Big Complex Biomedical Data: Towards a Taxonomy of Data

Andreas Holzinger[1(✉)], Christof Stocker[1], and Matthias Dehmer[2]

[1] Research Unit Human-Computer Interaction, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, 8036 Graz, Austria
{a.holzinger,c.stocker}@hci4all.at
[2] Institute for Bioinformatics and Translational Research, UMIT Tyrol, Hall in Tirol, Austria
matthias.dehmer@umit.at

Abstract. Professionals in the Life Sciences are faced with increasing masses of complex data sets. Very few data is structured, where traditional information retrieval methods work perfectly. A large portion of data is weakly structured; however, the majority falls into the category of unstructured data. To discover previously unknown knowledge from this data, we need advanced and novel methods to deal with the data from two aspects: time (e.g. information entropy) and space (e.g. computational topology). In this paper we show some examples of biomedical data and discuss a taxonomy of data with the specifics on medical data sets.

Keywords: Complex data · Biomedical data · Weakly-structured data · Information · Knowledge · Human-Computer Interaction · Data visualization · Biomedical informatics · Life sciences

1 Introduction

Data exploration has recently been hailed as the *fourth paradigm* in the investigation of nature, after empiricism, theory and computation [1]. Whether in astronomy or the life sciences, the flood of data requires sophisticated methods of handling. For example, researchers in bioinformatics collect, process and analyze masses of data, or in computational biology, they simulate biological systems, metabolic pathways, the behavior of a cell or how a protein is built [2]. In clinical medicine, the end users are confronted with increased volumes of highly complex, noisy, high-dimensional, multivariate and often weakly-structured data [3]. The field of biomedical informatics concerns the information processing by both humans and computers, dealing with biomedical complexity [4] to support decision making which is still a central topic in biomedical informatics [5].

Whereas Human-Computer Interaction (HCI) concentrates on human intelligence, and Knowledge Discovery in Data Mining (KDD) concentrates on machine intelligence, the grand challenge is to combine these diverse fields to support

M.S. Obaidat and J. Filipe (Eds.): ICETE 2012, CCIS 455, pp. 3–18, 2014.
DOI: 10.1007/978-3-662-44791-8_1

the expert end users in learning to interactively analyze information properties thus enabling them to visualize the relevant parts of their data. In other words, to enable effective human control over powerful machine intelligence and to integrate statistical methods with information visualization, to support human insight and decision making [6]. The broad application of business enterprise hospital information systems amasses large amounts of medical documents, which must be reviewed, observed, and analyzed by human experts [7]. All essential documents of the patient records contain a certain portion of data which has been entered in non-standardized format (aka *free text*). Although text can easily be *created* by the end users, the support of automatic analysis is extremely difficult [8–10].

2 Look at Your Data

Each observation can be seen as a data point in an n-dimensional Euclidian vector space $\mathbb{R}^n$. An n-dimensional vector is given by

$$\boldsymbol{x}_i = [x_{i_1}, \ldots, x_{i_n}],\ i_1, \ldots, i_n \in \mathcal{I}, \tag{1}$$

where $\mathcal{I}$ is an index set.

In an arbitrarily high dimensional space, methods from algebraic topology have proved to be compelling, because topological data abstractions let us investigate structures in a semantic context [11]; this can be seen as one step towards sensemaking [12]. The *global character* of the data requires that the domain expert is able to extract information about the phenomena represented by the data (Fig. 1). This expert asks a question, forms a hypothesis and transforms data into knowledge; which can be seen as a transfer from the *computational space* into the *cognitive space* [13] of 2D or 3D representations developing in time:

$$\mathbb{R}^n + t \rightarrow \mathbb{R}^2 + t \text{ or } \mathbb{R}^3 + t \tag{2}$$

The time t is an important, yet often neglected dimension in medicine [14]. The expert in Fig. 1 looks for interesting data. Interest is a human construct, a perspective on relationships between data, and is influenced by emotion, personal likings and previous experience. Interest is similar to beauty, which is in the eye of the beholder [15]. It is difficult to make knowledge discovery automatic, we need human intelligence for sensemaking. For example, fitness functionality cannot be formulated generally; hence automatic algorithms may not find a solution alone.

3 Seeing the World in Data

Current technological developments offer the opportunity to collect, store and process all kinds of data in an unprecedented way, in great detail and very large scale [16]. Although, we are aware that data is not information and information is not knowledge, we are able to perceive the fascinating perspectives of our world

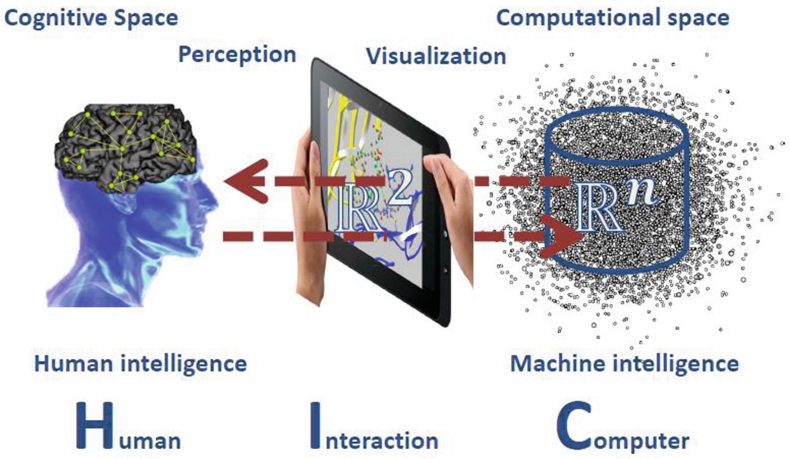

Fig. 1. Human-Computer Interaction bridging the cognitive space with the computational space.

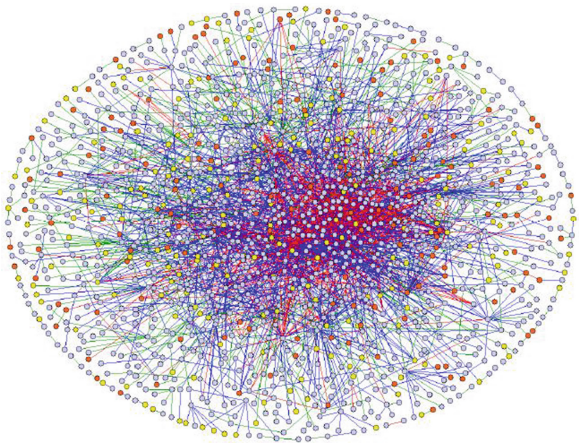

Fig. 2. First visualization of a human PPI structure; Experts gain knowledge of it, e.g. to understand complex processes, thereby understand illnesses [20].

in data. Let us look into the microscopic dimension (Fig. 2): Protein-protein interaction (PPI) [17,18] plays a fundamental role in all biological processes. A systematic analysis of PPI networks enables us to understand cellular organization, processes and function. This is big, complex, noisy data, consequently it is a great challenge to effectively analyse these massive data sets for biologically meaningful protein complex detection [19]. Moreover, this calls for novel techniques to infer biological networks as those are erroneous (measurement errors) and, hence, deterministic techniques can not be applied, see [18].

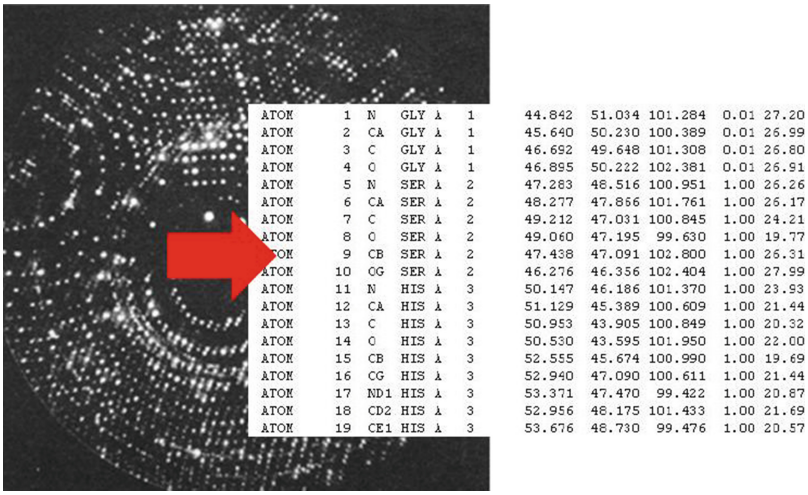

Fig. 3. Structures of protein complexes, determined by X-ray crystallography, and stored in the PDB [23].

The mathematical investigation of PPI networks starts with the inferred relational structure represented by a finite graph $G = (V, E)$, with a set of nodes V and edges E, where $E \subseteq V \times V$.

Proteins interact with each other to perform cellular functions or processes. These interacting patterns form the PPI network [21]

$$V \times V = \{(v_i, v_j) \mid v_i \in V, v_j \in V, i \neq j\} \tag{3}$$

Protein structures are studied for example with crystallographic methods (Fig. 3). Once the atomic coordinates of the protein structure have been determined, a table of these coordinates is deposited into a Protein Data Base (PDB), an international repository for 3D structure files. Scientific achievements coming from molecular biology greatly depend on computational applications and data management to explore lab results [22].

In Fig. 3, we see the structure and the data, representing the mean positions of the entities within the substance and their chemical relationships.

The structural information, stored in the PDB contains: a running number, atom type, residue name, the chain identification, the number of the residue in the chain, the triplet of coordinates. The PDB data files are downloaded from the database as input files for protein analysis and visualization.

Our quest is that an expert can gain knowledge from this data; for example by providing an interactive visualization of this data (Fig. 4): The Tumor Necrosis Factor (TNF - upper part) is interacting with the extra cellular domain of its receptor (lower part). The residues at the macromolecular interface are visualized in a "balland-stick" representation. The covalent bonds are represented as sticks between atoms, which are represented as balls. The rest of the two chains is

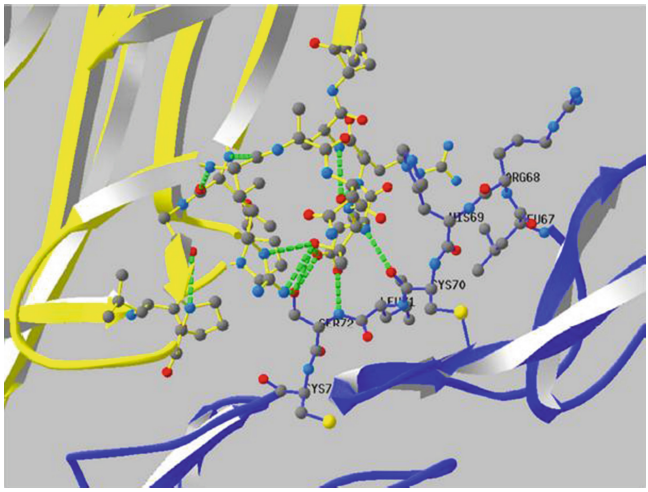

Fig. 4. Gaining knowledge from the data by interactive visualization [24].

represented as ribbons. Residue names and numbers of the TNF receptor are labelled, hydrogen bonds are represented by dotted lines (circled in Fig. 4).

Such complex network theory can be traced back to the first work on graph theory, developed by Leonhard Euler in 1736. However, stimulated by works as from Barabási, Albert and Jeong [25], research on complex networks has only recently been applied to biomedical informatics. As an extension of classical graph theory, complex network research focuses on the characterization, analysis, modeling and simulation of complex systems involving many elements and connections, examples including the internet, gene regulatory networks, PPI-networks, social relationships, the Web, and many more. Attention is given not only to the identification of special patterns of connectivity, such as the shortest average path between pairs of nodes [26], but also to the evolution of connectivity and the growth of networks, an example from biology being the evolution of PPI networks in different species (as shown in Fig. 2).

In order to understand complex biological systems, the three following key concepts have to be considered:

(i) emergence: the discovery of links between elements of a system as the study of individual elements (genes, proteins, metabolites) to explain the whole systems behavior;
(ii) robustness: biological systems maintain their main functions even under perturbations imposed by the environment; and
(iii) modularity: vertices sharing similar functions are highly connected.

Due to the ready availability of various network visualization tools [27], network theories can be applied to biomedical informatics.

4 Taxonomy of Data

Let us list some definitions first:

Definition 1. *Let a* ***relational system*** *be a pair* $(A, \{R_1, R_2, ..., R_n\})$*, where* A *is a set of elements, and* $R_1, R_2, ..., R_n$ *are relations defined on* A*.*

Definition 2. *Let an* ***attribute*** *be a homomorphism* $\mathcal{H}$ *from a relational system* $(A, \{R_1, R_2, ..., R_n\})$ *into a relational system* $(B, \{S_1, S_2, ..., S_n\})$*.*

The set A is a set of (visual) elements and the set B is either a set of (visual) elements or a set of attribute values such as the set $\mathbb{R}$, $\mathbb{Z}$ or a set of strings. The homomorphism $\mathcal{H}$ guarantees that every relation an attribute induces on elements has identical structural properties as its characterizing relations.

Dastani [28] described a special type of visual attributes which concerns various uses of topological properties of the space, i.e. perceptual structures that are constituted by perceivable topological relations, for example used in network visualizations (inside, outside, overlap, ...). This goes back to Egenhofer [29], who distinguished between spatial/nonspatial perceptual structures that are constituted by characterizing the relations of spatial and non-spatial attributes, and topological structures that are based on two or more topological attributes (Fig. 5). He used the nine-intersection model [30], which provides a framework and a relation algebra, for the description of topological relations between objects of area type, line, and point. This is based on the principles of algebraic topology, a branch of mathematics which deals with the manipulation of symbols that represent geometric configurations and their relationships to each other [31]. The data model is based on primitive objects, called cells, defined for different spatial dimensions: A 0-cell is a node (0-dimensional object); a 1-cell is the link between two 0-cells; a 2-cell is an area described by a closed sequence of three non-intersecting 1-cells and a face f is any cell that is contained in A. The relevant topological primitives include interior A°, boundary ∂A and exterior

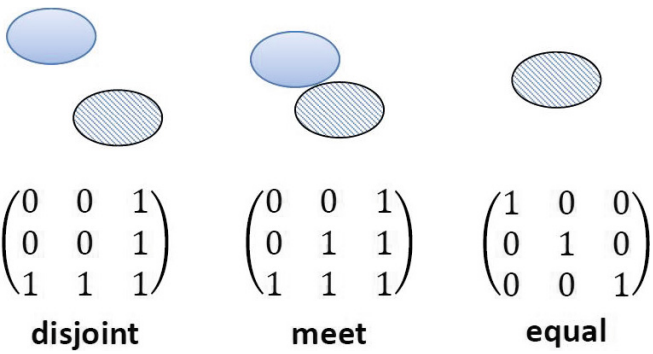

Fig. 5. Selected topological relations.

A^- of a cell; e.g., the boundary denoted by ∂A is the union of all r-faces $r - f$ where $0 \leq r \leq n$, i.e.

$$\partial A = \bigcup_{r=0}^{n-1} r - f \in A \tag{4}$$

The topological relation between two such geometric objects, A and B, is characterized by the binary values (empty, non-empty) of the 9-intersection, represented as a 3×3 matrix:

$$R(A,B) = \begin{pmatrix} A^\circ \cap B^\circ & A^\circ \cap \partial B & A^\circ \cap B^- \\ \partial A \cap B^\circ & \partial A \cap \partial B & \partial A \cap B^- \\ A^- \cap B^\circ & A^- \cap \partial B & A^- \cap B^- \end{pmatrix} \tag{5}$$

An important invariant is the number of components. Following the definition of Egenhofer and Franzosa [32] a component is based on the topological concepts separation and connectedness, i.e., for a set Y, a component is the largest connected (non-empty) subset of Y. Whenever any of the 9-set intersections is separated into disconnected subsets, these subsets are the components of this set intersection. Hence, any non-empty intersection may have several distinct components, each of which may be characterized by its own topological properties. This leads us to the definition of:

Weakly-structured Data. This must not be confused with weakly-structured information (e.g. [33], instead we follow the notions of topological relations (Fig. 5): Let $Y(t)$ be an ordered sequence of observed data, e.g., 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 [34,35].

Well-structured data has been seen to be the minority of data and an idealistic case when each data element has an associated defined structure, e.g., relational tables.

Ill-structured is a term often used for the opposite of well-structured, although this term originally was used in a different context of problem solving [36].

Semi-structured is a form of structured data that does not conform with the strict formal structure of tables and data models associated with relational databases, but contains tags or markers to separate both structure and content, i.e. these data are schema-less or self-describing; a typical example is a markup-language such as XML.

Non-structured 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, paragraphs. To be precise, unstructured data would mean completely randomized data which is usually called noise. Duda, Hart and Stork [37] define it as any property of data which is not due to the underlying model but instead to randomness (either in the real world, from the sensors or the measurement procedure). In Informatics, particularly, it can be considered as unwanted non-relevant data without meaning, or, even worse: with a not detected wrong meaning typical artifacts.

In addition to the above described structurization, data can also be standardized (e.g. numerical entries in laboratory reports) and nonstandardized (e.g. nonstandardized text often maybe inappropriately called "free text" in an electronic patient record, see e.g. [38].

Standardized data is a basis for accurate communication. In the medical domain, many different people work at different times in various locations. Data standards can ensure that information is presented in a form that facilitates interoperability of systems and a comparability of data for a common end user interpretation. It supports the reusability of the data, improves the efficiency of healthcare services and avoids errors by reducing duplicated efforts in data entry. Data standardization refers to (a) the data content; (b) terminologies used to represent the data; (c) how data is exchanged; and (d) how knowledge is applied; The last entry "*knowledge*" means e.g. clinical guidelines, protocols, decision support rules, checklists, standard operating procedures, etc. Technical elements for data sharing require standardization of identification, record structure, terminology, messaging, privacy etc. The most used standardized data set to date is the international Classification of Diseases (ICD), which was first adopted in 1900 for collecting statistics [39].

Non-standardized data as the majority of all data impedes data quality, data exchange and interoperability [40].

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 false knowledge discovery and overfitting. The possibility that important information may be contained in the complex interactions, along with the huge number of potential covariates that may be missed by simple methods, can be addressed by new and improved models and algorithms for classification and prediction [41].

5 Specifics of Medical Data

Biomedical data covers various structural dimensions, ranging from microscopic structures (e.g. DNA) to whole human populations (disease spreading). Clinical-medical data are defined and collected with a remarkable degree of uncertainty, variability and inaccuracy. Komaroff [42] stated that "*medical data is disturbingly soft*". Three decades later, the data still falls far short of the exactness that engineers prefer.

What did Komaroff mean with *soft*? The way patients define their sickness, questions and answers between clinicians and patients, physical examinations, diagnostic laboratory tests etc. Even the definitions of the diseases themselves are often ambiguous; some diseases cannot be defined by any available objective standard; other diseases do have an objective standard, but are variably interpreted.

Another complication inherent in the data is that most medical information is incomplete, with wide variation in the degree and type of missing information. In both the development and the application of statistical techniques, analysis of data with incomplete or missing information can be much more difficult than analysis of corresponding data with all the information available - interestingly this was known before the term medical informatics was defined [43].

Let us give a last example for the size aspect of medical data: In 1986, the INTERNIST-1 knowledge base (for diagnosis in internal medicine) contained 572 disorders, approx. 4,000 possible patient findings and links detailing the causal, temporal and probable interrelationships between the disorders [44]. Ten years ago, in 2002, a typical primary care doctor was kept informed of approximately 10,000 diseases and syndromes, 3,000 medications, and 1,100 laboratory tests [45]. In 2008, there were 18 million articles catalogued in the biomedical literature.

Working with big data requires certain issues to be addressed, such as data security, intellectual property and, particularly in the case of medical data, privacy issues [46].

6 Visualization of Data

How can visual representations of abstract data be used to amplify the acquisition of knowledge? [47].

Unfortunately, the creation of visualizations for complex data still remains more of a personal effort than a commercial enterprise. So many sophisticated visualization concepts have been developed, e.g. Parallel Coordinates [48], RadViz [49], or Glyphs [50], to mention only a few, but in business enterprise hospital information systems they are still not in use.

An interesting example is from the publication by Hey et al. [2] from the introduction to this paper, wherein from 30 essays on the emerging area of data-intensive science, all including visualizations of scientific results, only one is on visualization needs [51].

As a practical example, interactive computer simulations to teach complex concepts have become very popular [52]. The nature of such simulations ranges from compelling visualizations [53,54] to educational computer games [55,56]. A recent example is Foldit [57], where gamers can play cellular architect and build proteins. Scientists can crowdsource the data and design brand-new molecules in the lab. Such exploratory learning with interactive simulations is highly demanding from the perspective of *limited cognitive processing capabilities* and the research on interactive simulations [58,59] has revealed that learners *need further support and guidance*.

Learning in the area of physiology is difficult for medical students, because mostly they are lacking the mathematics necessary to understand the dynamics of complex mathematical rules related to physiological models.

In our application HAEMOSIM, we make complicated physiological data [60] interactively visible to medical learners (Fig. 6), so that they gain insight into the

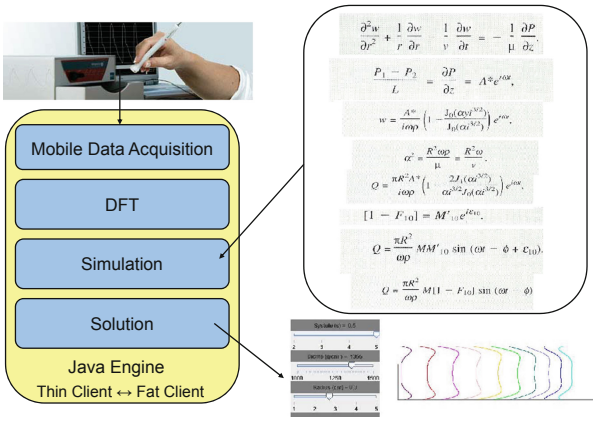

Fig. 6. Real data are used for the simulation of certain clinical relevant solutions and can be interactively displayed by a learner [68].

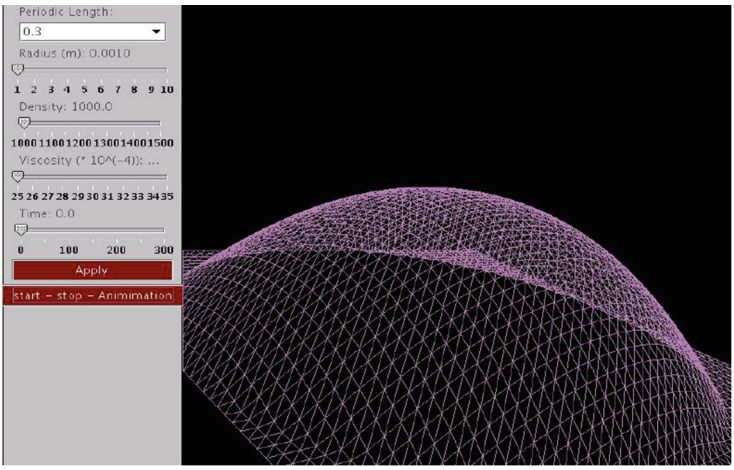

Fig. 7. The visualized data allows insights into medical contexts and sensemaking [68].

behavior of blood circulation dynamics, and to simulate certain defects (Fig. 7) and the dangers of diseases. The application simulates mathematical models [61–64] and presents these models in form of dynamic 2D and 3D visualizations. Special focus during the development was directed on usercentered design [65–67], for example, to understand the context and to adapt the various applets to the previous knowledge of the end users.

7 Conclusions and Future Outlook

Life sciences and human health are fundamentally biological, and biology is often described as *the* information science [69].

Consequently, research in computational biology may yield many beneficial results for medicine and health. A very intriguing question is to what extent randomness and stochasticity play a role. By adopting the computational thinking approach [70] to studying biological processes, we can improve our understanding and at the same time improve the design of algorithms [71].

The ability to define details of the interactions between small molecules and proteins promises unprecedented advances in the exploration of rational therapeutic strategies, for example, to combat infectious diseases and cancer. The opportunity to probe large macromolecular systems offers exciting opportunities for exploring the nature of PPIs and the mechanisms of trafficking of molecules to different regions of a cell, a process involving transport through membranes and diffusion over significant distances in the cytoplasm [72].

Following the quest "Science is to test ideas, engineering is to put these ideas into practice" [73], not only the scientific aspects will be challenging, but also the engineering ones, to support human intelligence with computational intelligence in the clinical domain. One challenge is in contextual computing; i.e. a medical professional may ask the business enterprise hospital information system: "Show me the similarities between patients with symptoms X and patients with symptoms Y". This brings us immediately back to the deep questions in computing [74], including: What is information? What is computable? What is intelligence? And most of all: (How) can we build complex systems in a simply?

Decision making is the key topic in medical informatics. For this we need to follow the three column approach: data - information - knowledge, with emphasis on the latter. Successful knowledge discovery and information retrieval systems will be those that bring the designer's model into harmony with the end user's mental model. We can conclude that combining HCI together with KDD will provide benefits to the medical domain. For this purpose, we must bridge Science and Engineering in order to answer fundamental questions on information quality [75] and to implement the findings on building information systems simply at the engineering level. A few important examples of future research aspects include:

1. Research on the physics of (time-oriented) information to contribute to fundamental research;
2. Considering temporal and spatial information; in networks, spatially distributed components raise fundamental issues on information exchange since available resources must be shared, allocated and reused. Information is exchanged in both space and time for decision making, therefore timeliness along with reliability and complexity constitute the main issues and are most often ignored;
3. We still lack measures and meters to define and appraise the amount of information embodied in structure and organization for example the entropy of a structure;
4. Considering information transfer: how we can assess, for example, the transfer of biological information;
5. Information and knowledge: In many scientific contexts we are dealing only with data without knowing precisely what these data are representing;
6. and most of all, we must gain value out of data making data valuable.

Concluding, we can say that the future in the life sciences will be definitely data-centric. This will apply equally to the medical clinical domain and health care. Mobile, ubiquitous computing, sensors everywhere, computational power and storage at very low cost will definitely produce an increasing avalanche of data and there definitely will be the danger of drowning in data, but starving for knowledge. Herbert Simon pointed out 40 years ago, when medical informatics was in its infancy: "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" [76].

Consequently, Human-Computer Interaction and Knowledge Discovery along with Biomedical Informatics are of increasing importance to effectively gain knowledge, to make sense out of the big data. This is our central quest the holy grail for the future. Let us put together all efforts to jointly make advances in this interesting, challenging and important area to benefit medicine, to benefit humans, to benefit us all.

However, even the best team is ineffective if there is no funding. A substantial budget is required to cover staff costs, premises and basic equipment, travel, computers and software, a scientific software portfolio, hosting, special equipment, literature, workshop organization, visiting researcher invitations, etc. In an environment of decreasing public budgets, external funding becomes increasingly important in order to sustain international competitiveness, quality and to maintain excellence [77].

References

1. Bell, G., Hey, T., Szalay, A.: Beyond the data deluge. Science **323**(5919), 1297–1298 (2009)
2. Hey, T., Tansley, S., Tolle, K.: The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond (2009)
3. Holzinger, A.: Weakly structured data in health-informatics: the challenge for human-computer interaction (2011)
4. Patel, V.L., Kahol, K., Buchman, T.: Biomedical complexity and error. J. Biomed. Inform. **44**(3), 387–389 (2011)
5. Holzinger, A.: Biomedical Informatics: Discovering Knowledge in Big Data. Springer, New York (2014)
6. 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)
7. Holzinger, A., Geierhofer, R., Modritscher, F., Tatzl, R.: Semantic information in medical information systems: utilization of text mining techniques to analyze medical diagnoses. J. Univ. Comput. Sci. **14**(22), 3781–3795 (2008)
8. Gregory, J., Mattison, J.E., Linde, C.: Naming notes - transitions from free-text to structured entry. Meth. Inf. Med. **34**(1–2), 57–67 (1995)

9. Holzinger, A., Kainz, A., Gell, G., Brunold, M., Maurer, H.: Interactive computer assisted formulation of retrieval requests for a medical information system using an intelligent tutoring system. World Conference on Educational Multimedia, Hypermedia and Telecommunications ED-MEDIA 2000, pp. 431–436. AACE, Charlottesville (2000)
10. Lovis, C., Baud, R.H., Planche, P.: Power of expression in the electronic patient record: structured data or narrative text? Int. J. Med. Inf. **58**, 101–110 (2000)
11. Pascucci, V., Tricoche, X., Hagen, H., Tierny, J.: Topological Methods in Data Analysis and Visualization: Theory, Algorithms, and Applications. Springer, Heidelberg (2011)
12. Blandford, A., Attfield, S.: Interacting with information. Synth. Lect. Hum. Centered Inf. **3**(1), 1–99 (2010)
13. Kaski, S., Peltonen, J.: Dimensionality reduction for data visualization (applications corner). IEEE Signal Process. Mag. **28**(2), 100–104 (2011)
14. 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)
15. Beale, R.: Supporting serendipity: using ambient intelligence to augment user exploration for data mining and web browsing. Int. J. Hum. Comput. Stud. **65**(5), 421–433 (2007)
16. Yau, N.: Seeing the World in Data, pp. 246–248. Princeton Architectural Press, New York (2011)
17. Pržulj, N., Higham, D.J.: Modelling protein-protein interaction networks via a stickiness index. J. Roy. Soc. Interface **3**(10), 711–716 (2006)
18. Emmert-Streib, F., Dehmer, M. (eds.): Analysis of Microarray Data: A Network-Based Approach. Wiley VCH Publishing, Chichester (2010)
19. Shi, L., Lei, X., Zhang, A.: Protein complex detection with semi-supervised learning in protein interaction networks. Proteome Sci. **9**(Suppl. 1), S5 (2011)
20. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., Wanker, E.E.: A human protein-protein interaction network: a resource for annotating the proteome. Cell **122**(6), 957–968 (2005)
21. Zhang, A.: Protein Interaction Networks: Computational Analysis. Cambridge University Press, Cambridge (2009)
22. Arrais, J.P., Lopes, P., Oliveira, J.L.: Challenges storing and representing biomedical data. In: Holzinger, A., Simonic, K.-M. (eds.) USAB 2011. LNCS, vol. 7058, pp. 53–62. Springer, Heidelberg (2011)
23. Wiltgen, M., Holzinger, A.: Visualization in Bioinformatics: Protein Structures with Physicochemical and Biological Annotations, pp. 69–74. Czech Technical University (CTU), Prague (2005)
24. 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)
25. Barabási, A.L., Albert, R., Jeong, H.: Mean-field theory for scale-free random networks. Physica A: Stat. Mech. Appl. **272**(1–2), 173–187 (1999)
26. Newman, M.: The structure and function of complex networks. SIAM Rev. **45**, 167–256 (2003)

27. Costa, L., Rodrigues, F., Cristino, A.: Complex networks: the key to systems biology. Genet. Mol. Biol. **31**(3), 591–601 (2008)
28. Dastani, M.: The role of visual perception in data visualization. J. Vis. Lang. Comput. **13**, 601–622 (2002)
29. Egenhofer, M.: Reasoning about binary topological relations. In: Günther, O., Schek, H.-J. (eds.) SSD 1991. LNCS, vol. 525, pp. 141–160. Springer, Heidelberg (1991)
30. Egenhofer, M., Herring, J.: Categorizing binary topological relations between regions, lines, and points in geographic databases. Technical Report, Department of Surveying Engineering, University of Maine (1990)
31. Aleksandrov, P.: Elementary Concepts of Topology. Dover Publications, New York (1961)
32. Egenhofer, M., Franzosa, R.: On the equivalence of topological relations. Int. J. Geogr. Inf. Syst. **9**(2), 133–152 (1995)
33. Stuckenschmidt, H., van Harmelen, F.: Information Sharing on the Semantic Web. Advanced Information and Knowledge Processing. Springer, Heidelberg (2005)
34. Kapovich, I., Myasnikov, A., Schupp, P., Shpilrain, V.: Generic-case complexity, decision problems in group theory, and random walks. J. Algebra **264**(2), 665–694 (2003)
35. de Silva, V., Carlsson, G.: Topological estimation using witness complexes. In: Proceedings of Eurographics Symposium on Point-Based Graphics, pp. 157–166 (2004)
36. Simon, H.A.: The structure of ill structured problems. Artif. Intell. **4**(3–4), 181–201 (1973)
37. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, New York (2000)
38. Kreuzthaler, M., Bloice, M., Faulstich, L., Simonic, K., Holzinger, A.: A comparison of different retrieval strategies working on medical free texts. J. Univ. Comput. Sci. **17**(7), 1109–1133 (2011)
39. Ahmadian, L., van Engen-Verheul, M., Bakhshi-Raiez, F., Peek, N., Cornet, R., de Keizer, N.F.: The role of standardized data and terminological systems in computerized clinical decision support systems: Literature review and survey. Int. J. Med. Inf. **80**(2), 81–93 (2011)
40. Batini, C., Scannapieco, M.: Data Quality: Concepts, Methodologies and Techniques. Springer, Heidelberg (2006)
41. Richman, J.S.: Multivariate Neighborhood Sample Entropy: A Method for Data Reduction and Prediction of Complex Data, pp. 297–408. Elsevier, Amsterdam (2011)
42. Komaroff, A.L.: The variability and inaccuracy of medical data. Proc. IEEE **67**(9), 1196–1207 (1979)
43. Walsh, J.E.: Analyzing medical data: some statistical considerations. IRE Trans. Med. Electron. **ME–7**(4), 362–366 (1960)
44. Miller, R., McNeil, M., Challinor, S., Masarie Jr, F., Myers, J.: The internist-1/quick medical reference project-status report. West. J. Med. **145**(6), 816 (1986)
45. Davenport, T., Glaser, J.: Just-in-time delivery comes to knowledge management. Harvard Bus. Rev. **80**(7), 107–111 (2002)
46. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute, Washington (DC) (2011)
47. Card, S.K., Mackinlay, J.D., Shneiderman, B.: Information Visualization: Using Vision to Think, pp. 1–34. Morgan Kaufmann, San Francisco (1999).

48. Inselberg, A.: Parallel Coordinates: Visual Multidimensional Geometry and Its Applications (foreword by Ben Shneiderman). Springer, Heidelberg (2009)
49. Novakova, L., Stepankova, O.: Radviz and identification of clusters in multidimensional data. In: 13th International Conference on Information Visualisation, pp. 104–109 (2009)
50. Meyer-Spradow, J., Stegger, L., Doering, C., Ropinski, T., Hinirchs, K.: Glyph-based spect visualization for the diagnosis of coronary artery disease. IEEE Trans. Visual Comput. Graphics **14**(6), 1499–1506 (2008)
51. Fox, P., Hendler, J.: Changing the equation on scientific data visualization. Science **331**(6018), 705–708 (2011)
52. de Jong, T.: Computer simulations - technological advances in inquiry learning. Science **312**(5773), 532–533 (2006)
53. Chittaro, L.: Information visualization and its application to medicine. Artif. Intell. Med. **22**(2), 81–88 (2001)
54. Johnson, C.R., MacLeod, R., Parker, S.G., Weinstein, D.: Biomedical computing and visualization software environments. Commun. ACM **47**(11), 64–71 (2004)
55. Ebner, M., Holzinger, A.: Successful implementation of user-centered game based learning in higher education an example from civil engineering. Comput. Educ. **49**(3), 873–890 (2007)
56. Kickmeier-Rust, M.D., Peirce, N., Conlan, O., Schwarz, D., Verpoorten, D., Albert, D.: Immersive Digital Games: The Interfaces for Next-Generation E-Learning?, pp. 647–656. Springer, Heidelberg (2007)
57. Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popovic, Z., Players, F.: Predicting protein structures with a multiplayer online game. Nature **466**(7307), 756–760 (2010)
58. Mayer, R.E., Hegarty, M., Mayer, S., Campbell, J.: When static media promote active learning: annotated illustrations versus narrated animations in multimedia instruction. J. Exp. Psychol. Appl. **11**(4), 256–265 (2005)
59. Holzinger, A., Kickmeier-Rust, M., Albert, D.: Dynamic media in computer science education; content complexity and learning performance: is less more? Educ. Technol. Soc. **11**(1), 279–290 (2008)
60. Hessinger, M., Holzinger, A., Leitner, D., Wassertheurer, S.: Haemodynamic models for education in physiology. Math. Comput. Simul. Simul. News Eur. **16**(2), 64–68 (2006)
61. McDonald, D.: The relation of pulsatile pressure to flow in arteries. J. Physiol. **127**, 533–552 (1955)
62. Womersley, J.R.: Method for the calculation of velocity, rate of flow and viscous drag in arteries when the pressure gradient is known. J. Physiol. **127**(3), 553–563 (1955)
63. Pedley, T.: The Fluid Mechanics of Large Blood Vessels. Cambridge University Press, Cambridge (1980)
64. Leitner, D., Wassertheurer, S., Hessinger, M., Holzinger, A.: A lattice boltzmann model for pulsative blood flow in elastic vessels. New Comput. Med. Inf. Health Care **123**(4), 64–68 (2006). Special Edition of Springer e&i
65. Holzinger, A., Ebner, M.: Interaction and Usability of Simulations & Animations: A Case Study of the Flash Technology, pp. 777–780. IOS Press, Zurich (2003)
66. Holzinger, A.: Application of rapid prototyping to the user interface development for a virtual medical campus. IEEE Softw. **21**(1), 92–99 (2004)
67. Holzinger, A.: Usability engineering for software developers. Commun. ACM **48**(1), 71–74 (2005)

68. Holzinger, A., Kickmeier-Rust, M.D., Wassertheurer, S., Hessinger, M.: Learning performance with interactive simulations in medical education: lessons learned from results of learning complex physiological models with the haemodynamics simulator. Comput. Educ. **52**(2), 292–301 (2009)
69. Schrödinger, E.: What Is Life? The Physical Aspect of the Living Cell. Dublin Institute for Advanced Studies at Trinity College, Dublin (1944)
70. Wing, J.M.: Computational thinking. Commun. ACM **49**(3), 33–35 (2006)
71. Fisher, J., Harel, D., Henzinger, T.: Biology as reactivity. Commun. ACM **54**(10), 72–82 (2011)
72. Vendruscolo, M., Dobson, C.M.: Protein dynamics: moore's law in molecular biology. Curr. Biol. **21**(2), R68–R70 (2011)
73. Holzinger, A.: Process Guide for Students for Interdisciplinary Work in Computer Science/Informatics, 2nd edn. BoD, Norderstedt (2010)
74. Wing, J.M.: Computational thinking and thinking about computing. Philos. Trans. Roy. Soc. A: Math. Phys. Eng. Sci. **366**(1881), 3717–3725 (2008)
75. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge discovery and interactive data mining in bioinformatics - state-of-the-art, future challenges and research directions. BMC Bioinform. **15**(Suppl 6), I1 (2014)
76. Simon, H.: Designing Organizations for an Information-Rich World, pp. 37–72. The Johns Hopkins Press, Baltimore (1971)
77. 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)