Big Data Calls for Machine Learning

Andreas Holzinger, Medical University Graz, Graz, Austria

Glossary

Big data Is a buzzword to indicate the flood of data, particularly as medicine of today is increasingly turning into a data science. Big data is necessary for automatic machine learning approaches to learn effectively.

Data fusion Is the process of integrating various data and knowledge representations of the same data sets into one consistent, accurate, and useful representation.

Data integration Is the combination of data from different sources and providing a unified view, whereas data fusion is integration of multiple data representing the same real-world object into one consistent representation.

Dimensionality Dimensionality of data is high, when the number of features is larger than the number of observations by magnitudes. A good example for high-dimensional data is *omics data, for example, genetic data.

Features Measurable properties of an observation and are key for learning and understanding. Machine learning is often called feature engineering. A synonym for feature is dimension, because an object with *n* features can be represented as a multidimensional point in an *n*-dimensional space by a feature vector. A central task is dimensionality reduction, which is the process of mapping an *n*-dimensional point into a lower *k*-dimensional (sub)space.

Knowledge discovery Includes exploratory analysis and modeling of data and the organized process to identify valid, novel, useful, and understandable patterns from these data sets.

Machine learning Is the ability of algorithms to learn from data and extract knowledge without explicitly being handcrafted. Probabilistic ML is extremely useful for the biomedical domain, where most problems involve dealing with uncertainty, because the inverse probability allows to infer unknowns, learn from data, and make predictions.

***omics data** Derived from various sources, for example, genomics, proteomics, metabolomics, lipidomics, transcriptomics, epigenetics, microbiomics, fluxomics, phenomics, foodomics, cytomics, embryomics, exposonomics, and phytochemomics (all *omics technologies).

Nonstructured data or unstructured data Is an imprecise definition often used for data expressed in natural language, when no specific *structure* has been defined. Yet, this is not true: text has also some structure, words, sentences, and paragraphs. Unstructured data would mean completely randomized data that are usually called noise. The correct term should be unstructured information.

Subspace clustering Is a technique to overcome the curse of dimensionality, that is, to discover relevant and interesting groups of data by selecting relevant dimensions for each cluster separately. This calls for a doctor in the loop, because what is interesting cannot be discovered automatically.

Uncertain data Is a challenge in the medical domain, since the aim is to identify which covariates out of millions are associated with a specific outcome such as a disease state. Often, the number of covariates is orders of magnitude larger than the number of observations, involving the risks of modeling artifacts or losing information.

Weakly structured data This is not to be confused with weakly structured information, instead we follow the notions from algebraic topology: let Y(t) be an ordered sequence of observed data, for example, of individual patient data sampled at different points $t \in T$ over a time sequence. We call the observed data Y(t) weakly structured if and only if the trajectory of Y(t) resembles a random walk – which is quite often the case.

https://doi.org/10.1016/B978-0-12-801238-3.10877-3

Abbreviations

aML Automatic machine learning
HCI Human–computer interaction
ICD International Classification of Diseases
iML Interactive machine learning
KDD Knowledge discovery from data
ML Machine learning
SNOMED CT Standardized Nomenclature of Medicine Clinical Terms
TDA Topological data analysis

Life Sciences Are Turning Into Data Science

Modern medicine turns increasingly into a data-intensive science. A grand challenge is to deal with increasingly large data sets (big data) in arbitrarily high dimensions and the increasing amounts of unstructured information in electronic patient records (natural language and textual information). The so-called "big data challenge" is driven by the trend toward precision medicine (P4 medicine: predictive, preventive, participatory, and personalized). The grand vision of personalized medicine is to model the complexity of patients to tailor medical decision-making, treatment, and therapies exactly to the needs and requirements of the individual. Such approaches result in a sheer explosion in the amount of generated data sets, in particular *omics data (e.g., from genomics, proteomics, and metabolomics). Applying sophisticated machine learning algorithms to make these data useful, usable, and accessible to the medical professional is a commandment of our time.

However, also traditional data are a challenge, that is, image data, physiological data, and time series data, along with natural language (textual data) in patient records and all sorts of physical entities in both time and structure. The biggest problem in the medical domain is not in the size of the data (S4 challenge: structure of data, size of data, speed of data, and source of data), but it is in "dirty data," that is, noisy data, incomplete data, missing data, not applicable data, not relevant data, unknown data, weakly structured data, and most of all uncertainty of the data. Here, probabilistic machine learning can help, because the inverse probability allows to infer unknowns, learn from data, and make predictions. However, to apply the machine learning machinery, a firm understanding of the health data ecosystem is of eminent importance. The relevant and interesting (yet interesting is genuinely human and difficult to define) data are called features. To emphasize the importance of understanding the data, machine learning is often called feature engineering.

Big Data: Little Data

Machine learning deals with the problem of extracting features from data to solve predictive tasks, including decision support, forecasting, ranking, classifying (e.g., in cancer diagnosis), or detecting anomalies (e.g., virus mutations). The challenge is to discover relevant structural patterns and/or temporal patterns (knowledge) in such data, which are often hidden and not accessible to the human expert. The problem is that a majority of the data sets in the biomedical domain are weakly structured and nonstandardized, and most data are in dimensions much higher than 3, and despite human experts are excellent in pattern recognition for dimensions ≤ 3, such data make manual analysis simply impossible.

Moreover, biomedical data sets are genuinely full of uncertainty, incompleteness, and probabilities, and many problems in the medical domain are computationally hard, which makes the application of fully automated approaches difficult or even impossible, or at least the quality of results from automatic approaches might be questionable. Moreover, the complexity of sophisticated machine learning algorithms has detained nonexperts from the application of such solutions. Consequently, the integration of the knowledge of a domain expert can sometimes be indispensable, and the interaction of a domain expert with the data would greatly enhance the knowledge discovery process pipeline.

Hence, interactive machine learning (iML) puts the "human-in-the-loop" to enable what neither a human nor a computer could do on their own. This idea is supported by a synergistic combination of methodologies of two areas that offer ideal conditions toward unraveling such problems: human–computer interaction (HCI) and knowledge discovery/data mining (KDD), with the goal of supporting human intelligence with machine intelligence to discover novel, previously unknown insights into data and to help to make decisions, and decision-making is the core essence of medical informatics.

Taxonomy of Data: Medical Perspective

In the medical domain, we can distinguish between four large data pools, considering the **context** in which the data have been produced:

(1) Biomedical research data, for example, clinical trial data and *omics data, for example, from genomic sequencing technologies (next-generation sequencing, NGS, e.g.), microarrays, transcriptomic technologies, proteomic and metabolomic technologies, lipidomics, transcriptomics, epigenetics, microbiomics, fluxomics, phenomics, cytomics, connectomics, environomics, exposomics, exonomics, foodomics, toponomics.
(2) All these data are important for biomarker discovery and drug design to support the goal of reaching a level of precision medicine.
(3) Clinical data, for example, electronic patient records (EPR), clinicians' documentations, medical terminologies (ontologies, e.g., ICD and SNOMED CT), medical surveys, laboratory tests, clinical and physiological parameters, and all types of biomedical signals, for example, ECG, EEG, and EOG.
(4) Health business data (e.g., costs, utilization, management data, logistics, accounting, billing, resource planning, and prediction).
(5) Private patient data, produced by various customers and stakeholders outside the clinical context (e.g., wellness data, sport data, insurance data, and ambient assisted living data).

The US Department of Health and Human Services (HHS) created a taxonomy of health data with the following seven dimensions:

(1) Demographics and socioeconomic data: age, race, sex, education, etc.
(2) Health status data: health status of the patient, for example, morbidities, problems, complaints, disabilities, diagnoses, and symptoms
(3) Health resources data: characteristics and capacity of the health system, operating figures, performance ratios, etc.
(4) Health-care utilization data: characteristics (e.g., time, duration, tests, procedures, and treatment) about medical care visits like discharge, stay, and use of health-care services
(5) Health-care financing and expenditure data: costs, charges, insurance status, etc.
(6) Health-care outcomes of current and past prevention, treatments, etc.
(7) Other data: *omics data, environmental exposures, etc.

Taxonomy of Data: Informatics Perspective

Most of our computers are Von Neumann machines, consequently at the lowest physical layer; data are represented as patterns of electric on-/off-states (1/0, H/L, and high/low); we speak of a bit, which is also known as Bit, the basic indissoluble information unit according to Shannon. Do not confuse this Bit with the IEC 60027-2 symbol bit – in small letters – which is used as an SI dimension prefix (e.g., 1 kbit = 1024 bit and 1 byte = 8 bit). We can determine various levels of data structures:

(1) Physical level: in a Von Neumann system, bit; in a quantum system, qubit
Note: Regardless of its physical realization (e.g., voltage or mechanical state or black/white), a bit is always logically either 0 or 1 (analog to a light switch). A qubit has similarities to a classical bit but is overall very different: A classical bit is a scalar variable with the single value of either 0 or 1, so the value is unique, deterministic, and unambiguous. A qubit is more general in the sense that it represents a state defined by a pair of complex numbers (a,b), which express the probability that a reading of the value of the qubit will give a value of 0 or 1. Thus, a qubit can be in the state of 0, 1, or some mixture – referred to as a superposition – of the 0 and 1 states. The weights of 0 and 1 in this superposition are determined by (a,b) in the following way: $\text{qubit} \hat{=} (a,b) \hat{=} a \cdot 0\_\text{bit} + b \cdot 1\_\text{bit}$. Please be aware that this model of quantum computation is not the only one.
(2) Logical level: (1) primitive data types, including (a) Boolean data type (true/false); (b) numerical data type (e.g., integer (Z) and floating-point numbers (reals)); (2) composite data types, including (a) array, (b) record, (c) union, (d) set (stores values without any particular order and no repeated values), and (e) object (contains others); (3) string and text types, including (a) alphanumeric characters and (b) alphanumeric strings (sequence of characters to represent words and text)
(3) Abstract level: including abstract data structures, for example, queue (FIFO), stack (LIFO), set (no order and no repeated values), lists, hash table, arrays, trees, and graphs
(4) Technical level: application data formats, for example, text, vector graphics, pixel images, audio signals, video sequences, and multimedia
(5) Hospital level: narrative (textual and natural language) patient record data (structured/unstructured and standardized/non-standardized), *omics data (genomics, proteomics, metabolomics, microarray data, fluxomics, and phenomics), numerical

measurements (physiological data, time series, lab results, vital signs, blood pressure, CO_2 partial pressure, temperature, etc.), recorded signals (ECG, EEG, ENG, EMG, EOG, and EP, e.g., from sleep monitoring, ambient assisted living, and gait sampling), graphics (sketches, drawings, handwriting, markers on images, etc.), audio signals (heart beat in the ICU, extracting knowledge from audio signals, hearing aids, sound therapy, etc.), images (from video cameras, radiological images from X-ray, MR, CT, PET, etc.), etc.

Data Preprocessing, Data Integration, and Data Fusion

Data preprocessing is a required first step before any machine learning machinery can be applied, because the algorithms learn from the data and the learning outcome for problem solving heavily depends on the proper data needed to solve a particular problem – which are called features. These features are key for learning and understanding, and therefore, machine learning is often considered as feature engineering. Data preprocessing, however, inflicts a heavy danger; for example, during the preprocessing, data can be inadvertently modified; for example, "interesting" data may be removed. Consequently, for discovery purposes, it would be wise to have a look at the original raw data first and maybe do a comparison between nonprocessed and preprocessed data.

Data integration is a hot topic generally and in health informatics specifically, and solutions can bridge the gap between clinical and biomedical research (bench vs. bed). This is becoming even more important due to the increasing amounts of heterogeneous, complex patient-related data sets, resulting from various sources including picture archiving and communication systems (PACS) and radiological information systems (RIS), hospital information systems (HIS), laboratory information systems (LIS), physiological and clinical data repositories, and all sorts of *omics data from laboratories, using samples from biobanks. The latter include large collections of DNA sequence data, proteomic and metabolic data, resulting from sophisticated high-throughput analytic technologies. Along with classical patient records containing large amounts of unstructured information (N.B., avoid the term unstructured data) and semistructured information, integration efforts incorporate enormous problems but at the same time offers new possibilities for translational research.

While data integration is on combining data from different sources and providing users with a unified view on these data (e.g., combining research results from different bioinformatics repositories), data fusion is matching various data sets that represent one and the same object into a single, **consistent representation.** In health informatics, these unified views are particularly important in high dimensions, for example, for integrating heterogeneous descriptions of the same set of genes. The main expectation is that fused data are more informative than the original inputs. Capturing all information describing a biological system is the implicit objective of all *omics methods; however, genomics, transcriptomics, proteomics, metabolomics, etc. need to be combined to approach this goal: valuable information can be obtained using various analytic techniques such as nuclear magnetic resonance, liquid chromatography, or gas chromatography coupled to mass spectrometry. Each method has inherent advantages and disadvantages but is complementary in terms of biological information, consequently combining multiple data sets, provided by different analytic platforms of utmost importance for discovery of new knowledge. For each platform, the relevant information is extracted in the first step. The obtained latent variables are then fused and further analyzed. The influence of the original variables is then calculated back and interpreted. There is plenty of open future research to include all possible sources of information.

The Curse of Dimensionality

A big problem is caused by the exponential increase in volume, when increasing dimensions within the Euclidean space. In such high-dimensional spaces, the intuitive concept of proximity or similarity gets lost. The ratio of a data object's nearest neighbor over its farthest neighbor is nearly 1, that is, suddenly all data objects are seemingly equidistant from each other, and so features get irrelevant. The fight against the curse of dimensionality is central in data science and essential for machine learning. There are several approaches to tackle the curse of dimensionality. Two of the most important ones are feature selection and feature extraction. These methods are summarized under the umbrella term dimension reduction.

The most useful subspace analysis technique in the medical domain is subspace clustering. The principle is illustrated in Fig. 1: automatic subspace clustering algorithms search for clusters not in the whole data space but within smaller subsets of dimensions in which discriminating clusters may be found. The goal is to understand data in terms of (a) groups of similar records (clusters) and (b) the underlying relationship to the dimensions (subspaces). However, here, a doctor in the loop must help to find out what is interesting and what is relevant.

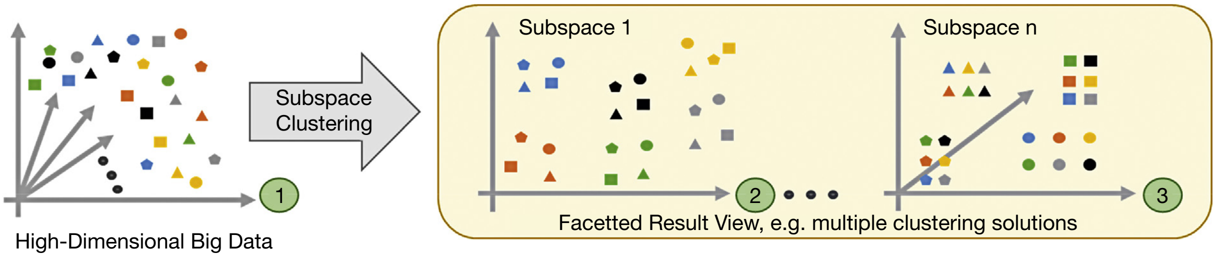
Subspace Clustering
1
High-Dimensional Big Data
Subspace 1
Subspace n
2
3
Facetted Result View, e.g. multiple clustering solutions

Fig. 1 Subspace clustering: automatic machine learning algorithms compute multiple, alternative solutions in different subspaces, that is, clustering by color (subspace 1) or by shape (subspace n); however, a doctor in the loop has finally to check the relevance of the data, that is, to answer the question "what is interesting," which cannot be done by any machine as "interest" is a genuinely human concept.

Example for Complex Data: Text

In hospital practice, the major medical documentation is only available in text format, and the amount of this unstructured information is immensely increasing. This is due to the fact that only the text of the findings in the patient record is **legally binding** – not the image nor any other related data. Text is the written form of natural language. The major problem in natural language understanding is the recognition of the context, that is, extracting knowledge from very few examples. Humans are very capable in the explorative learning of patterns from relatively few samples, while automatic machine learning methods need big data – and long processing time. Exactly, this is a challenge in the medical domain: often, we do not have big data, that is, large sets of training data, for example, with rare events, for example, in rare diseases. Rare diseases are often life-threatening and require a rapid intervention – the lack of much data makes automatic approaches nearly impossible. An example for such a rare disease with only few available data sets is cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL), a disease that is prevalent in 5 per 100,000 persons and is therefore the most frequent monogenic inherited apoplectic stroke in Central Europe.

Moreover, in clinical medicine, time is a crucial factor. A medical doctor needs the results quasi in real time or at least in a very short time (<5 min), for example, in emergency medicine or intensive care.

Particularly in the patient admission, human agents have the advantage to perceive the total situation at a glance. This aptitude results from the ability of transfer learning, where knowledge can be transferred from one situation to another situation, in which model parameters, that is, learned features or contextual knowledge, are transferred.

Consequently, the integration of a human (agent) into the whole machine learning loop can be beneficial in real-world situations, because the experience of a medical professional can help to reduce a search space of exponential possibilities drastically by heuristic selection of samples, thereby helping to solve NP-hard problems efficiently – or at least optimize them acceptably for a human end user to make a decision within a short time. Sometimes, it is better to have a good solution quickly, than a perfect solution never.

However, even the "simple" representation of natural language data presents many major challenges. It is difficult to automatically interpret even well-edited texts as well as a human expert would understand it. However, there have been advances in natural language processing (NLP), for example, the so-called "bag of words" methods, in which a document is treated as a collection of words occurring with some frequency; this works because they do not obscure this inherent meaning when presented to the analyst.

The first mechanized methods were developed in the 1960s for information retrieval, and the work of Salton et al. on identifying salient terms in a corpus, indexing, and constructing high-dimensional signature vectors that represent a corpus' topics or articles remains key to most of the current tools for analyzing big text data. A challenge is in mapping back the high-dimensional vectors into 2-D (or 3-D) representations to support visualizations that end users may understand and work on.

Future Outlook

Life sciences and human health are fundamentally biological, and biology is, according to Erwin Schrödinger, described as *the* information science. A very intriguing question is to what extent randomness and stochasticity play a role. By adopting the computational thinking approach to studying biological processes, we can improve our understanding and at the same time improve the design of algorithms. For example, the ability to define details of the interactions between small molecules and proteins promises

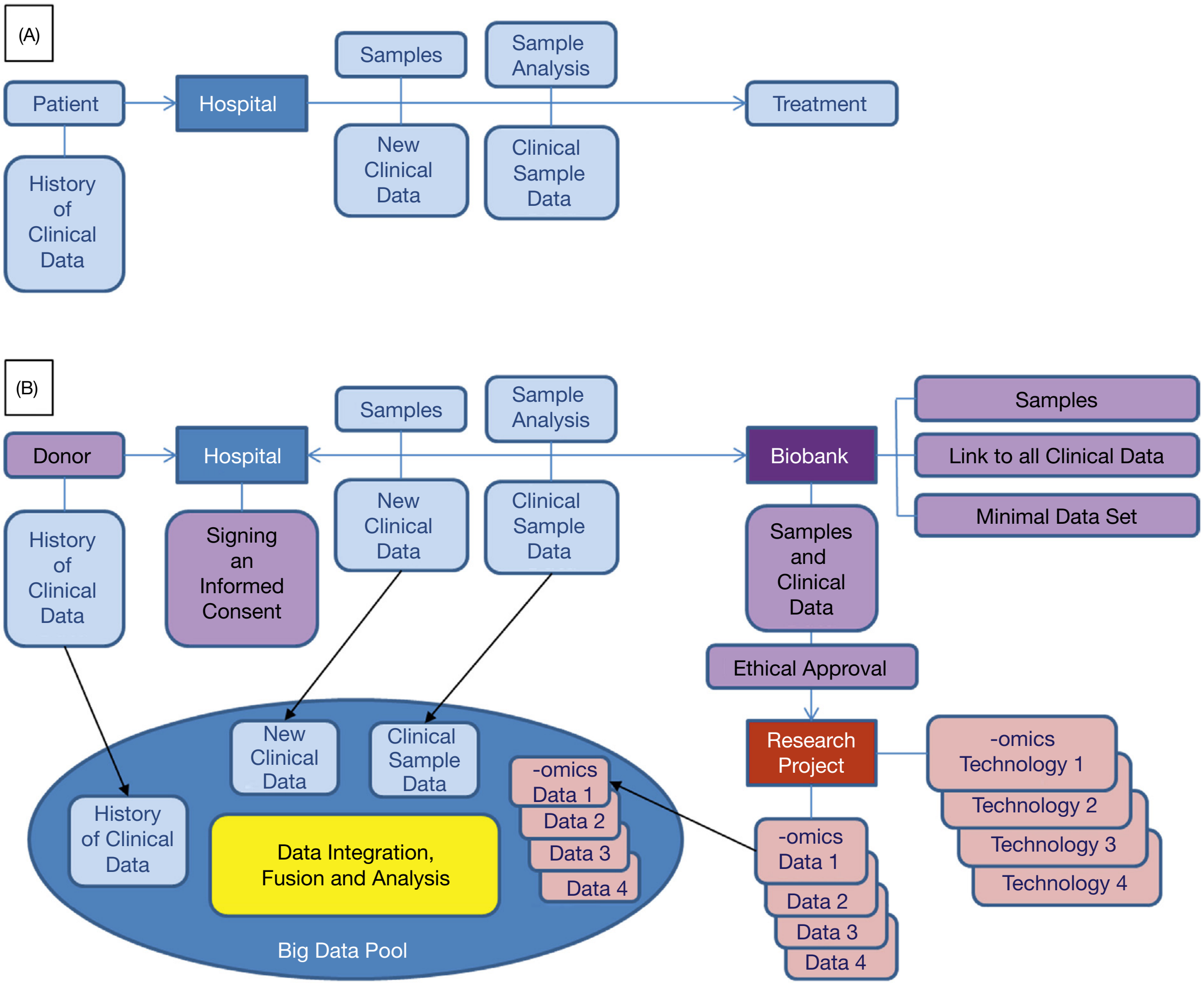

Fig. 2 (A) Typical traditional workflow in the hospital; (B) data integration, data fusion, and analysis as main future challenges in the cross-disciplinary workflows of hospitals and biobanks, as an underlying basis for future precision medicine.

unprecedented advances in the exploration of rational medical strategies, for example, for the therapy of infectious diseases and ultimately cancer. For this, the integration of *omics data and data from the health record is necessary, which remains open as a grand future challenge (see Fig. 2).

Further Reading

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Heidelberg: Springer.

Blanchet, L., & Smolinska, A. (2016). Data fusion in metabolomics and proteomics for biomarker discovery. In K. Jung (Ed.), *Statistical analysis in proteomics* (pp. 209–223). New York: Springer. https://doi.org/10.1007/978-1-4939-3106-4_14.

Bleiholder, J., & Naumann, F. (2008). Data fusion. *ACM Computing Survey (CSUR), 41*(1), 1–41. https://doi.org/10.1145/1456650.1456651.

Dehmer, M., Emmert-Streib, F., Pickl, S., & Holzinger, A. (Eds.). (2016). *Big data of complex networks*. Boca Raton, London, New York: CRC Press Taylor & Francis Group.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge (MA): MIT Press.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences, 20*(11), 818–829.

Holzinger, A. (2014). *Biomedical informatics: Discovering knowledge in big data*. New York: Springer. https://doi.org/10.1007/978-3-319-04528-3.

Holzinger, A. (2016a). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Inform., 3*(2), 119–131. https://doi.org/10.1007/s40708-016-0042-6.

Machine learning for health informatics: State-of-the-Art and future challenges. In Holzinger, A. (Ed.), *Lecture notes in artificial intelligence LNAI 9605*, (2016). Cham: Springer International. https://doi.org/10.1007/978-3-319-50478-0.

Holzinger, A., Dehmer, M., & Jurisica, I. (2014a). Knowledge discovery and interactive data mining in bioinformatics. State-of-the-art, future challenges and research directions. *BMC Bioinformatics, 15*(S6), I1. https://doi.org/10.1186/1471-2105-15-S6-I1.

Holzinger, A., & Jurisica, I. (2014). *Knowledge discovery and data mining in biomedical informatics: State-of-the-Art and future challenges, LNCS 8401*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-43968-5.

Holzinger, A., Stocker, C., & Dehmer, M. (2014b). Big complex biomedical data: Towards a taxonomy of data. In *Communications in computer and information science CCIS 455* (pp. 3–18). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-44791-8_1.

Hund, M., Boehm, D., Sturm, W., Sedlmair, M., Schreck, T., Ullrich, T., Keim, D. A., Majnaric, L., & Holzinger, A. (2016). Visual analytics for concept exploration in subspaces of patient groups: Making sense of complex datasets with the Doctor-in-the-loop. *Brain Informatics, 3*(4), 233–247. https://doi.org/10.1007/s40708-016-0043-5.

Huppertz, B., & Holzinger, A. (2014). Biobanks—A source of large biological data sets: open problems and future challenges. In *Lecture notes in computer science LNCS 8401* (pp. 317–330). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-43968-5_18.

Jeanquartier, F., Jean-Quartier, C., Schreck, T., Cemernek, D., & Holzinger, A. (2016). Integrating open data on cancer in support to tumor growth analysis. In E. M. Renda, M. Bursa, A. Holzinger, & S. Khuri (Eds.), *Lecture notes in computer science LNCS 9832* (pp. 49–66). Cham: Springer. https://doi.org/10.1007/978-3-319-43949-5_4.

Keogh, E., & Mueen, A. (2011). Curse of dimensionality. In *Encyclopedia of machine learning* (pp. 257–258). Springer.

Lafon, S., Keller, Y., & Coifman, R. R. (2006). Data fusion and multicue data matching by diffusion maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(11), 1784–1797. https://doi.org/10.1109/TPAMI.2006.223.

Lee, S., & Holzinger, A. (2016). Knowledge discovery from complex high dimensional data. In S. Michaelis, N. Piatkowski, & M. Stolpe (Eds.), *Challenges and algorithms, lecture notes in artificial intelligence, LNAI: vol. 9580. Solving large scale learning tasks* (pp. 148–167). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-41706-6_7.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Cambridge (MA): MIT press.

Salton, G., Wong, A., & Yang, C. S. (1975). Vector-Space Model for automatic indexing. *Communications of the ACM, 18*(11), 613–620.

Thomas, J. J., & Cook, K. A. (2005). *Illuminating the path: The research and development agenda for visual analytics*. New York: IEEE Computer Society Press.

Relevant Websites

http://hci-kdd.org/open-data-sets.
http://www.illc.uva.nl/EuroWordNet.
https://www.hhs.gov.