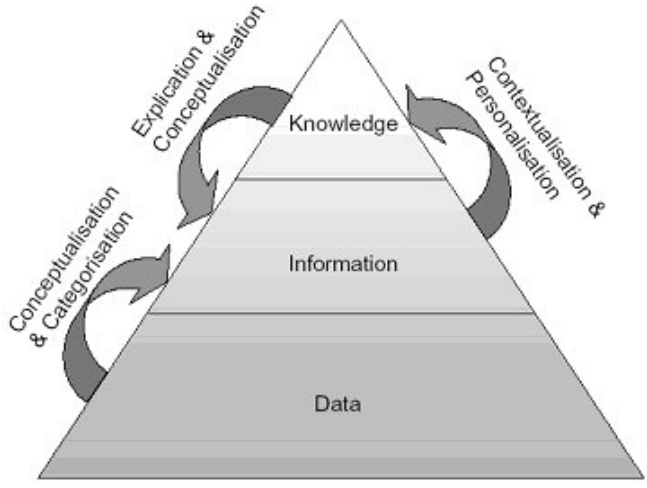

Figure 2: Data-Information-Knowledge pyramid

For a better understanding, a real example from the field of curriculum description and mapping is given below:

Floor of pyramid	Example
Data	A learning unit (typically one lecture of practice) is defined by several descriptive characteristics, i.e. title, section, range (in hours), type, annotation, keywords, medical discipline, etc. A total of 1,342 learning units were described in the General Medicine study programme at the Faculty of Medicine of Masaryk University.
Information	The Internal Medicine section contains the most learning units (446), and Diagnostic Sciences and Neurosciences (226) the least. Ratio[1] = 1.97
Knowledge	The average number of learning outcomes () listed per learning unit does not correspond to the proportion of learning units represented in each learning section. Internal Medicine 5.5 learning outcomes per learning unit Diagnostic Sciences and Neurosciences 7.7 learning outcomes per learning unit Ratio[1] = 0.73 Moreover, the total teaching range (in hours) does correspond to the number of learning units in each section. Internal Medicine 2,492.1 learning outcomes per learning unit Diagnostic Sciences and Neurosciences 1,225.5 learning outcomes per learning unit Ratio[1] = 2.03

1 Internal Medicine / Diagnostic Sciences and Neurosciences

Although the topic of **artificial intelligence** (AI) is not the main focus of this publication, given the current rapid development of this approach, it is appropriate to mention it here in passing. The application of AI in medicine has two main branches: (i) virtual component represented by **machine learning** techniques also called deep learning (i.e., electronic medical records where specific algorithms are used to identify subjects with a family history of a hereditary disease or an augmented risk of a chronic disease), (ii) physical component includes physical objects, medical devices and increasingly sophisticated robots (i.e., robot companion for the aging population with cognitive decline or limited mobility) [9].

Especially machine learning techniques, among all fields of human interest, represent a promising approach for extracting knowledge and high added value of the vast amount of high-granularity data. For the illustration, (Figure 3) demonstrates the most relevant data-driven related areas used for power systems data processing and analysis [10]. Generally speaking, the proper usage of AI tools that follow human behaviour based on collected data and are under human supervision can be the starting point for developing data-driven services.

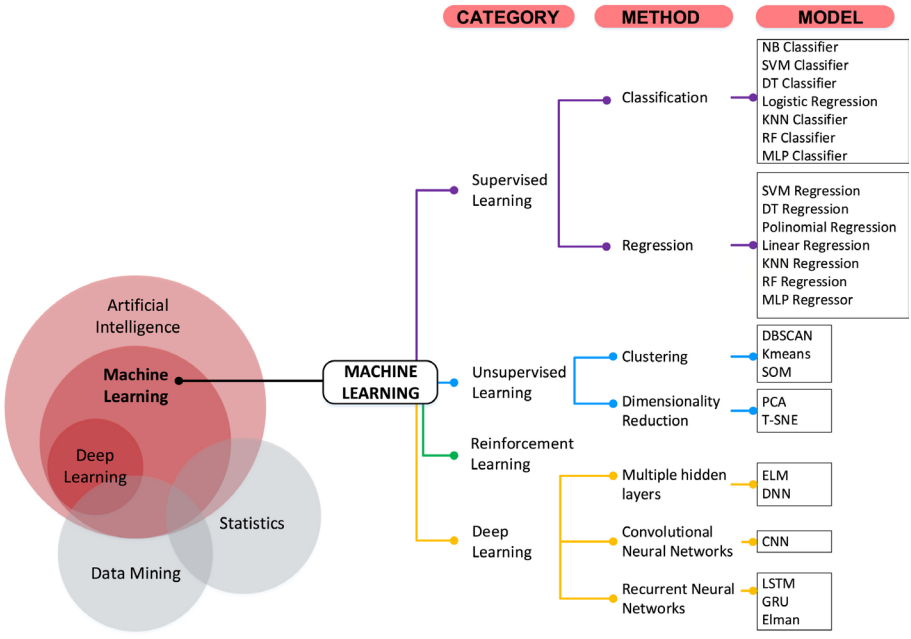

Figure 3: Data-driven techniques classification in the context of machine learning categories for power systems analysis

OPEN COMMUNITY

Open science is a collection of actions designed to make scientific processes more transparent and results more accessible. Its goal is to build a more replicable and robust science; it does so using new technologies, altering incentives, and changing attitudes. The most well-known initiative from the Open Science agenda is Open Access to scientific publications or Open/FAIR Data approach towards managing research data (in Europe, mostly connected for now with the initiative of European Open Science Cloud – EOSC) [11,12].

Open data, according to § 3 par. 5 of Act No. 106/1999 Coll. on Free Access to Information (Czech Freedom of Information Act) and in line with the directive (EU) 2019/1024 of the European Parliament and of the Council on open data and the re-use of public sector information, is "… information published in a manner that allows remote access in an open and machine-readable format, the manner and purpose of subsequent use of which is not restricted by the obliged entity publishing it and which is registered in the national catalogue of open data". In accordance with this definition and the recommended practices, each open dataset must comply with the rules set out below, independently on the field of human interest:

1. It is accessible as a data file in a machine-readable and open format with complete and up-to-date database content or aggregated statistics.

2. It is provided with non-restrictive terms of use, with the terms of use being set as follows, depending on the nature of the content of the dataset:

3. A CC BY 4.0 public license by which the copyright holder allows free use of his works provided that the user of the work credits him as the author.

4. The CC0 public license, which serves as a means of waiving the database rights of the database creator.

5. It is registered in the National Catalogue of Open Data as a direct link to the dataset.

6. It is accompanied by clear documentation.

7. It is available for download without technical barriers (registration, access restrictions, CAPTCHA, etc.).

8. It is prepared with the aim of making it as easy as possible for programmers, etc., to be machine-processed.

9. It is provided with a curator contact for feedback (bugs, extension requests, etc.).

FAIR data refers to a set of principles, focused on ensuring that research objects are reusable, and actually will be reused, and so become as valuable as is possible [13]. Making data available according to FAIR data principles require being in compliance with these attributes, which are composed of several recommendations clustered around the acronym FAIR:

— **F**indable – metadata, registration, global persistent identifiers
  — Data should be easily discoverable, allowing researchers to locate and identify the data of interest. This is achieved through persistent identifiers, standardised metadata, and comprehensive data descriptions.

— **A**ccessible – standards for machine-readability, authentication and authorisation infrastructure
  — Data should be readily accessible to both humans and machines. It should be available through well-defined access protocols, with minimal barriers to access, such as login requirements or subscription fees.

— **I**nteroperable – semantic description of data and metadata, ontologies, standards
  — Data should be structured and organised to facilitate integration and interoperability with other data sources. This involves using standardised data formats, adopting common vocabularies and ontologies, and providing clear data and metadata specifications.

— **R**eusable – clear licensing, data provenance (reproducibility) [14]
  — Data should be designed and documented in a manner that enables its reuse for different purposes. This includes providing detailed information about the data's provenance, methodology, and context, as well as clear licensing and usage terms.

FAIR data are not in conflict with open data, these two terms are usually linked together. To simplify it, all data can be FAIR, but only some data can be open (personal data handling restrictions, commercialisation aspects, etc.). Therefore, health data will usually be more FAIR than open data due to the extensive need to handle sensitive data.

DATA ANALYSIS AND DELIVERY

Data analytics is the application of computer systems to analyse large data sets to support decisions. This interdisciplinary field has adopted aspects from many other scientific disciplines, such as statistics, machine learning, pattern recognition, system theory, operations research, and artificial intelligence. Such an approach allows us to find relevant information, structures, and patterns, gain new insights, identify causes and effects, predict future developments, or suggest optimal decisions. We need models and algorithms to collect, preprocess, analyse, and evaluate data [15].

BUSINESS INTELLIGENCE

— **as a process** can be defined as the process of turning data into information and then into knowledge. Knowledge is typically obtained about customer (i.e. representative of the Ministry of Health of the Czech Republic, health insurance company, open data community, etc.) needs, customer decision-making processes, the competition, conditions in the industry, and general economic, technological, and cultural trends [16].

— **as a process and a product** can be used to refer to an organised and systematic process by which organisations acquire, analyse and disseminate information from both internal and external information sources significant for their business activities and for decision-making [17].

— **as a process, a product and technologies** encompass a set of tools, techniques, and processes to help harness this wide array of data and allow decision-makers to convert it to useful information and knowledge [18].

— **in general**, it allows managers to make informed and intelligent decisions regarding the functioning of their organisation [19].

DATA-DRIVEN APPROACH

People in academia, government and business sectors often make their decisions based on their subjective opinions and habits. A data-driven approach helps to bring objectivity and facts into decision-making. It must be emphasised, however, that data alone do not have the necessary and required informative and telling value. Context and correct interpretation must always be considered. The context and assumptions represent external aspects out of the control of any decision-maker, but the premises and the knowledge of the company depend on available data [20]. One of the definitions introduces **data-driven decision-making** as the practice of basing decisions on data analysis rather

than purely intuition [21]. On the government or the public sector level, **evidence-based policymaking** is similar in that decisions are based on factual data [22]. In both cases, the following applies: the more data are available, the more people, stakeholders and institutions can construct their perceptions and decisions of reality.

METHODOLOGIES FOR DATA PROCESSING

Setting standards in data mining primarily results in methodological guidelines on how to achieve this goal. It has become essential because of the increased demand for methodologies and tools to help analyse and understand data and, last but not least, to make data more interoperable for further easier exchange or processing. Many related standards and recommendations in this area have already been set up and published (Figure 4) [23]. SEMMA (Sample, Explore, Modify, Model, Assess), 5A (Assess, Access, Analyse, Act, and Automate), and CRISP-DM (CRoss-Industry Standard Process for Data Mining) are considered to be the most frequently used methodologies [24]. Each phase is crucial: it is not just about processing and visualising the data. It is good to remember that the lifecycle of long-standing and field-tested methodologies has its meaning and importance, and needs to be thoroughly addressed. Nowadays, we are inundated with data on a daily basis due to the routine operation of Internet applications and various information systems. All of them involve the generation of data, backups and archives, be it telecommunications, banking transactions or scientific research. The different phases of the chosen methodology will provide the space for a proper understanding and addressing of the differences that characterise the individual domains. In general, the process guides the interpreter of a given problem through theoretically very well-described steps, which always contain a set of detailed actions by the selected algorithm.

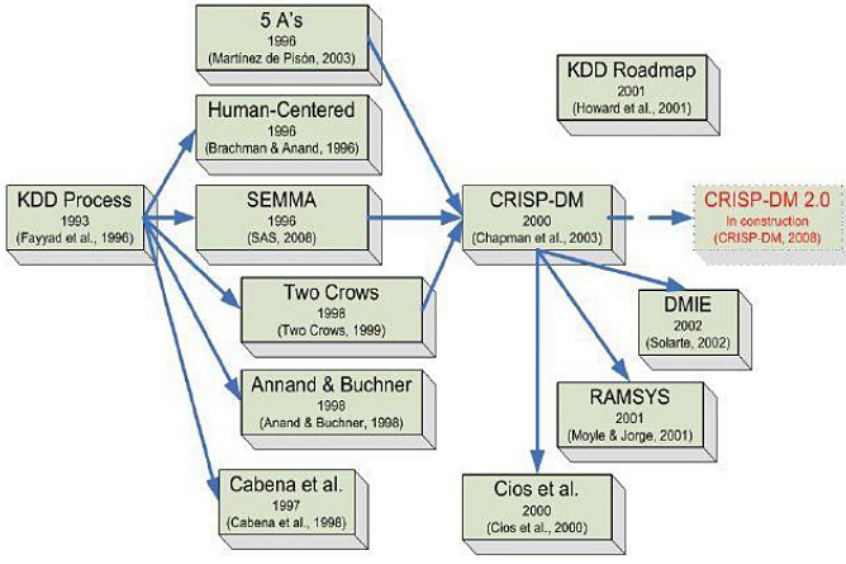

Figure 4: Evolution of DM & KDD process models and methodologies

Each chapter of this book is based on the most proven CRISP-DM methodology in practice, which consists of six loosely linked steps (Figure 5). Although these steps may seem trivial and simple, one or even more of them are often forgotten – and this is one of the main reasons why the complete process of each methodology needs to be very thoroughly adopted, addressed, and then correctly applied in real life:

1. what to solve (Business understanding) – understanding the problem, formulating the task,
2. where to get data (Data understanding),
3. how to prepare data (Data preparation),
4. how to analyse data (Data modelling),
5. what we found (Evaluation) – understanding the results,
6. how to use the results (Deployment).

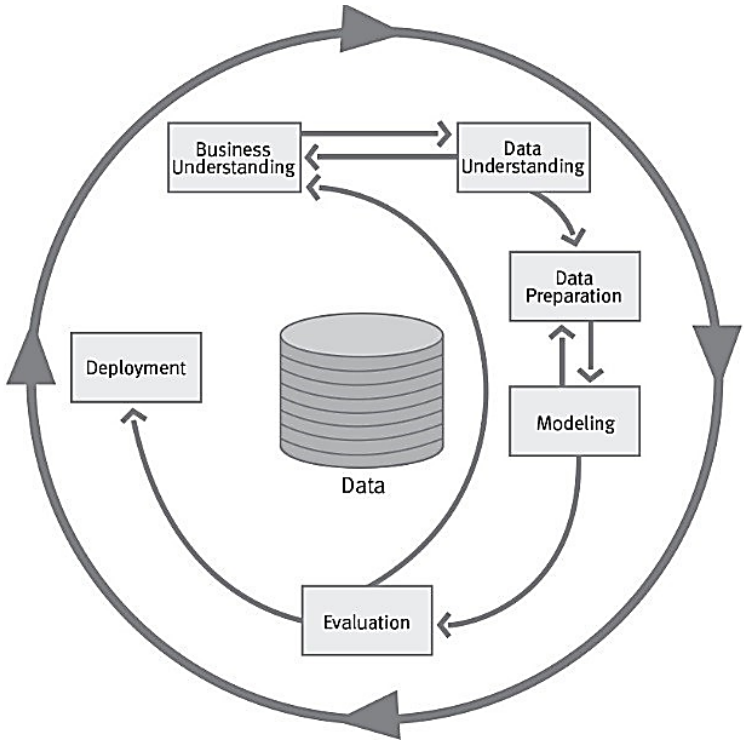

Figure 5: Diagram of CRISP-DM

1. **Business understanding** is a step towards understanding of particular domain that always hides the specifics of a given expertise. In the case of this book, this is the domain of healthcare, medicine, education, health literacy, and health informatics. Examples include working with thesauri or vocabularies closely related to, for example, selected nomenclature (International Classification of Diseases), a specific vocabulary for an individual medical domain (MeSH, SNOMED). The legislation is also a very important agenda, especially the nowadays very accentuated and sensitive issue of data protection (GDPR, General Data Protection Regulation). When working with data, it is necessary to think about protecting an individual's data and preventing their identification at all times. Finally, it is essential to recognise whether descriptive attributes (columns characterising a given row like an instance) are present. If the solver has thoroughly mastered this methodology step, he may need to pay more attention to it in subsequent tasks than in the beginning. In any case, it is crucial to highlight that, due to legal, privacy and sensitivity considerations, not all health data can be open; more often,

however, it should be FAIR. Last but not least, an understanding of the data lifecycle and how data are created must be emphasised. Who enters it into the system and how, what automatic checks and validation rules are implemented in the information system, and how long is the update interval? These processes are not always set up; therefore, the selected case studies focus on a thorough understanding of the data lifecycle.

2. Data understanding is directly related to the first of the steps. An example would be correctly selecting, understanding and using the above-mentioned vocabularies. Business understanding logically translates into a subsequent data understanding and its proper processing. This is not only in terms of content correctness but also economics (how much will it cost?) or personnel (do I have enough experts?). With a detailed check and subsequent knowledge of all the dataset's characteristics, it is possible to process and correctly present the data (regardless of the form - table, graph, or interactive visualisation). The quality of the input data determines the possibilities of descriptive statistics and the use of advanced analytical methods. Unfortunately, the input data may need to be completed or burdened with noise. This is a random error, but it affects the quality of the data. This is where the power of the Open and FAIR Data concept comes into full play, which always strictly requires a complete metadata description and, therefore, complete input information for anyone working with the data. In addition, of course, the technical treatment of the dataset is also essential, e.g. the data types or codebook structure used. This apparent small and technical detail can often cause considerable processing, mapping, and transformation complications.

3. The next step is to **prepare the data** in the form of pre-processing. Here we are talking about a significantly individualised approach directly related to the quality of the input data. Data cleaning, selecting relevant attributes, filling in missing values (e.g. using suitable open datasets), conversion or unification of data types, mapping multiple datasets to one, or transformations are standard and not entirely trivial techniques.

4. Modelling is used in basic descriptive statistics, analytical or machine learning methods (e.g. decision trees, clustering or association rules) etc. A combination of multiple ways is typically used to provide a comprehensive view of the available dataset. Undoubtedly, it also includes visualisation of the results obtained through basic or advanced statistical and analytical techniques. It must be emphasised here that the final form, i.e., how the results of the solution to the problem are also presented, can profoundly affect how the outputs are understood and subsequently used in practice. Simplicity, clarity, and comprehensibility are essential from a long-term

perspective and experience, of course always related to the named target groups for whom the outputs are primarily intended.

5. Despite its crucial importance, the **evaluation** phase is often neglected and should be addressed more thoroughly. Here it is not only the researcher's detailed manual or machine validation which should be automatically included in the entire CRISP-DM process. Feedback should also be provided by an expert who is ideally not part of the research team but belongs to the target group. Such experts are specialists in a particular field and can provide a relevant and valid review. This process should also include evaluating the results achieved against the exploratory/research questions and an overall performance assessment against the set brief. It is logical that, as in other phases, the outcome of the evaluation may be the need for more or less intervention in the previous steps. This may even mean a redesign or a more fundamental change in the solution of the whole task.
6. The **deployment** or application of the output in practice is the last phase of this methodology. It is not just a technical matter of, for example, releasing a new version of a web application or an interactive infographic. Compared to the results obtained in the modelling phase, there may be slight modifications resulting from the possibility of implementing the results in practice. For example, we can talk about publishing an open dataset, a single summary data table in the given context, a more detailed static presentation with graphs, explanations, and conclusions, or a comprehensive interactive data visualisation, including selections of custom views of the data using filters.

Moreover, a complex research data lifecycle model [25] can significantly help re-use data in a different context, such as research and development or policymaking, and it also provides more interoperability in various case studies on these data. This model consists of the following six stageTable 1 1. This concept also follows the research data management toolkit for life sciences[2], which introduces best practices and guidelines to help make data more manageable in accordance with the Horizon Europe Programme Guide recommendations.

2 https://rdmkit.elixir-europe.org/data_life_cycle

Table 1: The research data lifecycle model.

Stage	Activities
Creating data	• design research • plan data management (formats, storage etc.) • plan consent for sharing • locate existing data • collect data (experiment, observe, measure, simulate) • capture and create metadata
Processing data	• enter data, digitise, transcribe, translate • check, validate and clean data • anonymise data where necessary • describe data • manage and store data
Analysing data	• interpret data • derive data • produce research outputs • author publications • prepare data for preservation
Preserving data	• migrate data to best format • migrate data to suitable medium • back-up and store data • create metadata and documentation • archive data
Giving access to data	• distribute data • share data • control access • establish copyright • promote data
Re-using data	• follow-up research • new research • undertake research reviews • scrutinise findings • teach and learn

REFERENCES

[1] Weber S. Getting the knowledge. Library and Information Update. 2002;1:52–3.

[2] Calzada Prado J, Marzal MA. Incorporating data literacy into information literacy programs: Core competencies and contents. Libri. 2013;63(2):123–34.

[3] Institute of Health Literacy. Health literacy [Internet]. Prague: Ministry of Health of the Czech Republic; 2023 [cited 20 Jul 2023]. Available from: https://www.nzip.cz/clanek/226-zdravotni-gramotnost.

[4] World Health Organization. Improving health literacy [Internet]. [cited 16 Jul 2023]. Available from:
https://www.who.int/activities/improving-health-literacy.

[5] Fayyad UM, Piatetsky-Shapiro G, Smyth P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In: Simoudis E, Han J, Fayyad UM (eds). KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Washington: AAAI Press; 1996. p. 82–8.

[6] Fayyad UM, Stolorz P. Data mining and KDD: Promise and challenges. Future generation computer systems 1997;13(2–3):99–115.

[7] Davenport TH, Prusak L. Working Knowledge: How Organizations Manage What They Know. Brighton: Harvard Business Press; 1998.

[8] Zimmermann A, Lorenz A, Specht M. The Use of an Information Brokering Tool in an Electronic Museum Environment [Internet]. 2003 [cited 16 Jul 2023]. Available from:http://www.archimuse.com/mw2003/papers/zimmermann/zimmermann.html.

[9] Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism. 2017;69S:S36–S40.

[10] Barja-Martinez S, Aragüés-Peñalba M, Munné-Collado I, et al. Artificial intelligence techniques for enabling Big Data services in distribution networks: A review. Renew Sust Energ Rev. 2021;150:111459.

[11] Spellman B, Gilbert E, Corker KS. Open science: What, why, and how [Internet]. PsyArXiv; 2017. Available from: https://psyarxiv.com/ak6jr/.

[12] European Commission. Open Science [Internet]. European Commission; 2020 [cited 16 Jul 2023]. Available from: https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en.

[13] Mons B, Neylon C, Velterop J, et al. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Inform Serv Use. 2017;37(1):49–56.

[14] Wilkinson M, Dumontier M, Aalbersberg I, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.

[15] Runkler TA. Data analytics: Models and Algorithms for Intelligent Data Analysis. Wiesbaden: Springer Fachmedien Wiesbaden; 2020.

[16] Shollo A, Kautz K. Towards an understanding of business intelligence. In: ACIS 2010 Proceedings; 2010. 86.

[17] Lönnqvist A, Pirttimäki V. The measurement of business intelligence. Inf Syst Manag. 2006;23(1):32–40.

[18] Clark TD, Jones MC, Armstrong CP. The dynamic structure of management support systems: theory development, research focus, and direction. MIS Q. 2007;31(3):579–615.

[19] Foley E, Guillemett MG. What is business intelligence? Int J Bus Intell Res. 2010;1(4):1–28.

[20] Diván MJ. Data-driven decision making. In: Khatri SK, Kapur RK, Rana AS, Sanjay PK (eds). 2017 International Conference on Infocom Technologies and Unmanned Systems (ICTUS). Los Alamitos: IEEE; 2017. p. 50–6.

[21] Provost F, Fawcett T. Data science and its relationship to big data and data-driven decision making. Big Data. 2013;1(1):51–9.

[22] Luthfi A, Janssen M. Open data for evidence-based decision-making: Data-driven government resulting in uncertainty and polarization. Int J Adv Sci Eng Inform Technol. 2019;9(3):1071–8.

[23] Marbán O, Mariscal G, Segovia J. A data mining & knowledge discovery process model. In: Ponce J, Karahoca A (eds). Data Mining and Knowledge Discovery in Real Life Applications. London: IntechOpen; 2009. p. 1–16.

[24] Komenda M. Towards a Framework for Medical Curriculum Mapping [Ph.D. Thesis]. Brno: Masaryk University, Faculty of Informatics; 2016. Available from: https://is.muni.cz/th/hcl4g/.

[25] Ball A. Review of Data Management Lifecycle Models. Bath: University of Bath; 2012.