ON KNOWLEDGE DISCOVERY AND INTERACTIVE INTELLIGENT VISUALIZATION OF BIOMEDICAL DATA

Challenges in Human–Computer Interaction & Biomedical Informatics

Andreas Holzinger
Research Unit Human-Computer Interaction, Institute for Medical Informatics,Statistics & Documentation,
Medical University Graz, Austria
andreas.holzinger@medunigraz.at

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

Abstract: Biomedical Informatics can be defined as "the interdisciplinary field that studies and pursues the effective use of biomedical data, information and knowledge for scientific inquiry, problem solving, and decision making, motivated by efforts to improve human health." However, professionals in the life sciences are facing an increasing quantity of highly complex, multi-dimensional and weakly structured data. While researchers in Human-Computer Interaction (HCI) and Knowledge Discovery in Databases (KDD) have for long been working independently to develop methods that can support expert end users to identify, extract and understand information out of this data, it is obvious that an interdisciplinary approach to bring these two fields closer together can yield synergies in the application of these methods to weakly structured complex medical data sets. The aim is to support end users to learn how to interactively analyse information properties and to visualize the most relevant parts – in order to gain knowledge, and finally wisdom, to support a smarter decision making. The danger is not only to get overwhelmed by increasing masses of data, moreover, there is the risk of modelling artifacts.

1 INTRODUCTION

Data exploration has recently been hailed as the *fourth paradigm* in the investigation of nature, after empiricism, theory and computation (Bell, Hey & Szalay, 2009). 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 (Hey, Tansley & Tolle, 2009).

In clinical medicine, the end users are confronted with increased volumes of highly complex, noisy, high-dimensional, multivariate and often weakly-structured data (Holzinger, 2011c).

The field of biomedical informatics concerns the information processing by both humans and computers, dealing with biomedical complexity (Patel, Kahol & Buchman, 2011) to support decision making which is still a central topic in biomedical informatics (Shortliffe, 2011).

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 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 (Holzinger, 2011a). 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 (Holzinger et al., 2008a). 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 (Gregory, Mattison & Linde, 1995), (Holzinger et al., 2000), (Lovis, Baud & Planche, 2000).

In: Markus Helfert, Chiara Fancalanci, Joaquim Filipe, Eds, (2012)
Proceedings of the International Conference on Data Technologies and
Applications DATA 2012, Rome (Italy), pp. 5-16.

2 LOOK AT YOUR DATA

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

$$\boldsymbol{x}_i = [x_{i1}, \; \dots, \; x_{in}] \qquad (1)$$

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 (Pascucci et al., 2011); this can be seen as one step towards sensemaking (Blandford & Attfield, 2010).

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* (Kaski & Peltonen, 2011) of 2D or 3D representations developing in time:

$$\mathbb{R}^n + t \rightarrow \mathbb{R}^2 + t \; or \; \mathbb{R}^3 + t \qquad (2)$$

The time t is an important, yet often neglected dimension in medicine (Simonic et al., 2011).

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 (Beale, 2007). 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.

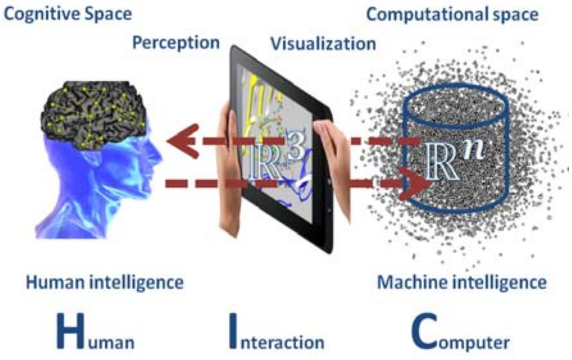


Figure 1: Human–Computer Interaction bridging the cognitive space with the computational space

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 (Yau, 2011). 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 in data. Let us start with some enthralling macroscopic dimensions: the night sky. In Fig. 2, you can see Omega Centauri, the most massive globular star cluster in our Milky Way galaxy. The core is 17,000 light-years away with a diameter of 450 light-years (Gratton et al., 2011). Globular clusters are dense, gravitationally bound collections of millions of stars that share a common age and chemical composition. Most galaxies are surrounded by systems of multiple globular clusters that swarm about them like bees around a hive (West et al., 2004).

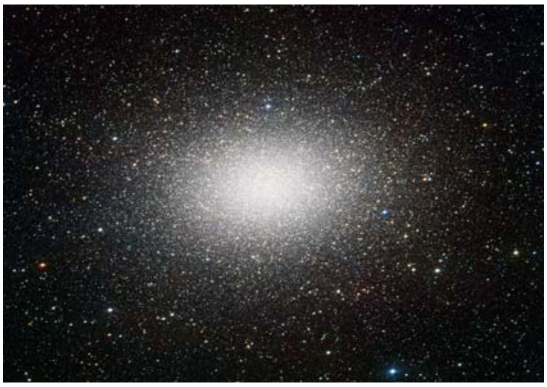


Figure 2: Globular star cluster Omega Centauri. Image available at the Eropean Southern Observatory (ESO) http://www.eso.org/public/images/eso1119b

Let us now look into the microscopic dimension (Fig. 3): Protein-protein interaction (PPI) 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 (Shi, Lei & Zhang, 2011).

The computational investigation of PPIs starts with the network structure represented by a 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 (Zhang, 2009)

$$V \times V = \{(v_i, v_j) | v_i \in V, v_j \in V, i \neq j\}. \qquad (3)$$

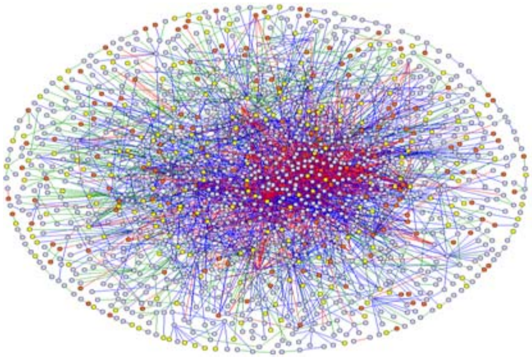
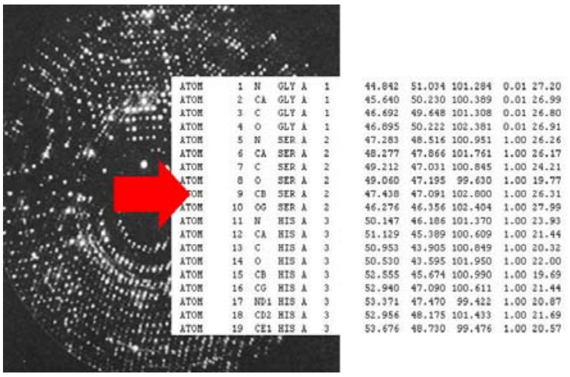


Figure 3: First visualization of a human PPI structure; Experts gain knowledge of it, e.g. to understand complex processes, thereby understand illnesses (Stelzl et al., 2005)

Protein structures are studied for example with crystallographic methods (Fig. 4). 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 (Arrais, Lopes & Oliveira, 2011).

In Fig. 4, we see the structure and the data, representing the mean positions of the entities within the substance and their chemical relationships.



```
ATOM      1  N   GLY A   1      44.842  51.034 101.284  0.01 27.20
ATOM      2  CA  GLY A   1      45.640  50.230 100.389  0.01 26.99
ATOM      3  C   GLY A   1      46.692  49.648 101.308  0.01 26.80
ATOM      4  O   GLY A   1      46.895  50.222 102.381  0.01 26.91
ATOM      5  N   SER A   2      47.283  48.516 100.951  1.00 26.26
ATOM      6  CA  SER A   2      48.277  47.866 101.761  1.00 26.17
ATOM      7  C   SER A   2      49.212  47.031 100.845  1.00 24.21
ATOM      8  O   SER A   2      49.060  47.195  99.630  1.00 19.77
ATOM      9  CB  SER A   2      47.438  47.091 102.800  1.00 26.31
ATOM     10  OG  SER A   2      46.276  46.356 102.404  1.00 27.99
ATOM     11  N   HIS A   3      50.147  46.186 101.370  1.00 23.93
ATOM     12  CA  HIS A   3      51.129  45.389 100.609  1.00 21.44
ATOM     13  C   HIS A   3      50.953  43.905 100.849  1.00 20.32
ATOM     14  O   HIS A   3      50.530  43.595 101.950  1.00 22.00
ATOM     15  CB  HIS A   3      52.555  45.674 100.990  1.00 19.69
ATOM     16  CG  HIS A   3      52.940  47.090 100.611  1.00 21.44
ATOM     17  ND1 HIS A   3      53.371  47.470  99.422  1.00 20.87
ATOM     18  CD2 HIS A   3      52.956  48.175 101.433  1.00 21.69
ATOM     19  CE1 HIS A   3      53.676  48.730  99.476  1.00 20.57
```

Figure 4: Structures of protein complexes, determined by X-ray crystallography, and stored in the PDB (Wiltgen & Holzinger, 2005)

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. 5): 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 "ball-and-stick" representation. The covalent bonds are represented as sticks between atoms, which are represented as balls. The rest of the two chains is represented as ribbons. Residue names and numbers of the TNF receptor are labelled, hydrogen bonds are represented by dotted lines (circled in Fig. 5).

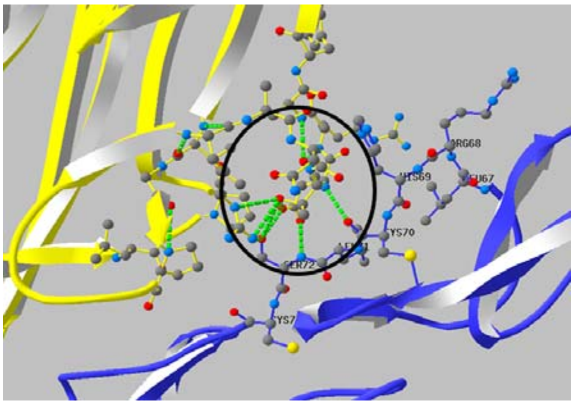


Figure 5: Gaining knowledge from the data by interactive visualization (Wiltgen, Holzinger & Tilz, 2007)

A non-natural structure is the *virtual "cosmos"*, which has been visualized in a number of ways. After six weeks of observation, Matthew Hurst mapped a visualization of the Blogosphere: "By showing only the links in the graph, we can get a far better look at the structure than if we include all the nodes" (Hurst, 2007). The most densely populated areas represent the most active portions of the blogosphere (Fig. 6). Here, we are looking at the core of the Blogosphere: The dark edges show the reciprocal links (where A has cited B and B has cited A), the lighter edges indicate a-reciprocal links. The larger, denser area of the graph is that part of the Blogosphere generally characterized by socio-political discussion (the periphery contains some topical groupings).

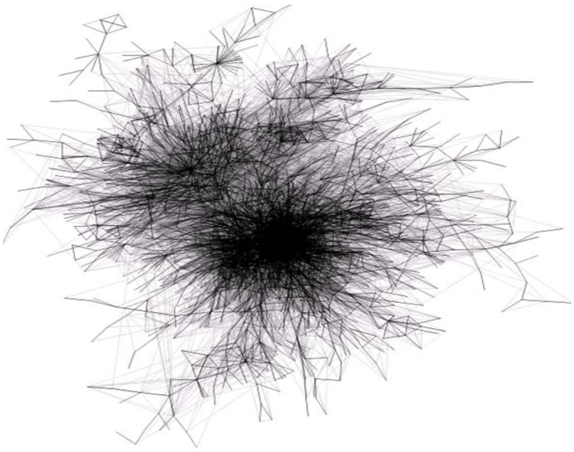


Figure 6: A visualization of the blogosphere (Hurst, 2007)

Natural structures can also be used in a completely different context: a final example should further demonstrate this: Fig. 7 shows the principle of viral marketing. The idea is to spread indirect messages which suggest spreading them farther. If you press the “Like-button” in Facebook, a process starts, which is similar to an epidemic in medicine, an illness spreading through a population. Consequently, Aral (2011) calls it *behavior contagion* and this is of importance for research to know how behavior can spread. We can mine masses of social network data in order to gain knowledge about the contagion of information. This is of particular interest for the health area, due to its remarkable similarity to the epidemic spreading of diseases (Risau-Gusman, 2012).

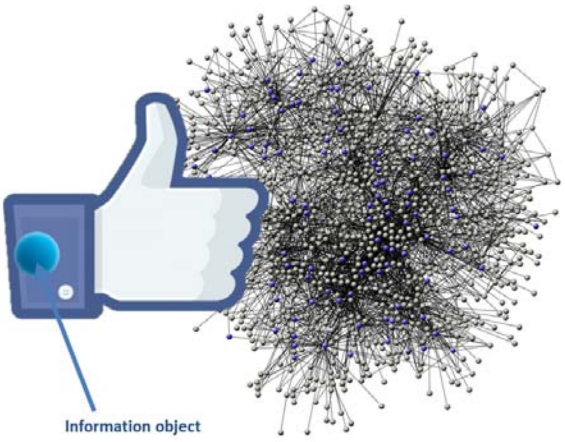


Figure 7: The principle of viral marketing – similar to contagious processes (Aral, 2011)

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 (1999), 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 (Newman, 2003), 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. 3).

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 system’s 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 (Costa, Rodrigues & Cristino, 2008), network theories can be applied to biomedical informatics.

4 TAXONOMY OF DATA

Let us list some definitions first:

Def. 1: A *relational system* is a pair $\langle A: R_1, \dots R_n \rangle$ where A is a set of elements, and $R_1, \dots R_n$ are relations defined on A.

Def. 2: An *attribute* is a homomorphism $\mathcal{H}$ from a relational system $\langle A: R_1, \dots R_n \rangle$ into a relational system $\langle B: S_1, \dots S_n \rangle$;

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 (2002) 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 (1991), who distinguished between spatial/non-spatial 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. 8). He used the nine-intersection model (Egenhofer & Herring, 1990), 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 (Aleksandrov, 1961). 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^o, boundary ∂A and exterior 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:

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

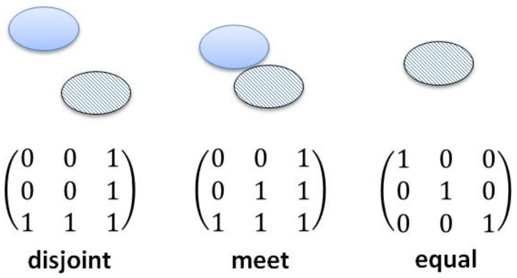

$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}$
disjoint
$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$
meet
$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
equal

Figure 8: Selected topological relations

An important invariant is the number of components. Following the definition of Egenhofer and Franzosa (1995) 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. (Stuckenschmidt & Harmelen, 2005), instead we follow the notions of topological relations (Fig 8): 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* (Kapovich et al., 2003), (de Silva & Carlsson, 2004).

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 (Simon, 1973).

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 (2000) 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 non-standardized (e.g. non-standardized text – wrongly called “free text” in an electronic patient record, see e.g. (Kreuzthaler et al., 2011).

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

iv) 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 (Ahmadian et al., 2011).

Non-standardized data, as the majority of all data impedes data quality, data exchange and interoperability (Batini & Scannapieco, 2006).

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 (Richman, 2011).

This concept has developed over the years from a basic idea. To represent a set of discrete symbols with associated probabilities, we postulate a box containing one colored ball: yellow, blue or red. If one blindly removes the ball, we are dealing with uncertainty and may ask: Is the ball red? NO. Is the ball yellow? NO. THEN it must be blue, so we need a minimum of 2 questions to provide the right answer. Because it is a binary decision (YES/NO) the maximum number of (binary) questions required to reduce the uncertainty is $\log_2(N)$, where N is the number of possible outcomes. If there are N events with equal probability p then: $N = 1/p$. If we have only 1 black ball, then: $\log_2(1) = 0$ which means there is no uncertainty. Shannon (1948) used this idea to propose a measure of uncertainty in a discrete distribution based on the Boltzmann entropy of classical statistical mechanics. He called it the information entropy H and defined it as:

$log\frac{1}{p} = -\log(p)$ wherein p is the probability of the event occurring. If this p is not identical for all events then H is a weighted average of all probabilities:

$$H = \sum_{i=1}^{N} p_i log_2(p_i) \qquad (6)$$

This measure of uncertainty has many important properties in line with the intuitive notion of randomness (Rao et al., 2004):

1) It is always positive;

2) It vanishes if and only if the event is certain;

3) Entropy is increased by the addition of an independent component, and decreased by conditioning;

For practical use it is important to know that highly structured data contain low entropy; ideally, when everything is in order the entropy is zero.

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 (1979) 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 (Walsh, 1960).

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 (Miller et al., 1986). 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 (Davenport & Glaser, 2002). 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 (Manyika et al., 2011).

6 VISUALIZATION OF DATA

How can visual representations of abstract data be used to amplify the acquisition of knowledge? (Card, Mackinlay & Shneiderman, 1999).

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 (Inselberg, 2009), RadViz (Novakova & Stepankova, 2009), or Glyphs (Meyer-Spradow et al., 2008), 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. (2009) 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 (Fox & Hendler, 2011).

Fig. 9 shows the User Interface Model for Infovis (UIMI), developed by Ren et al. (2010). In this approach, the developers can construct a model by answering four questions. The answers to this questions are then used to construct three declarative models of data, visualization and control.

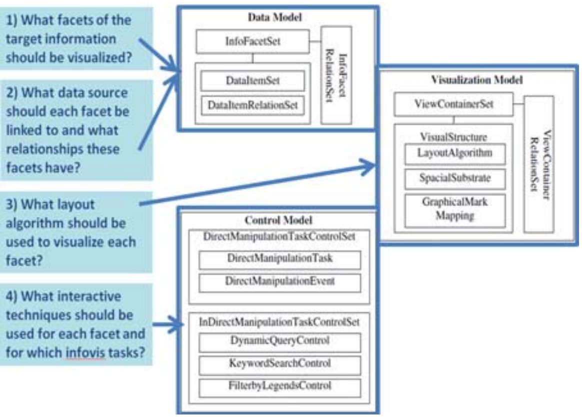

1) What facets of the target information should be visualized?
2) What data source should each facet be linked to and what relationships these facets have?
3) What layout algorithm should be used to visualize each facet?
4) What interactive techniques should be used for each facet and for which infovis tasks?
Data Model
InfoFacetSet
DataItemSet
DataItemRelationSet
InfoFacet RelationSet
Visualization Model
ViewContainerSet
VisualStructure
LayoutAlgorithm
SpacialSubstrate
GraphicalMark Mapping
ViewContainer RelationSet
Control Model
DirectManipulationTaskControlSet
DirectManipulationTask
DirectManipulationEvent
InDirectManipulationTaskControlSet
DynamicQueryControl
KeywordSearchControl
FilterbyLegendsControl

Figure 9: The conceptual model of UIMI (Ren et al., 2010)

A final practical example shall demonstrate this: Interactive computer simulations to teach complex concepts have become very popular (de Jong, 2006). The nature of such simulations ranges from compelling visualizations (Chittaro, 2001; Johnson et al., 2004) to educational computer games (Ebner & Holzinger, 2007; Kickmeier-Rust et al., 2007). A recent example is Foldit (Cooper et al., 2010), 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 (Mayer et al., 2005; Holzinger, Kickmeier-Rust & Albert, 2008b) 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 (Hessinger et al., 2006) interactively visible to medical learners (Fig. 10), so that they gain insight into the behavior of blood circulation dynamics, and to simulate certain defects (Fig. 11) and the dangers of diseases. The application simulates mathematical models (McDonald, 1955; Womersley, 1955; Pedley, 1980; Leitner et al., 2006) and presents these models in form of dynamic 2D and 3D visualizations. Special focus during the development was directed on user-centered design (Holzinger & Ebner, 2003; Holzinger, 2004; Holzinger, 2005), for example, to understand the context and to adapt the various applets to the previous knowledge of the end users.

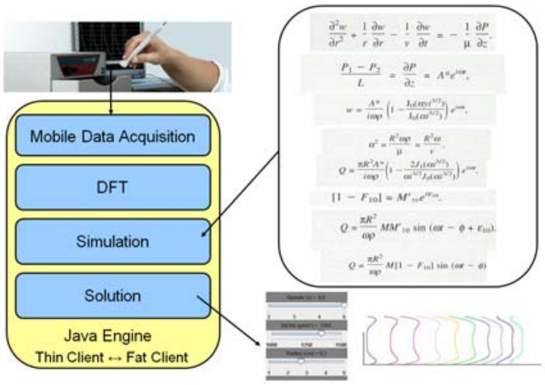
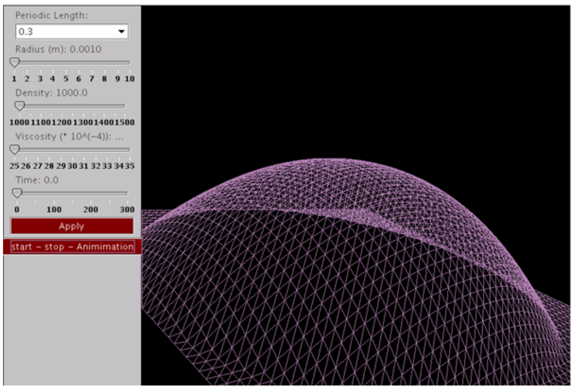


Figure 10: Real data are used for the simulation of certain clinical relevant solutions and can be interactively displayed by a learner (Holzinger et al. 2009)


Periodic Length: 0.3

Radius (m): 0.0010

Density: 1000.0

Viscosity (* 10^(−4)): ...

Time: 0.0

Figure 11: The visualized data allows insights into medical contexts and sensemaking (Holzinger et al., 2009)

6 CONCLUSION AND FUTURE OUTLOOK

Life sciences and human health are fundamentally biological, and biology is often described as *the* information science (Schrödinger, 1944).

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 (Wing, 2006) to studying biological processes, we can improve our understanding and at the same time improve the design of algorithms (Fisher, Harel & Henzinger, 2011).

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 (Vendruscolo & Dobson, 2011).

Following the quest “Science is to test ideas, engineering is to put these ideas into practice” (Holzinger, 2010), 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 (Wing, 2008), 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 (Holzinger & Simonic, 2011) 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 re-used. 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" (Simon, 1971).

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 (Holzinger, 2011b).

What price health? (*Nature,* 458, 7234, 7)

ACKNOWLEDGEMENTS

I am grateful for support, comments and help from Gig Searle, Rudi Freund and the hci4all.at team.

REFERENCES

Ahmadian, L., van Engen-Verheul, M., Bakhshi-Raiez, F., Peek, N., Cornet, R., & de Keizer, N. F. (2011). The role of standardized data and terminological systems in computerized clinical decision support systems: Literature review and survey. *International Journal of Medical Informatics, 80*(2), 81-93.

Aleksandrov, P. S. (1961). *Elementary concepts of topology*. New York: Dover Publications.

Aral, S. (2011). Identifying Social Influence: A Comment on Opinion Leadership and Social Contagion in New Product Diffusion. *Marketing Science, 30*(2), 217-223.

Arrais, J., Lopes, P., & Oliveira, J. (2011). Challenges Storing and Representing Biomedical Data. *Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058*, 53-62.

Barabási, A.-L., Albert, R., & Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications, 272*(1-2), 173-187.

Batini, C., & Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques*. Berlin, Heidelberg, New York: Springer.

Beale, R. (2007). Supporting serendipity: Using ambient intelligence to augment user exploration for data mining and Web browsing. *International Journal of Human-Computer Studies, 65*(5), 421-433.

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the data deluge. *Science, 323*(5919), 1297-1298.

Blandford, A., & Attfield, S. (2010). Interacting with Information. *Synthesis Lectures on Human-Centered Informatics, 3*(1), 1-99.

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999). In *Information Visualization: Using Vision to Think* (pp. 1-34). San Francisco: Morgan Kaufmann.

Chittaro, L. (2001). Information visualization and its application to medicine. *Artificial Intelligence in Medicine, 22*(2), 81-88.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., et al. (2010). Predicting protein structures with a multiplayer online game. *Nature, 466*(7307), 756-760.

Costa, L. F., Rodrigues, F. A., & Cristino, A. S. (2008). Complex networks: the key to systems biology. *Genetics and Molecular Biology, 31*(3), 591–601.

Dastani, M. (2002). The Role of Visual Perception in Data Visualization. *Journal of Visual Languages and Computing, 13,* 601-622.

Davenport, T. H., & Glaser, J. (2002). Just-in-time delivery comes to knowledge management. *Harvard Business Review, 80*(7), 107-111.

de Jong, T. (2006). Computer simulations - Technological advances in inquiry learning. *Science, 312*(5773), 532-533.

de Silva, V., & Carlsson, G. (2004). Topological estimation using witness complexes. *Proceedings of Eurographics Symposium on Point-Based Graphics*, 157-166.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification. Second Edition*. New York et al.: Wiley.

Ebner, M., & Holzinger, A. (2007). Successful Implementation of User-Centered Game Based Learning in Higher Education – an Example from Civil Engineering. *Computers & Education, 49*(3), 873-890.

Egenhofer, M. (1991). Reasoning about binary topological relations. In O. Günther & H.-J. Schek (Eds.), *Advances in Spatial Databases, Lecture Notes in Computer Sciences LNCS 525* (pp. 141-160). Berlin, Heidelberg: Springer

Egenhofer, M., & Franzosa, R. (1995). On the equivalence of topological relations. *International Journal of Geographical Information Systems, 9*(2), 133-152.

Egenhofer, M. J., & Herring, J. (1990). Categorizing binary topological relations between regions, lines, and points in geographic databases. *Technical Report, Department of Surveying Engineering, University of Maine.*

Fisher, J., Harel, D., & Henzinger, T. A. (2011). Biology as reactivity. *Communications of the ACM, 54*(10), 72-82.

Fox, P., & Hendler, J. (2011). Changing the equation on scientific data visualization. *Science, 331*(6018), 705-708.

Gratton, R. G., Johnson, C. I., Lucatello, S., D'Orazi, V., & Pilachowski, C. (2011). Multiple populations in omega Centauri: a cluster analysis of spectroscopic data. *Astronomy & Astrophysics, 534*.

Gregory, J., Mattison, J. E., & Linde, C. (1995). Naming Notes - Transitions from Free-Text to Structured Entry. *Methods of Information in Medicine, 34*(1-2), 57-67.

Hessinger, M., Holzinger, A., Leitner, D., & Wassertheurer, S. (2006). Haemodynamic Models for Education in Physiology. *Mathematics and Computers in Simulation: Simulation News Europe, 16*(2), 64-68.

Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: data-intensive scientific discovery*. Redmond (WA): Microsoft Research.

Holzinger, A. (2004). Application of Rapid Prototyping to the User Interface Development for a Virtual Medical Campus. *IEEE Software, 21*(1), 92-99.

Holzinger, A. (2005). Usability Engineering for Software Developers. *Communications of the ACM, 48*(1), 71-74.

Holzinger, A. (2010). *Process Guide for Students for Interdisciplinary Work in Computer Science/Informatics. Second Edition*. Norderstedt: BoD.

Holzinger, A. (2011a). Interacting with Information: Challenges in Human-Computer Interaction and Information Retrieval (HCI-IR). In *IADIS Multiconference on Computer Science and Information Systems (MCCSIS), Interfaces and Human-Computer Interaction* (pp. 13-17). Rome: IADIS.

Holzinger, A. (2011b). *Successful Management of Research and Development*. Norderstedt: BoD.

Holzinger, A. (2011c). *Weakly Structured Data in Health-Informatics: The Challenge for Human-Computer Interaction.* Paper presented at the Proceedings of INTERACT 2011 Workshop: Promoting and supporting healthy living by design.

Holzinger, A., & Ebner, M. (2003). Interaction and Usability of Simulations & Animations: A case study of the Flash Technology. In M. Rauterberg, M. Menozzi & J. Wesson (Eds.), *Human-Computer Interaction Interact 2003* (pp. 777-780). Zurich, Amsterdam: IOS Press.

Holzinger, A., Geierhofer, R., Modritscher, F., & Tatzl, R. (2008a). Semantic Information in Medical Information Systems: Utilization of Text Mining Techniques to Analyze Medical Diagnoses. *Journal of Universal Computer Science, 14*(22), 3781-3795.

Holzinger, A., Kainz, A., Gell, G., Brunold, M., & Maurer, H. (2000). *Interactive Computer Assisted Formulation of Retrieval Requests for a Medical Information System using an Intelligent Tutoring System.* Paper presented at the World Conference on Educational Multimedia, Hypermedia and Telecommunications ED-MEDIA 2000, Montreal.

Holzinger, A., Kickmeier-Rust, M., & Albert, D. (2008b). Dynamic Media in Computer Science Education; Content Complexity and Learning Performance: Is Less More? *Educational Technology & Society, 11*(1), 279-290.

Holzinger, A., Kickmeier-Rust, M. D., Wassertheurer, S., & Hessinger, M. (2009). Learning performance with interactive simulations in medical education: Lessons learned from results of learning complex physiological models with the HAEMOdynamics SIMulator. *Computers & Education, 52*(2), 292-301.

Holzinger, A., & Simonic, K.-M. (Eds.). (2011). *Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058*. Heidelberg, Berlin, New York: Springer.

Hurst, M. (2007). Data Mining: Text Mining, Visualization and Social Media.Data accessed: from http://datamining.typepad.com/data_mining/2007/01/the_blogosphere.html

Inselberg, A. (2009). *Parallel coordinates: visual multidimensional geometry and its applications (foreword by Ben Shneiderman)*. Dordrecht, Heidelberg, London, New York: Springer.

Johnson, C. R., MacLeod, R., Parker, S. G., & Weinstein, D. (2004). Biomedical computing and visualization software environments *Communications of the ACM, 47*(11), 64-71.

Kapovich, I., Myasnikov, A., Schupp, P., & Shpilrain, V. (2003). Generic-case complexity, decision problems in group theory, and random walks. *Journal of Algebra, 264*(2), 665-694.

Kaski, S., & Peltonen, J. (2011). Dimensionality Reduction for Data Visualization (Applications Corner). *IEEE Signal Processing Magazine, 28*(2), 100-104.

Kickmeier-Rust, M. D., Peirce, N., Conlan, O., Schwarz, D., Verpoorten, D., & Albert, D. (2007). Immersive Digital Games: The Interfaces for Next-Generation E-Learning? In C. Stephanidis (Ed.), *Universal Access in Human-Computer Interaction. Applications and Services (Lecture Notes in Computer Science 4556)* (pp. 647-656). Heidelberg, Berlin, New York: Springer.

Komaroff, A. L. (1979). The variability and inaccuracy of medical data. *Proceedings of the IEEE, 67*(9), 1196-1207.

Kreuzthaler, M., Bloice, M. D., Faulstich, L., Simonic, K. M., & Holzinger, A. (2011). A Comparison of Different Retrieval Strategies Working on Medical Free Texts. *Journal of Universal Computer Science, 17*(7), 1109-1133.

Leitner, D., Wassertheurer, S., Hessinger, M., & Holzinger, A. (2006). A Lattice Boltzmann Model for Pulsative Blood Flow in Elastic Vessels. . *New Computing in Medical Informatics & Health Care. Special Edition of Springer e&i, 123*(4), 64-68.

Lovis, C., Baud, R. H., & Planche, P. (2000). Power of expression in the electronic patient record: structured data or narrative text? *International Journal of Medical Informatics, 58*, 101-110.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al. (2011). *Big data: The next frontier for innovation, competition, and productivity*. Washington (DC): McKinsey Global Institute.

Mayer, R. E., Hegarty, M., Mayer, S., & Campbell, J. (2005). When Static Media Promote Active Learning: Annotated Illustrations Versus Narrated Animations in Multimedia Instruction. *Journal of Experimental Psychology: Applied, 11*(4), 256-265.

McDonald, D. A. (1955). The relation of pulsatile pressure to flow in arteries. *Journal of Physiology, 127*(533-552).

Meyer-Spradow, J., Stegger, L., Doering, C., Ropinski, T., & Hinirchs, K. (2008). Glyph-Based SPECT Visualization for the Diagnosis of Coronary Artery Disease. *IEEE Transactions on Visualization and Computer Graphics, 14*(6), 1499-1506.

Miller, R. A., McNeil, M. A., Challinor, S. M., Masarie Jr, F. E., & Myers, J. D. (1986). The INTERNIST-1/quick medical REFERENCE project—Status report. *Western Journal of Medicine, 145*(6), 816.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM review*, 167-256.

Novakova, L., & Stepankova, O. (2009, 15-17 July 2009). *RadViz and Identification of Clusters in Multidimensional Data.* Paper presented at the 13th International Conference on Information Visualisation.

Pascucci, V., Tricoche, X., Hagen, H., & Tierny, J. (2011). *Topological Methods in Data Analysis and Visualization: Theory, Algorithms, and Applications*. Berlin, Heidelberg: Springer.

Patel, V. L., Kahol, K., & Buchman, T. (2011). Biomedical Complexity and Error. *Journal of Biomedical Informatics, 44*(3), 387-389.

Pedley, T. (1980). *The fluid mechanics of large blood vessels.* . Cambridge (MA): Cambridge University Press.

Rao, M., Chen, Y. M., Vemuri, B. C., & Wang, F. (2004). Cumulative residual entropy: A new measure of information. *IEEE Transactions on Information Theory, 50*(6), 1220-1228.

Ren, L., Tian, F., Zhang, X., & Zhang, L. (2010). DaisyViz: A model-based user interface toolkit for interactive information visualization systems. *Journal of Visual Languages & Computing, 21*(4), 209-229.

Richman, J. S. (2011). Multivariate Neighborhood Sample Entropy: A Method for Data Reduction and Prediction of Complex Data. In *Methods in Enzymology, Volume 487* (pp. 297-408). Amsterdam: Elsevier.

Risau-Gusman, S. (2012). Influence of network dynamics on the spread of sexually transmitted diseases. *Journal of the Royal Society Interface, 9*(71), 1363-1372.

Schrödinger, E. (1944). *What Is Life? The Physical Aspect of the Living Cell*. Dublin: Dublin Institute for Advanced Studies at Trinity College.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*, 379-423.

Shi, L., Lei, X., & Zhang, A. (2011). Protein complex detection with semi-supervised learning in protein interaction networks. *Proteome Science, 9*(Suppl 1), S5.

Shortliffe, E. H. (2011). Biomedical Informatics: Defining the Science and its Role in Health Professional Education. In A. Holzinger & K.-M. Simonic (Eds.), *Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058* (pp. 711-714). Heidelberg, New York: Springer.

Simon, H. A. (1971). Designing Organizations for an Information-Rich World. In M. Greenberger (Ed.), *Computers, Communication, and the Public Interest* (pp. 37-72). Baltimore (MD): The Johns Hopkins Press.

Simon, H. A. (1973). The structure of ill structured problems. *Artificial Intelligence, 4*(3-4), 181-201.

Simonic, K. M., Holzinger, A., Bloice, M., & Hermann, J. (2011). Optimizing Long-Term Treatment of Rheumatoid Arthritis with Systematic Documentation. In *Proceedings of Pervasive Health - 5th International Conference on Pervasive Computing Technologies for Healthcare* (pp. 550-554). Dublin: IEEE.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., et al. (2005). A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. *Cell, 122*(6), 957-968.

Stuckenschmidt, H., & Harmelen, F. v. (2005). *Information Sharing on the Semantic Web. Series: Advanced Information and Knowledge Processing*. Heidelberg, Berlin, New York: Springer.

Vendruscolo, M., & Dobson, C. M. (2011). Protein Dynamics: Moore's Law in Molecular Biology. *Current Biology, 21*(2), R68-R70.

Walsh, J. E. (1960). Analyzing Medical Data: Some Statistical Considerations. *Medical Electronics, IRE Transactions on, ME-7*(4), 362-366.

West, M. J., Cote, P., Marzke, R. O., & Jordan, A. (2004). Reconstructing galaxy histories from globular clusters. *Nature, 427*(6969), 31-35.

Wiltgen, M., & Holzinger, A. (2005). Visualization in Bioinformatics: Protein Structures with Physicochemical and Biological Annotations. In J. Zara & J. Sloup (Eds.), *Central European Multimedia and Virtual Reality Conference (available in EG Eurographics Library)* (pp. 69-74). Prague: Czech Technical University (CTU).

Wiltgen, M., Holzinger, A., & Tilz, G. P. (2007). Interactive Analysis and Visualization of Macromolecular Interfaces Between Proteins. In A. Holzinger (Ed.), *HCI and Usability for Medicine and Health Care. Lecture Notes in Computer Science (LNCS 4799)* (pp. 199-212). Berlin, Heidelberg, New York: Springer.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM, 49*(3), 33-35.

Wing, J. M. (2008). Computational thinking and thinking about computing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 366*(1881), 3717-3725.

Womersley, J. R. (1955). Method for the calculation of velocity, rate of flow and viscous drag in arteries when the pressure gradient is known. *The Journal Of Physiology, 127*(3), 553-563.

Yau, N. (2011). Seeing the World in Data. In M. Lima (Ed.), *Visual Complexity: Mapping Patterns of Information* (pp. 246-248). New York: Princeton Architectural Press.

Zhang, A. (2009). *Protein Interaction Networks: Computational Analysis*. Cambridge, New York et al.: Cambridge University Press.