Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions

Andreas Holzinger[1] and Igor Jurisica[2]

[1] Medical University Graz, Institute for Medical Informatics, Statistics and Documentation
Research Unit HCI, Austrian IBM Watson Think Group,
Auenbruggerplatz 2/V, A-8036 Graz, Austria
a.holzinger@hci4all.at
[2] Princess Margaret Cancer Centre, University Health Network, IBM Life Sciences Discovery Centre, and TECHNA Institute for the Advancement of Technology for Health,
TMDT 11-314, 101 College Street, Toronto, ON M5G 1L7, Canada
juris@ai.utoronto.ca

Abstract. Biomedical research is drowning in data, yet starving for knowledge. Current challenges in biomedical research and clinical practice include information overload – the need to combine vast amounts of structured, semi-structured, weakly structured data and vast amounts of unstructured information – and the need to optimize workflows, processes and guidelines, to increase capacity while reducing costs and improving efficiencies. In this paper we provide a very short overview on interactive and integrative solutions for knowledge discovery and data mining. In particular, we emphasize the benefits of including the end user into the "interactive" knowledge discovery process. We describe some of the most important challenges, including the need to develop and apply novel methods, algorithms and tools for the integration, fusion, pre-processing, mapping, analysis and interpretation of complex biomedical data with the aim to identify testable hypotheses, and build realistic models. The HCI-KDD approach, which is a synergistic combination of methodologies and approaches of two areas, Human–Computer Interaction (HCI) and Knowledge Discovery & Data Mining (KDD), offer ideal conditions towards solving these challenges: with the goal of supporting human intelligence with machine intelligence. There is an urgent need for integrative and interactive machine learning solutions, because no medical doctor or biomedical researcher can keep pace today with the increasingly large and complex data sets – often called "Big Data".

Keywords: Knowledge Discovery, Data Mining, Machine Learning, Biomedical Informatics, Integration, Interaction, HCI-KDD, Big Data.

A. Holzinger, I. Jurisica (Eds.): Knowledge Discovery and Data Mining, LNCS 8401, pp. 1–18, 2014.

1 Introduction and Motivation

Cinical practice, healthcare and biomedical research of today is drowning in data, yet starving for knowledge as Herbert A. Simon (1916–2001) pointed it out 40 years ago: *"A wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it* [1]."

The central problem is that biomedical data models are characterized by significant **complexity** [2-5] making manual analysis by the end users difficult, yet often impossible. Hence, current challenges in clinical practice and biomedical research include **information overload** – an often debated phenomenon in medicine for a long time [6-10].

There is the pressing need to combine vast amounts of diverse data, including structured, semi-structured and weakly structured data and unstructured information [11]. Interestingly, many powerful computational tools advancing in recent years have been developed by separate communities following different philosophies: Data mining and machine learning researchers tend to believe in the power of their statistical methods to identify relevant patterns – mostly automatic, without human intervention. There is, however, the danger of modelling artefacts when end user comprehension and control are diminished [12-15]. Additionally, mobile, ubiquitous computing and automatic medical sensors everywhere, together with low cost storage, will even accelerate this avalanche of data [16].

Another aspect is that, faced with unsustainable health care costs worldwide and enormous amounts of under-utilized data, medicine and health care needs more efficient practices; experts consider health information technology as key to increasing efficiency and quality of health care, whilst decreasing the costs [17].

Moreover, we need more research on methods, algorithms and tools to harness the full benefits towards the concept of **personalized medicine** [18]. Yet, we also need to substantially expand automated data capture to further **precision medicine** [19] and truly enable evidence-based medicine [20].

To capture data and task diversity, we continue to expand and improve individual knowledge discovery and data mining approaches and frameworks that let the end users gain insight into the nature of massive data sets [21-23].

The trend is to move individual systems to integrated, ensemble and interactive systems (see Figure 1).

Each type of data requires different, optimized approach; yet, we cannot interpret data fully without linking to other types. Ensemble systems and integrative KDD are part of the answer. Graph-based methods enable linking typed and annotated data further. Rich ontologies [24-26] and aspects from the Semantic Web [27-29] provide additional abilities to further characterize and annotate the discoveries.

2 Glossary and Key Terms

Biomedical Informatics: similar to medical informatics (see below) but including the optimal use of *biomedical data,* e.g. from the "–omics world" [30];

Data Mining: methods, algorithms and tools to extract patterns from data by combining methods from computational statistics [31] and machine learning: *"Data mining is about solving problems by analyzing data present in databases* [32]";

Deep Learning: is a machine learning method which models high-level abstractions in data by use of architectures composed of multiple non-linear transformations [33].

Ensemble Machine Learning: uses multiple learning algorithms to obtain better predictive performance as could be obtained from any standard learning algorithms [34]; A tutorial on ensemble-based classifiers can be found in [35].

Human–Computer Interaction: involves the study, design and development of the interaction between end users and computers (data); the classic definition goes back to Card, Moran & Newell [36], [37]. Interactive user-interfaces shall, for example, empower the user to carry out visual data mining;

Interactome: is the whole set of molecular interactions in a cell, i.e. genetic interactions, described as biological networks and displayed as graphs. The term goes back to the work of [38].

Information Overload: is an often debated, not clearly defined term from decision making research, when having to many alternatives to make a satisfying decision [39]; based on, e.g. the theory of cognitive load during problem solving [40-42].

Knowledge Discovery (KDD): Exploratory analysis and modeling of data and the organized process of identifying valid, novel, useful and understandable patterns from these data sets [21].

Machine Learning: the classic definition is "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" [43].

Medical Informatics: in the classical definition: "… scientific field that deals with the storage, retrieval, and optimal use of medical information, data, and knowledge for problem solving and decision making" [44];

Usability Engineering: includes methods that shall ensure that integrated and interactive solutions are useable and useful for the end users [45].

Visual Data Mining: An interactive combination of visualization and analysis with the goal to implement workflow that enables integration of user's expertise [46].

3 State-of-the-Art of Interactive and Integrative Solutions

Gotz et al. (2014) [47] present in a very recent work an interesting methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. They start with the evidence that the medical conditions of patients often evolve in *complex* and *unpredictable* ways and that variations between patients in both their progression and eventual outcome can be dramatic. Consequently, they state that understanding the patterns of events observed within a population that most correlate with differences in outcome is an important task. Their approach for **interactive pattern mining** supports ad hoc visual exploration of patterns mined from retrospective clinical patient data and combines three issues: visual query capabilities to interactively specify episode definitions; pattern mining techniques to help discover important intermediate events within an episode; and interactive visualization techniques that help uncover event patterns that most impact outcome and how those associations change over time.

Pastrello et al. (2014) [48] emphasize that first and foremost it is important to integrate the large volumes of heterogeneous and distributed data sets and that interactive data visualization is essential to obtain meaningful hypotheses from the diversity of various data (see Figure 1). They see **network analysis** (see e.g. [49]) as a key technique to integrate, visualize and extrapolate relevant information from diverse data sets and emphasize the huge challenge in integrating different types of data and then focus on systematically exploring network properties to gain insight into network functions. They also accentuate the role of the *interactome* in connecting data derived from different experiments, and they emphasize the importance of network analysis for the recognition of interaction context-specific features.

A previous work of Pastrello et al. (2013) [50] states that, whilst high-throughput technologies produce massive amounts of data, individual methods yield data, specific to the technique and the specific biological setup used. They also emphasize that at first the **integration of diverse data sets** is necessary for the qualitative analysis of information relevant to build hypotheses or to discover knowledge. Moreover, Pastrello et al. are of the opinion that it is useful to integrate these data sets by use of pathways and protein interaction networks; the resulting network needs to be able to focus on either a large-scale view or on more detailed small-scale views, depending on the research question and experimental goals. In their paper, the authors illustrate a workflow, which is useful to integrate, analyze, and visualize data from different sources, and they highlight important features of tools to support such analyses.

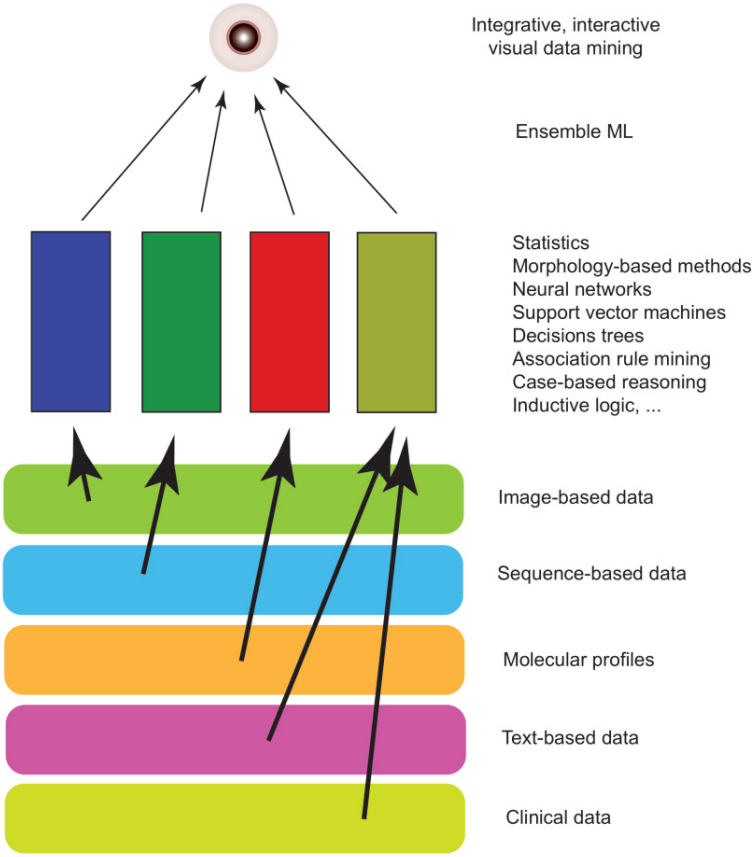

Fig. 1. Integrative analysis requires systematically combining various data sets and diverse algorithms. To support multiple user needs and enable integration of user's expertise, it is essential to support visual data mining.

An example from Neuroimaging provided by Bowman et al. (2012) [51], shows that electronic data capture methods will significantly advance the populating of large-scale neuroimaging databases: As these archives grow in size, a particular challenge is in the examination of and interaction with the information that these resources contain through the development of user-driven approaches for data exploration and data mining. In their paper they introduce the visualization for neuroimaging (INVIZIAN) framework for the graphical rendering of, and the dynamic interaction with the contents of large-scale neuroimaging data sets. Their system graphically displays brain surfaces as points in a coordinate space, thereby enabling the classification of clusters of neuroanatomically similar MRI-images and data mining.

Koelling et al. (2012) [52] present a web-based tool for visual data mining colocation patterns in multivariate bioimages, the so-called Web-based Hyperbolic Image Data Explorer (WHIDE). The authors emphasize that bioimaging techniques rapidly develop toward higher resolution and higher dimension; the increase in dimension is achieved by different techniques, which record for each pixel an *n*-dimensional intensity array, representing local abundances of molecules, residues or interaction patterns. The analysis of such Multivariate Bio-Images (MBIs) calls for new approaches to support end users in the analysis of both feature domains: space (i.e. sample morphology) and molecular colocation or interaction. The approach combines principles from computational learning, dimension reduction and visualization within, freely available via: http://ani.cebitec.uni-bielefeld.de/BioIMAX (login: whidetestuser; Password: whidetest).

An earlier work by Wegman (2003) [53], emphasizes that data mining strategies are usually applied to "opportunistically" collected data sets, which are frequently in the focus of the discovery of structures such as clusters, trends, periodicities, associations, correlations, etc., for which a visual data analysis is very appropriate and quite likely to yield insight. On the other hand, Wegman argues that data mining strategies are often applied to large data sets where standard visualization techniques may not be appropriate, due to the limits of screen resolution, limits of human perception and limits of available computational resources. Wegman thus envisioned Visual Data Mining (VDM) as a possible successful approach for attacking high-dimensional and large data sets.

4 Towards Finding Solutions: The HCI-KDD Approach

The idea of the HCI-KDD approach is in combining 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-preprocessing, computational statistics, machine learning and artificial intelligence [54].

In Figure 2 it can be seen how the concerted HCI-KDD approach may provide contributions to research and development for finding solutions to some challenges mentioned before. However, before looking at further details, one question may arise: What is the difference between Knowledge Discovery and Data Mining? The paradigm "Data Mining (DM)" has an established tradition, dating back to the early days of databases, and with varied naming conventions, e.g., "data grubbing", "data fishing" [55]; the term "Information Retrieval (IR)" was coined even earlier in 1950 [56, 57], whereas the term "Knowledge Discovery (KD)" is relatively young, having its roots in the classical work of Piatetsky-Shapiro (1991) [58], and gaining much popularity with the paper by Fayyad et al. (1996) [59]. Considering these definitions, we need to explain the difference between Knowledge Discovery and Data Mining itself: Some researchers argue that there is *no* difference, and to emphasize this it is often called "Knowledge Discovery and Data Mining (KDD)", whereas the original definition by Fayyad was "Knowledge Discovery from Data (KDD)", which makes also sense but separates it from Data Mining (DM). Although it makes sense to differentiate between these two terms, we prefer the first notion: "Knowledge

Discovery and Data Mining (KDD)" to emphasize that *both* are of equal importance and necessary in combination. This orchestrated interplay is graphically illustrated in Figure 2: Whilst KDD encompasses the whole *process* workflow ranging from the very physical data representation (left) to the human aspects of information processing (right), data mining goes *in depth* and includes the algorithms for particularly finding patterns in the data. Interaction is prominently represented by HCI in the left side.

Within this "big picture" seven research areas can be identified, numbered from area 1 to area 7:

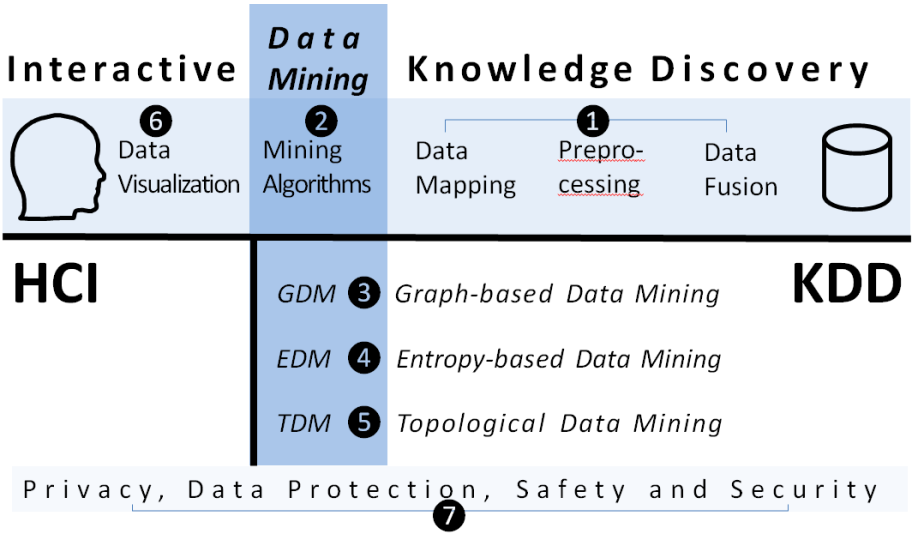

Fig. 2. The big picture of the HCI-KDD approach: KDD encompasses the whole **horizontal** process chain from data to information and knowledge; actually from physical aspects of raw data, to human aspects including attention, memory, vision, interaction etc. as core topics in HCI, whilst DM as a **vertical**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

4.1 Area 1: Data Integration, Data Pre-processing and Data Mapping

In this volume three papers (#4, #8 and #15) are addressing research area 1:

In paper #4 *"On the Generation of Point Cloud Data Sets: Step one in the Knowledge Discovery Process"* Holzinger et al. [60] provide some answers to the question "How do you get a graph out of your data?" or more specific "How to get **point cloud data sets** from natural images?". The authors present some solutions, open problems and a future outlook when mapping continuous data, such as natural images, into discrete point cloud data sets (PCD). Their work is based on the assumption that geometry, topology and graph theory have much potential for the analysis of arbitrarily high-dimensional data.

In paper #8 "*A Policy-based Cleansing and Integration Framework for Labour and Healthcare Data"* Boselli et al. [61] report on a **holistic data integration strategy** for large amounts of health data. The authors describe how a model based cleansing framework is extended to address such integration activities. Their combined approach facilitates the rapid prototyping, development, and evaluation of data preprocessing activities. They found, that a combined use of formal methods and visualization techniques strongly empower the data analyst, which can effectively evaluate how cleansing and integration activities can affect the data analysis. The authors show also an example focusing on labour and healthcare data integration.

In paper #15 *"Intelligent integrative knowledge bases: bridging genomics, integrative biology and translational medicine",* Nguyen et al. [62] present a perspective for data management, statistical analysis and knowledge discovery related to human disease, which they call an intelligent integrative knowledge base (I2KB). By building a bridge between patient associations, clinicians, experimentalists and modelers, I2KB will facilitate the emergence and propagation of **systems medicine** studies, which are a prerequisite for large-scaled clinical trial studies, efficient diagnosis, disease screening, drug target evaluation and development of new therapeutic strategies.

In paper #18 *"Biobanks – A Source of large Biological Data Sets: Open Problems and Future Challenges"*, Huppertz & Holzinger [63] are discussing Biobanks in light of a source of large biological data sets and present some open problems and future challenges, amongst them **data integration and data fusion** of the heterogeneous data sets from various data banks. In particular the fusion of two large areas, i.e. the business enterprise hospital information systems with the biobank data is essential, the grand challenge remains in the extreme heterogeneity of data, the large amounts of weakly structured data, in data complexity, and the massive amount of unstructured information and the associated lack of data quality.

4.2 Area 2: Data Mining Algorithms

Most of the papers in this volume are dealing with data mining algorithms, in particular:

In paper #3 *"Darwin or Lamarck? Future Challenges in Evolutionary Algorithms for Knowledge Discovery and Data Mining"* Katharina Holzinger et al. [64] are discussing the differences between evolutionary algorithms, beginning with some background on the **theory of evolution** by contrasting the original ideas of Charles Darwin and Jean-Baptiste de Lamarck; the authors provide a discussion on the analogy between biological and computational sciences, and briefly describe some fundamentals of various algorithms, including Genetic Algorithms, but also new and promising ones, including Invasive Weed Optimization, Memetic Search, Differential Evolution Search, Artificial Immune Systems, and Intelligent Water Drops.

In paper #5 *"Adapted Features and Instance Selection for Improving Co-Training",* Katz et al. [65] report on the importance of high quality, labeled data as it is essential for successfully applying machine learning to real-world problems. Because often the amount of labeled data is insufficient and labeling that data is time consuming, Katz et al. propose co-training algorithms, which use unlabeled data in order to improve classification. The authors propose simple and effective strategies for improving the basic co-training framework, i.e.: the manner in which the features set is partitioned and the method of selecting additional instances. Moreover, they present a study over 25 datasets, and prove that their proposed strategies are especially effective for **imbalanced datasets.**

In paper #6 *"Knowledge Discovery & Visualization of Clusters for Erythromycin Related Adverse Events in the FDA Drug Adverse Event Reporting System",* Yildirim et al. [66] present a study to discover hidden knowledge in the reports of the public release of the Food and Drug Administration (FDA)'s Adverse Event Reporting System (FAERS) for the antibiotic Erythromycin. This is highly relevant, due to the fact that bacterial infections can cause significant morbidity, mortality and high costs of treatment and are known as a significant health problem in the world. The authors used **cluster analysis** and the DBSCAN algorithm. Medical researchers and pharmaceutical companies may utilize these results and test these relationships along with their clinical studies.

In paper #10 *"Resources for Studying Statistical Analysis of Biomedical Data and R"*, Kobayashi [67] introduces some online resources to help medical practitioners with little or no background in **predictive statistics**, to learn basic statistical concepts and to implement data analysis methods on their personal computers by using R, a high-level open source computer language that requires relatively little training. This offers medical practitioners an opportunity to identify effectiveness of treatments for patients using summary statistics, so to offer patients more personalized medical treatments based on predictive analytics. Some open problems emphasized by Kobayashi include Privacy Preserving Data Mining (PPDM) algorithms and High Speed Medical Data Analysis.

In paper #11 *"A Kernel-based Framework for Medical Big-Data Analytics",* Windridge & Bober [68] point out that issues of incompleteness and heterogeneity are problematic and that data in the biomedical domain can be as diverse as handwritten notes, blood pressure readings, and MR scans, etc., and typically very little of this data will be co-present for each patient at any given time interval. Windridge & Bober therefore advocate a **kernel-based framework** as being most appropriate for handling these issues, using the neutral point substitution method to accommodate missing inter-modal data, and advocates for the pre-processing of image based MR data a **deep learning** solution for contextual areal segmentation, with edit-distance based kernel measurement, used to characterize relevant morphology. Moreover, the authors promote the use of **Boltzmann machines.**

In paper #16 *"Biomedical Text Mining: Open Problems and Future Challenges"* Holzinger et al. [69] provide a short, concise overview of some selected text mining methods, focusing on **statistical methods** (Latent Semantic Analysis, Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation, Hierarchical Latent Dirichlet Allocation, Hierarchical Latent Dirichlet Allocation, Principal Component Analysis), but also introduces relatively new and promising text mining methods including **graph-based** approaches and **topological text mining.** Although in our modern graphic-driven multimedia world, the importance of text is often debated, it should not be underestimated, as particularly in the medical domain "free text" is a very important type of data for medical communication; however, the increasing volumes of this unstructured information makes manual analysis nearly impossible, and calls for machine learning approaches for text mining.

4.2.1 Area 3: Graph Based Data Mining

In paper #14 *"Multi-touch Graph-Based Interaction for Knowledge Discovery on Mobile Devices: State-of-the-Art and Future Challenges"* Holzinger et al. [70] provide an overview on **graph-based knowledge representation:** Graphs are most powerful tools to map structures within a given data set and to recognize relationships between specific data objects. Many advantages of graph-based data structures can be found in the applicability of methods from network analysis, topology and data mining (e.g. small world phenomenon, cluster analysis). Moreover, Holzinger et al. present graph-based approaches for multi-touch interaction on mobile devices (tablets, smartphones), which is particularly important in the medical domain, as a conceptual graph analysis may provide novel insights on hidden patterns in data, hence support interactive knowledge discovery. Amongst the open problems the authors list the question "Which structural properties possess the multi-touch interaction graphs?", which calls for investigating graph classes beyond small world and random networks.

In paper #13 *"Sparse Inverse Covariance Estimation for Graph Representation of Feature Structure"*, Lee [71] states that higher dimensionality makes it challenging to understand complex systems. The author reports on structure learning with the Gaussian Markov random field, by identifying conditional independence structure of features in a form that is easy to visualize and understand. The learning is based on a convex optimization problem, called the sparse inverse covariance estimation, for which many efficient algorithms have been developed in the past. When dimensions are much larger than sample sizes, **structure learning** requires to consider statistical stability, in which connections to data mining arise in terms of discovering common or rare sub-graphs as patterns. Lee discusses the outcome of structure learning, which can be visualized as graphs provide a perceivable way to investigate complex feature spaces. He identifies two major open challenges for solving the sparse inverse covariance estimation problem in high-dimensions: development of efficient optimization algorithms and consideration of statistical stability of solutions.

4.2.2 Area 4: Entropy Based Data Mining

In paper #12 *"On Entropy-based Data Mining",* Holzinger et al. [72], start with some basics on information entropy as **measure for the uncertainty of data.** Then the authors provide a taxonomy of various entropy methods, whereby describing in more detail: Approximate Entropy, Sample Entropy, Fuzzy Entropy, and particularly **Topological Entropy** for finite sequences. Holzinger et al. state that entropy measures have successfully been tested for analysing short, sparse and noisy time series data, but that they have not yet been applied to weakly structured data in combination with techniques from computational topology, which is a hot and promising research route.

4.2.3 Area 5: Topological Data Mining

In paper #19 *"Topological Data Mining in a Nutshell"* [73] Holzinger presents a nutshell-like overview on some basics of **topology and data** and discusses some issues on why this is important for knowledge discovery and data mining: Humans are very good at pattern recognition in dimensions of lower or equal than 3, this suggests that computer science should develop methods for exploring this capacity, whereas computational geometry and topology have much potential for the analysis of arbitrarily high-dimensional data sets. Again, both together could be powerful beyond imagination.

4.3 Area 6: Data Visualization

In paper #2 *"Visual Data Mining: Effective Exploration of the Biological Universe",* Otasek et al. [74] present their experiences with Visual Data Mining (VDM), supported by interactive and scalable network visualization and analysis, which enables effective exploration within multiple biological and biomedical fields. The authors discuss large networks, such as the protein interactome and transcriptional regulatory networks, which contain hundreds of thousands of objects and millions of relationships. The authors report on the involved workflows and their experiences with biological researchers on how they can discover knowledge and new theories from their complex data sets.

In paper #7 *"On Computationally-enhanced Visual Analysis of Heterogeneous Data and its Application in Biomedical Informatics"*, Turkay et al. [75] present a concise overview on the state-of-the-art in interactive data visualization, relevant for knowledge discovery, and particularly focus on the issue of integrating computational tools into the workplace for the analysis of heterogeneous data. Turkay et al. emphasize that seamlessly integrated concepts are rare, although there are several solutions that involve a tight integration between computational methods and visualization. Amongst the open problems, the most pressing one is the application of sophisticated visualization techniques, seamlessly integrated into the (bio)-medical workplace, useable and useable to the medical professional.

In paper #9 *"Interactive Data Exploration using Pattern Mining"* van Leeuwen [76] reports on challenges in exploratory data mining to provide insight in data, i.e. to develop principled methods that allow both user-specific and task-specific information to be taken into account, by directly involving the user into the discovery process. The author states that pattern mining algorithms will need to be combined with techniques from visualization and human-computer interaction. As ultimate goal van Leeuwen states to make pattern mining practically more useful, by enabling the user to interactively explore the data and identify interesting structures.

4.4 Area 7: Privacy, Data Protection, Safety and Security

In the biomedical domain it is mandatory to consider aspects of privacy, data protection, safety and security, and a fair use of data sets, and one paper is particularly dealing with these topics:

In paper #17 Kieseberg et al. [77] discuss concerns of the disclosure of research data, which raises considerable privacy concerns, as researchers have the responsibility to protect their (volunteer) subjects and must adhere to respective policies. The authors provide an overview on the most important and well-researched approaches to deal with such concerns and discuss open research problems to stimulate further investigation: One solution for this problem lies in the protection of sensitive information in medical data sets by applying appropriate anonymization techniques, due to the fact that the underlying data set should always be made available to ensure the quality of the research done and to prevent fraud or errors.

5 Conclusion and Future Outlook

Some of the most important challenges in clinical practice and biomedical research include the need to develop and apply novel tools for the effective integration, analysis and interpretation of complex biomedical data with the aim to identify testable hypothesis, and build realistic models. A big issue is the limited time to make a decision, e.g. a medical doctor has in average five minutes to make a decision [78], [79].

Data and requirements also evolve over time – we need approaches that seamlessly and robustly handle *change.*

The algorithms must also handle incomplete, noisy, even contradictory/ambiguous information, and they have to support multiple viewpoints and contexts.

Solutions need to be interactive, seamlessly integrating diverse data sources, and able to scale to ultra-high dimensions, support multimodal and rapidly evolving representations.

Major future research areas in HCI-KDD in the biomedical field include graph-based analysis and pattern discovery, streaming data mining, integrative and interactive visual data mining. Thus, solutions will need to use heuristics, probabilistic and data-driven methods, with rigorous train-test-validate steps. Especially the last point highlights the need for **open data.**

It is paramount importance that the data is broadly available in usable formats – without relevant reliable and clean data there is no data mining; without accessible data we cannot assure correctness; without data, we cannot train and validate machine learning systems. It is alarming to see an exponential trend in number of retracted papers per year, and especially since the majority of them are fraud – 21.3% being attributed to error and 67.4% to (suspected) fraud [80]: A detailed review of over 2,000 biomedical research articles indexed by PubMed as retracted by May, 2012 revealed that only 21.3% of retractions were attributable to error [80]. In contrast, 67.4% of retractions were attributable to misconduct, including fraud or suspected fraud (43.4%), or duplicate publication (14.2%), and even plagiarism (9.8%) [80]. Incomplete, uninformative or misleading retraction announcements have led to a previous underestimation of the role of fraud in the ongoing retraction epidemic. Machine learning and data mining also plays a significant role in identifying outliers, errors, and thus could contribute to 'cleaning up' science from fraud and errors.

Concluding, there are a lot of open problems and future challenges in dealing with massive amounts of heterogeneous, distributed, diverse, highly dynamic data sets, complex, high-dimensional and weakly structured data and increasingly large amounts of unstructured and non-standardized information. The limits of our human capacities makes it impossible to deal manually with such data, hence, efficient machine learning approaches becomes indispensable.

Acknowledgements. We would like to thank the HCI-KDD network of excellence for valuable comments and our institutional colleagues for appreciated discussions.

References

1. Simon, H.A.: Designing Organizations for an Information-Rich World. In: Greenberger, M. (ed.) Computers, Communication, and the Public Interest, pp. 37–72. The Johns Hopkins Press, Baltimore (1971)
2. Dugas, M., Hoffmann, E., Janko, S., Hahnewald, S., Matis, T., Miller, J., Bary, C.V., Farnbacher, A., Vogler, V., Überla, K.: Complexity of biomedical data models in cardiology: the Intranet-based AF registry. Computer Methods and Programs in Biomedicine 68(1), 49–61 (2002)
3. Akil, H., Martone, M.E., Van Essen, D.C.: Challenges and opportunities in mining neuroscience data. Science 331(6018), 708–712 (2011)
4. Holzinger, A.: Biomedical Informatics: Computational Sciences meets Life Sciences. BoD, Norderstedt (2012)
5. Holzinger, A.: Biomedical Informatics: Discovering Knowledge in Big Data. Springer, New York (2014)
6. Berghel, H.: Cyberspace 2000: Dealing with Information Overload. Communications of the ACM 40(2), 19–24 (1997)
7. Noone, J., Warren, J., Brittain, M.: Information overload: opportunities and challenges for the GP's desktop. Medinfo 9(2), 1287–1291 (1998)

8. Holzinger, A., Geierhofer, R., Errath, M.: Semantic Information in Medical Information Systems - from Data and Information to Knowledge: Facing Information Overload. In: Procedings of I-MEDIA 2007 and I-SEMANTICS 2007, pp. 323–330 (2007)
9. Holzinger, A., Simonic, K.-M., Steyrer, J.: Information Overload - stößt die Medizin an ihre Grenzen? Wissensmanagement 13(1), 10–12 (2011)
10. Holzinger, A., Scherer, R., Ziefle, M.: Navigational User Interface Elements on the Left Side: Intuition of Designers or Experimental Evidence? In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) INTERACT 2011, Part II. LNCS, vol. 6947, pp. 162–177. Springer, Heidelberg (2011)
11. 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(suppl. 6), I1 (2014)
12. Shneiderman, B.: Inventing Discovery Tools: Combining Information Visualization with Data Mining. In: Jantke, K.P., Shinohara, A. (eds.) DS 2001. LNCS (LNAI), vol. 2226, pp. 17–28. Springer, Heidelberg (2001)
13. Shneiderman, B.: Inventing Discovery Tools: Combining Information Visualization with Data Mining. Information Visualization 1(1), 5–12 (2002)
14. Shneiderman, B.: Creativity support tools. Communications of the ACM 45(10), 116–120 (2002)
15. Shneiderman, B.: Creativity support tools: accelerating discovery and innovation. Communications of the ACM 50(12), 20–32 (2007)
16. Butler, D.: 2020 computing: Everything, everywhere. Nature 440(7083), 402–405 (2006)
17. Chaudhry, B., Wang, J., Wu, S.Y., Maglione, M., Mojica, W., Roth, E., Morton, S.C., Shekelle, P.G.: Systematic review: Impact of health information technology on quality, efficiency, and costs of medical care. Ann. Intern. Med. 144(10), 742–752 (2006)
18. Chawla, N.V., Davis, D.A.: Bringing Big Data to Personalized Healthcare: A Patient-Centered Framework. J. Gen. Intern. Med. 28, S660–S665 (2013)
19. Mirnezami, R., Nicholson, J., Darzi, A.: Preparing for Precision Medicine. N. Engl. J. Med. 366(6), 489–491 (2012)
20. Sackett, D.L., Rosenberg, W.M., Gray, J., Haynes, R.B., Richardson, W.S.: Evidence based medicine: what it is and what it isn't. BMJ: British Medical Journal 312(7023), 71 (1996)
21. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM 39(11), 27–34 (1996)
22. Jurisica, I., Mylopoulos, J., Glasgow, J., Shapiro, H., Casper, R.F.: Case-based reasoning in IVF: prediction and knowledge mining. Artificial Intelligence in Medicine 12(1), 1–24 (1998)
23. Yildirim, P., Ekmekci, I.O., Holzinger, A.: On Knowledge Discovery in Open Medical Data on the Example of the FDA Drug Adverse Event Reporting System for Alendronate (Fosamax). In: Holzinger, A., Pasi, G. (eds.) HCI-KDD 2013. LNCS, vol. 7947, pp. 195–206. Springer, Heidelberg (2013)
24. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies 43(5-6), 907–928 (1995)
25. Pinciroli, F., Pisanelli, D.M.: The unexpected high practical value of medical ontologies. Computers in Biology and Medicine 36(7-8), 669–673 (2006)
26. Eiter, T., Ianni, G., Polleres, A., Schindlauer, R., Tompits, H.: Reasoning with rules and ontologies. In: Barahona, P., Bry, F., Franconi, E., Henze, N., Sattler, U. (eds.) Reasoning Web 2006. LNCS, vol. 4126, pp. 93–127. Springer, Heidelberg (2006)

27. Tjoa, A.M., Andjomshoaa, A., Shayeganfar, F., Wagner, R.: Semantic Web challenges and new requirements. In: Database and Expert Systems Applications (DEXA), pp. 1160–1163. IEEE (2005)
28. d'Aquin, M., Noy, N.F.: Where to publish and find ontologies? A survey of ontology libraries. Web Semantics: Science, Services and Agents on the World Wide Web 11, 96–111 (2012)
29. Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., Doherty, D., Forsberg, K., Gao, Y., Kashyap, V., Kinoshita, J., Luciano, J., Marshall, M.S., Ogbuji, C., Rees, J., Stephens, S., Wong, G.T., Wu, E., Zaccagnini, D., Hongsermeier, T., Neumann, E., Herman, I., Cheung, K.H.: Methodology - Advancing translational research with the Semantic Web. BMC Bioinformatics 8 (2007)
30. Shortliffe, E.H., Barnett, G.O.: Biomedical data: Their acquisition, storage, and use. Biomedical informatics, pp. 39–66. Springer, London (2014)
31. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York (2009)
32. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco (2011)
33. Arel, I., Rose, D.C., Karnowski, T.P.: Deep Machine Learning - A New Frontier in Artificial Intelligence Research [Research Frontier]. IEEE Computational Intelligence Magazine 5(4), 13–18 (2010)
34. Dietterich, T.G.: Ensemble methods in machine learning. Multiple classifier systems, pp. 1–15. Springer (2000)
35. Rokach, L.: Ensemble-based classifiers. Artif. Intell. Rev. 33(1-2), 1–39 (2010)
36. Card, S.K., Moran, T.P., Newell, A.: The keystroke-level model for user performance time with interactive systems. Communications of the ACM 23(7), 396–410 (1980)
37. Card, S.K., Moran, T.P., Newell, A.: The psychology of Human-Computer Interaction. Erlbaum, Hillsdale (1983)
38. Sanchez, C., Lachaize, C., Janody, F., Bellon, B., Roder, L., Euzenat, J., Rechenmann, F., Jacq, B.: Grasping at molecular interactions and genetic networks in Drosophila melanogaster using FlyNets, an Internet database. Nucleic Acids Res. 27(1), 89–94 (1999)
39. McNeil, B.J., Keeler, E., Adelstein, S.J.: Primer on Certain Elements of Medical Decision Making. N. Engl. J. Med. 293(5), 211–215 (1975)
40. Sweller, J.: Cognitive load during problem solving: Effects on learning. Cognitive Science 12(2), 257–285 (1988)
41. Stickel, C., Ebner, M., Holzinger, A.: Useful Oblivion Versus Information Overload in e-Learning Examples in the Context of Wiki Systems. Journal of Computing and Information Technology (CIT) 16(4), 271–277 (2008)
42. Workman, M.: Cognitive Load Research and Semantic Apprehension of Graphical Linguistics. In: Holzinger, A. (ed.) USAB 2007. LNCS, vol. 4799, pp. 375–388. Springer, Heidelberg (2007)
43. Mitchell, T.M.: Machine learning, p. 267. McGraw-Hill, Boston (1997)
44. Shortliffe, E.H., Perrault, L.E., Wiederhold, G., Fagan, L.M.: Medical Informatics: Computer Applications in Health Care and Biomedicine. Springer, New York (1990)
45. Holzinger, A.: Usability engineering methods for software developers. Communications of the ACM 48(1), 71–74 (2005)
46. Keim, D.A.: Information visualization and visual data mining. IEEE Transactions on Visualization and Computer Graphics 8(1), 1–8 (2002)

47. Gotz, D., Wang, F., Perer, A.: A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. J. Biomed. Inform. (in print, 2014)
48. Pastrello, C., Pasini, E., Kotlyar, M., Otasek, D., Wong, S., Sangrar, W., Rahmati, S., Jurisica, I.: Integration, visualization and analysis of human interactome. Biochemical and Biophysical Research Communications 445(4), 757–773 (2014)
49. Dehmer, M.: Information-theoretic concepts for the analysis of complex networks. Applied Artificial Intelligence 22(7-8), 684–706 (2008)
50. Pastrello, C., Otasek, D., Fortney, K., Agapito, G., Cannataro, M., Shirdel, E., Jurisica, I.: Visual Data Mining of Biological Networks: One Size Does Not Fit All. PLoS Computational Biology 9(1), e1002833 (2013)
51. Bowman, I., Joshi, S.H., Van Horn, J.D.: Visual systems for interactive exploration and mining of large-scale neuroimaging data archives. Frontiers in Neuroinformatics 6(11) (2012)
52. Kolling, J., Langenkamper, D., Abouna, S., Khan, M., Nattkemper, T.W.: WHIDE–a web tool for visual data mining colocation patterns in multivariate bioimages. Bioinformatics 28(8), 1143–1150 (2012)
53. Wegman, E.J.: Visual data mining. Stat. Med. 22(9), 1383–1397 (2003)
54. 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)
55. Lovell, M.C.: Data Mining. Review of Economics and Statistics 65(1), 1–12 (1983)
56. Mooers, C.N.: Information retrieval viewed as temporal signalling. In: Proc. Internatl. Congr. of Mathematicians, August 30-September 6, p. 572 (1950)
57. Mooers, C.N.: The next twenty years in information retrieval; some goals and predictions. American Documentation 11(3), 229–236 (1960)
58. Piatetsky-Shapiro, G.: Knowledge Discovery in Real Databases - A report on the IJCAI-89 Workshop. AI Magazine 11(5), 68–70 (1991)
59. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. Ai Magazine 17(3), 37–54 (1996)
60. Holzinger, A., Malle, B., Bloice, M., Wiltgen, M., Ferri, M., Stanganelli, I., Hofmann-Wellenhof, R.: On the Generation of Point Cloud Data Sets: the first step 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)
61. Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M.: A Policy-based Cleansing and Integration Framework for Labour and Healthcare Data. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 141–168. Springer, Heidelberg (2014)
62. Nguyen, H., Thompson, J.D., Schutz, P., Poch, O.: Intelligent integrative knowledge bases: bridging genomics, integrative biology and translational medicine. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 255–270. Springer, Heidelberg (2014)
63. 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, vol. 8401, pp. 317–330. Springer, Heidelberg (2014)

64. 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)
65. Katz, G., Shabtai, A., Rokach, L.: Adapted Features and Instance Selection for Improving Co-Training. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 81–100. Springer, Heidelberg (2014)
66. Yildirim, P., Bloice, M., Holzinger, A.: Knowledge Discovery & Visualization of Clusters for Erythromycin Related Adverse Events in the FDA Drug Adverse Event Reporting System. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 101–116. Springer, Heidelberg (2014)
67. Kobayashi, M.: Resources for Studying Statistical Analysis of Biomedical Data and R. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 183–195. Springer, Heidelberg (2014)
68. 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)
69. Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., Verspoor, K.: Biomedical Text Mining: Open Problems and Future Challenges. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 271–300. Springer, Heidelberg (2014)
70. 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)
71. Lee, S.: Sparse Inverse Covariance Estimation for Graph Representation of Feature Structure. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 227–240. Springer, Heidelberg (2014)
72. Holzinger, A., Hortenhuber, M., Mayer, C., Bachler, M., Wassertheurer, S., Pinho, A., Koslicki, D.: On Entropy-based Data Mining. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 209–226. Springer, Heidelberg (2014)
73. Holzinger, A.: Topological Data Mining in a Nutshell. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 331–356. Springer, Heidelberg (2014)
74. Otasek, D., Pastrello, C., Holzinger, A., Jurisica, I.: Visual Data Mining: Effective Exploration ofthe Biological Universe. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 19–33. Springer, Heidelberg (2014)
75. Turkay, C., Jeanquartier, F., Holzinger, A., Hauser, H.: On Computationally-enhanced Visual Analysis of Heterogeneous Data and its Application in Biomedical Informatics. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 117–140. Springer, Heidelberg (2014)
76. van Leeuwen, M.: Interactive Data Exploration using Pattern Mining. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 169–182. Springer, Heidelberg (2014)
77. Kieseberg, P., Hobel, H., Schrittwieser, S., Weippl, E., Holzinger, A.: Protecting Anonymity in the Data-Driven Medical Sciences. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 301–316. Springer, Heidelberg (2014)

78. Gigerenzer, G.: Gut Feelings: Short Cuts to Better Decision Making. Penguin, London (2008)
79. Gigerenzer, G., Gaissmaier, W.: Heuristic Decision Making. In: Fiske, S.T., Schacter, D.L., Taylor, S.E. (eds.) Annual Review of Psychology, vol. 62, pp. 451–482. Annual Reviews, Palo Alto (2011)
80. Fang, F.C., Steen, R.G., Casadevall, A.: Misconduct accounts for the majority of retracted scientific publications. Proc. Natl. Acad. Sci. U.S.A 109(42), 17028–17033 (2012)