Extravaganza Tutorial on Hot Ideas for Interactive Knowledge Discovery and Data Mining in Biomedical Informatics

Andreas Holzinger[1,2]

[1] Research Unit Human–Computer Interaction, Institute for Medical Informatics,
Statistics & Documentation, Medical University Graz, Austria
a.holzinger@hci4all.at

[2] Institute for Information Systems and Computer Media
Graz University of Technology, Austria

Abstract. Biomedical experts are confronted with "Big data", driven by the trend towards precision medicine. Despite the fact that humans are excellent at pattern recognition in dimensions of ≤ 3, most biomedical data is in dimensions much higher than 3, making manual analysis often impossible. Experts in daily routine are decreasingly capable of dealing with such data. Efficient, useable and useful computational methods, algorithms and tools to interactively gain insight into such data are a commandment of the time. A synergistic combination of methodologies of two areas may be of great help here: Human–Computer Interaction (HCI) and Knowledge Discovery/Data Mining (KDD), with the goal of supporting human intelligence with machine learning. Mapping higher dimensional data into lower dimensions is a major task in HCI, and a concerted effort including recent advances from graph-theory and algebraic topology may contribute to finding solutions. Moreover, much biomedical data is sparse, noisy and time-dependent, hence entropy is also amongst promising topics. This tutorial gives an overview of the HCI-KDD approach and focuses on 3 topics: graphs, topology and entropy. The goal of this intro tutorial is to motivate and stimulate further research.

Keywords: Knowledge Discovery, Data Mining, HCI-KDD, Graph-based Text Mining, Topological Data Mining, Entropy-based Data Mining.

1 Introduction and Motivation

Experts in the life sciences have to deal with large amounts of complex, high-dimensional, heterogenous, noisy, and weakly structured data sets and massive sets of unstructured information from various sources [1], [2]. "Big Data" [3] in the medical domain is driven by the trend towards precision P4-medicine (Predictive, Preventive, Participatory, Personalized) and has resulted in an explosion in the amount of generated data sets, in particular "-omics" data, for

D. Ślęzak et al. (Eds.): BIH 2014, LNAI 8609, pp. 502–515, 2014.

example from genomics, proteomics, metabolomics, etc. [4]. Within such data, relevant *structural* patterns and/or *temporal* patterns (”knowledge”) are often hidden and not accessible to the expert. The progressively trend towards data intensive science, which is nearly a reverse of the classical hypothetico-deductive approach, makes optimization of discovery tools imperative [5], and calls for visual data mining approaches [6]. This paper is organized as follows: In section 2 some key terms are briefly explained. In section 3 the basic idea of the HCI-KDD approach is presented, along with the seven research areas involved, however, in the following we concentrate briefly on only three of them: In section 4 on graph-based data mining, in section 5 on topological data mining and in section 6 on entropy-based data mining, concluding by emphasizing that the *combination* of such approaches may bring added values. In the limited space given, such vast topics can only be touched, so the goal of this tutorial is to provide a coarse overview, to motivate and stimulate further research and to encourage to test crazy ideas.

2 Glossary and Key Terms

Algebraic Topology: is concerned with computations of homologies and homotopies in topological spaces [7].

Alpha Shapes: family of piecewise linear simple curves in the Euclidean plane associated with the shape of a finite set of points [8]; i.e. α-shapes are a generalization of the convex hull of a point set: Let $\mathbf{S}$ be a finite set in $\mathbb{R}^3$ and α a real number $0 \leq \alpha \leq \infty$; the u-shape of $\mathbf{S}$ is a polytope that is neither necessarily convex nor necessarily connected. For $\alpha \rightarrow \infty$ the α-shape is identical to the convex hull of $\mathbf{S}$ [9]; important e.g. in protein-related interactions [10].

Betti Number: can be used to distinguish topological spaces based on the connectivity of n-dimensional simplicial complexes: In dimension k, the rank of the k-th homology group is denoted β_k, useful in the presence of noisy shapes, because Betti numbers can be used as shape descriptor admitting dissimilarity distances stable under continuous shape deformations [11].

Graph mining: is the application of graph-based methods to structural data sets [12], a survey on graph mining can be found here [13].

Homomorphism: is a function that preserve the operators associated with the specified structure.

Homotopy: Given two maps $f, g : X \rightarrow Y$ of topological spaces, f and g are homotopic, $f \simeq g$, if there is a continuous map $H : X \times [0,1] \rightarrow Y$ so that $H(x,0) = f(x)$ and $H(x,1) = g(x)$ for all $x \in X$ [14].

Homology: (and cohomology) are algebraic objects associated to a manifold, which give one measure of the number of holes of the object. Computation of the homology groups of topological spaces is a central topic in topology; if the simplicial complex is small, the homology group computations can be done manually; to solve such problems generally a classic algorithm exists [15].

Human–Computer Interaction: study, design and development of the interaction between end users and computers; this classic definition goes back to the work of Alan Newell and Herbert Simon (refs), and HCI research has in the last decades focused almost exclusively on ergonomics of the user interface, while the HCI-KDD approach concentrates almost exclusively on human–data interaction.

Information Entropy: is a measure of the uncertainty in a random variable. This refers to the Shannon entropy, which quantifies the expected value of the information contained in a message.

Manifold: is a fundamental mathematical object which locally resembles a line, a plane, or space.

Network: Synonym for a graph, which can be defined as an ordered or unordered pair (N, E) of a set N of nodes and a set E of edges [16]. Engineers often mention: Data + Graph = Network, or call at least directed graphs as networks; however, in theory, there is no difference between a graph and a network.

Pattern discovery: subsumes a plethora of machine learning methods to detect complex patterns in data sets [17]; applications thereof are, for instance, graph mining [18] and string matching [19].

Persistent Homology: Persistent homology is an algebraic tool for measuring topological features of shapes and functions. It casts the multi-scale organization we frequently observe in nature into a mathematical formalism [20].

Simplicial Complex: is made up of simplices, e.g. a simplicial polytope has simplices as faces and a simplicial complex is a collection of simplices pasted together in any reasonable vertex-to-vertex and edge-to-edge arrangement. A graph is a 1-dim simplicial complex.

Small world networks: are generated based on certain rules with high clustering coefficient [16, 21] but the distances among the vertices are rather short in average, hence they are somewhat similar to random networks and they have been found in several classes of biological networks, see [22].

Topological Entropy: is a nonnegative real number that is a measure of the complexity of a dynamical system [23].

3 The HCI-KDD Approach

Humans are very good at pattern recognition in the low-dimensional space, although humans do not see in three spatial dimensions directly, but rather via sequences of planar projections integrated in a manner that is *sensed* if not comprehended. Humans spend a lot of their life time to learn how to infer three-dimensional spatial data from paired planar projections. Years of practice have tuned a remarkable ability to extract global structures from representations in lower dimension. On the other hand, computers can be used to deal with high-dimensional data, where we can make use of the benefits of computational topology [24], e.g. by replacing a set of point cloud data with a simplicial complex, which converts the data into global topological objects. To combine the most

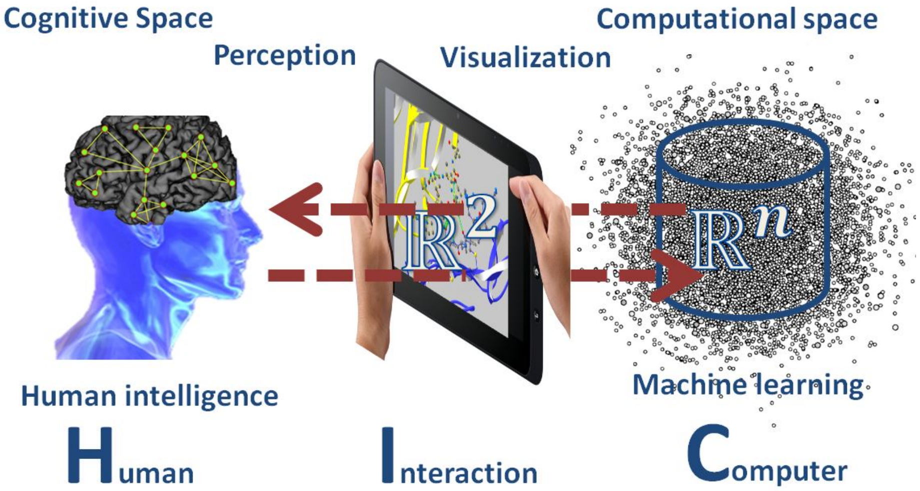

Fig. 1. This image, created originally by A. Holzinger as logo for his group, emphasizes the importance of the interaction between high-dimensional computational spaces in $\mathbb{R}^n$ and highlights the reality that current devices only allow data visualization in $\mathbb{R}^2$. Consequently, the major task of Human–Computer Interaction is to map data from high-dimensional spaces into lower-dimensional spaces, hence enabling interaction, which is the most difficult and challenging task in this field.

desirable of these formidable talents might highly benefit the knowledge discovery process [25]. The most critical and not easy endeavour is in interaction and visualization (see Figure 1).

The original idea of the HCI-KDD [26] approach (Figure 2) is in combining aspects of the best of two worlds: Human–Computer Interaction (HCI), with emphasis on perception, cognition, interaction, reasoning, decision making, human learning and human intelligence, and Knowledge Discovery/Data Mining (KDD), dealing with data processing, computational statistics, artificial intelligence and particularly with machine learning [27].

Whilst interactive knowledge discovery encompasses the horizontal process ranging from physical aspects of data (left in Figure 2) to the human aspects of information processing (right in Figure 2), data mining can be seen vertically and deals specifically with methods, algorithms and tools for finding patterns in the data. In the HCI-KDD approach, seven (the new magical number 7) essential research areas can be determined as outlined in Figure 2, including: Area 1: Data integration, data fusion and data mapping; Area 2: mining algorithms and Area 6: data visualization [28], [29], [30]. This tutorial focuses on three hot topics: ***Area 3: Graph-based Data Mining (GDM)*** [31], [32], [33],[34].

Area 4: Entropy-based Data Mining (EDM) [35], [36].
Area 5: Topological Data Mining (TDM) [37].

In the biomedical domain as in some other domains issues of Area 7: privacy, data protection, safety and security are mandatory [38].

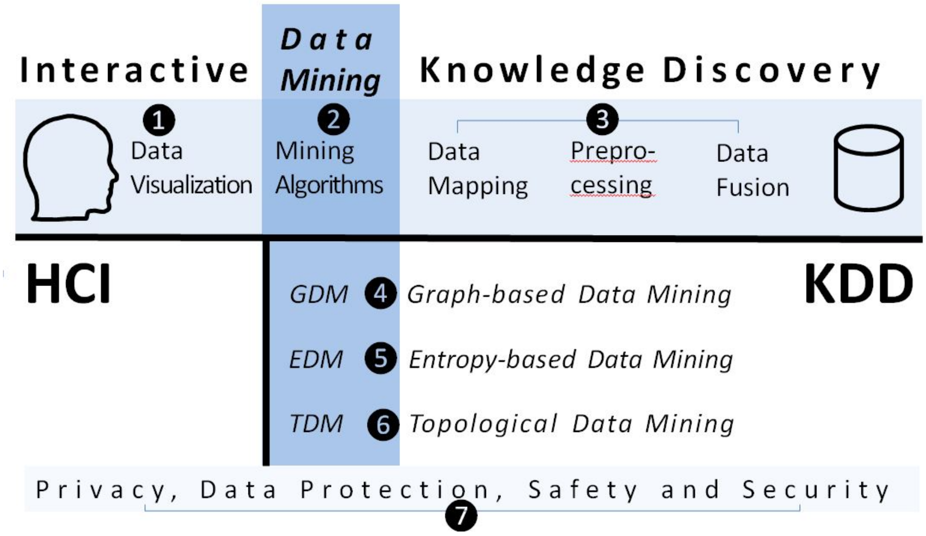
Interactive
Data Mining
Knowledge Discovery
1
Data Visualization
2
Mining Algorithms
3
Data Mapping
Prepro-cessing
Data Fusion
HCI
KDD
GDM 4 Graph-based Data Mining
EDM 5 Entropy-based Data Mining
TDM 6 Topological Data Mining
Privacy, Data Protection, Safety and Security
7

Fig. 2. The big picture of the HCI-KDD approach: KDD encompasses the whole horizontal process chain from data to information and knowledge; actually from physical aspects of raw data, to human aspects including attention, memory, vision, interaction etc. as core topics in HCI, whilst DM as a vertical subject focuses on the development of methods, algorithms and tools for data mining (Image taken from the hci4all.at website, as of May, 2014)

4 Graph-Based Data Mining

Graph-Theory [39] provides powerful tools to map data structures and to find novel connections between single data objects [16, 40]. The inferred graphs can be further analyzed by using graph-theoretical, statistical and machine learning techniques [41]. A mapping of already existing and in medical practice approved *knowledge spaces* as a conceptual graph (as e.g. demonstrated in [32] and a subsequent visual and graph-theoretical analysis can bring novel insights on hidden patterns in the data, which exactly is the goal of knowledge discovery. Another benefit of a graph-based data structure is in the applicability of methods from network topology and network analysis and data mining, e.g. small-world phenomenon [42, 43], and cluster analysis [44, 45]. However, the first question is ”How to get a graph?”, or simpler ”How to get point sets?”, because point cloud data sets (PCD) are used as primitives for such approaches. The answer to this question is not trivial (see [46]), apart from “naturally available” point clouds, e.g. from laser scanners, protein structures [47], or text mapped into a set of points (vectors) in $\mathbb{R}^n$. Sticking on the last example, graphs are intuitively more informative as example words/phrase representations [48], and graphs are the best studied data structures in computer science, with a strong relation to logical languages [49]. The beginning of graph-based data mining approaches was two decades ago, some pioneering work include [50–52]. According to [49] there are five theoretical bases of graph-based data mining approaches such as (1) subgraph categories, (2) subgraph isomorphism, (3) graph invariants, (4) mining measures and (5) solution methods. Furthermore, there are five groups of different graph-theoretical approaches for data mining such as (1) greedy search based approach, (2) inductive logic programming based approach, (3) inductive database based approach, (4) mathematical graph theory based approach and (5) kernel function based approach [53]. However, the main disadvantage of graph-theoretical text mining is the computational complexity of the graph representation, consequently the goal of future research in the field of graph-theoretical approaches for text mining is to develop efficient graph mining algorithms which implement effective search strategies and data structures [48].

In [54] a graph-theoretical approach for text mining is used to extract relation information between terms in "free-text" electronic health care records that are semantically or syntactically related. Another field of application is the text analysis of web and social media for detecting influenza-like illnesses [55].

Moreover there can be content-rich relationship networks among biological concepts, genes, proteins and drugs developed with topological text data mining like shown in [56]. According to [57] network medicine describes the clinical application field of topological text mining due to addressing the complexity of human diseases with molecular and phenotypic network maps.

5 Topological Data Mining

Closely related to graph-based methods are topological data mining methods; for both we need point cloud data sets - or at least distances - as input. A set of such primitives forms a space, and if we have finite sets equipped with proximity or similarity measure functions $sim_q \colon S^{q+1} \to [0,1]$, which measure how "close" or "similar" $(q+1)$-tuples of elements of S are, we speak about a *topological space*. A value of 0 means totally different objects, while 1 corresponds to equivalent items. Interesting are manifolds, which can be seen as a topological space, which is locally homeomorphic (that means it has a continuous function with an inverse function) to a real n-dimensional space. In other words: X is a d-manifold if every point of X has a neighborhood homeomorphic to $\mathbb{B}^d$; with boundary if every point has a neighborhood homeomorphic to $\mathbb{B}$ or $\mathbb{B}^d_+$ [58].

A topological space may be viewed as an abstraction of a metric space, and similarly, manifolds generalize the connectivity of d-dimensional Euclidean spaces $\mathbb{B}^d$ by being locally similar, but globally different. A d-dimensional chart at $p \in X$ is a homeomorphism $\phi : U \to \mathbb{R}^d$ onto an open subset of $\mathbb{R}^d$, where U is a neighborhood of p and open is defined using the metric. A d-dimensional manifold (d-manifold) is a topological space X with a d-dimensional chart at every point $x \in X$ [59].

For us also interesting are simplicial complexes (”simplicials”) which are spaces described in a very particular way, the basis is in Homology. The reason is that it is not possible to represent surfaces precisely in a computer system due to limited computational storage; thus, surfaces are sampled and represented with triangulations. Such a triangulation is called a simplicial complex, and is a combinatorial space that can represent a space. With such simplicial complexes, the topology of a space from its geometry can be separated. Zomorodian [59] compares it with the separation of syntax and semantics in logic.

Topological techniques originated in pure mathematics, but have been adapted to the study and analysis of data during the past two decades. The two most popular topological techniques in the study of data are *homology* and *persistence.* The connectivity of a space is determined by its cycles of different dimensions. These cycles are organized into groups, called homology groups. Given a reasonably explicit description of a space, the homology groups can be computed with linear algebra. Homology groups have a relatively strong discriminative power and a clear meaning, while having low computational cost. In the study of persistent homology the invariants are in the form of persistence diagrams or barcodes [60].

In data mining it is important to extract significant features, and exactly for this, topological methods are useful, since they provide robust and general feature definitions with emphasis on global information, for example Alpha Shapes [9].

A recent example for topological data mining is given by [61]: Topological text mining, which builds on the well-known vector space model, which is a standard approach in text mining [62]: a collection of text documents (corpus) is mapped into points (=vectors) in $\mathbb{R}^n$. Moreover, each word can be mapped into so-called term vectors, resulting in a very high dimensional vector space. If there are n words extracted from all the documents then each document is mapped to a point (*term vector*) in $\mathbb{R}^{\ltimes}$ with coordinates corresponding to the weights. This way the whole corpus can be transformed into a point cloud data set. Instead of the Euclidean metric the use of a similarity (proximity) measure is sometimes more convenient; the *cosine similarity measure* is a typical example: the cosine of the angle between two vectors (points in the cloud) reflects how “similar” the underlying weighted combinations of keywords are. Amongst the many different text mining methods (for a recent overview refer to [63]); topological approaches are promising, but need a lot of further research.

Due to finding meaningful topological patterns greater information depth can be achieved from the same data input [64]. However, with increasing complexity of the data to process also the need to find a scalable shape characteristic is greater [65]. Therefore methods of the mathematical field of topology are used for complex data areas like the biomedical field [65], [60]. Topology as the mathematical study of shapes and spaces that are not rigid [65], pose a lot of possibilities for the application in knowledge discovery and data mining, as topology is the study of connectivity information and it deals with qualitative geometric properties [66].

One of the main tasks of applied topology is to find and analyse higher dimensional topological structures in lower dimensional spaces (e.g. point cloud from vector space model as discussed in [64]). A common way to describe topological spaces is to first create simplicial complexes, because a simplicial complex structure on a topological space is an expression of the space as a union of simplices such as points, intervals, triangles, and higher dimensional analogues. Simplicial complexes provide an easy combinatorial way to define certain topological spaces [66]. A simplical complex K is defined as a finite collection of simplices such that $\sigma \in K$ and τ, which is a face of σ, implies $\tau \in K$, and $\sigma, \sigma' \in K$ implies $\sigma \cap \sigma'$ can either be a face of both σ and σ' or empty [67]. One way to create a simplical complex is to examine all subsets of points, and if any subsets of points are close enough, a p-simplex (e.g. line) is added to the complex with those points as vertices. For instance, a Vietoris-Rips complex of diameter ϵ is defined as $VR(\epsilon) = \sigma | diam(\sigma) \leq \epsilon$, where $diam(\epsilon)$ is defined as the largest distance between two points in σ [67]. A common way a analyse the topological structure is to use persistent homology, which identifies cluster, holes and voids therein. It is assumed that more robust topological structures are the one which persist with increasing ϵ. For detailed information about persistent homology, see [67], [68], [69].

6 Entropy-Based Data Mining

In the real medical world, we are confronted not only with complex and high-dimensional data sets, but usually with sparse, noisy, incomplete and uncertain data, where the application of traditional methods of knowledge discovery and data mining always entail the danger of modeling artifacts. Originally, information entropy was introduced by Shannon (1949), as a measure of *uncertainty in the data*. To date, there have emerged many different types of entropy methods with a large number of different purposes and applications. Here we mention only two:

Graph Entropy was described by [70] to measure structural information content of graphs, and a different definition, more focused on problems in information and coding theory, was introduced by Körner in [71]. Graph entropy is often used for the characterization of the structure of graph-based systems, e.g. in mathematical biochemistry, but also for any complex network [72]. In these applications the entropy of a graph is interpreted as its structural information content and serves as a complexity measure, and such a measure is associated with an equivalence relation defined on a finite graph; by application of Shannons Eq. 2.4 in [41] with the probability distribution we get a numerical value that serves as an index of the structural feature captured by the equivalence relation [41].

Topological Entropy (TopEn), was introduced by [73] with the purpose to introduce the notion of entropy as an invariant for continuous mappings: Let (X, T) be a topological dynamical system, i.e., let X be a nonempty compact

Hausdorff space and $T : X \rightarrow X$ a continuous map; the TopEn is a nonnegative number which measures the complexity of the system [74].

Hornero et al. [75] performed a complexity analysis of intracranial pressure dynamics during periods of severe intracranial hypertension. For that purpose they analyzed eleven episodes of intracranial hypertension from seven patients. They measured the changes in the intracranial pressure complexity by applying ApEn, as patients progressed from a state of normal intracranial pressure to intracranial hypertension, and found that a decreased complexity of intracranial pressure coincides with periods of intracranial hypertension in brain injury. Their approach is of particular interest to us, because they proposed classification based on ApEn tendencies instead of absolute values.

Pincus et al. took in [76] heart rate recordings of 45 healthy infants with recordings of an infant one week after an aborted sudden infant death syndrom (SIDS) episode. They then calculated the ApEn of these recordings and found a significant smaller value for the aborted SIDS infant compared to the healthy ones.

Holzinger et al. (2012) [77] experimented with point cloud data sets in the two dimensional space: They developed a model of handwriting, and evaluated the performance of entropy based slant and skew correction, and compared the results to other methods. This work is the basis for further entropy-based approaches, which are very relevant for advanced entropy-based data mining approaches.

7 Conclusion and Future Outlook

Discovering knowledge in complex, high-dimensional data sets needs a concerted effort of various topics, ranging from data preprocessing, data fusion, data integration and data mapping to interactive visualization within a low-dimensional space. For this reason, graph-based and topological methods are very useful, since they provide robust and general feature definitions and may support a "global information view". A promising area of future research is in graph-theoretical approaches for text mining, in particular to develop efficient graph mining algorithms which implement robust and efficient search strategies and data structures [48]. Such approaches could be combined with techniques from machine learning, e.g. multi-agents and evolutionary algorithms [78]. However, there remain many open questions, for example about the graph characteristics and the isomorphism complexity [49], to mention just only one. A further promising research route is to combine such methods with entropy-based approaches, which have extensively been applied for analyzing sparse and noisy time series data, but so far have not yet been applied to weakly structured data in combination with techniques from computational topology. Consequently, the inclusion of entropy measures for discovery of knowledge in high-dimensional biomedical data is a big future issue, opening a lot of challenging research routes [35].

The grand vision for the future is to effectively support human learning with machine learning. The HCI-KDD network of excellence is proactively supporting this vision in bringing together people with diverse background - sharing a common goal: finding solutions for dealing with big and complex data sets. A recent output of the network can be found here [79] (for more information please refer to www.hci4all.at).

Acknowledgments. The author is very grateful for the friendly support of Dominik Slezak, Jerzy Stefanowski, Juzhen Dong and Andrzej Skowron.

References

1. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge discovery and interactive data mining in bioinformatics - state-of-the-art, future challenges and research directions. BMC Bioinformatics 15, I1 (2014)
2. Holzinger, A.: Biomedical Informatics: Discovering Knowledge in Big Data. Springer, New York (2014)
3. Wu, X.D., Zhu, X.Q., Wu, G.Q., Ding, W.: Data mining with big data. IEEE Transactions on Knowledge and Data Engineering 26, 97–107 (2014)
4. Huppertz, B., Holzinger, A.: Biobanks – A source of large biological data sets: Open problems and future challenges. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 317–330. Springer, Heidelberg (2014)
5. Mattmann, C.A.: Computing: A vision for data science. Nature 493, 473–475 (2013)
6. Otasek, D., Pastrello, C., Holzinger, A., Jurisica, I.: Visual data mining: Effective exploration of the biological universe. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 19–33. Springer, Heidelberg (2014)
7. Hatcher, A.: Algebraic Topology. Cambridge University Press, Cambridge (2002)
8. Edelsbrunner, H., Kirkpatrick, D., Seidel, R.: On the shape of a set of points in the plane. IEEE Transactions on Information Theory 29, 551–559 (1983)
9. Edelsbrunner, H., Mucke, E.P.: 3-dimensional alpha-shapes. ACM Transactions on Graphics 13, 43–72 (1994)
10. Albou, L.P., Schwarz, B., Poch, O., Wurtz, J.M., Moras, D.: Defining and characterizing protein surface using alpha shapes. Proteins-Structure Function and Bioinformatics 76, 1–12 (2009)
11. Frosini, P., Landi, C.: Persistent betti numbers for a noise tolerant shape-based approach to image retrieval. Pattern Recognition Letters 34, 863–872 (2013)
12. Cook, D., Holder, L.B.: Mining Graph Data. Wiley Interscience (2007)
13. Chakrabarti, D., Faloutsos, C.: Graph mining: Laws, generators, and algorithms. ACM Computing Surveys (CSUR) 38, 2 (2006)
14. Whitehead, G.W.: Elements of homotopy theory. Springer (1978)
15. Munkres, J.R.: Elements of algebraic topology, vol. 2. Addison-Wesley Reading (1984)
16. Dorogovtsev, S., Mendes, J.: Evolution of networks: From biological nets to the Internet and WWW. Oxford University Press (2003)
17. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, vol. 2. Wiley, New York (2000)
18. Cook, D.J., Holder, L.B.: Graph-based data mining. IEEE Intelligent Systems and their Applications 15, 32–41 (2000)

19. Gusfield, D.: Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press (1997)
20. Edelsbrunner, H., Harer, J.: Persistent homology - a survey. Contemporary Mathematics Series, vol. 453, pp. 257–282. Amer Mathematical Soc., Providence (2008)
21. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393, 440–442 (1998)
22. Emmert-Streib, F., Dehmer, M.: Networks for systems biology: Conceptual connection of data and function. IET Systems Biology 5, 185–207 (2011)
23. Koslicki, D.: Topological entropy of dna sequences. Bioinformatics 27, 1061–1067 (2011)
24. Ghrist, R.: Barcodes: the persistent topology of data. Bulletin of the American Mathematical Society 45, 61–75 (2008)
25. Holzinger, A.: Human-computer interaction and knowledge discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In: Cuzzocrea, A., Kittl, C., Simos, D.E., Weippl, E., Xu, L. (eds.) CD-ARES 2013. LNCS, vol. 8127, pp. 319–328. Springer, Heidelberg (2013)
26. Holzinger, A.: On knowledge discovery and interactive intelligent visualization of biomedical data - challenges in human–computer interaction and biomedical informatics. In: DATA 2012, Rome, Italy, pp. 9–20 (2012)
27. Holzinger, A., Jurisica, I.: Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 1–18. Springer, Heidelberg (2014)
28. Holzinger, A., Bruschi, M., Eder, W.: On interactive data visualization of physiological low-cost-sensor data with focus on mental stress. In: Cuzzocrea, A., Kittl, C., Simos, D.E., Weippl, E., Xu, L. (eds.) CD-ARES 2013. LNCS, vol. 8127, pp. 469–480. Springer, Heidelberg (2013)
29. Wong, B.L.W., Xu, K., Holzinger, A.: Interactive visualization for information analysis in medical diagnosis. In: Holzinger, A., Simonic, K.-M. (eds.) USAB 2011. LNCS, vol. 7058, pp. 109–120. Springer, Heidelberg (2011)
30. Wiltgen, M., Holzinger, A., Tilz, G.P.: Interactive analysis and visualization of macromolecular interfaces between proteins. In: Holzinger, A. (ed.) USAB 2007. LNCS, vol. 4799, pp. 199–212. Springer, Heidelberg (2007)
31. Preuss, M., Dehmer, M., Pickl, S., Holzinger, A.: On terrain coverage optimization by using a network approach for universal graph-based data mining and knowledge discovery. In: Proceedings of the Active Media Technology - 10th International Conference, AMT 2014, Warsaw, Poland, August 11-14. LNCS, vol. 8610, Springer, Heidelberg (2014)
32. Holzinger, A., Ofner, B., Dehmer, M.: Multi-touch graph-based interaction for knowledge discovery on mobile devices: State-of-the-art and future challenges. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 241–254. Springer, Heidelberg (2014)
33. Holzinger, A., Malle, B., Aigner, R., Giuliani, N.: On graph extraction from image data. In: Slezak, D., Schaefer, G., Vuong, T.S., Kim, Y.S. (eds.) Active Media Technology AMT 2014. LNCS, vol. 8610, Springer, Heidelberg (2014)
34. Holzinger, A., Ofner, B., Stocker, C., Calero Valdez, A., Schaar, A.K., Ziefle, M., Dehmer, M.: On graph entropy measures for knowledge discovery from publication network data. In: Cuzzocrea, A., Kittl, C., Simos, D.E., Weippl, E., Xu, L. (eds.) CD-ARES 2013. LNCS, vol. 8127, pp. 354–362. Springer, Heidelberg (2013)

35. Holzinger, A., Hörtenhuber, M., Mayer, C., Bachler, M., Wassertheurer, S., Pinho, A.J., Koslicki, D.: On entropy-based data mining. In: Holzinger, A., Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. LNCS, vol. 8401, pp. 209–226. Springer, Heidelberg (2014)
36. Holzinger, A., Stocker, C., Bruschi, M., Auinger, A., Silva, H., Gamboa, H., Fred, A.: On applying approximate entropy to ECG signals for knowledge discovery on the example of big sensor data. In: Huang, R., Ghorbani, A.A., Pasi, G., Yamaguchi, T., Yen, N.Y., Jin, B. (eds.) AMT 2012. LNCS, vol. 7669, pp. 646–657. Springer, Heidelberg (2012)
37. Holzinger, A.: On topological data mining. In: Holzinger, A., Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. LNCS, vol. 8401, pp. 331–356. Springer, Heidelberg (2014)
38. Kieseberg, P., Hobel, H., Schrittwieser, S., Weippl, E., Holzinger, A.: Protecting anonymity in data-driven biomedical science. In: Holzinger, A., Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. LNCS, vol. 8401, pp. 301–316. Springer, Heidelberg (2014)
39. Harary, F.: Structural models. An introduction to the theory of directed graphs. Wiley (1965)
40. Strogatz, S.: Exploring complex networks. Nature 410, 268–276 (2001)
41. Dehmer, M., Mowshowitz, A.: A history of graph entropy measures. Information Sciences 181, 57–78 (2011)
42. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science 286, 509–512 (1999)
43. Kleinberg, J.: Navigation in a small world. Nature 406, 845–845 (2000)
44. Koontz, W., Narendra, P., Fukunaga, K.: A graph-theoretic approach to nonparametric cluster analysis. IEEE Transactions on Computers 100, 936–944 (1976)
45. Wittkop, T., Emig, D., Truss, A., Albrecht, M., Boecker, S., Baumbach, J.: Comprehensive cluster analysis with transitivity clustering. Nature Protocols 6, 285–295 (2011)
46. Holzinger, A., Malle, B., Bloice, M., Wiltgen, M., Ferri, M., Stanganelli, I., Hofmann-Wellenhof, R.: On the generation of point cloud data sets: Step one in the knowledge discovery process. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 57–80. Springer, Heidelberg (2014)
47. Canutescu, A.A., Shelenkov, A.A., Dunbrack, R.L.: A graph-theory algorithm for rapid protein side-chain prediction. Protein science 12, 2001–2014 (2003)
48. Jiang, C., Coenen, F., Sanderson, R., Zito, M.: Text classification using graph mining-based feature extraction. Knowledge-Based Systems 23, 302–308 (2010)
49. Washio, T., Motoda, H.: State of the art of graph-based data mining. ACM SIGKDD Explorations Newsletter 5, 59 (2003)
50. Cook, D.J., Holder, L.B.: Substructure discovery using minimum description length and background knowledge. J. Artif. Int. Res. 1, 231–255 (1994)
51. Yoshida, K., Motoda, H., Indurkhya, N.: Graph-based induction as a unified learning framework. Applied Intelligence 4, 297–316 (1994)
52. Dehaspe, L., Toivonen, H.: Discovery of frequent DATALOG patterns. Data Mining and Knowledge Discovery 3, 7–36 (1999)
53. Windridge, D., Bober, M.: A kernel-based framework for medical big-data analytics. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 197–208. Springer, Heidelberg (2014)

54. Zhou, X., Han, H., Chankai, I., Prestrud, A., Brooks, A.: Approaches to text mining for clinical medical records. In: Proceedings of the 2006 ACM symposium on Applied computing - SAC 2006, p. 235. ACM Press, New York (2006)
55. Corley, C.D., Cook, D.J., Mikler, A.R., Singh, K.P.: Text and structural data mining of influenza mentions in Web and social media. International journal of environmental research and public health 7, 596–615 (2010)
56. Chen, H., Sharp, B.M.: Content-rich biological network constructed by mining PubMed abstracts. BMC bioinformatics 5, 147 (2004)
57. Barabási, A., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. Nature Reviews Genetics 12, 56–68 (2011)
58. Cannon, J.W.: The recognition problem: what is a topological manifold? Bulletin of the American Mathematical Society 84, 832–866 (1978)
59. Zomorodian, A.: Chapman & Hall/CRC Applied Algorithms and Data Structures series. In: Computational Topology, pp. 1–31. Chapman and Hall, Boca Raton (2010), doi:10.1201/9781584888215-c3.
60. Epstein, C., Carlsson, G., Edelsbrunner, H.: Topological data analysis. Inverse Problems 27, 120201 (2011)
61. Wagner, H., Dlotko, P.: Towards topological analysis of high-dimensional feature spaces. Computer Vision and Image Understanding 121, 21–26 (2014)
62. Kobayashi, M., Aono, M.: Vector space models for search and cluster mining. In: Berry, M.W. (ed.) Survey of Text Mining: Clustering, Classification, and Retrieval, pp. 103–122. Springer, New York (2004)
63. Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., Verspoor, K.: Biomedical text mining: State-of-the-art, open problems and future challenges. In: Holzinger, A., Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. LNCS, vol. 8401, pp. 271–300. Springer, Heidelberg (2014)
64. Wagner, H., Dlotko, P., Mrozek, M.: Computational topology in text mining. In: Ferri, M., Frosini, P., Landi, C., Cerri, A., Di Fabio, B. (eds.) CTIC 2012. LNCS, vol. 7309, pp. 68–78. Springer, Heidelberg (2012)
65. Nicolau, M., Levine, A.J., Carlsson, G.: Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proceedings of the National Academy of Sciences of the United States of America 108, 7265–7270 (2011)
66. Carlsson, G.: Topology and Data. Bull. Amer. Math. Soc. 46, 255–308 (2009)
67. Zhu, X.: Persistent homology: An introduction and a new text representation for natural language processing. In: Rossi, F. (ed.) IJCAI. IJCAI/AAAI (2013)
68. Cerri, A., Fabio, B.D., Ferri, M., Frosini, P., Landi, C.: Betti numbers in multidimensional persistent homology are stable functions. Mathematical Methods in the Applied Sciences 36, 1543–1557 (2013)
69. Bubenik, P., Kim, P.T.: A statistical approach to persistent homology. Homology, Homotopy and Applications 9, 337–362 (2007)
70. Mowshowitz, A.: Entropy and the complexity of graphs: I. an index of the relative complexity of a graph. The Bulletin of Mathematical Biophysics 30, 175–204 (1968)
71. Körner, J.: Coding of an information source having ambiguous alphabet and the entropy of graphs. In: 6th Prague Conference on Information Theory, pp. 411–425 (1973)

72. Holzinger, A., Ofner, B., Stocker, C., Calero Valdez, A., Schaar, A.K., Ziefle, M., Dehmer, M.: On graph entropy measures for knowledge discovery from publication network data. In: Cuzzocrea, A., Kittl, C., Simos, D.E., Weippl, E., Xu, L. (eds.) CD-ARES 2013. LNCS, vol. 8127, pp. 354–362. Springer, Heidelberg (2013)
73. Adler, R.L., Konheim, A.G., McAndrew, M.H.: Topological entropy. Transactions of the American Mathematical Society 114, 309–319 (1965)
74. Adler, R., Downarowicz, T., Misiurewicz, M.: Topological entropy. Scholarpedia 3, 2200 (2008)
75. Hornero, R., Aboy, M., Abasolo, D., McNames, J., Wakeland, W., Goldstein, B.: Complex analysis of intracranial hypertension using approximate entropy. Crit. Care Med. 34, 87–95 (2006)
76. Pincus, S.M.: Approximate entropy as a measure of system complexity. Proceedings of the National Academy of Sciences 88, 2297–2301 (1991)
77. Holzinger, A., Stocker, C., Peischl, B., Simonic, K.M.: On using entropy for enhancing handwriting preprocessing. Entropy 14, 2324–2350 (2012)
78. Holzinger, K., Palade, V., Rabadan, R., Holzinger, A.: Darwin or lamarck? Future challenges in evolutionary algorithms for knowledge discovery and data mining. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 35–56. Springer, Heidelberg (2014)
79. Holzinger, A., Jurisica, I.: Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions. In: Holzinger, A., Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. LNCS, vol. 8401, pp. 1–18. Springer, Heidelberg (2014)