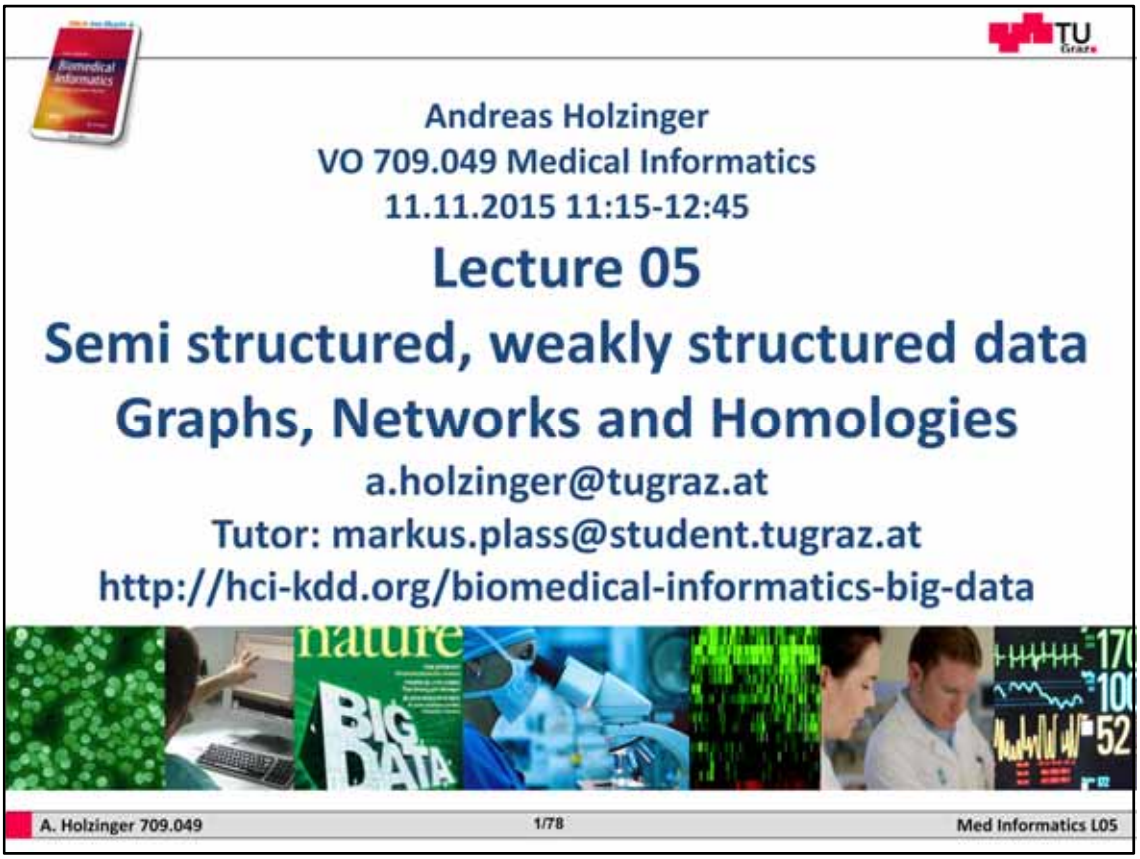



Status as of 08.11.2015 10:00

Dear Students, welcome to the 5th lecture of our course. Please remember from the last lecture the basic architecture of a hospital information system, the complexity of medical workflows, the challenges of data integration, data fusion, data curation; the building blocks of hospital information systems, databases, data warehouses, data marts; the difference between knowledge discovery and information retrieval; please remember the formal description of a information retrieval model – the best practice example is the Page-Rank Algorithm, see:
Hastie, T., Tibshirani, R. & Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, New York, Springer.
Or have a look to the reprint paper:
Brin, S. & Page, L. 2012. Reprint of: The anatomy of a large-scale hypertextual web search engine. Computer Networks, 56, (18), 3825-3833.
http://www.sciencedirect.com/science/article/pii/S1389128612003611
doi:10.1016/j.comnet.2012.10.007

Please always be aware of the definition of biomedical informatics (Medizinische Informatik):
Biomedical Informatics is the inter-disciplinary 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 (and well-being).

## Schedule

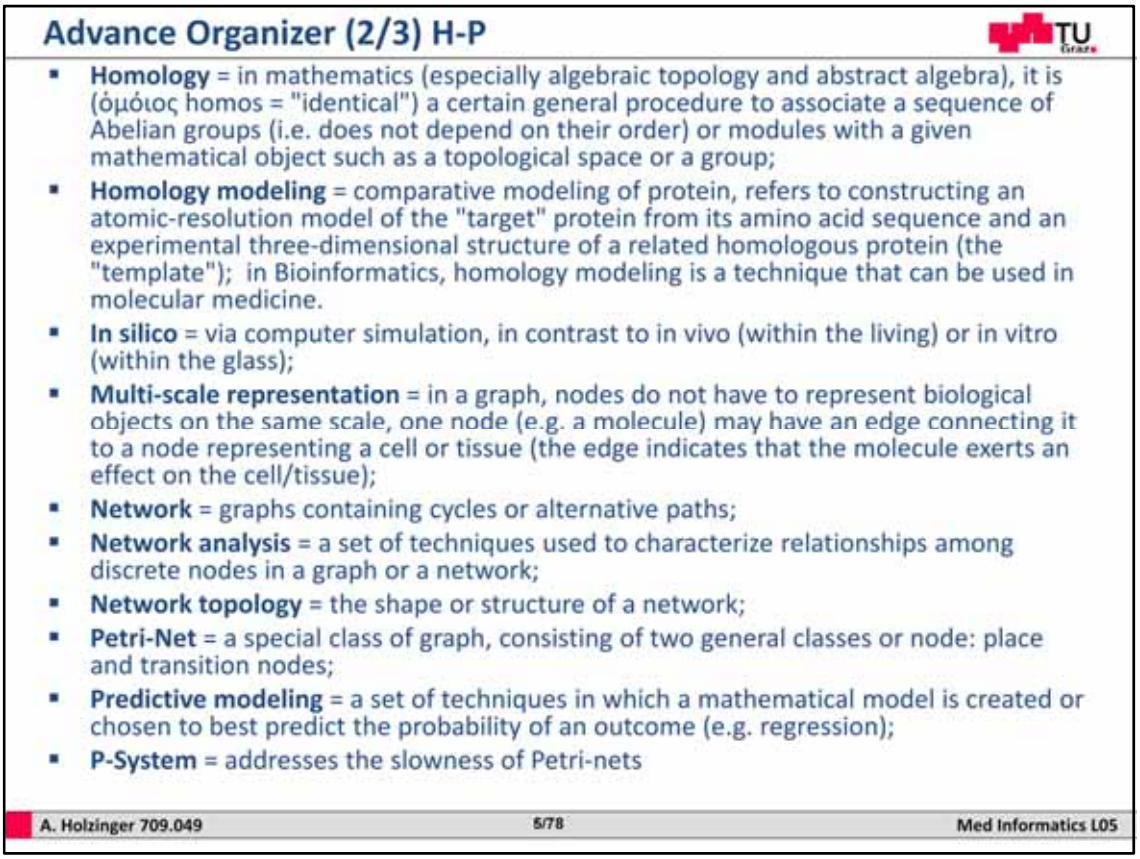
- 1. Intro: Computer Science meets Life Sciences, challenges, future directions
- 2. Back to the future: Fundamentals of Data, Information and Knowledge
- 3. Structured Data: Coding, Classification (ICD, SNOMED, MeSH, UMLS)
- 4. Biomedical Databases: Acquisition, Storage, Information Retrieval and Use
- **5. Semi structured and weakly structured data (structural homologies)**
- 6. Multimedia Data Mining and Knowledge Discovery
- 7. Knowledge and Decision: Cognitive Science & Human-Computer Interaction
- 8. Biomedical Decision Making: Reasoning and Decision Support
- 9. Intelligent Information Visualization and Visual Analytics
- 10. Biomedical Information Systems and Medical Knowledge Management
- 11. Biomedical Data: Privacy, Safety and Security
- 12. Methodology for Info Systems: System Design, Usability & Evaluation

## Keywords of the 5th Lecture

- Big data pools
- Complex networks
- Computational graph representation
- Electronic patient record (EPR)
- Homology modeling
- Macroscopic structures
- Medical documentation
- Metabolic network
- Microscopic structures
- Network metrics
- Structural data dimension
- Topological structures

## Advance Organizer (1/3) A-G

- **Adjacency matrix** = simplest form of computational graph representation, in which 0 or 1 denotes whether or not there is a directed edge from one node to another (in graph theory adjacent nodes in a graph are linked by an edge);
- **Artifacts** = not only a noise disturbance, which is contaminating and influencing the signal (surrogates) but also data which is wrong, however interpreted as to be reliable, consequently may lead to a wrong decision;
- **Computational graph representation** = e.g. by adjacency matrices
- **Data fusion** = data integration techniques that analyze data from multiple sources in order to develop insights in ways that are more efficient and potentially more accurate than if they were developed by analyzing a single source of data. Signal processing techniques can be used to implement some types of data fusion (e.g. combined sensor data in Ambient Assisted Living);
- **Global Distance Test (GDT)** = a measure of similarity between two protein structures with identical amino acid sequences but different tertiary structures. It is most commonly used to compare the results of protein structure prediction to the experimentally determined structure as measured by X-ray crystallography or protein NMRM;
- **Graph theory** = study of mathematical structures to model relations between objects from a certain collection;
- **Graphs** = a hypothetical structure consisting of a series of nodes connected by weighted edges (graphs can be directed/undirected and stoichometric/non-stoichometric regarding interaction classes);




In vivo (Latin for "within the living") is experimentation using a whole, living organism as opposed to a partial or dead organism, or an in vitro ("within the glass", i.e., in a test tube or petri dish) controlled environment.

## Advance Organizer (3/3) R-V

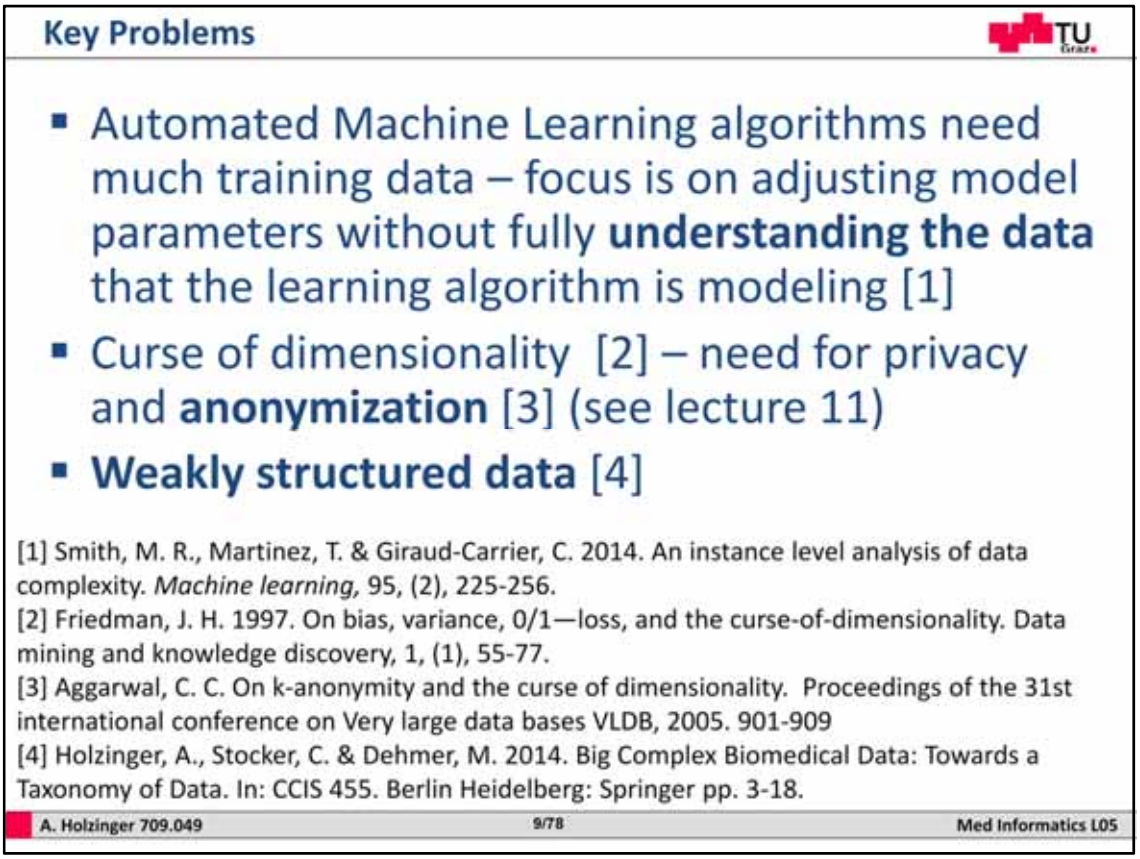
- **Radius of a graph** = average minimum path length (biological networks are not arranged in a regular or symmetrical pattern);
- **Scale-free Topology** = ensures that there are very short paths between any given pair of nodes, allowing rapid communication between otherwise distant parts of the network (e.g. the Web has such a topology);
- **Semi-structured data** = does not conform with the formal structure of tables/data models assoc. with relational databases, but at least contains tags/markers to separate semantic elements and enforce hierarchies of records and fields within the data; aka schemaless or self-describing structure; the entities belonging to the same class may have different attributes even though they are grouped together;
- **Spatial analysis** = a set of techniques, applied from statistics, which analyze the topological, geometric, or geographic properties encoded in a data set;
- **Structural homology** = similar structure but different function;
- **Supervised learning** = machine learning techniques that infer a function or relationship from a set of training data (e.g. classification and support vector machines);
- **Time series analysis** = set of techniques from both statistics and signal processing for analyzing sequences of data points, representing values at successive times, to extract meaningful characteristics from the data;
- **Time series forecasting** = use of a model to predict future values of a time series based on known past values of the same or other series (e.g. structural modeling); decomposition of a series into trend, seasonal, and residual components, which can be useful for identifying cyclical patterns in the data;
- **Unstructured data** = complete randomness, noise; (wrongly, text is called unstructured, but there is some structure, too, so text data is a kind of weakly structured data);
- **Vertex degree** = within a topology, the numbers of edges connecting to a node;

## Glossary

- ANSI = American National Standards Institute
- CD = cardiac development
- CDA = Clinical Document Architecture
- CHD = congenital heart disease
- CMM = Correlated motif mining
- DPI = Dossier Patient Integre´ = integrated patient record
- E = Edge
- EPR = Electronic Patient Record
- G(V,E) = Graph
- GI = gastrointestinal
- HER = Electronic Health Record
- HL7 = Health Level 7
- KEGG = Kyoto Encyclopedia of Genes and Genomes
- NP = nondeterministic polynomial time
- OWL = Web Ontology Language
- PPI = Protein-Protein Interaction
- SGML = Standard Generalized Markup Language
- TF= Transcription factor
- TG = Target Gene
- V = Vertex
- XML = Extensible Markup Language

## Learning Goals … at the end of the 5th lecture you …

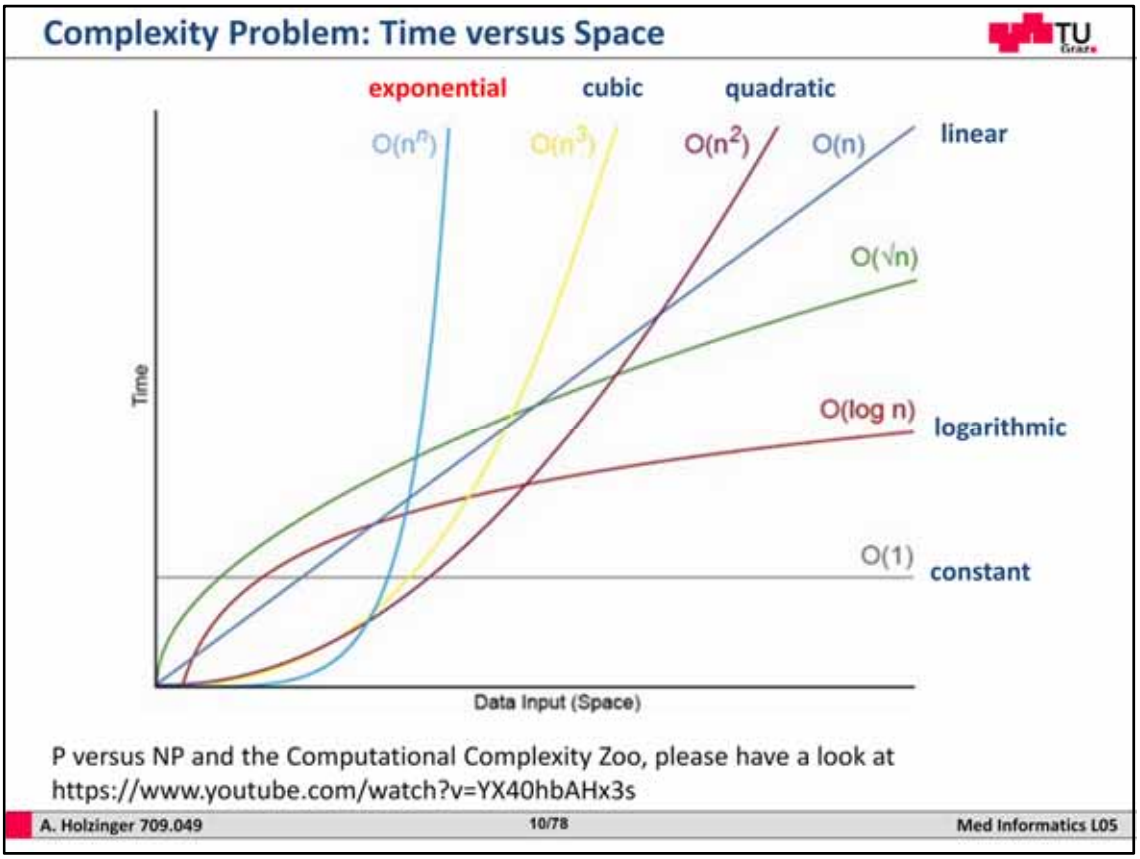
- … have an idea of the **complexity of data** in biomedical informatics
- … are aware of the various **contents** of Electronic Patient Records
- … have seen some application examples of **network structures** from both macro-cosmos and micro-cosmos and are fascinated about it;
- … have a rough overview about some basics of how to **get point clouds** out of data sets
- … have an understanding of the challenges of **network science**




It is widely acknowledged in machine learning that the performance of a learning algorithm is dependent on both its parameters and the training data. Yet, the bulk of algorithmic development has focused on adjusting model parameters without fully understanding the data that the learning algorithm is modeling. As such, algorithmic development for classification problems has largely been measured by classification accuracy, precision, or a similar metric on benchmark data sets. As most machine learning research is focused on the data set level, one is concerned with maximizing $p(h|t)$, where $h : X \rightarrow Y$ is a hypothesis or function mapping input feature vectors X to their corresponding label vectors Y , and $t = \{(x_i, y_i) : x_i \in X \wedge y_i \in Y \}$ is a training set.

One of the methods for privacy preserving data mining is that of anonymization, in which a record is released only if it is indistinguishable from k other entities in the data. We note that methods such as k-anonymity are highly dependent upon spatial locality in order to effectively implement the technique in a statistically robust way. In high dimensional space the data becomes sparse, and the concept of spatial locality is no longer easy to define from an application point of view. Aggarwal, C. C. On k-anonymity and the curse of dimensionality. Proceedings of the 31st international conference on Very large data bases VLDB, 2005. 901-909.

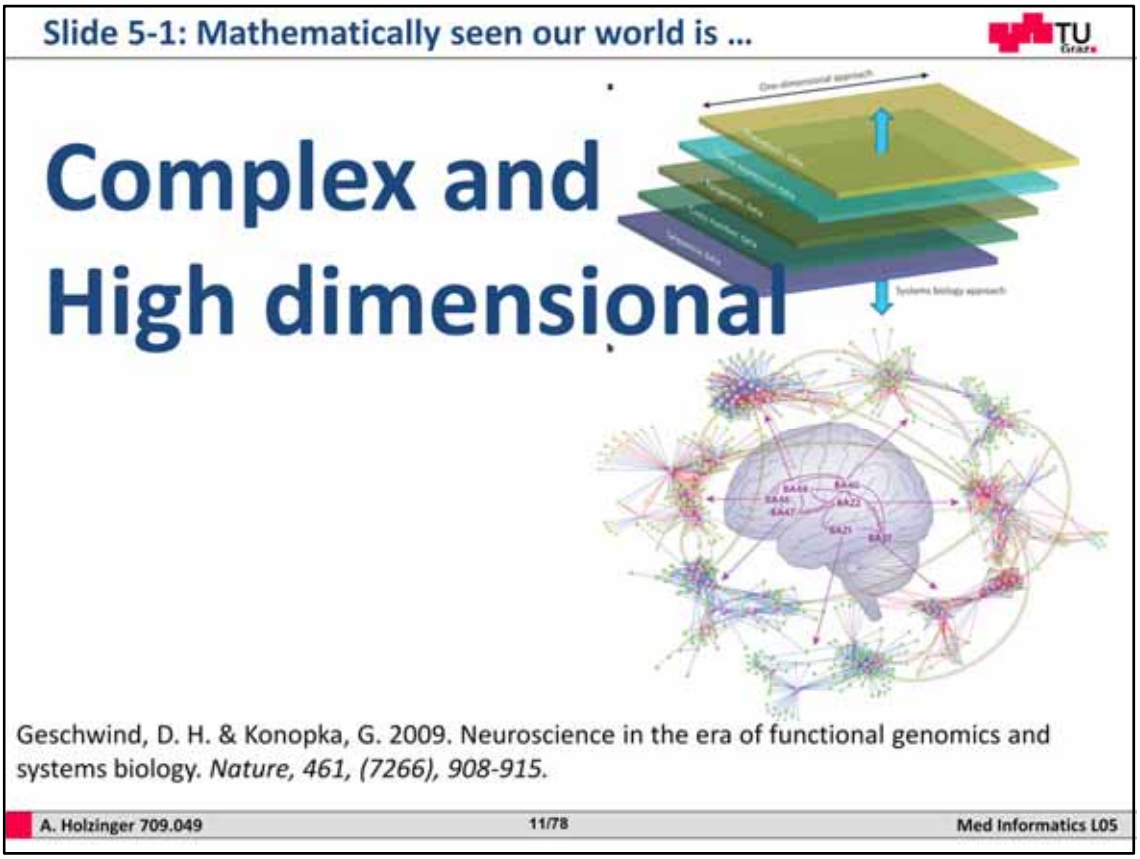
Holzinger, A., Stocker, C. & Dehmer, M. 2014. Big Complex Biomedical Data: Towards a Taxonomy of Data. In: Obaidat, M. S. & Filipe, J. (eds.) Communications in Computer and Information Science CCIS 455. Berlin Heidelberg: Springer pp. 3-18.




https://www.projectrhea.org/rhea/index.php/File:Complexitytable.png

P stands for "polynomial time". This the subset of problems that can be guaranteed to be solved in a polynomial amount of time related to their input length. Problems in P commonly operate on single inputs, lists, or matrices, and can occasionally apply to graphs. The typical types of operations they perform are mathematical operators, sorting, finding minimum and maximum values, determinates, and many others.
NP stands for "nondeterministic polynomial time". These problems are ones that can be solved in polynomial time using a nondeterministic computer. This concept is a little harder to understand, so another definition that is a consequence of the first is often used. NP problems are problems that can be checked, or "certified", in polynomial time. The output of an NP solving program is called a certificate, and the polynomial time program that checks the certificate for its validity is called the certification program.
NP-hard:
A problem is NP-hard if it as least as hard as the hardest problems known to be NP. This leads to two possibilities: either the problem is in NP and also considered NP-hard, or it is more difficult than any NP problem.
NP-complete:
This classification is the intersection of NP and NP-hard. If a problem is in NP and also NP-hard, then it is considered NP-complete. This class of problems is arguably the most interesting for its consequences on many other types of problems.

For those who want to go deeper into complexity theory, there is excellent MIT Open Courseware by Eric Demaine, http://erikdemaine.org/
https://www.youtube.com/watch?v=moPtwq_cVH8
You can do some own experimentation via http://www.algomation.com

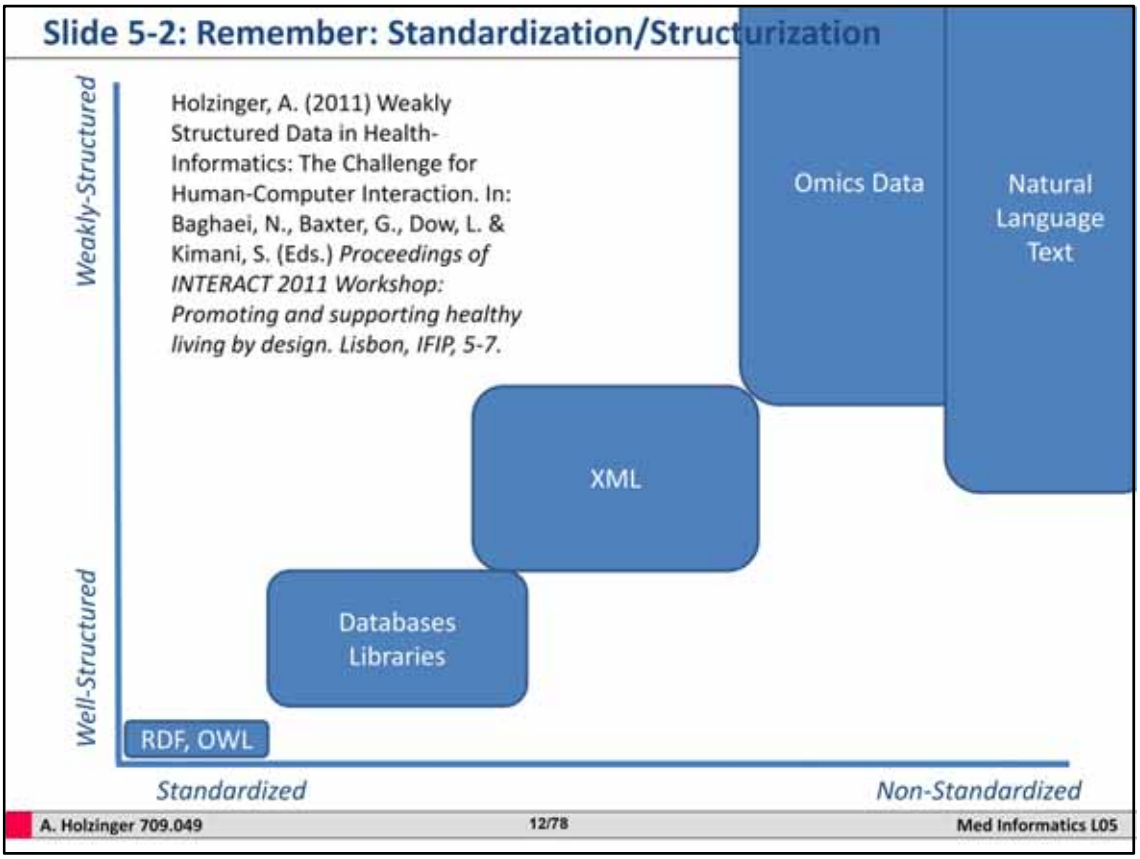



Key problems in dealing with data in the life sciences include:

- Complexity of our world
- High-dimensionality (curse of dimensionality (Catchpoole et al., 2010))
- Most of the data is weakly-structured and unstructured

A grand challenge in healthcare is the complexity of data, implicating two issues: structurization and standardization. As we have learned in lecture 2, very little of the data is structured. Most of our data is weakly structured (Holzinger, 2012). In the language of business there is often the use of the word "unstructured", but we have to use this word with care; unstructured would mean – in a strict mathematical sense – that we are talking about total randomness and complete uncertainty, which would mean noise, where standard methods fail or lead to the modeling of artifacts, and only statistical approaches may help. The correct term would be unmodeled data – or we shall speak about unstructured information. Please mind the differences.

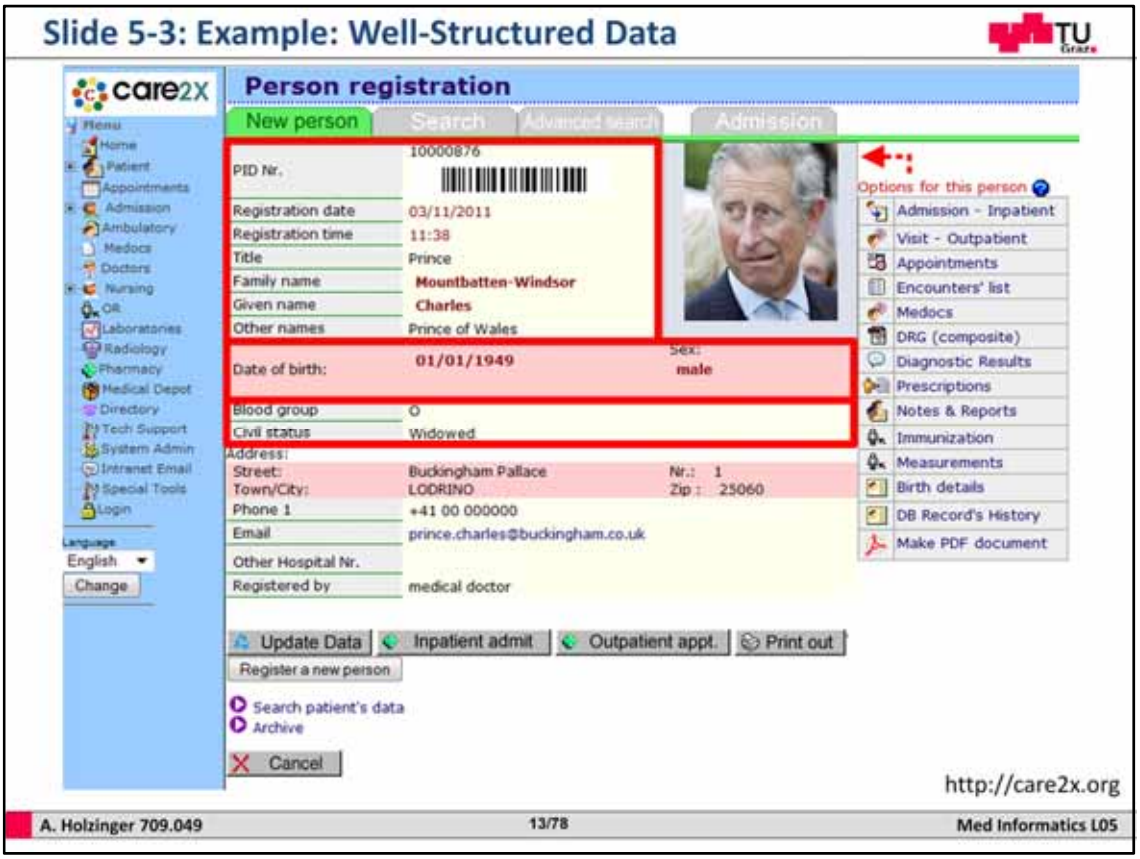
To the image above: Advances in genetics and genomics have accelerated the discovery-based (=hypotheses generating) research that provides a powerful complement to the direct hypothesis-driven molecular, cellular and systems sciences. For example, genetic and functional genomic studies have yielded important insights into neuronal function and disease. One of the most exciting and challenging frontiers in neuroscience involves harnessing the power of large-scale genetic, genomic and phenotypic data sets, and the development of tools for data integration and data mining (Geschwind & Konopka, 2009).



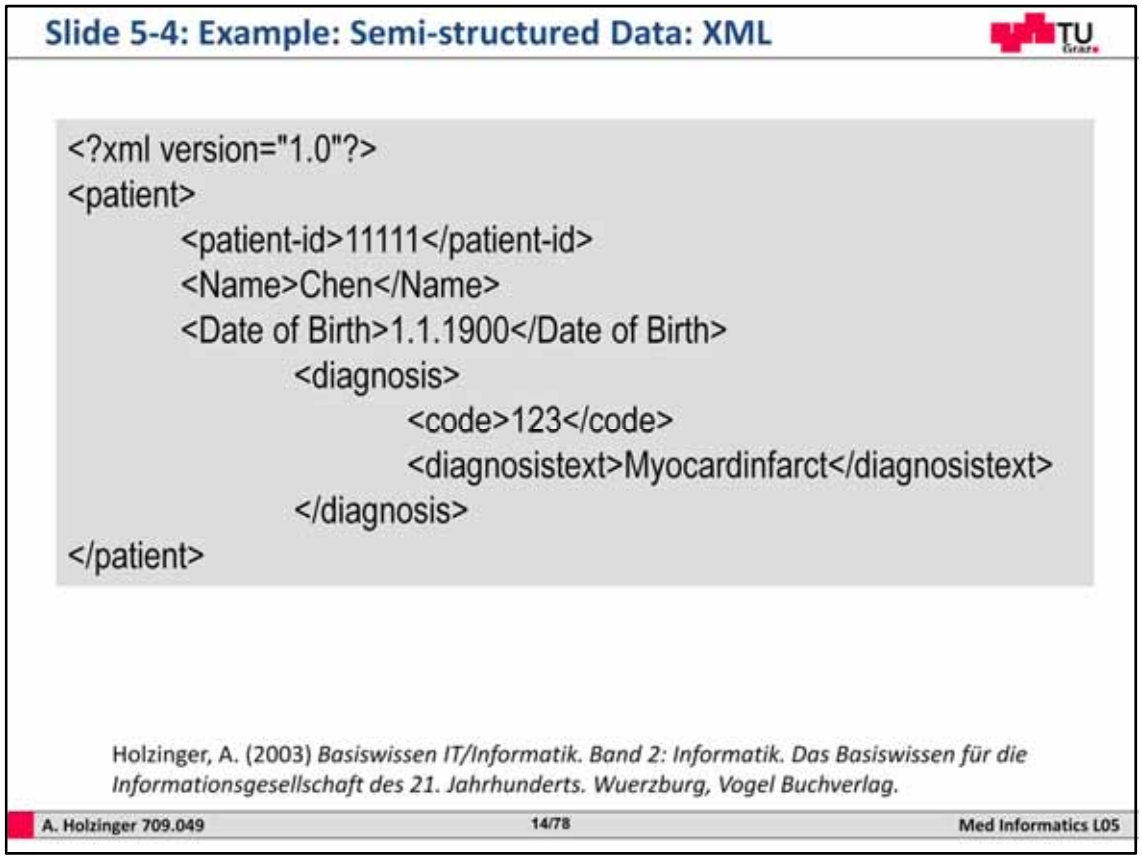

Do not confuse structure with standardization (see Slide 2-9). Data can be standardized (e.g. numerical entries in laboratory reports) and non-standardized. A typical example is non-standardized text – imprecisely called "Free-Text" or "unstructured data" in an electronic patient record (Kreuzthaler et al., 2011).

**Standardized data** is *the* 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 interpreted by all users with the same understanding. Moreover, standardized data facilitate comparability of data and interoperability of systems. It supports the reusability of the data, improves the efficiency of healthcare services and avoids errors by reducing duplicated efforts in data entry.

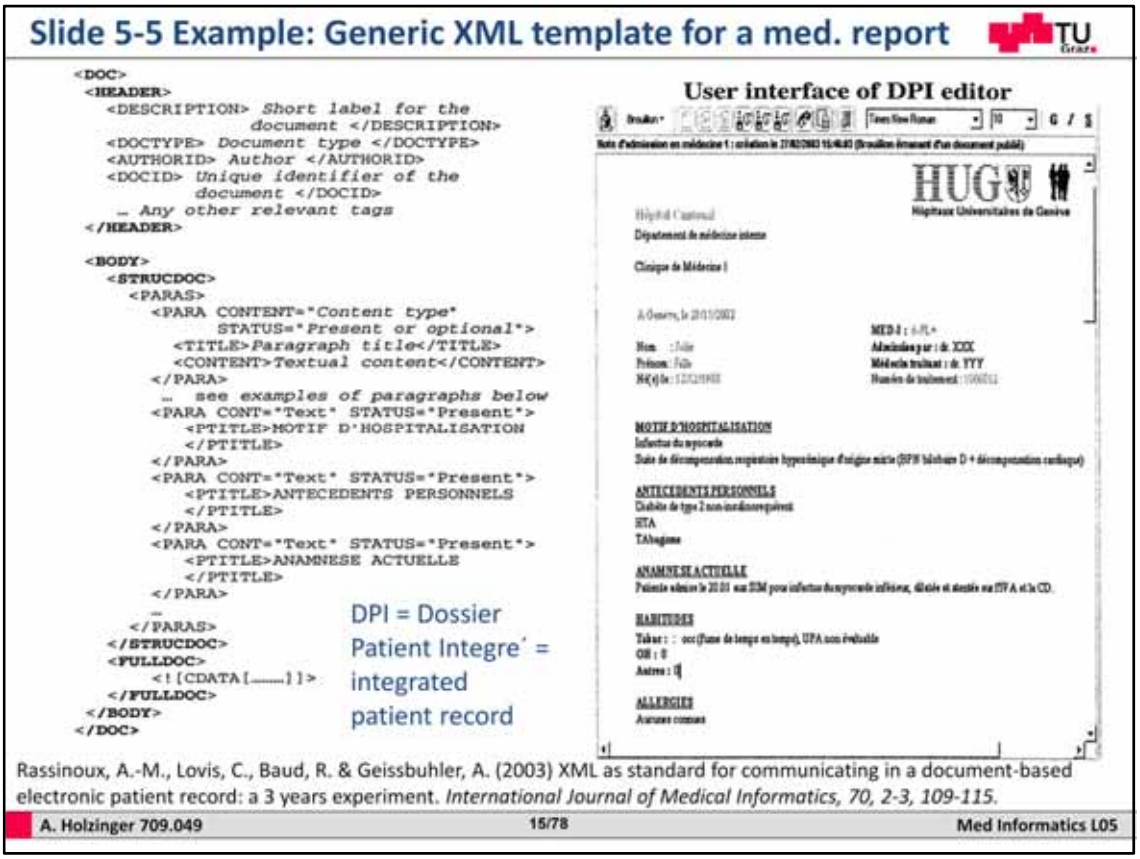
Data standardization refers to
a) the data content;
b) the terminologies that are used to represent the data;
c) how data is exchanged; and
iv) how knowledge, e.g. clinical guidelines, protocols, decision support rules, checklists, standard operating procedures are represented in the health information system (refer to IOM).
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), which we will discuss in →Lecture 3.
**Non-standardized data** is the majority of data and inhibit data quality, data exchange and interoperability.
**Well-structured data** is the minority of data and an idealistic case when each data element has an associated defined structure, relational tables, or the resource description framework RDF, or the Web Ontology Language OWL (see →Lecture 3).
Note: **Ill-structured** is a term often used for the opposite of well-structured, although this term originally was used in the 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 structure and content, i.e. are schema-less or self-describing; a typical example is a markup-language such as XML (see →Lecture 3 and 4).
**Weakly-Structured data** is the most of our data in the whole universe, whether it is in macroscopic (astronomy) or microscopic structures (biology) – see →Lecture 5.
**Non-structured data** or *unstructured data* is an imprecise definition used for *information* expressed in natural language, when no specific structure has been defined. This is an issue for debate: Text has also some structure: words, sentences, paragraphs. If we are very precise, unstructured data would meant that the data is complete randomized – which is usually called noise and is defined by (Duda, Hart & Stork, 2000) 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).



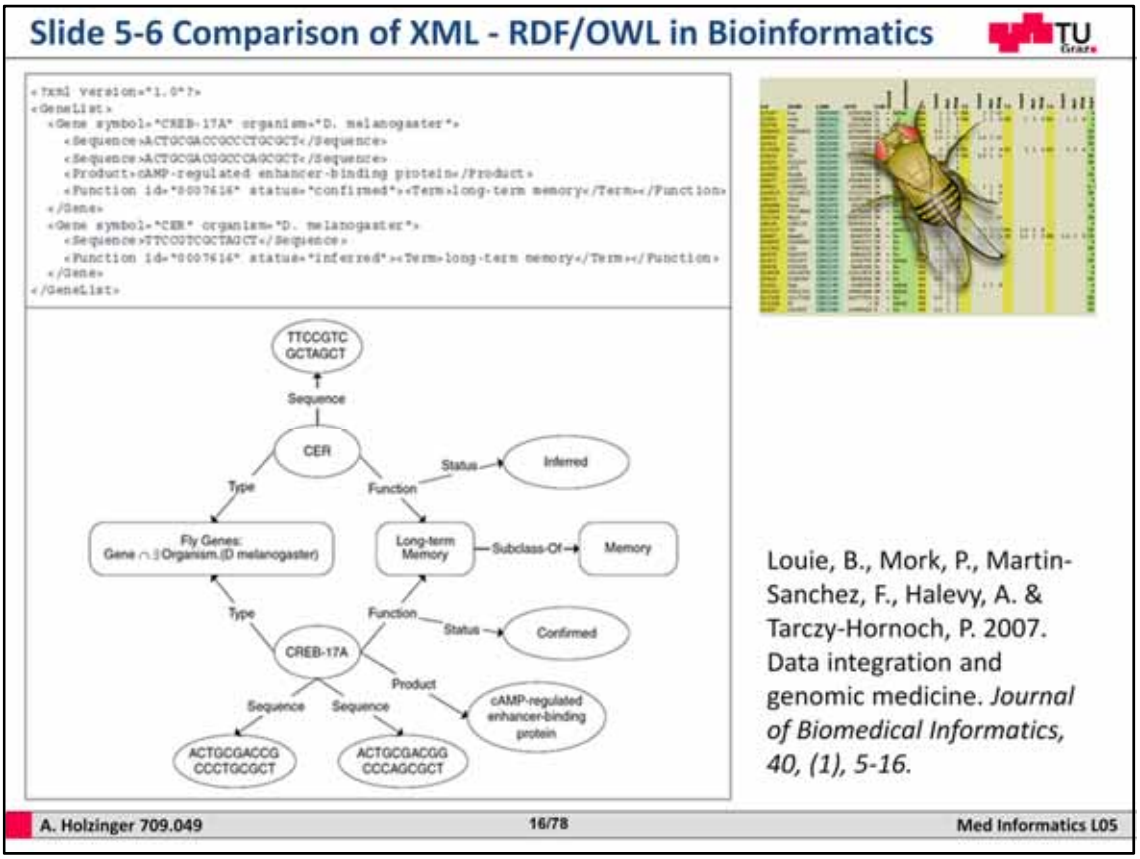

A look on the typical view of an hospital information system shows us the organization of well-structured data: Standardized and well-structured data is the basis for accurate communication. In the medical domain, many different people work at different times in various locations. Data standards can ensure that information is interpreted by all users with the same understanding. Moreover, standardized data facilitate comparability of data and interoperability of systems. It supports the reusability of the data, improves the efficiency of healthcare services and avoids errors by reducing duplicated efforts in data entry. Remember: Data standardization refers to a) the data content; b) the terminologies that are used to represent the data; c) how data is exchanged; and d) how knowledge, e.g. clinical guidelines, protocols, decision support rules, checklists, standard operating procedures are represented in the health information system.
Note: The opposite, i.e. non-standardized data is the majority of data and inhibit data quality, data exchange and interoperability.
Remark: Care2x is an Open Source Information System, see: http://care2x.org
See →Lecture 10 for more details.



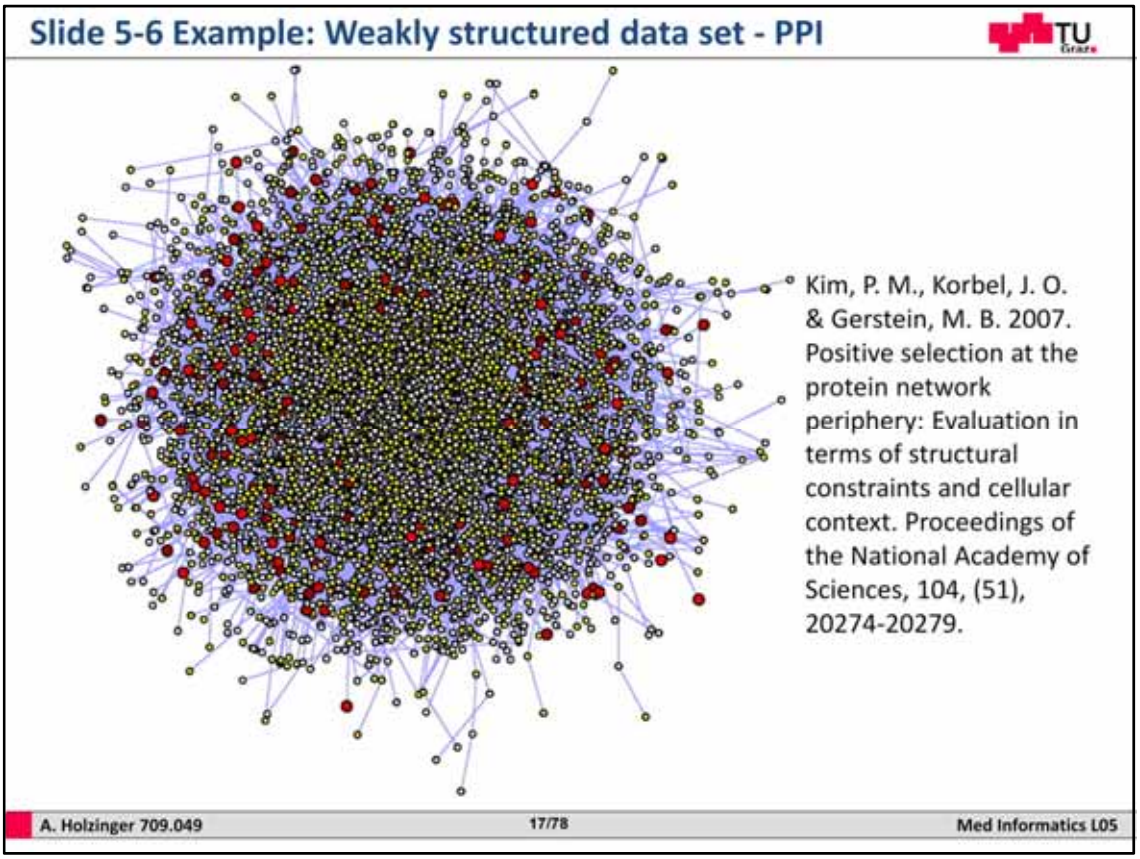

This is a Medical example for semi-structured data in XML (Holzinger, 2003). The eXtensible Markup Language (XML) is a flexible text format recommended by the W3C for data exchange and derived from SGML (ISO 8879), (Usdin & Graham, 1998).
XML is often classified as semi-structured, however this is in some way misleading, as the data itself is still structured, but in a flexible rather than a static way (Forster & Vossen, 2012). Such data does not conform to the formal structure of tables and data models as for example in relational databases, but at least contains tags/markers to separate semantic elements and enforce hierarchies of records and fields within these data.



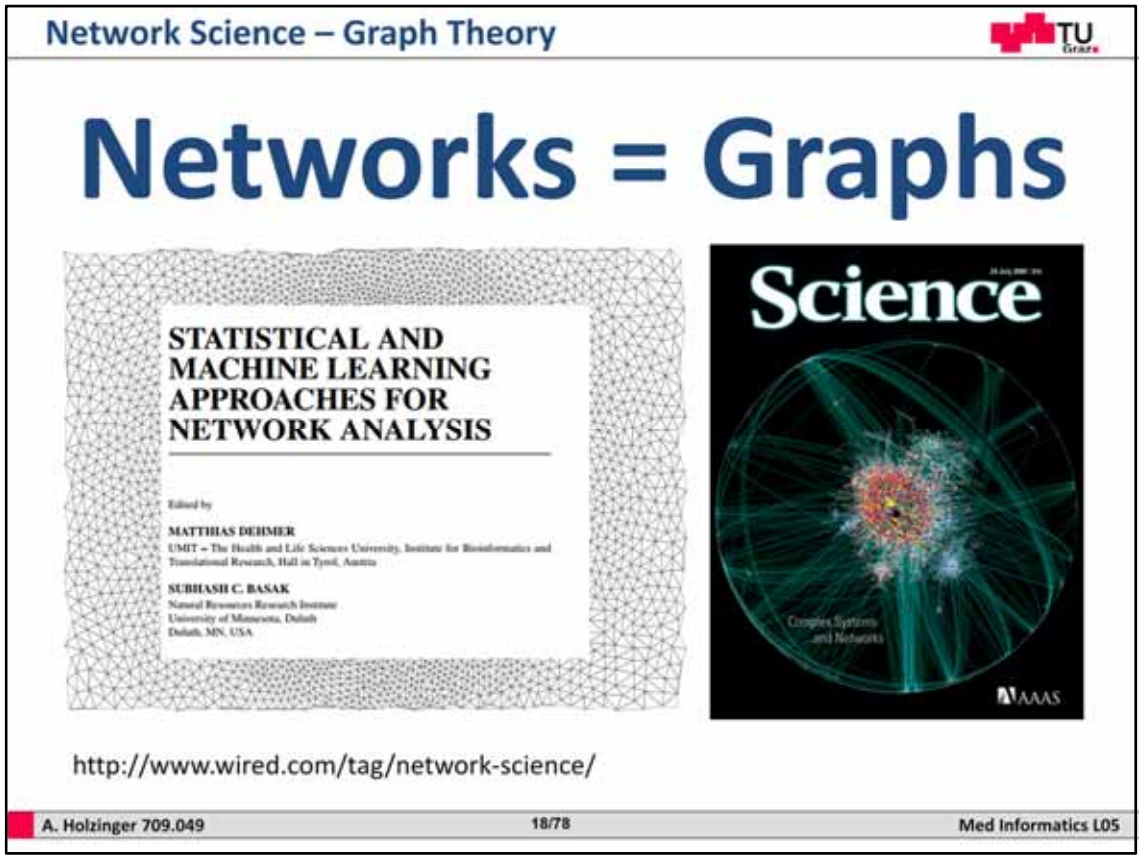

This example by (Rassinoux et al., 2003) shows how XML can be used in the hospital information system: The structure of any new document edited in the Patient Record (here: DPI) is based on a template defined in XML format (left). These templates play the role of DTDs or XML schemas as they precisely define the structure and content type of each paragraph, thus validating the document at the application level. Such a structure embeds a <HEADER> and a <BODY>. The header encapsulates the properties that are inherent to the new document and that will be useful to further classify it, according to various criteria, including: the patient identification, the document type, the identifier of its redactors and of the hospitalization stay or ambulatory consultation to which the document will be attached in the patient trajectory, etc. The body encapsulates the content, and is divided into two parts: The <STRUCDOC> part describes the semantic entities that compose the document. The <FULLDOC> part embeds the document itself with its page layout information, which can be stored either as a draft, a temporary text or as a definitive text. This format guarantees the storage of dynamic and controlled fields for data input, thus allowing the combination of free text and structured data entry in the document. Once the document is no longer editable, it is definitively saved into the RTF format. A CDATA section is utilized for storing the rough document whatever its format, as it permits to disregard blocks of text containing characters that would otherwise be regarded as markup (Rassinoux, Lovis, Baud & Geissbuhler, 2003).




On top in this slide you can see a sample XML describing genes from Drosophila melanogaster involved in long-term memory. Nested within the gene elements, are sub-elements related to the parent. The first gene includes two nucleic acid sequences, a protein product, and a functional annotation. Additional information is provided by attributes, such as the organism. This example illustrates the difficulty of modeling many-to-many relationships, such as the relationship between genes and functions. Information about functions must be repeated under each gene with that function. If we invert the nesting, then we must repeat information about genes with more than a single function. Below the XML we see the information about genes using both RDF and OWL. Both genes are instances of the class Fly Gene, which has been defined as the set of all Genes for the organism D. melanogaster. The functional information is represented using a hierarchical taxonomy, in which Long-Term Memory is a subclass of Memory (Louie et al., 2007).
Remark: Drosophila melanogaster is a model organism and shares many genes with humans. Although Drosophila is an insect whose genome has only about 14,000 genes (half of humans), a remarkable number of these have very close counterparts in humans; some even occur in the same order in the fly's DNA as in our own. This, plus the organism's more than 100-year history in the lab, makes it one of the most important models for studying basic biology and disease (see e.g. http://www.lbl.gov/Science-Articles/Archive/sabl/2007/Feb/drosophila.html)
Note: The relational data model requires preciseness: The data must be regular, complete and structured. However, in Biology the relationships are mostly un-precise. Genomic medicine is extremely data intensive and there is an increasing diversity in the type of data: DNA sequence, mutation, expression arrays, haplotype, proteomic etc. In bioinformatics many heterogeneous data sources are used to model complex biological systems (Rassinoux, Lovis, Baud & Geissbuhler, 2003), (Achard, Vaysseix & Barillot, 2001). The challenge in genomic medicine is to integrate and analyze these diverse and huge data sources to elucidate physiology and in particular disease physiology. XML is suited for describing semi-structured data, including a kind of natural modeling of biological entities, because it allows features as e.g. nesting (see Slide 5-6 on top). Still a key limitation of XML is, that it is difficult to model complex relationships; for example, there is no obvious way to represent many-to-many relationships, which are needed to model complex pathways. On top in Figure 5-9 we can see a sample XML, describing genes involved in the long-term memory of a sample specimen d. melanogaster. Nested within the gene elements, are sub-elements related to the parent. The first gene includes two nucleic acid sequences, a protein product, and a functional annotation. Additional information is provided by attributes, such as the organism. This example illustrates the difficulty of modeling many-to-many relationships, such as the relationship between genes and functions. Information about functions must be repeated under each gene with that function. If we invert the nesting (i.e., nesting genes inside function elements), then we must repeat information about genes with more than a single function. At the bottom in Slide 5-6 we see the same information about genes, but using RDF and OWL. Both genes are instances of the class Fly Gene, which has been defined as the set of all Genes for the organism D. melanogaster. The functional information is represented using a hierarchical taxonomy, in which Long-Term Memory is a subclass of Memory (Louie et al., 2007).

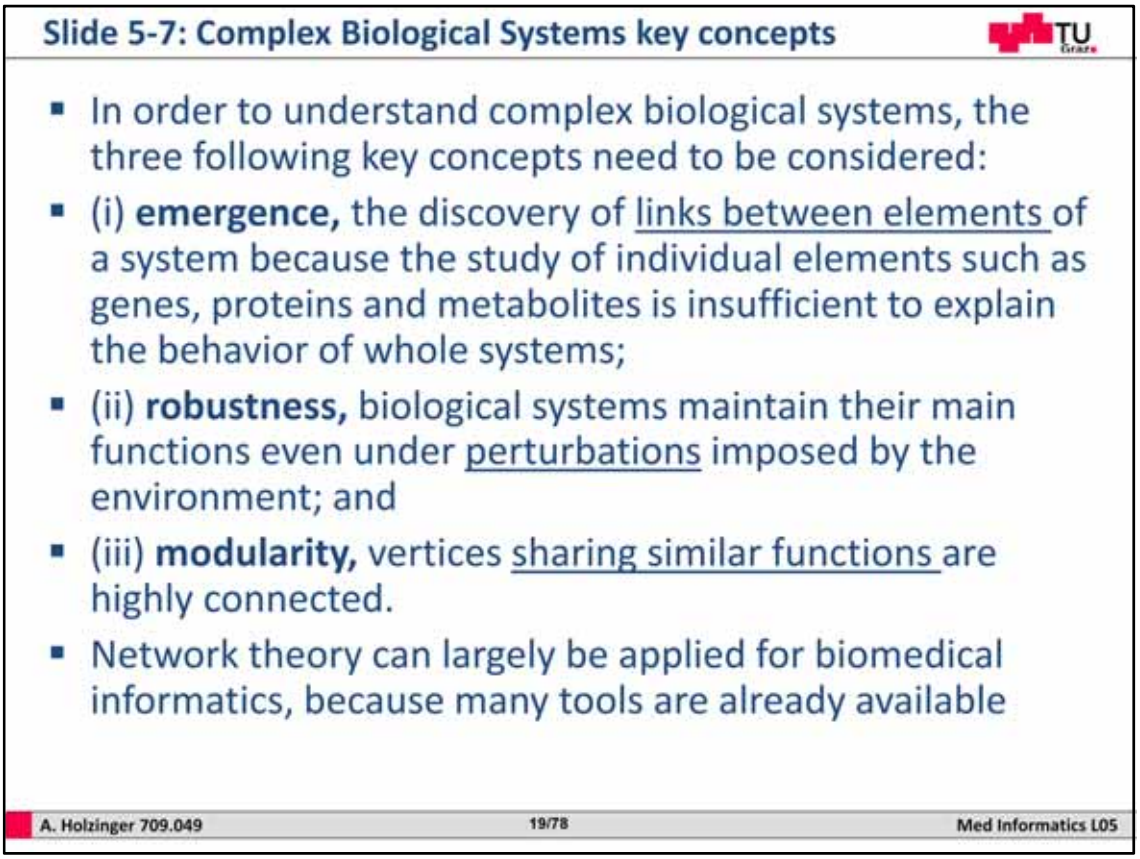



The human protein interaction network and its connection to positive selection. Proteins likely to be under positive selection are colored in shades of red (light red, low likelihood of positive selection; dark red, high likelihood) (6). Proteins estimated not to be under positive selection are in yellow, and proteins for which the likelihood of positive selection was not estimated are in white (6).




http://www.barabasilab.com/pubs/CCNR-ALB_Publications/200907-24_Science-Decade/200907-24_Science-CoverImage.gif

An emerging trend in many scientific disciplines is a strong tendency toward being transformed into some form of information science. One important pathway in this transition has been via the application of network analysis. The basic methodology in this area is the representation of the structure of an object of investigation by a graph representing a relational structure. It is because of this general nature that graphs have been used in many diverse branches of science including bioinformatics, molecular and systems biology, theoretical physics, computer science, chemistry, engineering, drug discovery, and linguistics, to name just a few. An important feature of the book "Statistical and Machine Learning Approaches for Network Analysis" is to combine theoretical disciplines such as graph theory, machine learning, and statistical data analysis and, hence, to arrive at a new field to explore complex networks by using machine learning techniques in an interdisciplinary manner. The age of network science has definitely arrived. Large-scale generation of genomic, proteomic, signaling, and metabolomic data is allowing the construction of complex networks that provide a new framework for understanding the molecular basis of physiological and pathological states. Networks and network-based methods have been used in biology to characterize genomic and genetic mechanisms as well as protein signaling. Diseases are looked upon as abnormal perturbations of criticalcellular networks. Onset, progression, and intervention in complex diseases such as
cancer and diabetes are analyzed today using network theory. Once the system is represented by a network, methods of network analysis can be applied to extract useful information regarding important system properties and toinvestigate its structure and function. Various statistical and machine learning methods
have been developed for this purpose and have already been applied to networks.
Dehmer, M. & Basak, S. C. 2012. Statistical and Machine Learning Approaches for Network Analysis, Wiley Online Library.




The concept of network structures is fascinating, compelling and powerful and applicable in nearly any domain at any scale.
Network theory can be traced back to graph theory, developed by Leonhard Euler in 1736 (see →Slide 5-8). However, stimulated by works e.g. from Barabási, Albert & Jeong (1999), research on complex networks has only recently been applied to biomedical informatics. As an extension of classical graph theory, see for example (Diestel, 2010), 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, protein-protein networks, social relationships and the Web and many more. Attention is given not only to try to identify special patterns of connectivity, such as the shortest average path between pairs of nodes (Newman, 2003), but also to consider the evolution of connectivity and the growth of networks, an example from biology being the evolution of protein-protein interaction networks in different species (→Slide 5-8). In order to understand complex biological systems, the three following key concepts need to be considered:
(i) emergence, the discovery of links between elements of a system because the study of individual elements such as genes, proteins and metabolites is insufficient to explain the behavior of whole systems;
(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. Network theory can largely be applied for biomedical informatics, because many tools are already available (Costa, Rodrigues & Cristino, 2008).

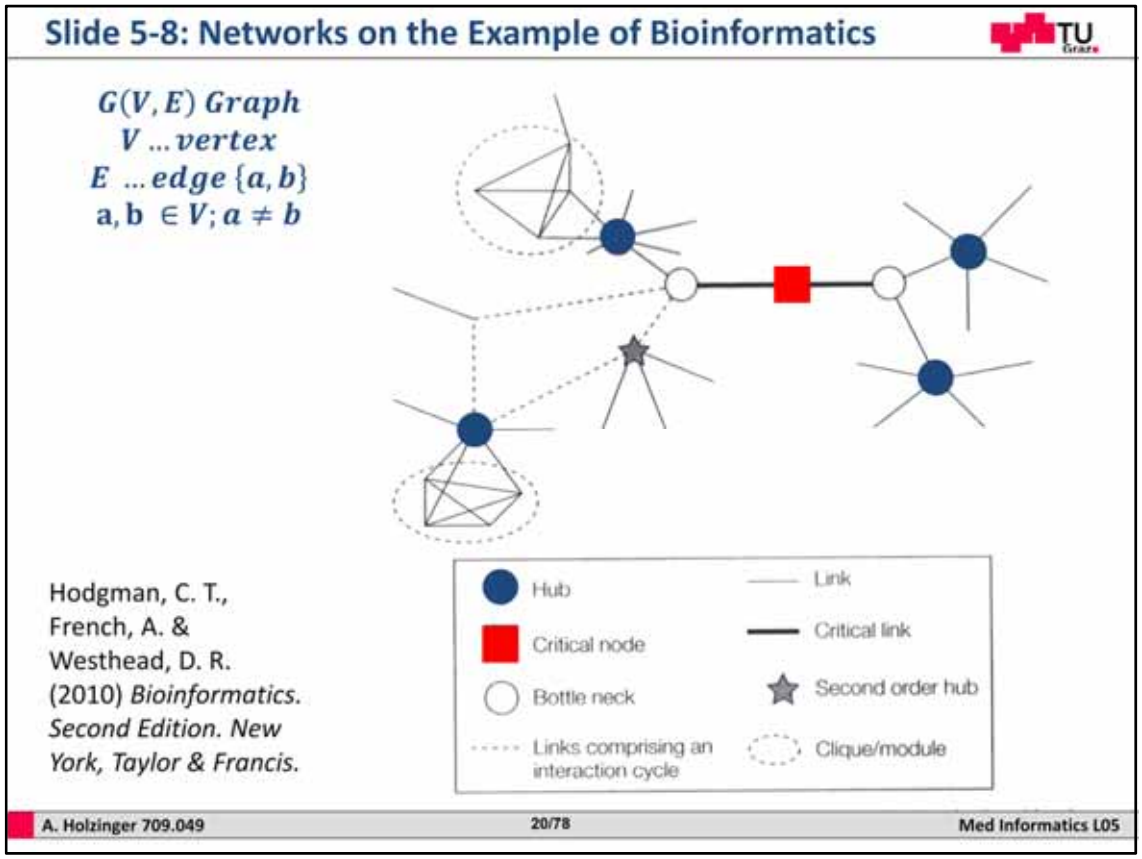
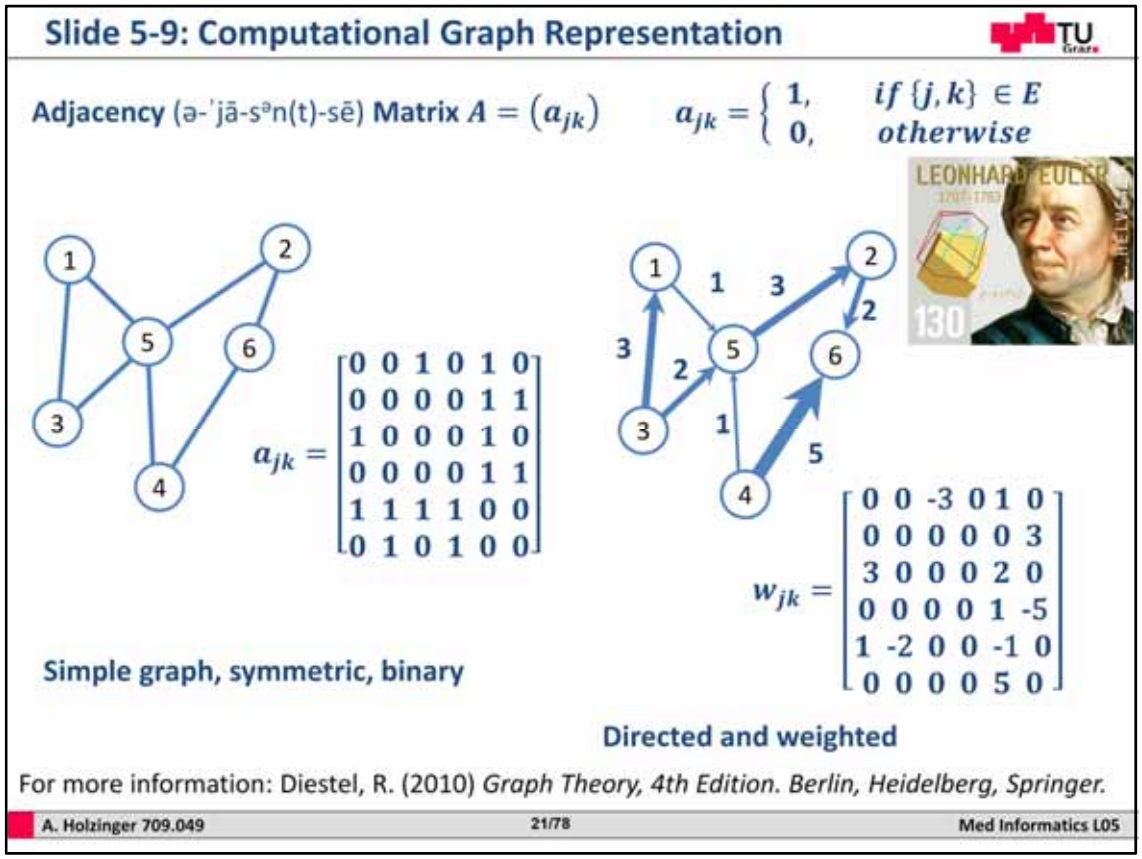
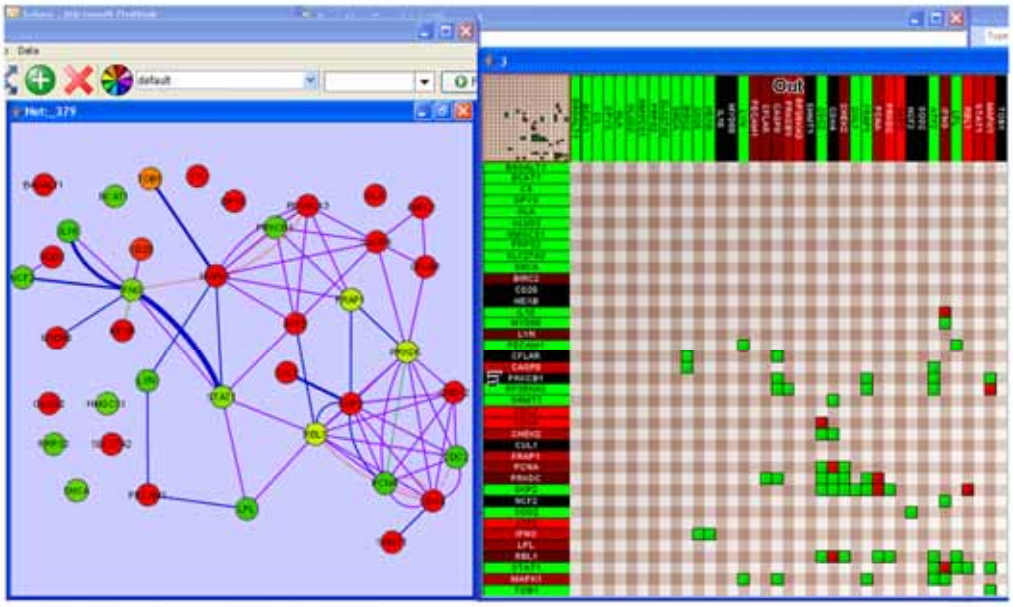



Figure 1 p.74 - preceding to the yeast protein network
A graph G(V,E) describes a structure which consists of nodes aka vertices V, connected by a set of pairs of distinct nodes (links), called edges E {a,b} with a,b∈V;a≠b.
Graphs containing cycles and/or alternative paths are referred to as networks. The vertexes and edges can have a range of properties defined as colors, which also may have quantitative values, referred to as weights. In this Slide we see the basic building block symbols of a biological network as used in bioinformatics. The blue dots are serving as network hubs, the red block is a critical node (on a critical link), the white balls are bottle necks, the stars second order hubs etc. (Hodgman, French & Westhead, 2010).




In order to represent network data in computers it is not comfortable to use sets; more practical are matrices. The simplest form of a graph representation is the so called adjacency matrix. In this Slide we see an undirected (left) and a directed graph and their respective adjacency matrices. If the graph is undirected, the adjacency matrix is symmetric, i.e., the elements aij = aji for any i and j.

## Slide 5-10: Example: Tool for Node-Link Visualization



Jean-Daniel Fekete http://wiki.cytoscape.org/InfoVis_Toolkit

Fekete, J.-D. The infovis toolkit. Information Visualization, INFOVIS 2004, 2004. IEEE, 167-174.

A. Holzinger 709.049 22/78 Med Informatics L05

This Tool is a nice example on the usefulness of adjacency matrices: The InfoVis Toolkit is an interactive graphics toolkit developed by Jean-Daniel Fekete at INRIA (The French National Institute for Computer Science and Control). The toolkit implements nine types of visualization: Scatter Plots, Time Series, Parallel Coordinates and Matrices for tables; Node-Link diagrams, Icicle trees and Tree maps for trees; Adjacency Matrices and Node-Link diagrams for graphs. Node-Link visualizations provides several variants (8 for graphs and 4 for trees). There are also a number of interactive controls and information displays, including dynamic query sliders, fisheye lenses, and excentric labels. Information about the InfoVis toolkit can be found at http://ivtk.sourceforge.net

The InfoVis Toolkit provides interactive components such as range sliders and tailored control panels required to configure the visualizations. These components are integrated into a coherent framework that simplifies the management of rich data structures and the design and extension of visualizations. Supported data structures include tables, trees and graphs. All visualizations can use fisheye lenses and dynamic labeling (Fekete, 2004).

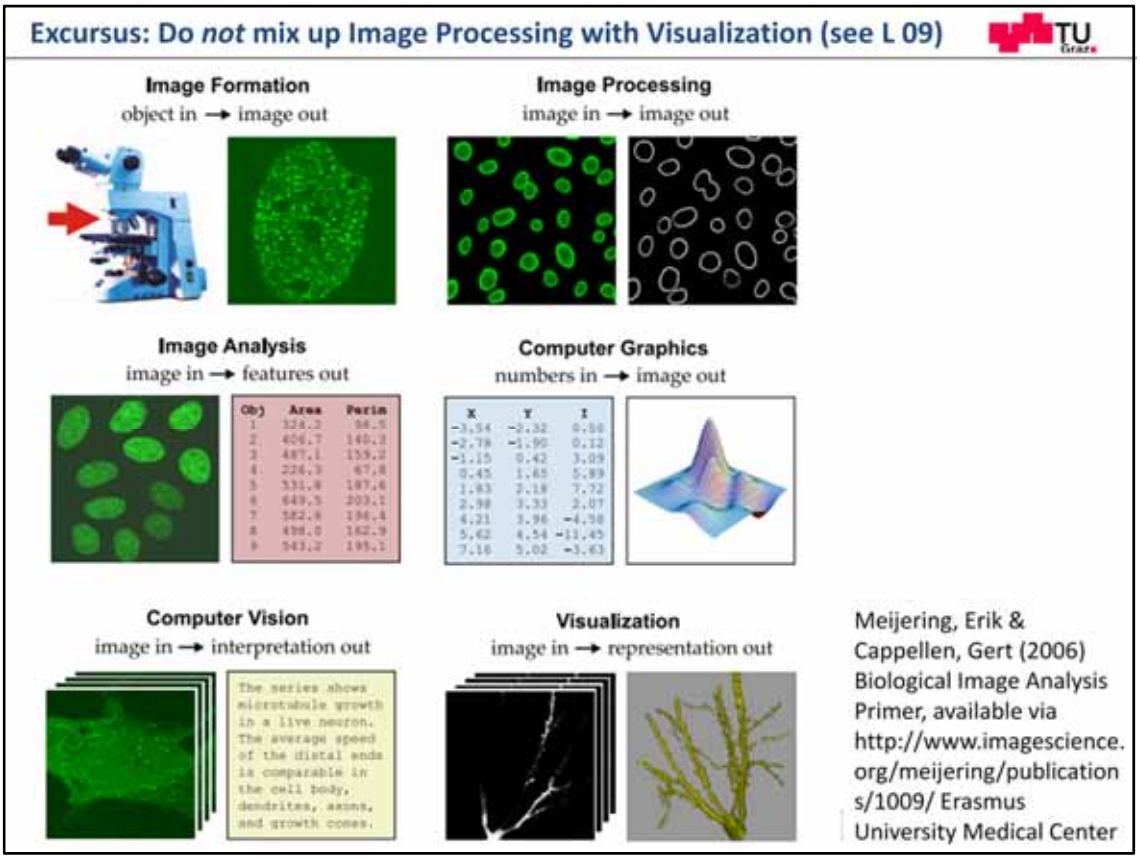
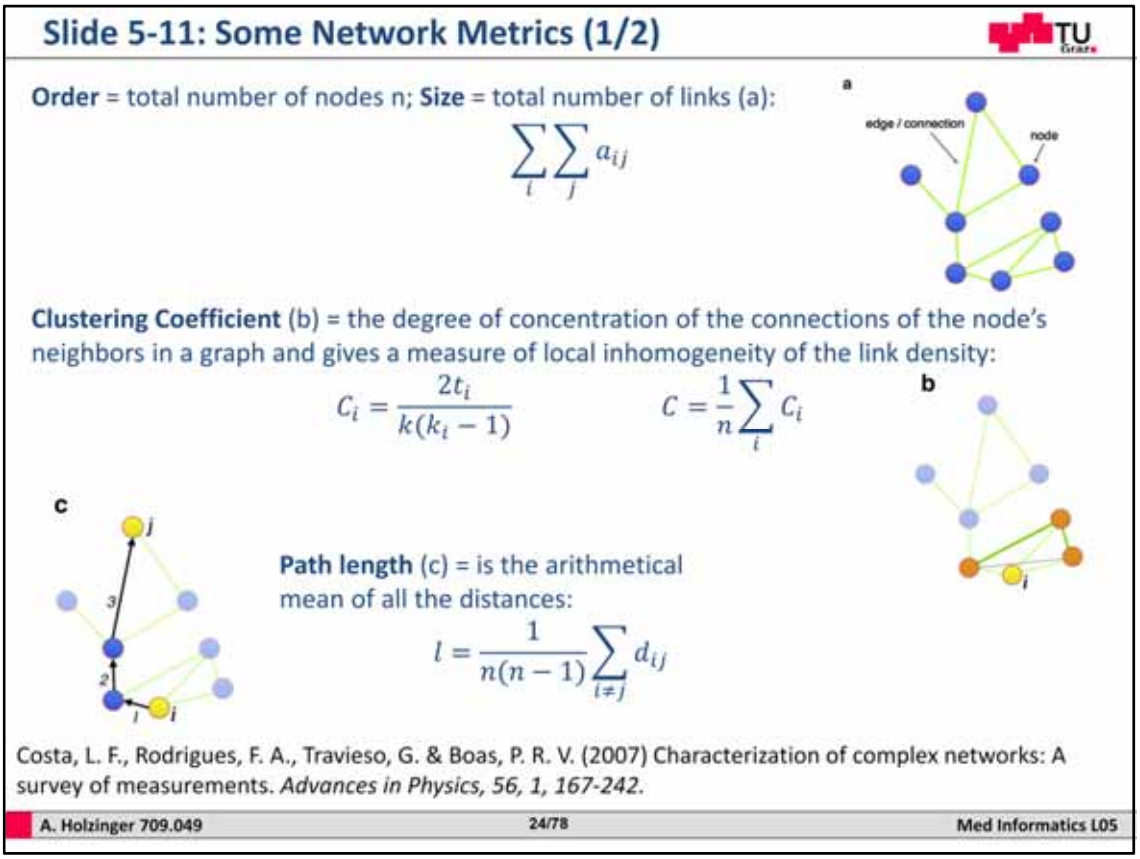
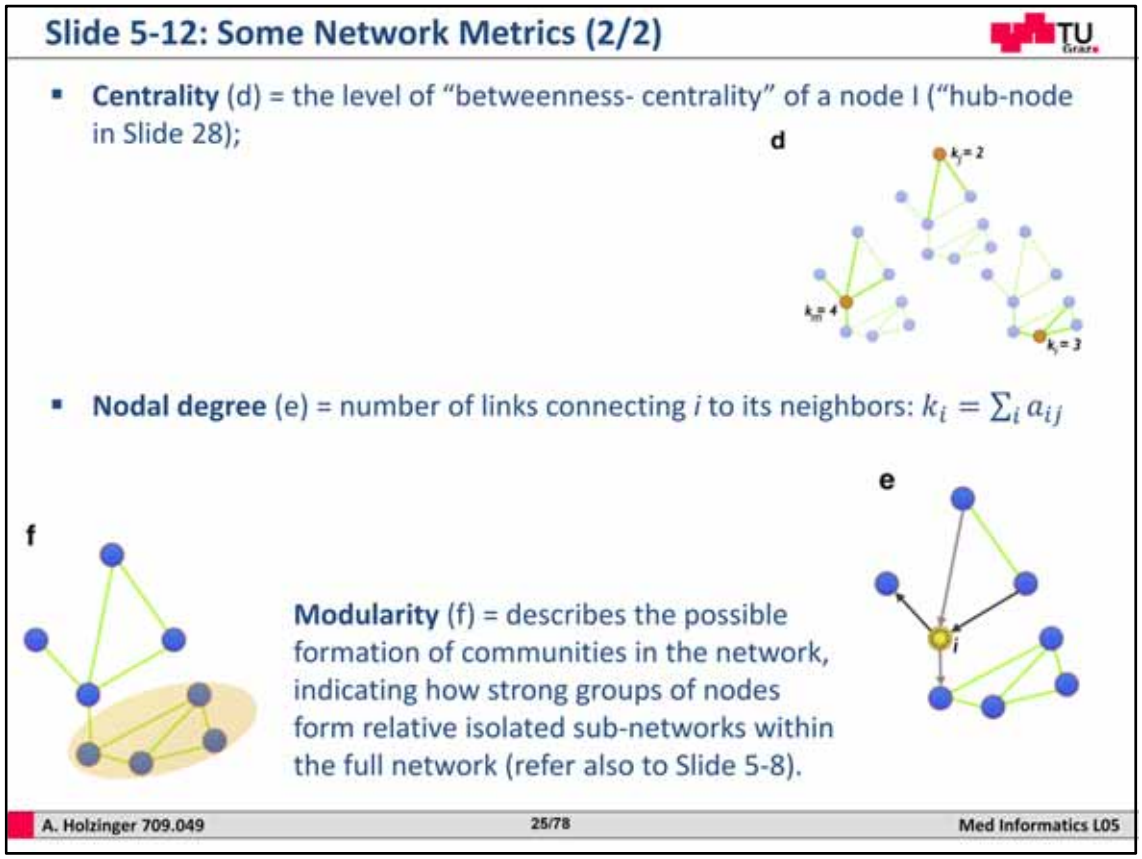


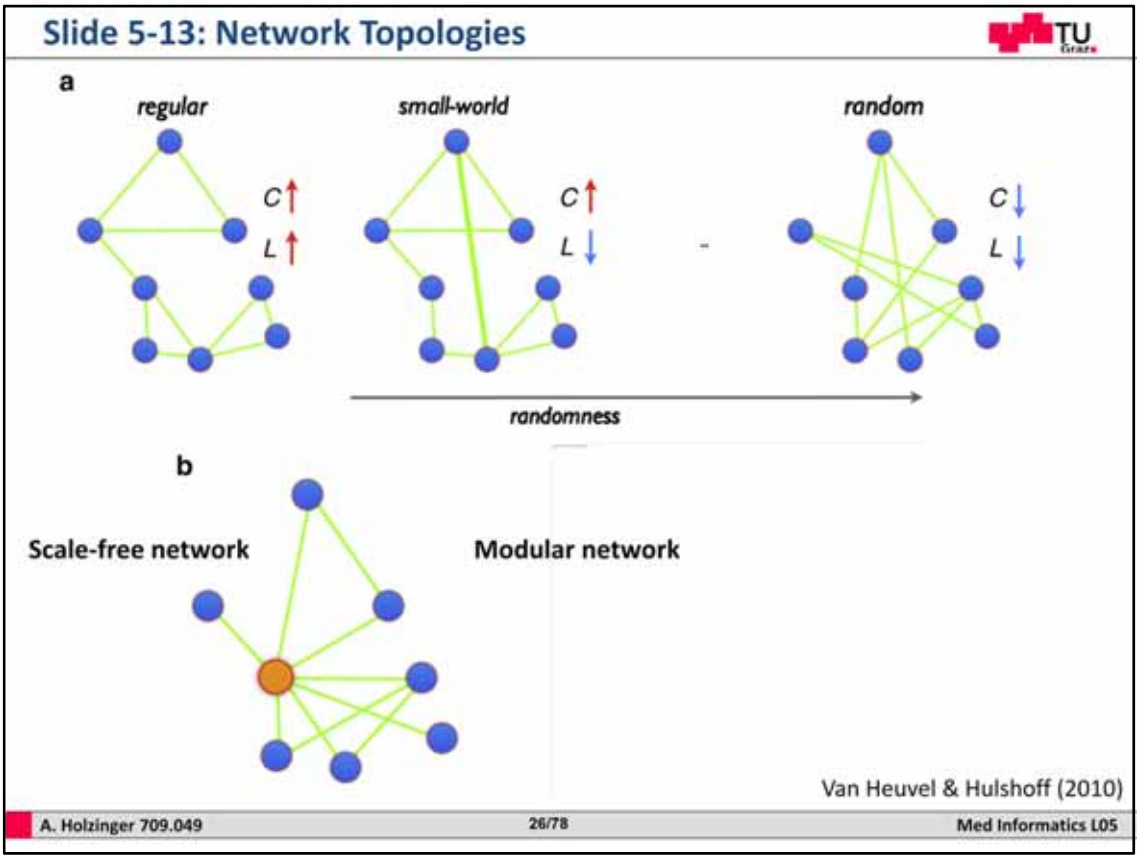

Illustration of the meaning of commonly used terms. The process of digital image formation in microscopy is described in other books. Image processing takes an image as input
and produces a modified version of it (in the case shown, the object contours are enhanced using
an operation known as edge detection, described in more detail elsewhere in this booklet). Image
analysis concerns the extraction of object features from an image. In some sense, computer graphics is the inverse of image analysis: it produces an image from given primitives, which could be
numbers (the case shown), or parameterized shapes, or mathematical functions. Computer vision
aims at producing a high-level interpretation of what is contained in an image. This is also known
as image understanding. Finally, the aim of visualization is to transform higher-dimensional image data into a more primitive representation to facilitate exploring the dat




The truly multi-disciplinary network science has led to a wide variety of quantitative measurements of their topological characteristics (Costa et al., 2007). The identification between a graph and an adjacency matrix makes all the powerful methods of linear algebra, graph theory and statistical mechanics available to us for investigating specific network characteristics :
Order (a in Figure Slide 5-11) = total number of nodes n
Size = total number of links:
∑_i∑_j a_ij
Clustering Coefficient (b in Slide 5-11) = the degree of concentration of the connections of the node's neighbors in a graph and gives a measure of local inhomogeneity of the link density, i.e. the level of connectedness of the graph. It is calculated as the ratio between the actual number ti of links connecting the neighborhood (the nodes immediately connected to a chosen node) of a node and the maximum possible number of links in that neighborhood:
C_i=(2t_i)/(k(k_i-1))
For the whole network, the clustering coefficient is the arithmetic mean:
C=1/n ∑_i C_i
Path length (c in Slide 5-11) = is the arithmetical mean of all the distances; The characteristic path length of node i provides information about how close node i is connected to all other nodes in the network and is given by the distance d(i,j) between node i and all other nodes j in the network. The Path length l provides important information about the level of global communication efficiency of a network:
l=1/(n(n-1)) ∑_(i≠j) d_ij
Note: Numerical methods, e.g. the Dijkstra's algorithm (1959) are used to calculate all the possible paths between any two nodes in a network.

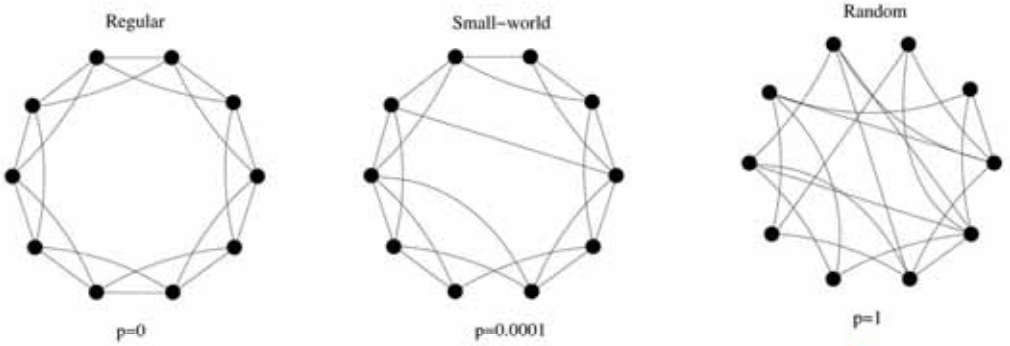



Centrality (d in Slide 5-12) = the level of "betweenness- centrality" of a node i; it indicates how many of the shortest paths between the nodes of the network pass through node i. A high "betweenness-centrality" indicates that this node is important in interconnecting the nodes of the network, marking a potential hub role (refer to →Slide 5-8) of this node in the overall network.
Nodal degree (e in Slide 5-12) = number of links connecting i to its neighbors. The degree of node i is defined as its total number of connections.
k_i=∑_i▒a_ij
The degree probability distribution P(k) describes the p(x) that a node is connected to k other nodes in the network.
Modularity (f in Slide 5-12) = describes the possible formation of communities in the network, indicating how strong groups of nodes form relative isolated sub-networks within the full network (refer also to →Slide 5-8)) .




Regular network (a in Slide 5-13) has a local character, characterized by a high clustering-coefficient (c in Slide 5-13) and a high path length (L, Slide 5-13). It takes a large number of steps to travel from a specific node to a node on the other end of the graph. A special case of a regular network is the:
Random network, where all connections are distributed randomly across the network; the result is a graph with a random organization (outer right in Slide 5-13). In contrast to the local character of the regular network, a random network has a more global character, with a low C and a much shorter path length L than the regular network. A particular case is the:
Small-world network (center of Slide 5-13) which are very robust and combine a high level of local and global efficiency. Watts & Strogatz (1998) showed that with a low probability p of randomly reconnecting a connection in the regular network, a so-called small-world organization arises. It has both a high C and a low L, combining a high level of local clustering with still a short average travel distance. Many networks in nature are small-world (e.g. internet, protein-networks, social networks, functional and structural brain network etc.), combining a high level of segregation with a high level of global information integration. In addition, such networks can have a heavy tailed connectivity distribution, in contrast to random networks in which the nodes roughly all have the same number of connections.
Scale-free networks (B in Slide 5-13) are characterized by a degree probability distribution that follows a power-law function, indicating that on average a node has only a few connections, but with the exception of a small number of nodes that are heavily connected. These nodes are often referred to as hub nodes (see →Slide 5-8) and they play a central role in the level of efficiency of the network, as they are responsible for keeping the overall travel distance in the network to a minimum. As these hub nodes play a key role in the organization of the network, scale-free networks tend to be vulnerable to specialized attack on the hub nodes.
Modular networks (c in Slide 5-13) show the formation of so-called communities, consisting of a subset of nodes that are mostly connected to their direct neighbors in their community and to a lesser extend to the other nodes in the network. Such networks are characterized by a high level of modularity of the nodes.

## Slide 5-14: Small-World Networks



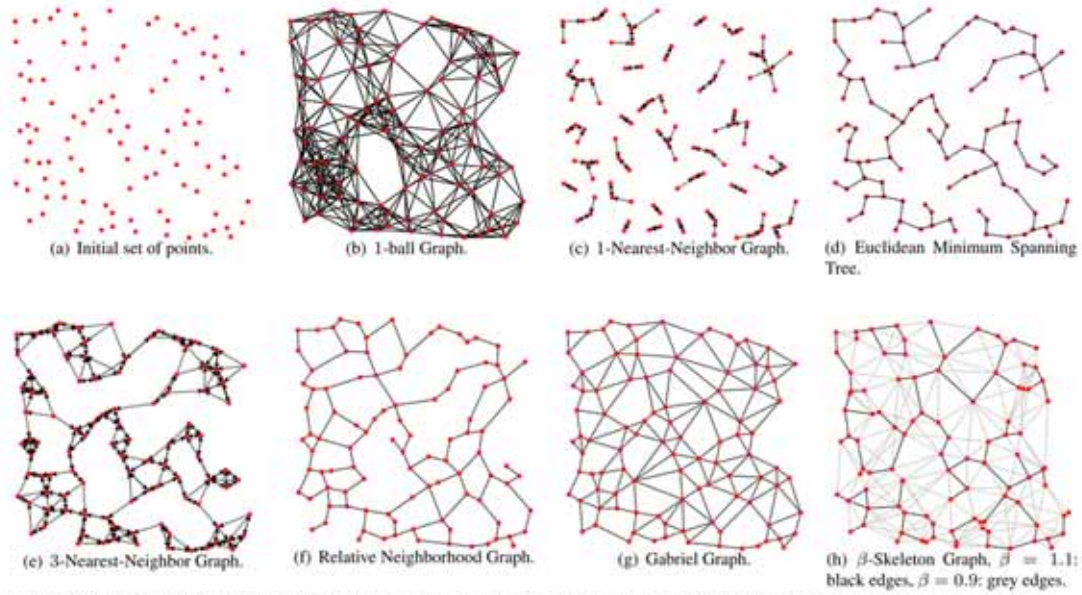

Increasing randomness

21.000 citations ...

Watts, D. J. & Strogatz, S. (1998) Collective dynamics of small-world networks. *Nature, 393, 6684, 440-442.*

Milgram, S. 1967. The small world problem. *Psychology today, 2, (1), 60-67.*

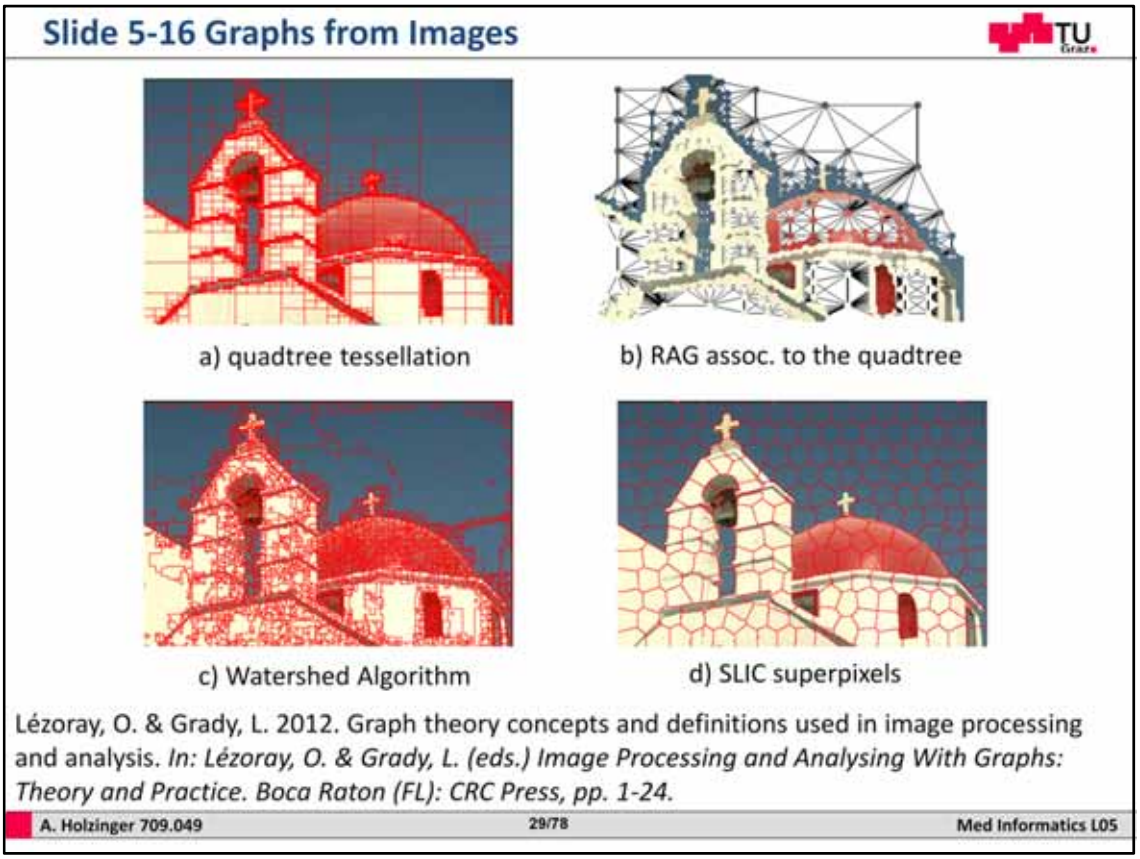
## Slide 5-15 Graphs from Point Cloud Data Sets




Lézoray, O. & Grady, L. 2012. Graph theory concepts and definitions used in image processing and analysis. *In: Lézoray, O. & Grady, L. (eds.) Image Processing and Analysing With Graphs: Theory and Practice. Boca Raton (FL): CRC Press, pp. 1-24.*

A. Holzinger 709.049 28/78 Med Informatics L05

There are many ways to construct a proximity graph representation from a set of data points that are embedded in R^d.
Let us consider a set of data points {x_1,...,x_n} ∈ R^d .
To each data point we associate a vertex of a proximity graph G to define a set of vertices V = {v1,v2,...,vn}. Determining the edge set E of the proximity graph G requires defining the neighbors of each vertex vi according to its embedding xi.
Consequently, a proximity graph is a graph in which two vertices are connected by an edge iff the data points associated to the vertices satisfy particular geometric requirements. Such particular geometric requirements are usually based on a metric measuring the distance between two data points. A usual choice of metric is the Euclidean metric. Look at the slide:
a) is our initial set of points in the plane R^2
b) ε-ball graph vi ~ vj if xj ∈ B(vi; ε)
c) k-nearest-neighbor graph (k-NNG): vi ~ vj if the distance between xi and xj is among the k-th smallest distances from xi to other data points. The k-NNG is a directed graph since one can have xi among the k-nearest neighbors of xj but not vice versa.
d) Euclidean Minimum Spanning Tree (EMST) graph is a connected tree sub-graph that contains all the vertices and has a minimum sum of edge weights. The weight of the edge between two vertices is the Euclidean distance between the corresponding data points.
e) Symmetric k-nearest-neighbor graph (Sk-NNG): vi ~ vj if xi is among the k-nearest neighbors of y or vice versa.
f) Mutual k-nearest-neighbor graph (Mk-NNG): vi ~ vj if xi is among the k-nearest neighbors of y and vice versa. All vertices in a mutual k-NN graph have a degree upper-bounded by k, which is not usually the case with standard k-NN graphs.
g) Relative Neighborhood Graph (RNG): vi ~ vj iff there is no vertex in
B(vi ; D(vi,vj)) ∩B(vj ; D(vi,vj)) .
h) Gabriel Graph (GG)
i) The β-Skeleton Graph (β-SG):
For details please refer to (Lézoray & Grady, 2012), or to a classical graph theory book, e.g. (Harary, 1969), (Bondy & Murty, 1976), (Golumbic, 2004), (Diestel, 2010)




Slide 5-16: Graphs from Images
In this slide we see the examples of
a) a real image with the quadtree tessellation,
b) the region adjacency graph associated to the quadtree partition,
c) Irregular tessellation using image-dependent superpixel Watershed Segmentation (Vincent & Soille, 1991)
d) irregular tessellation using image-dependent SLIC superpixels (Lucchi et al., 2010)
SLIC = Simple Linear Iterative Clustering)

## Slide 5-18 Example Watershed Algorithm

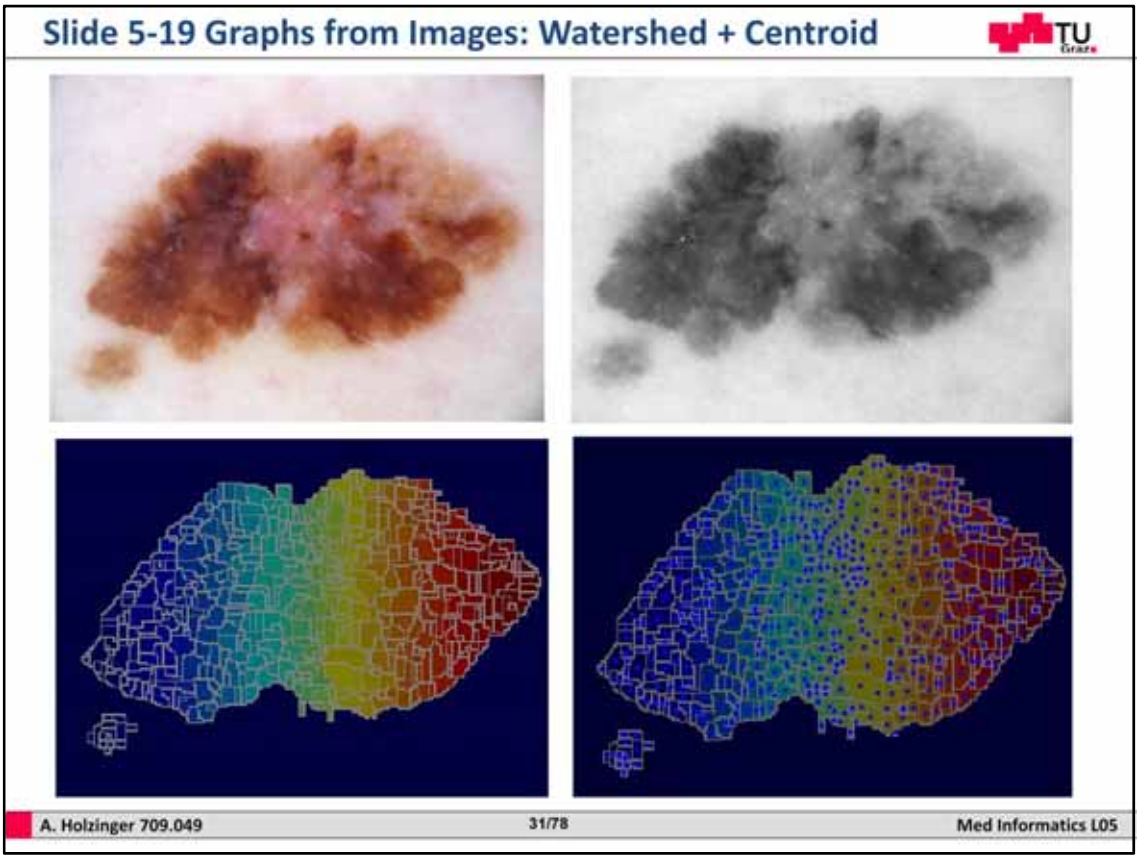
**Algorithm 4.2** Watershed transform w.r.t. topographical distance based on image integration via the Dijkstra-Moore shortest paths algorithm.

```
1: procedure ShortestPathWatershed;
2: INPUT: lower complete digital grey scale image G = (V,E,im) with cost function cost.
3: OUTPUT: labelled image lab on V.
4: #define WSHED 0                    (* label of the watershed pixels *)
5: (* Uses distance image dist. On output, dist[v] = im[v], for all v ∈ V. *)
6:
7: for all v ∈ V do                   (* Initialize *)
8:    lab[v] ← 0 ; dist[v] ← ∞
9: end for
10: for all local minima m_i do
11:    for all v ∈ m_i do
12:       lab[v] ← i ; dist[v] ← im[v]      (* initialize distance with values of minima *)
13:    end for
14: end for
15: while V ≠ ∅ do
16:    u ← GetMinDist(V)              (* find u ∈ V with smallest distance value dist[u] *)
17:    V ← V\{u}
18:    for all v ∈ V with (u,v) ∈ E do
19:       if dist[u] + cost[u,v] < dist[v] then
20:          dist[v] ← dist[u] + cost(u,v)
21:          lab[v] ← lab[u]
22:       else if lab[v] ≠ WSHED and dist[u] + cost[u,v] = dist[v] and lab[v] ≠ lab[u] then
23:          lab[v] = WSHED
24:       end if
25:    end for
26: end while
```

Meijster, A. & Roerdink, J. B. A proposal for the implementation of a parallel watershed algorithm. Computer Analysis of Images and Patterns, 1995. Springer, 790-795.

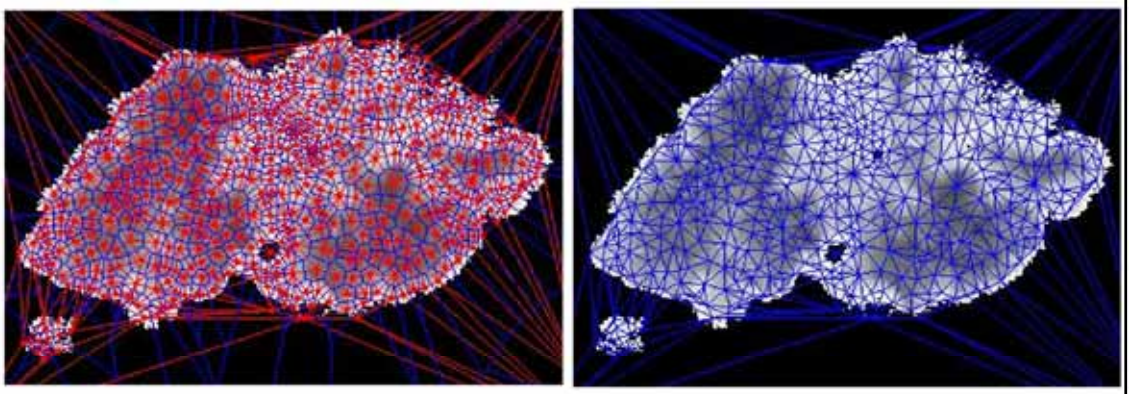
A. Holzinger 709.049 30/78 Med Informatics L05

A straightforward implementation of the original Vincent-Soille algorithm is difficult if plateaus occur. Therefore, an alternative approach was proposed by (Meijster & Roerdink, 1995), in which the image is first transformed to a directed valued graph with distinct neighbor values, called the components graph of f. On this graph the watershed transform can be computed by a simplied version of the Vincent-Soille algorithm, where fifo queues are no longer necessary, since there are no plateaus in the graph (Roerdink & Meijster, 2000).




The original natural digital image is first transformed into grey-scale, then the Watershed algorithm is applied and then the centroid function calculated, the results are representative point sets in the plane.

## Slide 5-20 Graphs from Images: Voronoi <> Delaunay



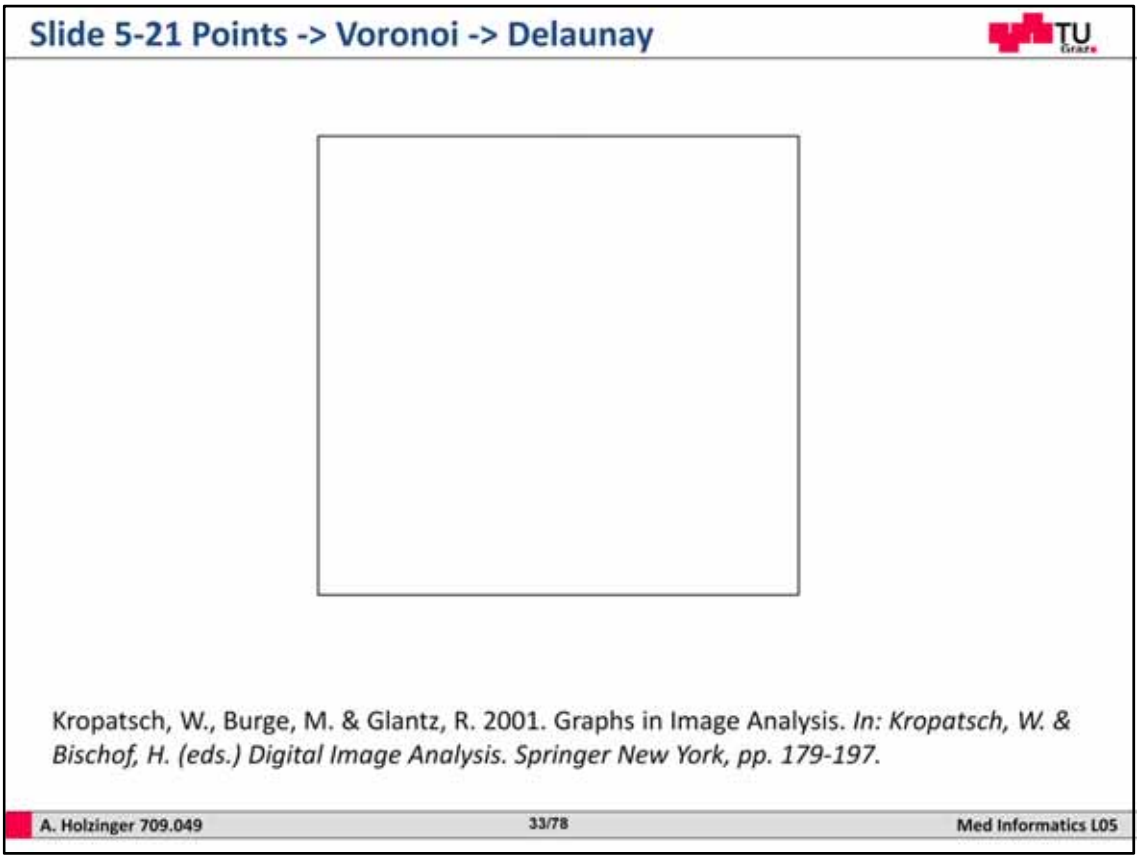
Holzinger, A., Malle, B. & Giuliani, N. 2014. On Graph Extraction from Image Data. In: Slezak, D., Peters, J. F., Tan, A.-H. & Schwabe, L. (eds.) Brain Informatics and Health, BIH 2014, Lecture Notes in Artificial Intelligence, LNAI 8609. Heidelberg, Berlin: Springer, pp. 552-563.

For Voronoi please refer to: Aurenhammer, F. 1991. Voronoi Diagrams - A Survey of a fundamental geometric data structure. *Computing Surveys,* 23, (3), 345-405.

For Delaunay please refer to: Lee, D.-T. & Schachter, B. J. 1980. Two algorithms for constructing a Delaunay triangulation. Intl. Journal of Computer & Information Sciences, 9, (3), 219-242.

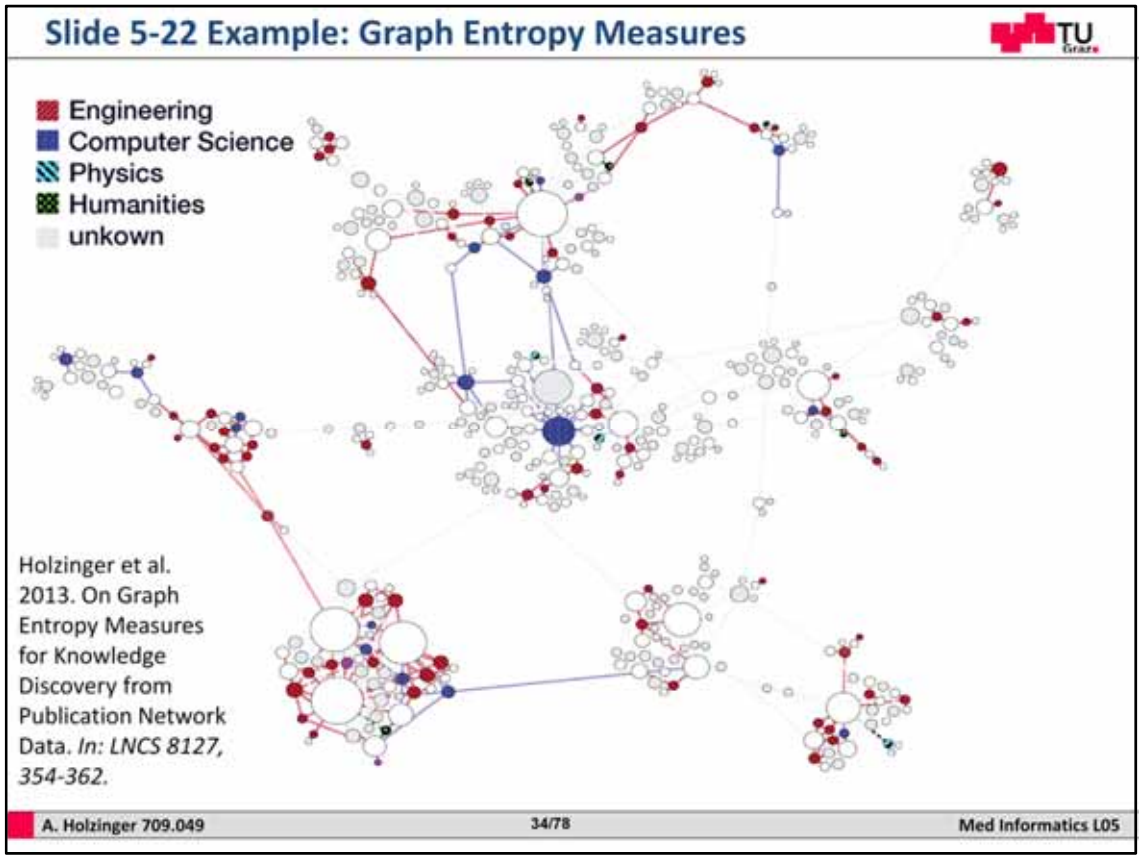
A. Holzinger 709.049 32/78 Med Informatics L05

The Delaunay Triangulation (DT): vi ~ vj iff there is a closed ball B(• ; r) with vi and vj on its boundary and no other vertex vk contained in it. The dual to the DT is the Voronoi irregular tessellation where each Voronoi cell is defined by the set {x ∈ Rn | D(x,vk) ≤ D(x,vj) for all vj = vk}. In such a graph, ∀ vi,deg (vi)=3. (Lézoray & Grady, 2012)

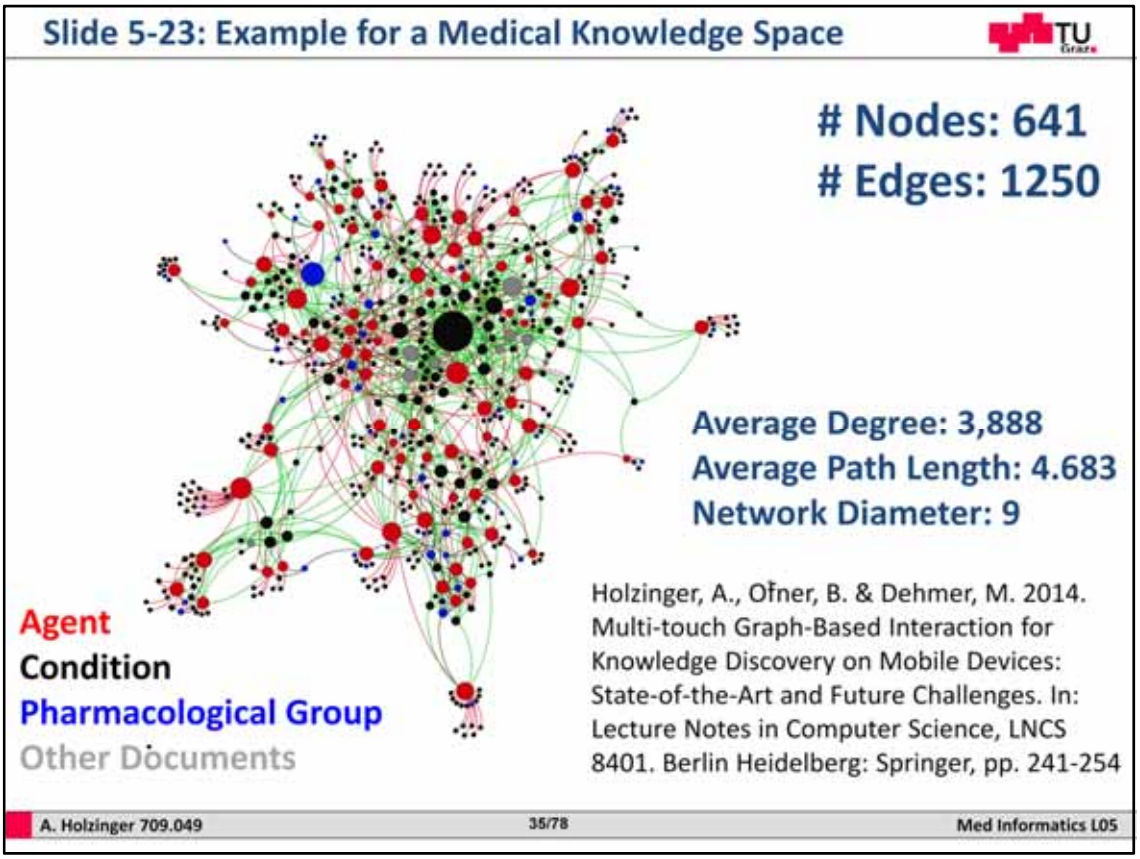



This animation shows the construction of a Delaunay graph: First the red points on the plane are drawn, then we insert the blue edges and the blue vertices on the Voronoi graph, finally he red edges drawn build the Delaunay graph (Kropatsch, Burge & Glantz, 2001).

http://oldwww.prip.tuwien.ac.at/research/research-areas/structure-and-topology/graphs-in-image-analysis/graphs-in-image-analysis/use-of-graphs-in-image-analysis/voronoi-graph-and-delaunay-graph




In this Slide we see the evaluated information-theoretic network measures on publication networks. Here from the excellence network of RWTH Aachen University. Those measures can be understood as graph complexity measures which evaluate the structural complexity based on the corresponding concept. A possible useful interpretation of these measures helps to understand the differences in subgraphs of a cluster. For example one could apply community detection algorithms and compare entropy measures of such detected communities. Relating these data to social measures (e.g. balanced score card data) of sub-communities could be used as indicators of collaboration success or lack thereof. The node size shows the node degree whereas the node color shows the betweenness centrality, darker color means higher centrality (Holzinger et al., 2013a).

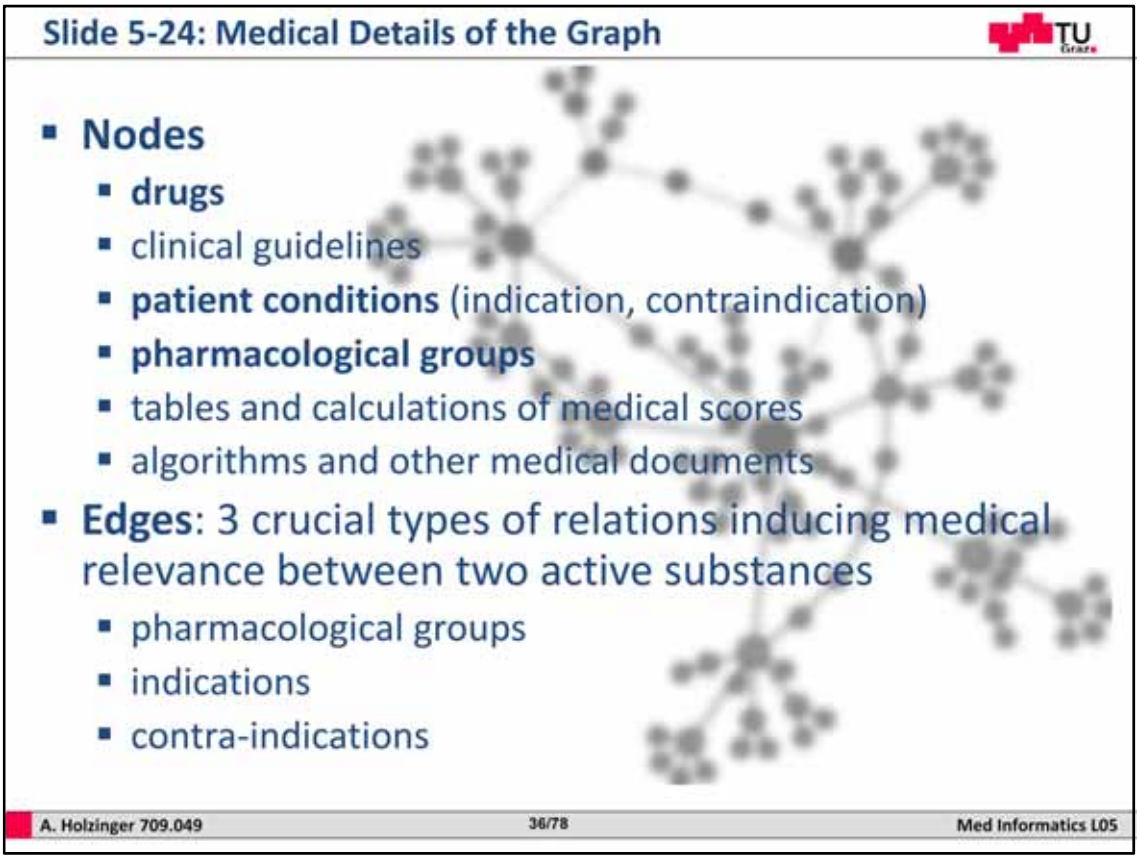



A further example shall demonstrate the usefulness of graph theory and network analysis: This graph shows the medical knowledge space of a standard quick reference guide for emergency doctors and paramedics in the German speaking area. It has been subsequently developed, tested in the medical real world and constantly improved for 20 years by Dr. med. Ralf Müller, emergency doctor at Graz-LKH University Hospital and is practically in the pocket of every emergency and family doctor and paramedics in the German speaking area (Holzinger et al., 2013b).

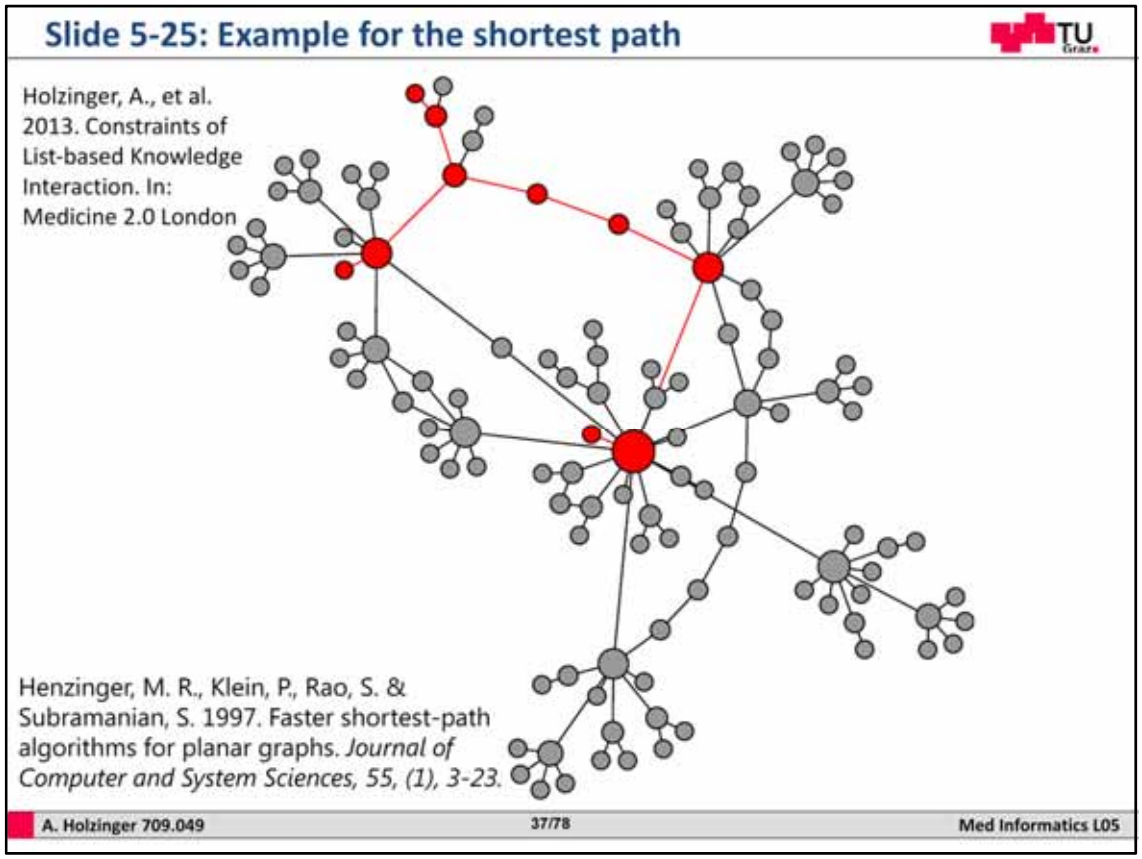
Up to know we know that Graphs and Graph-Theory are powerful tools to map data structures and to find novel connections between single data objects (Strogatz, 2001), (Dorogovtsev & Mendes, 2003). The inferred graphs can be further analyzed by using graph-theoretical and statistical and machine learning techniques (Dehmer, Emmert-Streib & Mehler, 2011). A mapping of the already existing and in the medical practice approved "knowledge space" as a conceptual graph and the subsequent visual and graph-theoretical analysis may provide novel insights on hidden patterns in the data. Another benefit of the graph-based data structure is in the applicability of methods from network topology and network analysis and data mining, e.g. small-world phenomenon (Barabasi & Albert, 1999), (Kleinberg, 2000), and cluster analysis (Koontz, Narendra & Fukunaga, 1976), (Wittkop et al., 2011).

The graph-theoretic data of the graph seen in this Slide include:

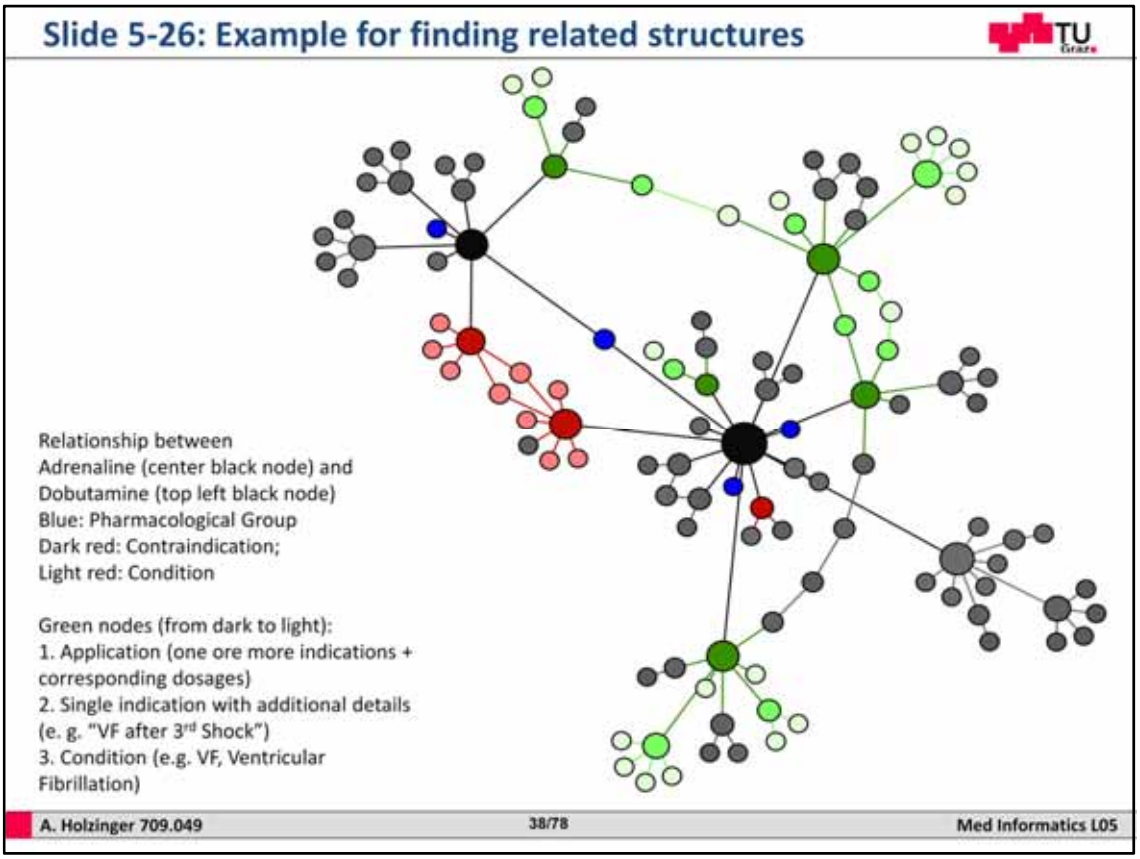
Number of nodes = 641, number of edges = 1250, red are agents, black are conditions, blue are pharmacological groups, grey are other documents. The average degree of this graph = 3.888, the average path length = 4.683, the network diameter = 9.



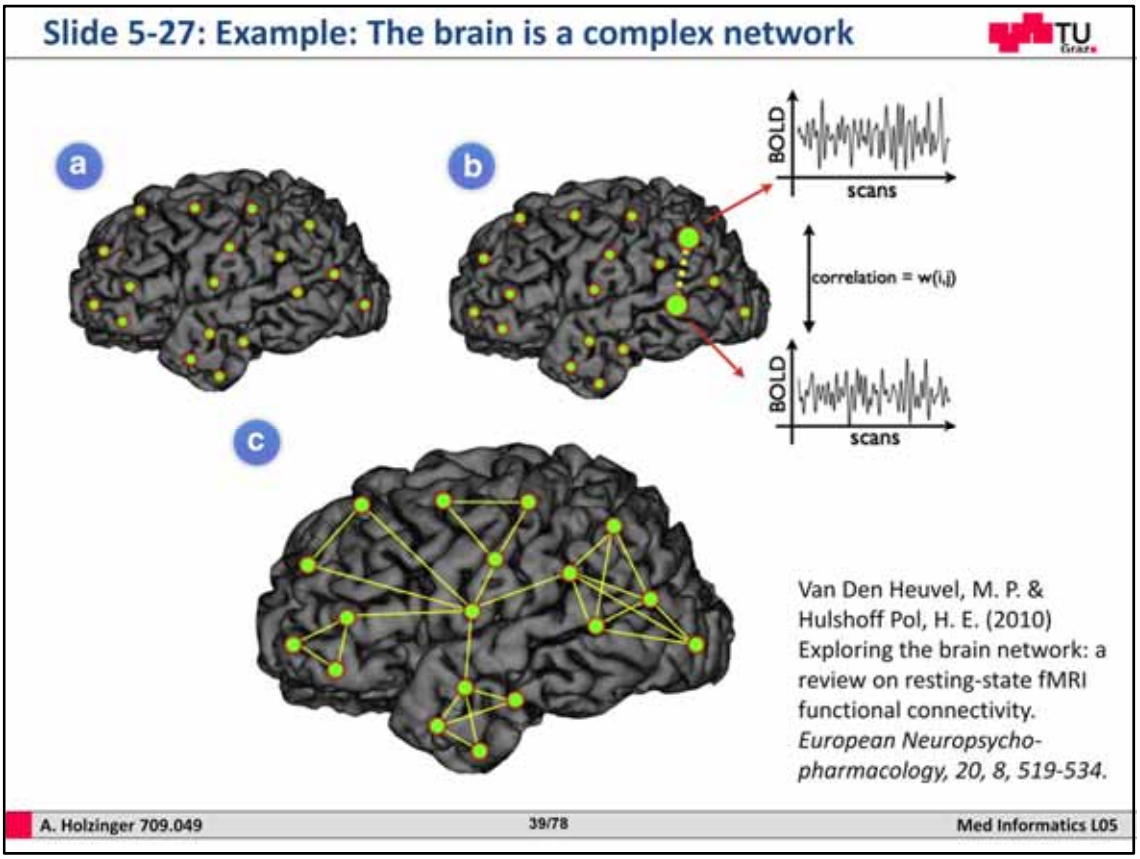

The nodes of the sample graph represent: drugs, clinical guidelines, patient conditions (indication, contraindication), pharmacological groups, tables and calculations of medical scores, algorithms and other medical documents; and the edges represent 3 crucial types of relations inducing medical relevance between two active substances, i.e.: pharmacological groups, indications and contra-indications. The following example will demonstrate the usefulness of this approach.



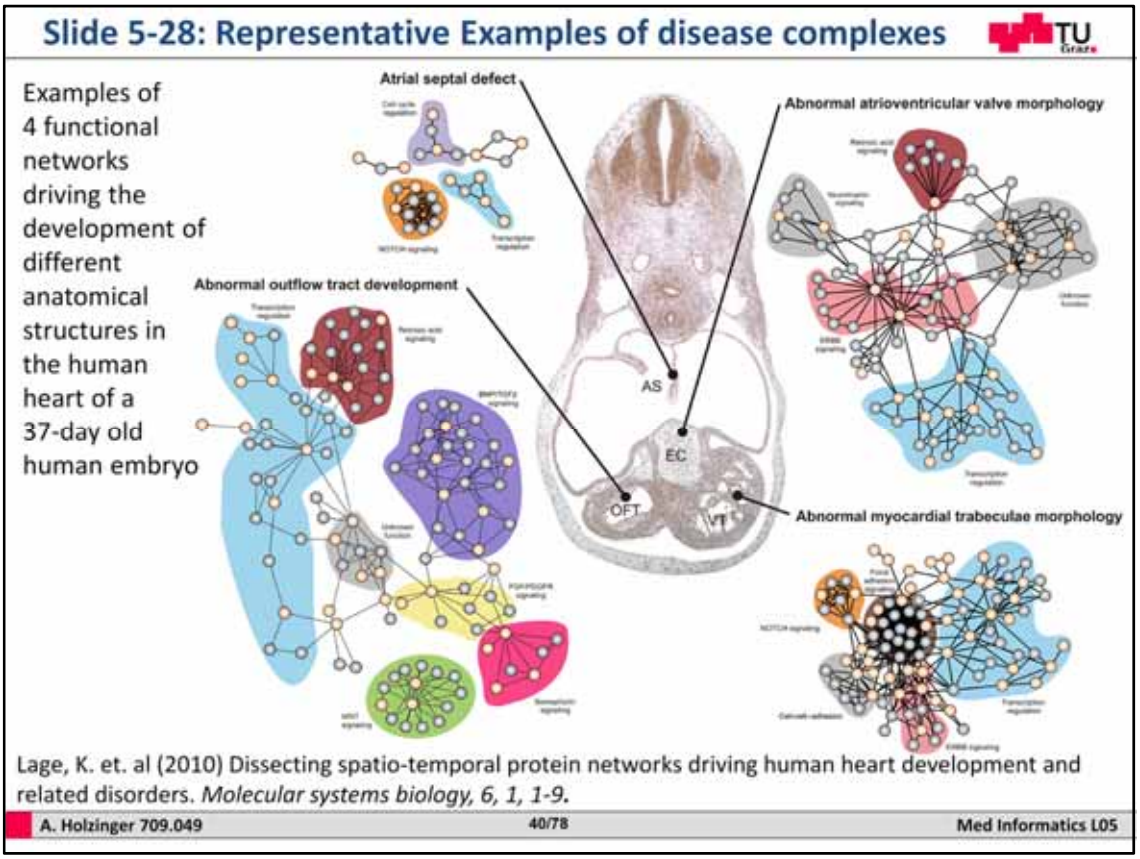

This example shows us how convenient we can find which path between two nodes is the shortest as well as the navigation way between these nodes. Computing shortest paths is a fundamental and ubiquitous problem in network analysis. We can, e.g. apply the Dijkstra-algorithm, solves the shortest path problem for a graph with non-negative edge path costs, producing a shortest path tree. This algorithm is often used in routing and as a subroutine in other graph algorithms: For a given node, the algorithm finds the path with lowest cost (i.e. the shortest path) between that node and every other node(Henzinger et al., 1997).




Here we see the relationship between Adrenaline (center black node) and Dobutamine (top left black node), Blue: Pharmacological Group, Dark red: Contraindication; Light red: Condition, the Green nodes (from dark to light) are:
1. Application (one ore more indications + corresponding dosages)
2. Single indication with additional details (e. g. "VF after 3rd Shock")
3. Condition (e.g. VF, Ventricular Fibrillation)

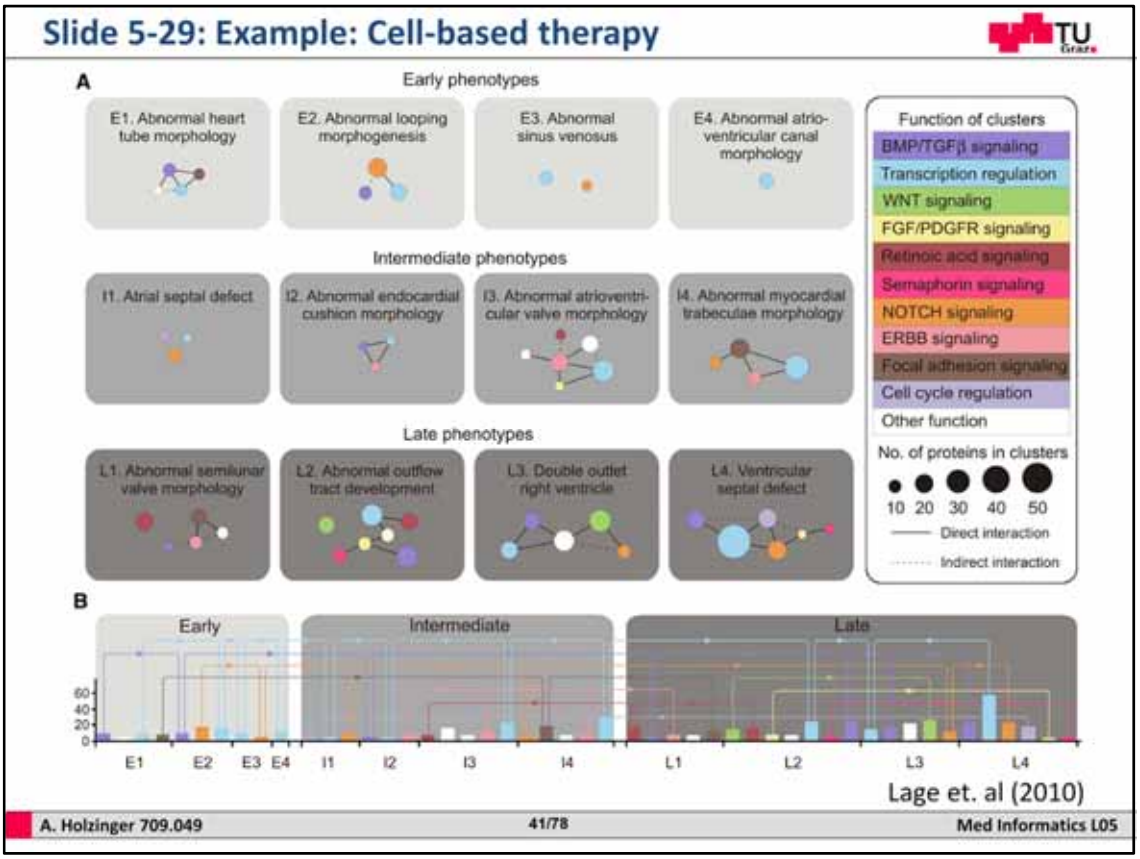



Our brain forms one integrative complex network, linking all brain regions and sub-networks together (Van Den Heuvel & Hulshoff Pol, 2010). Examining the organization of this network provides insights in how our brain works. Graph theory provides a framework in which the topology of complex networks can be examined; thus can reveal novelties about both the local and global organization of functional brain networks. In the slide we can see how the modeling of the functional brain by a graph works: edges are the connections between regions that are functionally linked. First, the collection of nodes is defined (A), second the existence of functional connections between the nodes in the network needs to be defined, resulting in a connectivity matrix (B). Finally, the existence of a connection between two points can be defined as whether their level of functional connectivity exceeds a certain predefined threshold (C) (Van Den Heuvel & Hulshoff Pol, 2010).

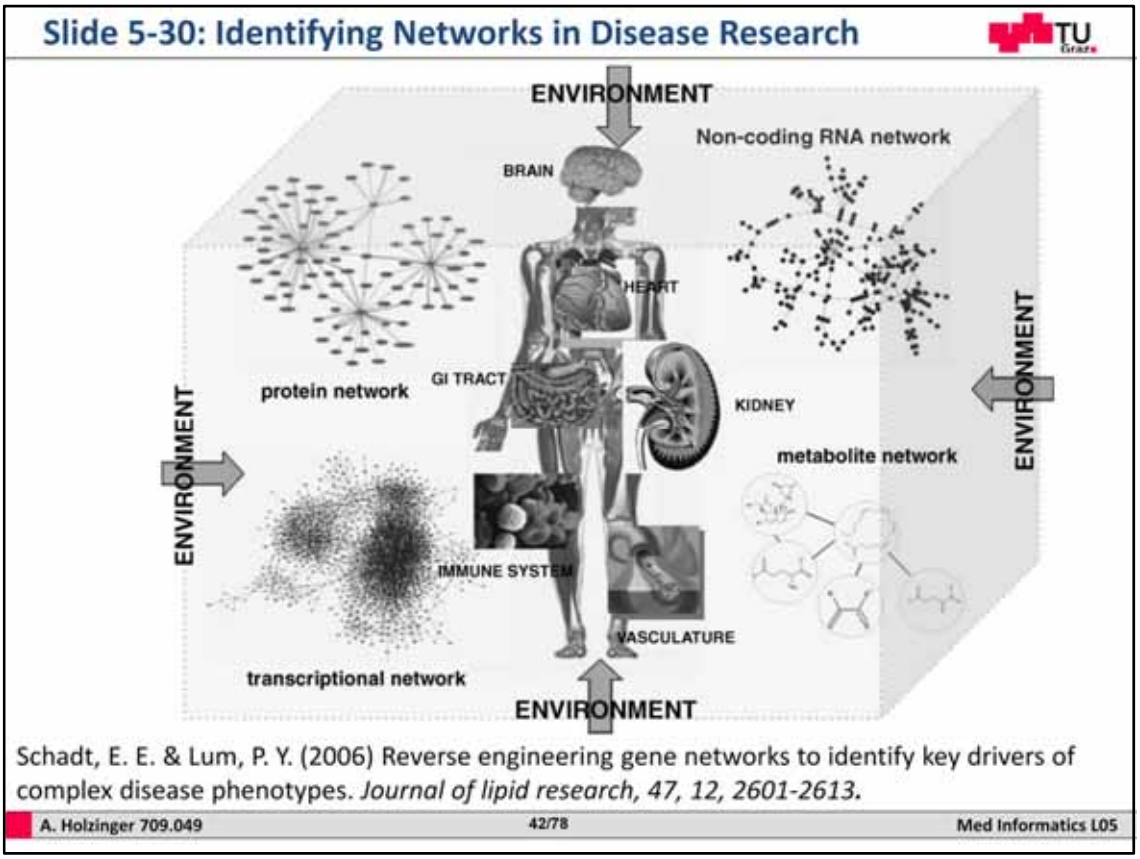



Development of the human heart starts 2 weeks after fertilization, with the formation of the cardiac crescent and the subsequent formation and looping of the primitive heart tube. Insight into the biology of molecular networks is an important field, as anomalies in these systems underlie a wide spectrum of polygenetic human disorders, ranging from schizophrenia to congenital heart disease (CHD). Understanding the functional architecture of networks that organize the development of organs, see e.g. (Chien, Domian & Parker, 2008), lays the foundation of novel approaches in regenerative medicine, since manipulation of such systems is necessary for success of tissue engineering technologies and stem cell therapy.

Lage et al. (2010) developed a framework for gaining new insights into the systems biology of the protein networks driving organ development and related polygenic human disease phenotypes, exemplified with heart development and CHD. In the Slide we see examples of four functional networks driving the development of different anatomical structures in the human heart. These four networks are constructed by analyzing the interaction patterns of four different sets of cardiac development (CD): proteins corresponding to the morphological groups 'atrial septal defects,' 'abnormal atrioventricular valve morphology,' 'abnormal myocardial trabeculae morphology,' and 'abnormal outflow tract development'. CD proteins from the relevant groups are shown in orange and their interaction partners are shown in gray. Functional modules annotated by literature curation are indicated with a colored background. Centrally in the Figure is a haematoxylin-eosin stained frontal section of the heart from a 37-day human embryo, where tissues affected by the four networks are marked; AS (developing atrial septum), EC (endocardial cushions, which are anatomical precursors to the atrioventricular valves), VT (developing ventricular trabeculae), and OFT (developing outflow tract).

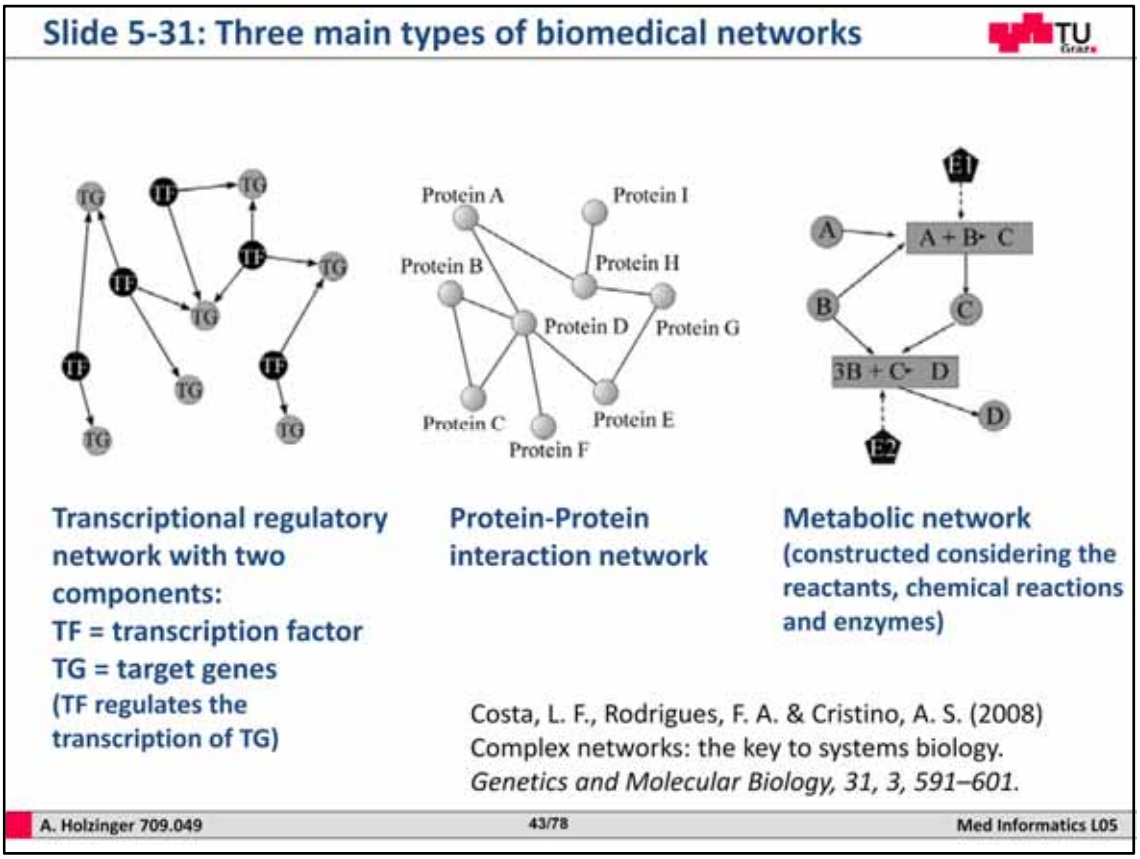



In this Slide we see an overview of the modular organization of heart development: (A) Protein interaction networks are plotted at the resolution of functional modules. Each module is color coded according to functional assignment as determined by literature curation. The amount of proteins in each module is proportional to the area of its corresponding node. Edges indicate direct (lines) or indirect (dotted lines) interactions between proteins from the relevant modules. (B) Recycling of functional modules during heart development. The bars represent functional modules and recycling is indicated by arrows. The bars follow the color code of (A) and the height of the bars represent the number of proteins in each module, as shown left on the y axis (Lage et al., 2010).
Note: Phenotype = an organism's observable characteristics (traits), e.g. morphology, biochemical/physiological properties, behaviour, etc. Phenotypes result from the expression of an organism's genes as well as the influence of environmental factors and the interactions between them. Genotype = inherited instructions within its genetic code.

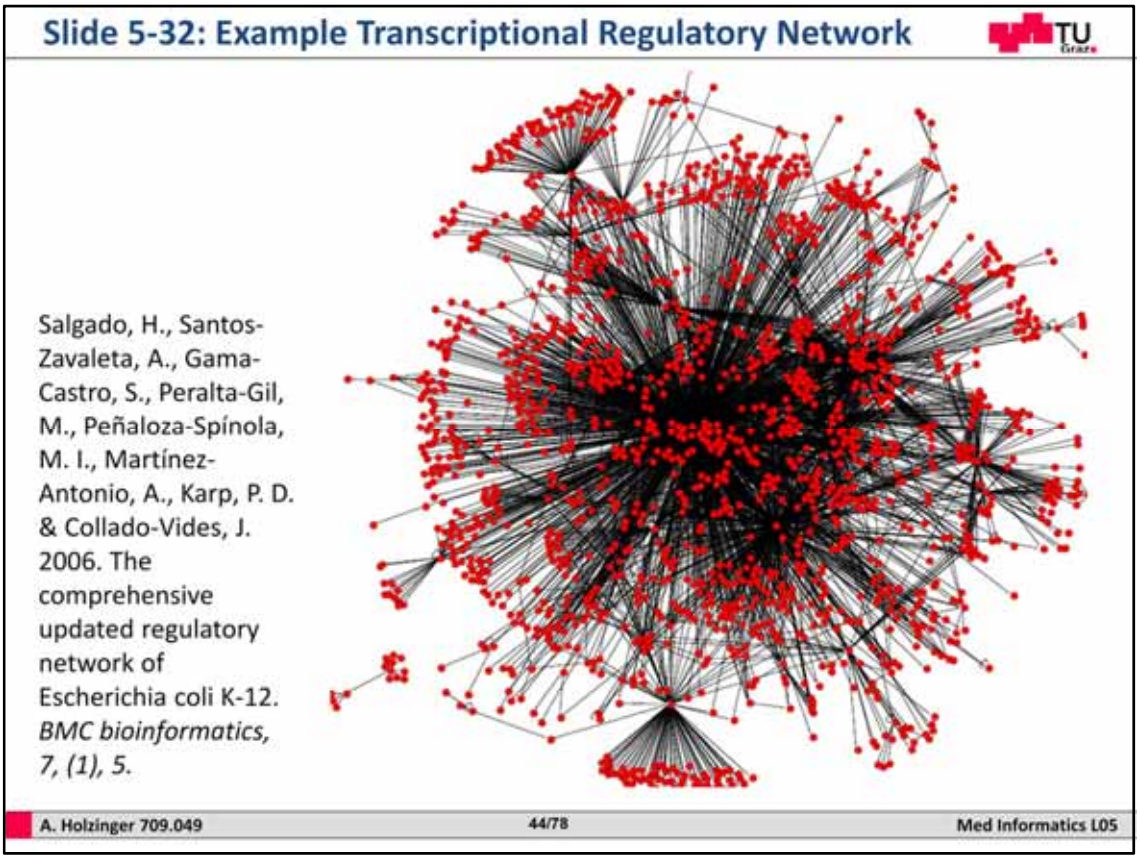



Diseases (e.g. obesity, diabetes, atherosclerosis etc.) result from multiple genetic and environmental factors, and importantly, interactions between genetic and environmental factors. This Slide shows the vast networks of molecular interactions. It can be seen that the gastrointestinal (GI) tract, vasculature, immune system, heart and brain are all potentially involved in either the onset of diseases such as atherosclerosis or in comorbidities such as myocardial infarction and stroke brought on by such diseases. Further, the risks of comorbidities for diseases such as atherosclerosis are increased by other diseases, such as hypertension, which may, in turn, involve other organs, such as kidney. The role that each organ and tissue type plays in a given disease is largely determined by genetic background and environment, where different perturbations to the genetic background (perturbations corresponding to DNA variations that affect gene function, which, in turn, leads to disease) and/or environment (changes in diet, levels of stress, level of activity, and so on) define the subtypes of disease manifested in any given individual. Although the physiology of diseases such as atherosclerosis is beginning to be better understood, what have not been fully exploited to data are the vast networks of molecular interactions within the cells.

We see clearly in the Slide that there is a diversity of molecular networks functioning in any given tissue, including genomics networks, networks of coding and noncoding RNA, protein interaction networks, protein state networks, signaling networks, and networks of metabolites. Further, these networks are not acting in isolation within each cell, but instead interact with one another to form complex, giant molecular networks within and between cells that drive all activity in the different tissues, as well as signaling between tissues. Variations in DNA and environment lead to changes in these molecular networks, which, in turn, induce complicated physiological processes that can manifest as disease. Despite this vast complexity, the classic approach to elucidating genes that drive disease has focused on single genes or single linearly ordered pathways of genes thought to be associated with disease. This narrow approach is a natural consequence of the limited set of tools that were available for querying biological systems; such tools were not capable of enabling a more holistic approach, resulting in the adoption of a reductionist approach to teasing apart pathways associated with complex disease phenotypes. Although the emerging view that complex biological systems are best modeled as highly modular, fluid systems exhibiting a plasticity that allows them to adapt to a vast array of conditions, the history of science demonstrates that this view, although long the ideal, was never within reach, given the unavailability of tools adequate to carrying out this type of research. The explosion of large-scale, high-throughput technologies in the biological sciences over the past 15 to 20 years has motivated a rapid paradigm shift away from reductionism in favor of a systems-level view of biology (Schadt & Lum, 2006).

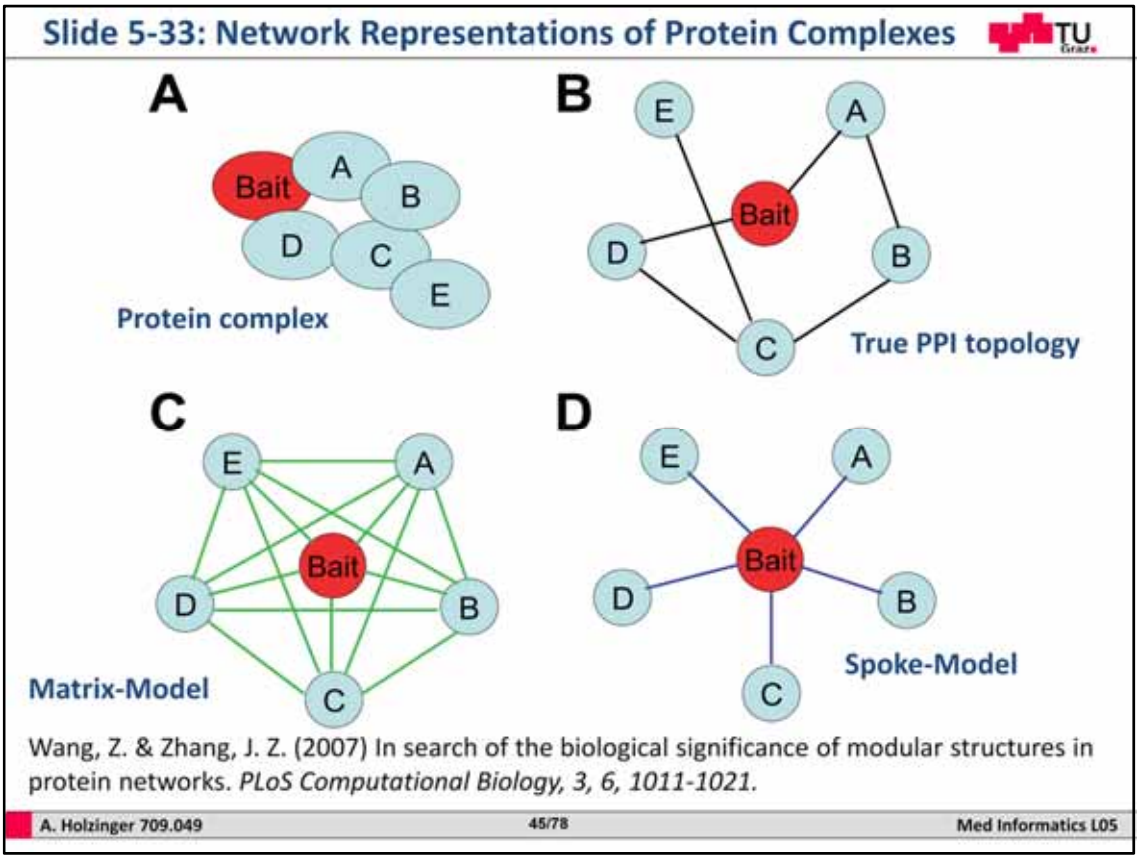



The three main types of biological networks: (a) a transcriptional regulatory network has two components: transcription factor (TF) and target genes (TG), where TF regulates the transcription of TGs; (b) protein-protein interaction networks: two proteins are connected if there is a docking between them; (c) a metabolic network is constructed considering the reactants, chemical reactions and enzymes.




The extreme complexity of the E. coli transcriptional regulatory network. In this graphical representation, nodes are genes, and edges represent regulatory interactions. The network was reconstructed using data from the RegulonDB (Salgado et al. 2006). This figure highlights the extreme complexity in regulatory networks. To obtain a deeper understanding of regulatory complexity, scientists must first discover biologically relevant organizational principles to unravel the hidden architecture governing these networks (see Nature Education: http://www.nature.com/scitable/content/the-extreme-complexity-of-the-e-coli-14457504)

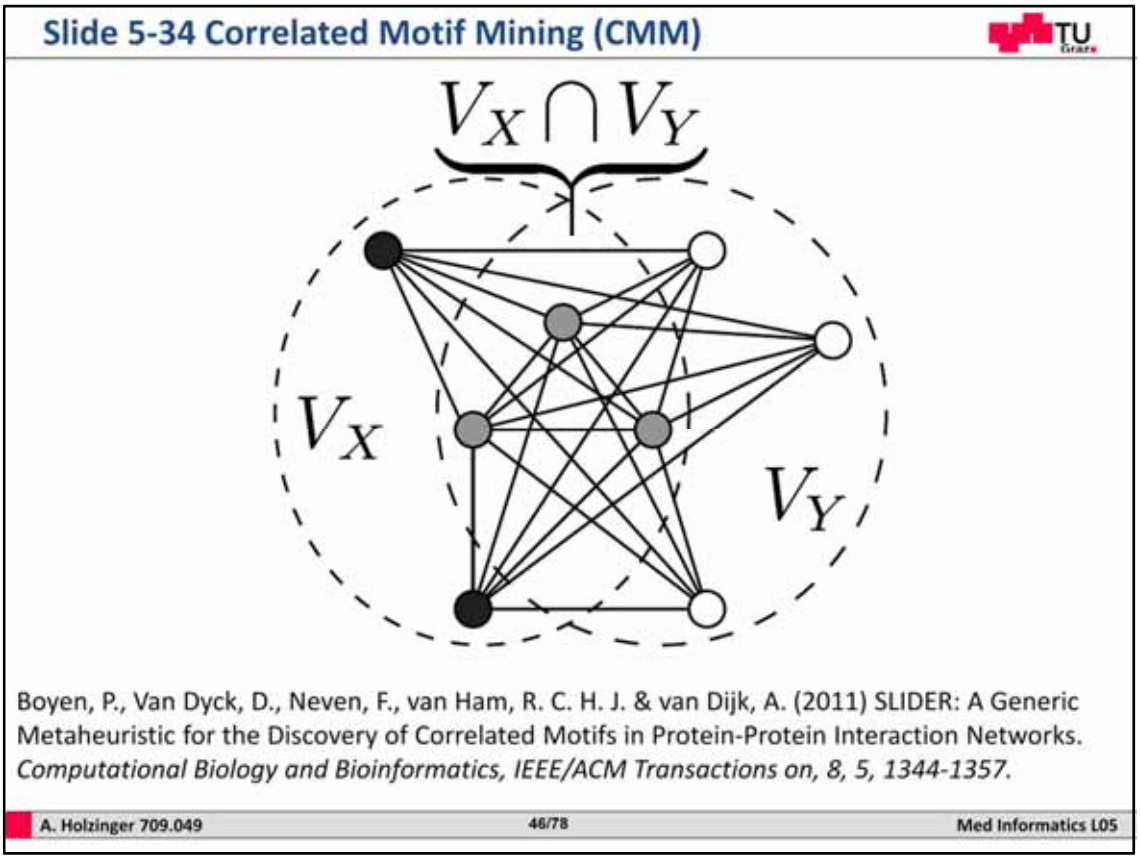
The complexity of organisms arises rather as a consequence of elaborated regulations of gene expression than from differences in genetic content in terms of the number of genes. The transcription network is a critical system that regulates gene expression in a cell. Transcription factors (TFs) respond to changes in the cellular environment, regulating the transcription of target genes (TGs) and connecting functional protein interactions to the genetic information encoded in inherited genomic DNA in order to control the timing and sites of gene expression during biological development. The interactions between TFs and TGs can be represented as a directed graph: The two types of nodes (TF and TG) are connected by arcs (see →Slide 5-31, arrows) when regulatory interaction occurs between regulators and targets. Transcriptional regulatory networks display interesting properties that can be interpreted in a biological context to better understand the complex behavior of gene regulatory networks. At a local network level, these networks are organized in substructures such as motifs and modules. Motifs represent the simplest units of a network architecture required to create specific patterns of inter-regulation between TFs and TGs. Three most common types of motifs can be found in gene regulatory networks:
(1) single input,
(2) multiple input and
(3) feed-forward loop
Target genes belonging to the same single and multiple input motifs tend to be co-expressed, and the level of co-expression is higher when multiple transcription factors are involved.
Modularity in the regulatory networks arises from groups of highly connected motifs that are hierarchically organized, in which modules are divided into smaller ones. The evolution of gene regulatory networks mainly occurs through extensive duplication of transcription factors and target genes with inheritance of regulatory interactions from ancestral genes while the evolution of motifs does not show common ancestry but is a result of convergent evolution (Costa, Rodrigues & Cristino, 2008) .




The interactions between proteins are essential to keep the molecular systems of living cells working properly. Protein-protein interaction (PPI) is important for various biological processes such as cell-cell communication, the perception of environmental changes, protein transport and modification. Complex network theory is suitable to study protein-protein interaction maps because of its universality and integration in representing complex systems. In complex network analysis each protein is represented as a node and the physical interactions between proteins are indicated by the edges in the network .

Many complex networks are naturally divided into communities or modules, where links within modules are much denser than those across modules (e.g. human individuals belonging to the same ethnic groups interact more than those from different ethnic groups). Cellular functions are also organized in a highly modular manner, where each module is a discrete object composed of a group of tightly linked components and performs a relatively independent task. It is interesting to ask whether this modularity in cellular function arises from modularity in molecular interaction networks such as the transcriptional regulatory network and PPI network.

The Slide shows a hypothetical protein complex (A). Binary protein-protein interactions (PPI) are depicted by direct contacts between proteins. Although five proteins (A, B, C, D, and E) are identified through the use of a bait protein (red), only A and D directly bind to the bait. (B) shows the true PPI network topology of the protein complex is shown in. (C) depicts the PPI network topology of the protein complex inferred by the ‘‘matrix’’ model, where all proteins in a complex are assumed to interact with each other. Finally (D) demonstrates the PPI network topology of the protein complex inferred by the ‘‘spoke’’ model, where all proteins in a complex are assumed to interact with the bait; but no other interactions are allowed (Wang & Zhang, 2007).




Correlated motif mining (CMM) is the challenge to find overrepresented pairs of patterns (motifs), in sequences of interacting proteins. Algorithmic solutions for CMM thereby provide a computational method for predicting binding sites for protein interaction. The task is basically to represent motifs X and Y (Figure 119) to truly represent an overrepresented consensus pattern in the sequences of the proteins in VX, respectively VY, in order to increase the likelihood that they correspond or overlap with a so called binding site—a site on the surface of the molecule that makes interactions between proteins from VX and VY possible through a molecular lock-and-key mechanism.
We call {X,Y} a (k_x k_y k_xy )-motif pair of a PPI network
G=(V,E,λ) if |V_x |= k_x,|V_y |=k_y and |V_x∩V_y | = k_xy
It is called complete if all vertices from V_x are connected with all vertices from V_y (Boyen et al., 2011).

In genetics, a sequence motif is a nucleotide or amino-acid sequence pattern that is widespread and has, or is conjectured to have, a biological significance. For proteins, a sequence motif is distinguished from a structural motif, a motif formed by the three dimensional arrangement of amino acids, which may not be adjacent. In a chain-like biological molecule, such as a protein or nucleic acid, a structural motif is a supersecondary structure, which appears also in a variety of other molecules. Motifs do not allow us to predict the biological functions because they are found in proteins and enzymes with dissimilar functions. Network motifs are connectivity-patterns (sub-graphs) that occur much more often than they do in random networks. Most networks studied in biology, ecology and other fields have been found to show a small set of network motifs; surprisingly, in most cases the networks seem to be largely composed of these network motifs, occurring again and again.

## Slide 5-35 Steepest Ascent Algorithm applied to CMM

**Input:** PPI-network $G = (V, E, \lambda)$, $\ell, d \in \mathbb{N}$, $d < \ell$
**Output:** $\{X^*, Y^*\}$ best correlated motif pair found in $G$

1: $\{X^*, Y^*\} \leftarrow \text{randomMotifPair}()$
2: $maxsup \leftarrow f(\{X^*, Y^*\}, G)$
3: $sup \leftarrow -\infty$
4: **while** $maxsup > sup$ **do**
5: $\quad \{X, Y\} \leftarrow \{X^*, Y^*\}$
6: $\quad sup \leftarrow maxsup$
7: $\quad$ **for all** $\{X', Y'\} \in N(\{X, Y\})$ **do**
8: $\quad\quad$ **if** $f(\{X', Y'\}, G) > maxsup$ **then**
9: $\quad\quad\quad \{X^*, Y^*\} \leftarrow \{X', Y'\}$
10: $\quad\quad\quad maxsup \leftarrow f(\{X', Y'\}, G)$

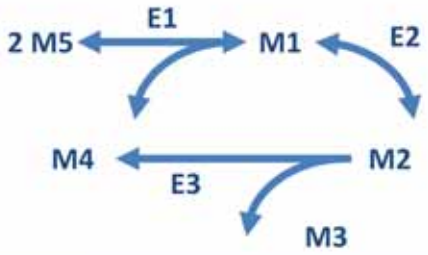
Boyen et al. (2011)

A. Holzinger 709.049 47/78 Med Informatics L05

The general steepest ascent algorithm with abstract neighbor function applied to CMM (SA-CMM).

Since the decision problem associated with CMM is in NP , we can efficiently check if a motif pair has higher support than another which makes it possible to tackle CMM as a search problem in the space of all possible (l,d)-motif pairs. If we add the assumption that similar motifs can be expected to get similar support, it has the typical form of a combinatorial optimization problem. In combinatorial optimization, the objective is to find a point in a discrete search space which maximizes a user-provided function f. A number of heuristic algorithms called metaheuristics are known to yield stable results, e.g. the steepest ascent algorithm (Aarts & Lenstra, 1997), illustrated as pseudocode in the Slide.

## Slide 5-36: Metabolic Network




| | M1 | M2 | M3 | M4 | M5 |
|---|---|---|---|---|---|
| M1 | 0 | 1 | 0 | 1 | 1 |
| M2 | 1 | 0 | 1 | 1 | 0 |
| M3 | 0 | 0 | 0 | 0 | 0 |
| M4 | 1 | 0 | 0 | 0 | 0 |
| M5 | 1 | 0 | 0 | 0 | 0 |

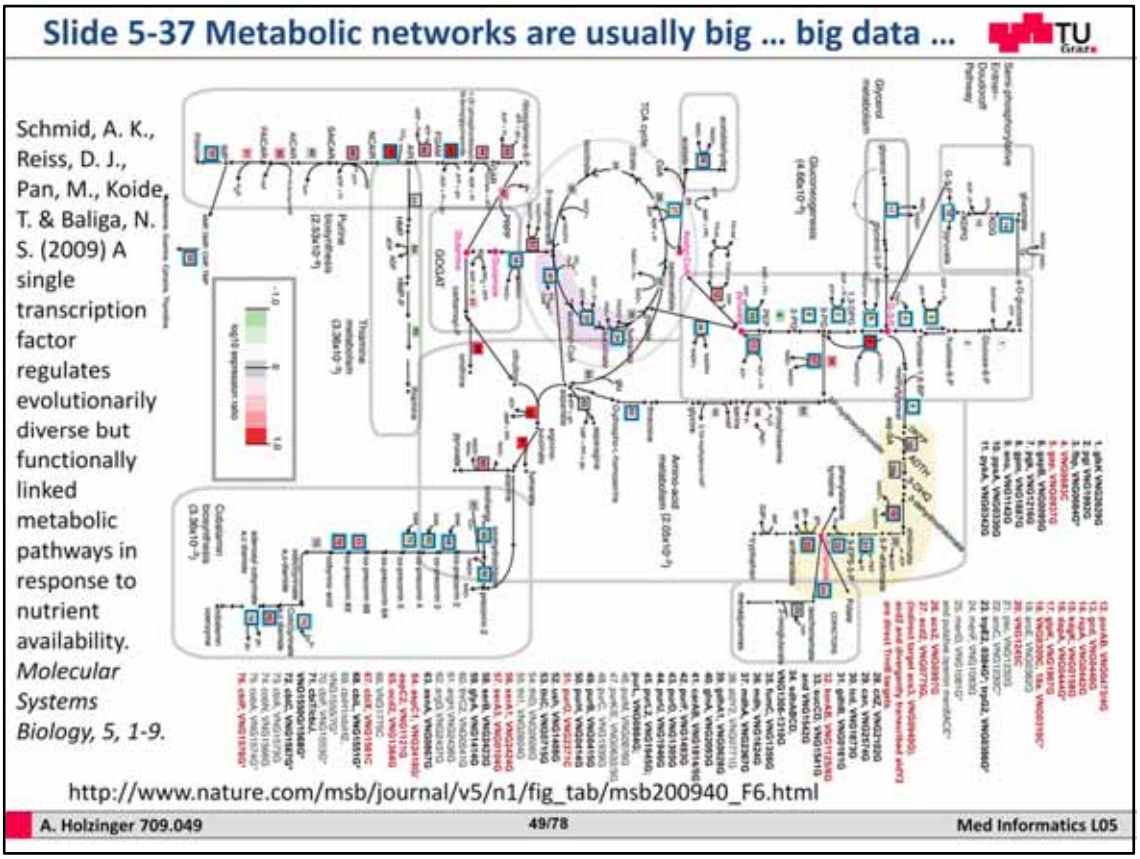
| | |
|---|---|
| M1 | M2 |
| M1 | M4 |
| M1 | M5 |
| M2 | M1 |
| M2 | M3 |
| M2 | M4 |
| M4 | M1 |
| M5 | M1 |

**Matrix contains many sparse elements - In this case it is computationally more efficient to represent the graph as an adjacency list**

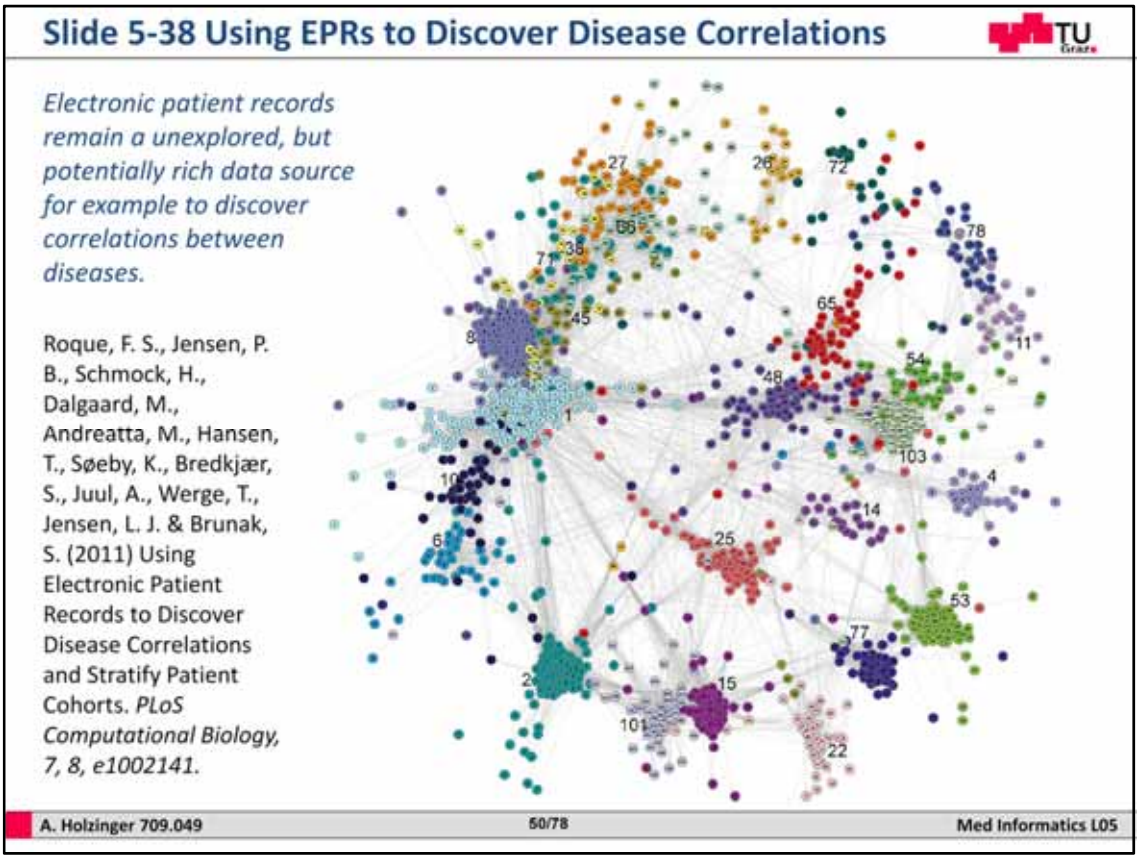
Hodgman, C. T., French, A. & Westhead, D. R. (2010) *Bioinformatics. Second Edition. New York, Taylor & Francis.*

A. Holzinger 709.049 48/78 Med Informatics L05

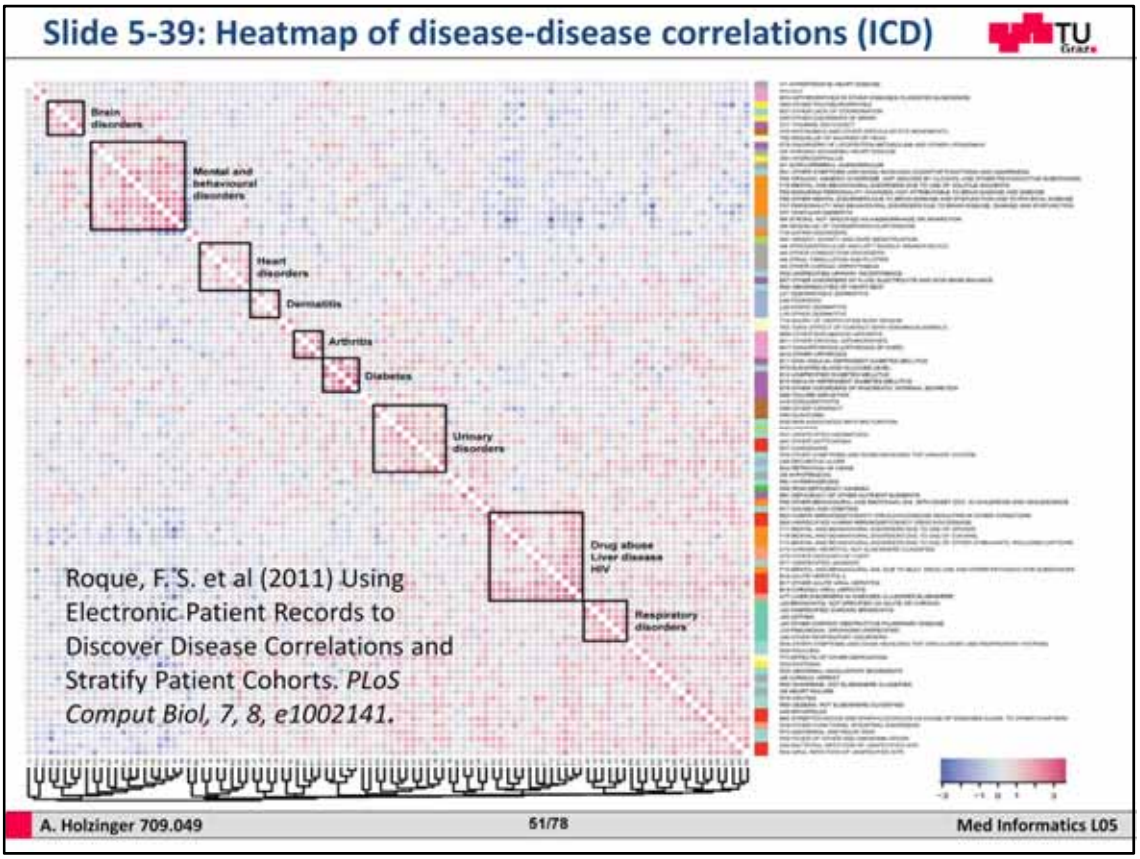
Metabolism is primarily determined by genes, environment and nutrition. It consists of chemical reactions catalyzed by enzymes to produce essential components such as amino acids, sugars and lipids, and also the energy necessary to synthesize and use them in constructing cellular components. Since the chemical reactions are organized into metabolic pathways, in which one chemical is transformed into another by enzymes and co-factors, such a structure can be naturally modeled as a complex network. In this way, metabolic networks are directed and weighted graphs, whose vertices can be metabolites, reactions and enzymes, and two types of edges that represent mass flow and catalytic reactions. One widely considered catalogue of metabolic pathways available on-line is the Kyoto Encyclopedia of Genes and Genomes (KEGG). In the Slide we see a simple metabolic network involving five metabolites M1-M5 and three enzymes E1-E3, of which the latter catalyzes an irreversible reaction (Hodgman, French & Westhead, 2010).




Such metabolic structures can be very large, as can be seen in this Slide. The enzyme-coding genes under TrmB (this is the thermococcus regulator of maltose binding) acts as a repressor for genes encoding glycolytic enzymes and as activator for genes encoding gluconeogenic enzymes control included in the metabolic pathways shown in the Slide (13 are unique to archaea and 35 are conserved across species from all three domains of life. Integrated analysis of the metabolic and gene regulatory network architecture reveals various interesting scenarios (Schmid et al., 2009).

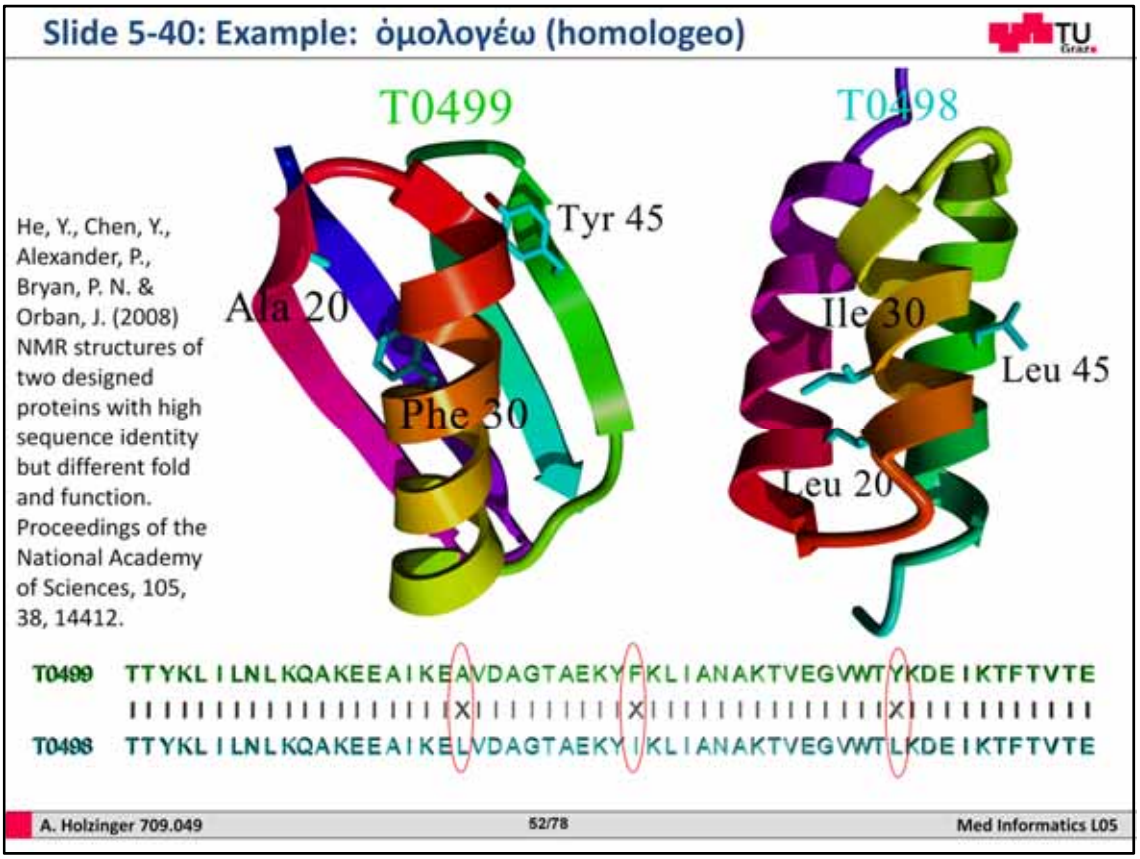



Electronic patient records (EPR remain an unexplored, but rich data source for discovering e.g. correlations between diseases. (Roque et al., 2011) describe a general approach for gathering phenotypic descriptions of patients from medical records in a systematic and non-cohort dependent manner: By extracting phenotype information from the "free-text" (= unstructured information) in such records they demonstrated that they can extend the information contained in the structured record data, and use it for producing fine-grained patient stratification and disease co-occurrence statistics. Their approach uses a dictionary based on the International Classification of Disease (ICD-10) ontology and is therefore in principle language independent. As a use case they show how records from a Danish psychiatric hospital lead to the identification of disease correlations, which subsequently can be mapped to systems biology frameworks.

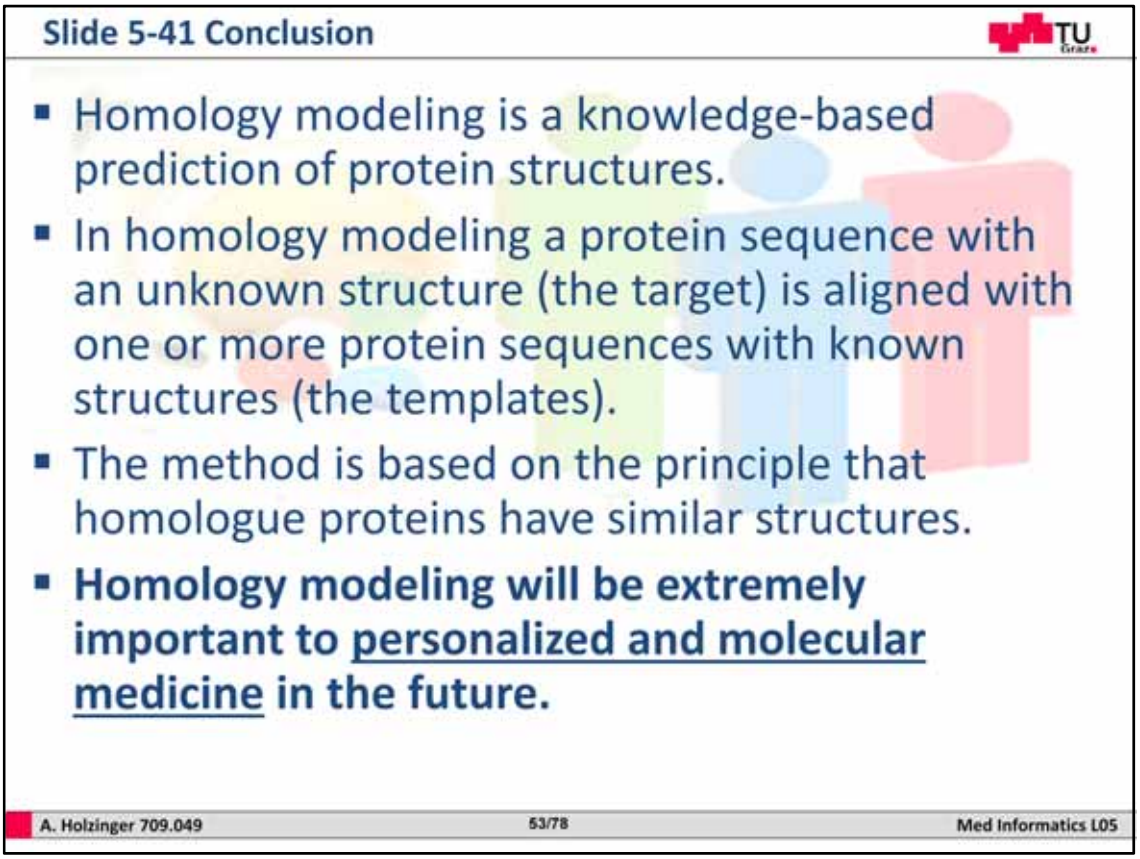



Disease-disease correlations. Heatmap of the most significant 100 ICD10 codes, based on ranking the list of 802 candidate pairs by their comorbidity scores. Chapter colors are highlighted next to the ICD10 codes. Diseases that occur often together have red color in the heatmap, while those with lower than expected co-occurrence are colored blue. The color label shows the log2 change of comorbidity between two diseases
when compared to the expected level.
doi:10.1371/journal.pcbi.1002141.g002

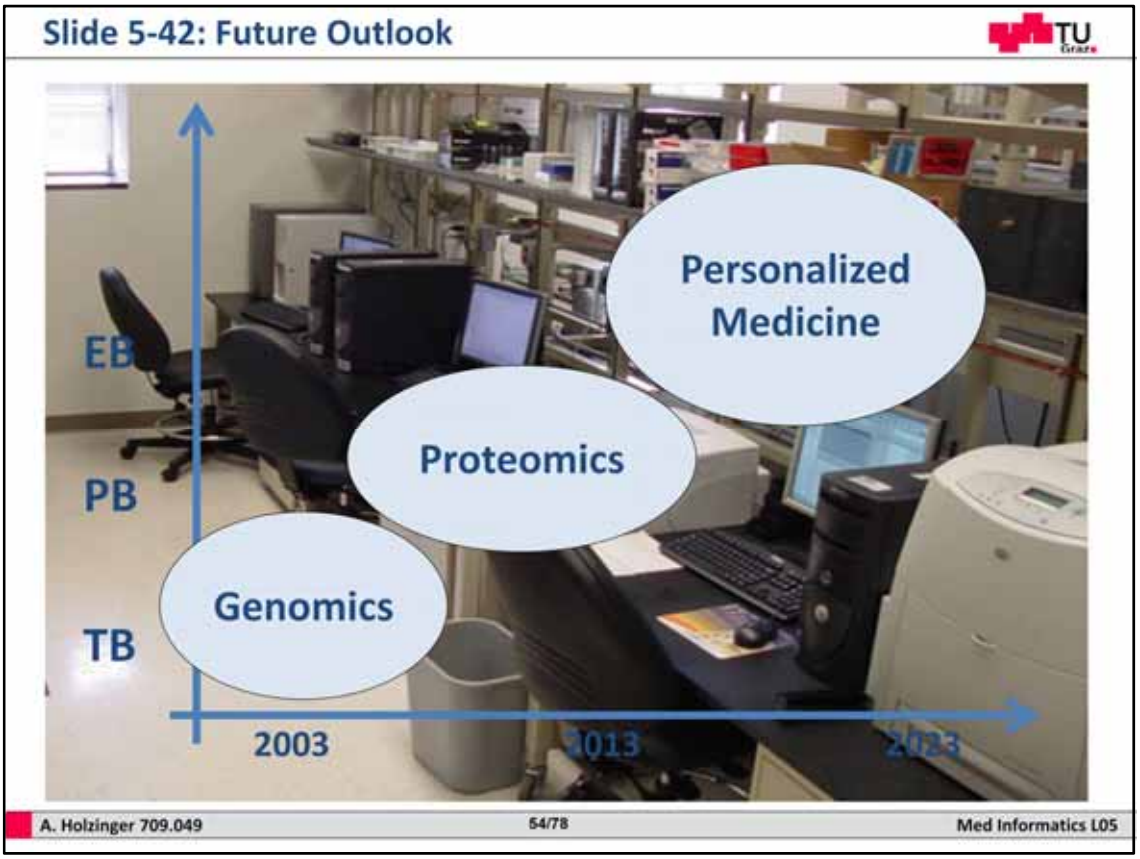
Roque et al. (2011) have used text mining to automatically extract clinically relevant terms from 5543 psychiatric patient records and mapped these to disease codes in the ICD10. They clustered patients together based on the similarity of their profiles. The result is a patient stratification, based on more complete profiles than the primary diagnosis, which is typically used. Figure 124 illustrates the general approach to capture correlations between different disorders. Several clusters of ICD10 codes relating to the same anatomical area or type of disorder can be identified along the diagonal of the heatmap, ranging from trivial correlations (e.g., different arthritis disorders), to correlations of cause and effect codes (e.g., stroke and mental/behavioural disorders), to social and habitual correlations (e.g. drug abuse, liver diseases and HIV).




Homology (plural: homologies) origins from Greek ὁμολογέω (homologeo) and means "to conform" (in German: übereinstimmen) and has its origins in Biology and Anthropology, where the word is used for a correspondence of structures in two life forms with a common evolutionary origin (Darwin, 1859).
In chemistry it is used for the relationship between the elements in the same group of the periodic table, or between organic compounds in a homologous series.
In mathematics homology is a formalism for talking in a quantitative and unambiguous manner about how a space is connected (Edelsbrunner & Harer, 2010).
Basically, homology is a concept that is used in many branches of algebra and topology. Historically, the term was first used in a topological sense by Henry Poincaré.
In Bioinformatics, homology modelling is a mature technique that can be used to address many problems in molecular medicine. Homology modelling is one of the most efficient methods to predict protein structures. With the increase in the number of medically relevant protein sequences, resulting from automated sequencing in the laboratory, and in the fraction of all known structural folds, homology modelling will be even more important to personalized and molecular medicine in the future. Homology modelling is a knowledge-based prediction of protein structures. In homology modelling a protein sequence with an unknown structure (the target) is aligned with one or more protein sequences with known structures (the templates).
The method of homology modelling is based on the principle that homologue proteins have similar structures. The prerequisite for successful homology modelling is a detectable similarity between the target sequence and the template sequences (more than 30%) allowing the construction of a correct alignment. Homology modelling is a knowledge-based structure prediction relying on observed features in known homologous protein structures. By exploiting this information from template structures the structural model of the target protein can be constructed (Wiltgen & Tilz, 2009).
Two well-known homology modelling programs, which are free for academic research, are
MODELLER (http://salilab.org/modeller) and
SWISSMODEL (http://swissmodel.expasy.org).
The slide shows the comparison of two proteins: The sequences of both proteins are 95% (53 of 56) identical (only residues 20, 30 and 45 differ), yet the structures are totally different.




All the areas we have touched in this lecture are extremely important towards the concept of personalized medicine and molecular medicine and will keep us busy within the next decades.
Data mining is maybe the most central and most important computational subject in this respect.




All these approaches are producing gigantic amounts of highly complex data sets!

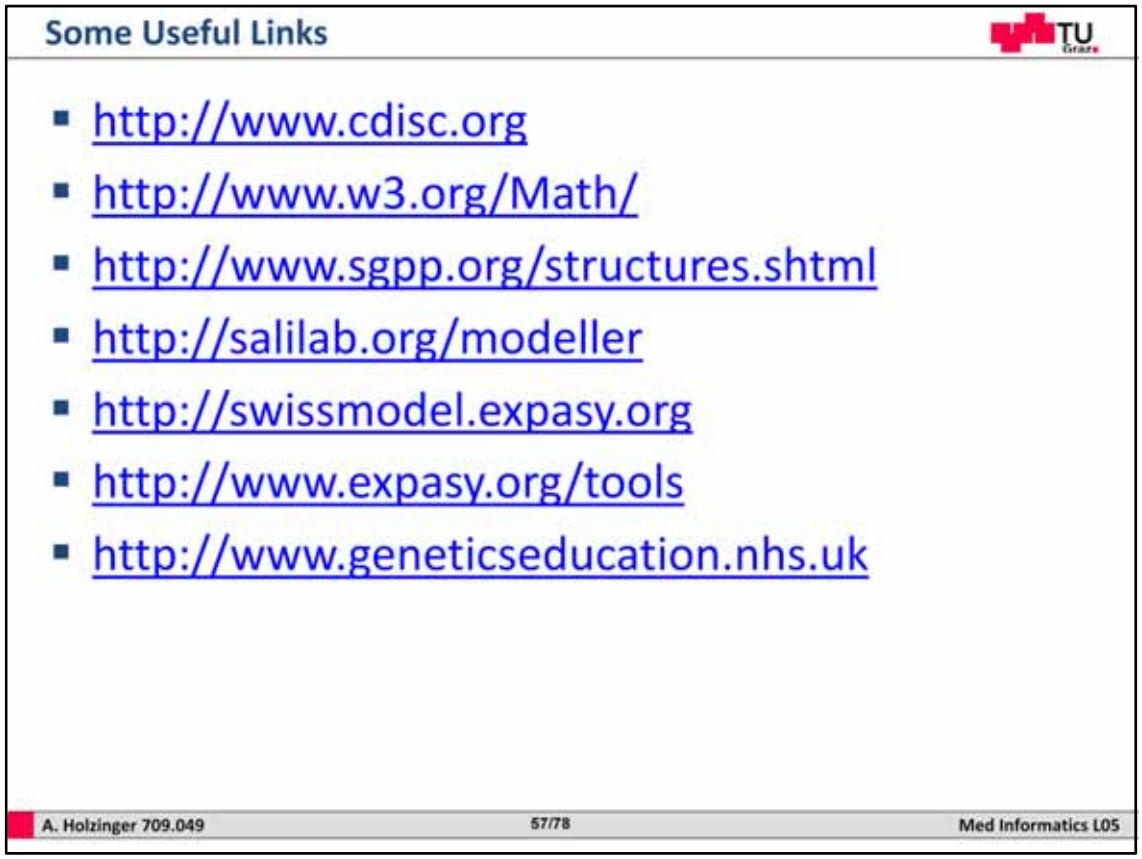
See the recent article in Science – doubling of data in proteomics every 18 months


My DEDICATION is to make data valuable … Thank you!
The Klein-Bottle is the symbol for geometry and topology.

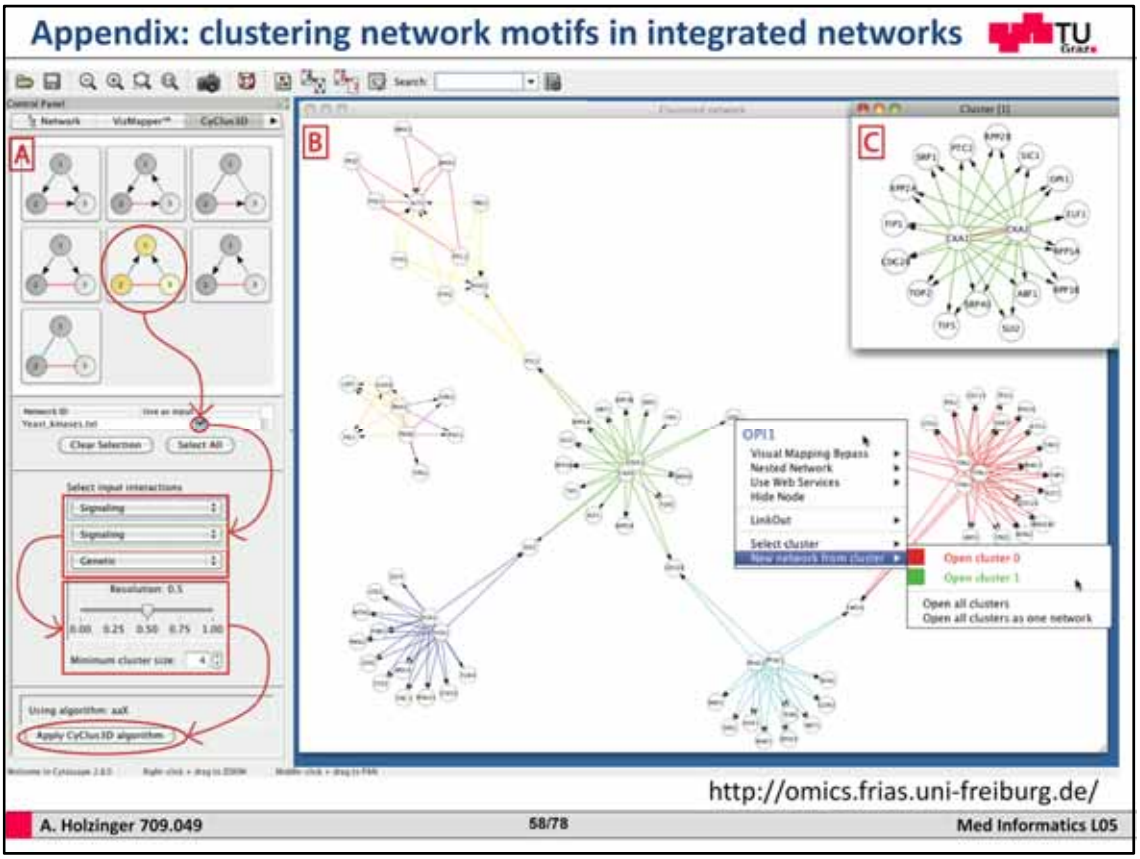
Topological data analysis (TDA) is a fast growing branch of applied mathematics and of enormous importance for data mining and knowledge discovery,
particularly from large, high-dimensional, incomplete and noisy dirty data.

## Sample Questions

- Which are the four main "big data" pools in the health care domain and what problems involved?
- What is the main problem in medical documentation?
- What is the advantage of an integrated Patient record?
- What are the advantages/disadvantages of XML/OWL for data in bioinformatics?
- What are the three key concepts in order to understand complex biological systems?
- What are the main symbols describing a network as used in Bioinformatics?
- How can networks represented computationally effectively?
- What are the main network metrics?
- What are the main network topologies used in Biomedical informatics?
- What is the Small-World Theory?
- Why is the study of networks relevant for medical professionals?
- Which are the three main types of biomedical networks?
- What is a Motif?
- What benefits can we gain from Correlated Motif Mining (CMM)?
- What is more efficient if a matrix contains many sparse elements?
- Why are structural homologies interesting for biomedical informatics?

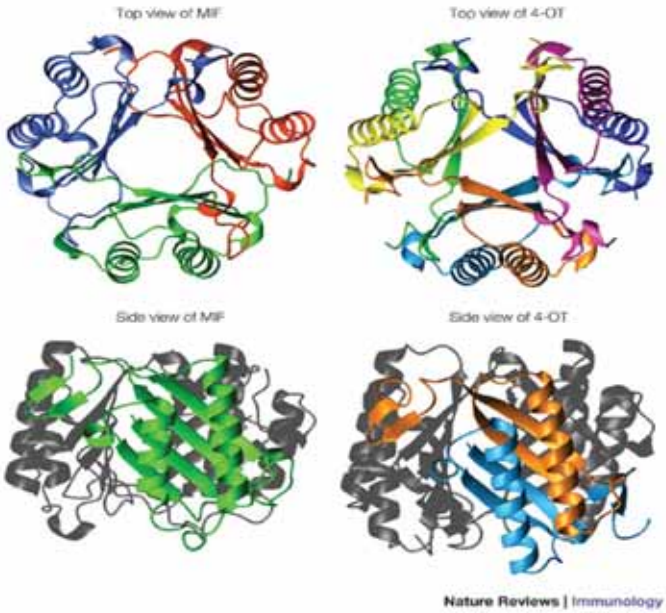



http://psychology.wikia.com/wiki/Information_retrieval




Network motifs in integrated molecular networks represent functional relationships between distinct data types. They aggregate to form dense topological structures corresponding to functional modules which cannot be detected by traditional graph clustering algorithms.

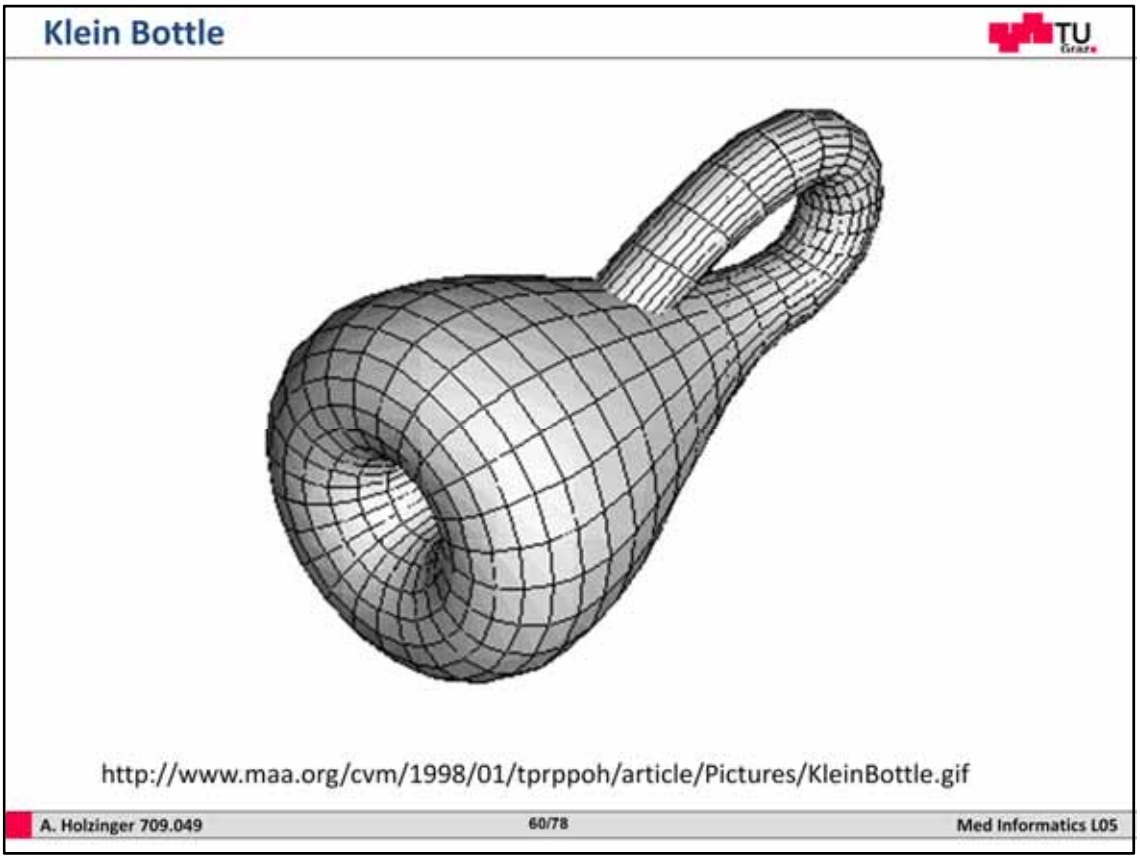
## Example from Immunology: Structural Homology




Calandra, T. & Roger, T. 2003. Macrophage migration inhibitory factor: a regulator of innate immunity. *Nat Rev Immunol, 3, **791-800.***

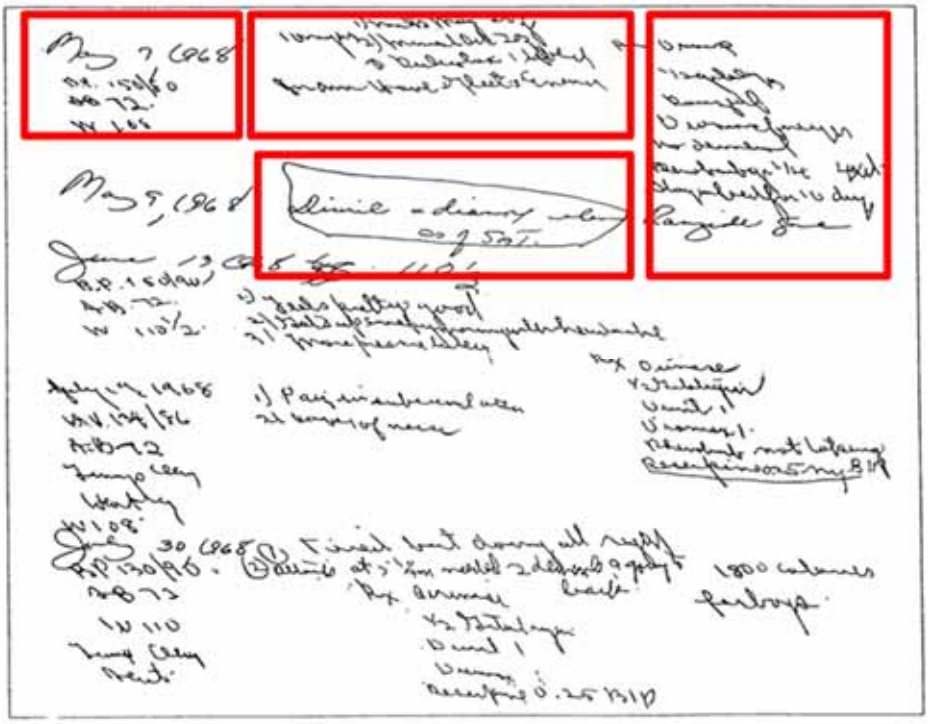


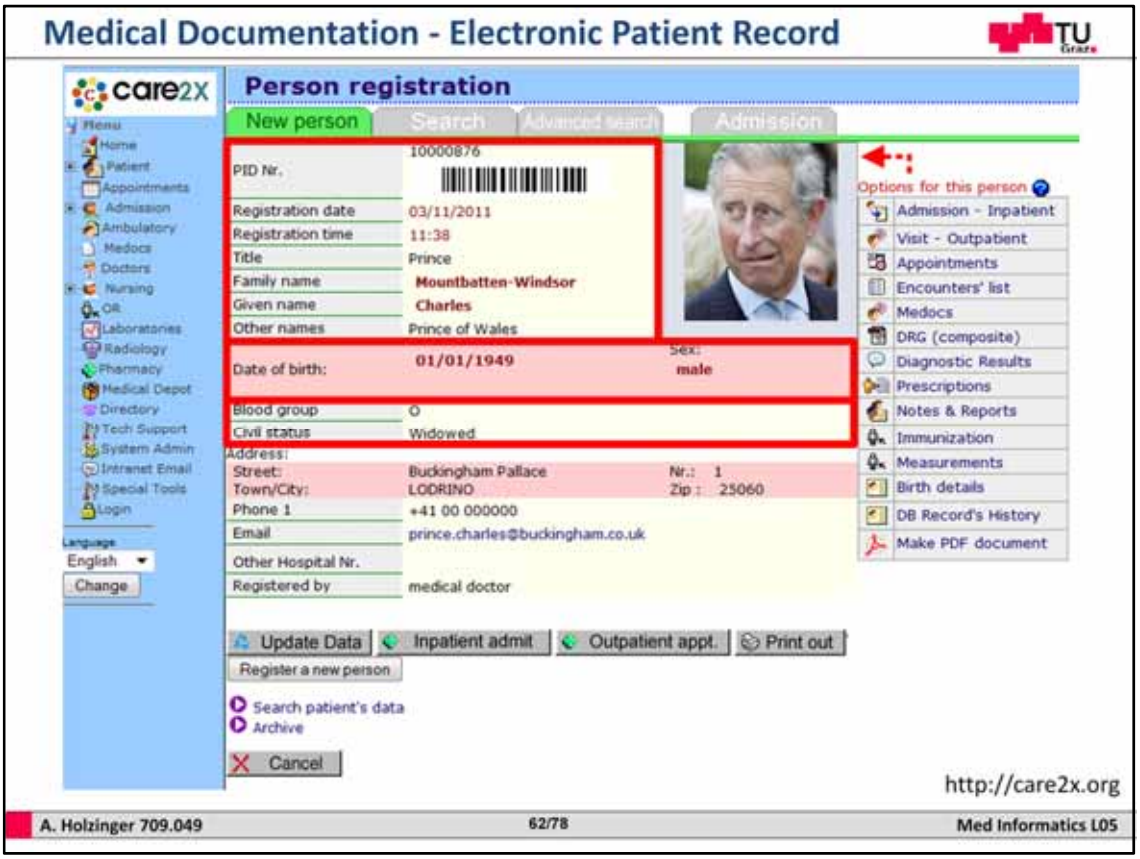
http://www.nature.com/nri/journal/v3/n10/fig_tab/nri1200_F2.html




http://www.maa.org/cvm/1998/01/tprppoh/article/Pictures/KleinBottle.gif

## Medical Documentation – Patient Record



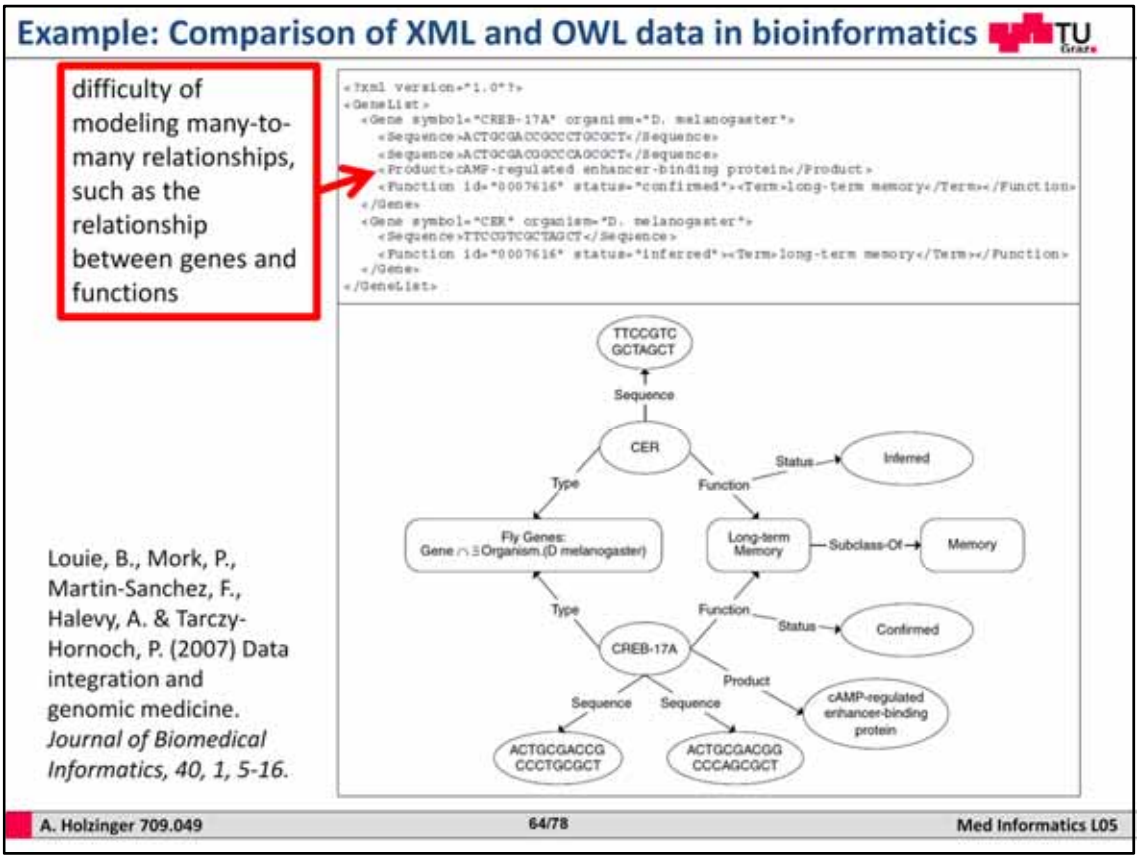

Medical Documentation - Electronic Patient Record
TU Graz
care2X
Person registration
New person
Search
Advanced search
Admission
Menu
Home
Patient
Appointments
Admission
Ambulatory
Medocs
Doctors
Nursing
OR
Laboratories
Radiology
Pharmacy
Medical Depot
Directory
Tech Support
System Admin
Intranet Email
Special Tools
Login
Language
English
Change
PID Nr. 10000876
Registration date 03/11/2011
Registration time 11:38
Title Prince
Family name Mountbatten-Windsor
Given name Charles
Other names Prince of Wales
Date of birth: 01/01/1949
Sex: male
Blood group O
Civil status Widowed
Address:
Street: Buckingham Pallace Nr.: 1
Town/City: LODRINO Zip: 25060
Phone 1 +41 00 000000
Email prince.charles@buckingham.co.uk
Other Hospital Nr.
Registered by medical doctor
Options for this person
Admission - Inpatient
Visit - Outpatient
Appointments
Encounters' list
Medocs
DRG (composite)
Diagnostic Results
Prescriptions
Notes & Reports
Immunization
Measurements
Birth details
DB Record's History
Make PDF document
Update Data
Inpatient admit
Outpatient appt.
Print out
Register a new person
Search patient's data
Archive
Cancel
http://care2x.org
A. Holzinger 709.049
62/78
Med Informatics L05

## Challenge is in Genomic medicine …

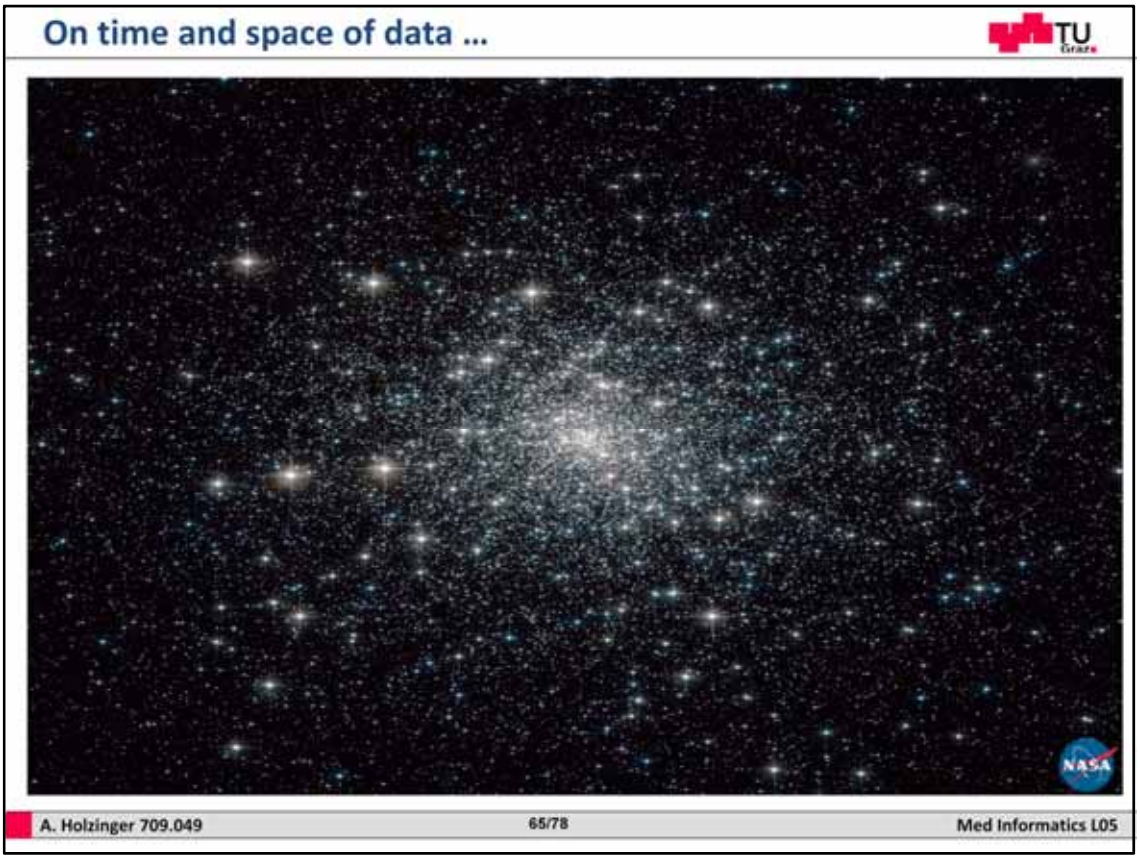

- … to integrate and analyze these diverse and voluminous data sources to elucidate both normal and disease physiology.
- XML is suited for describing semi-structured data including a natural modeling of biological entities, because it allows features as e.g. nesting …

A. Holzinger 709.049 63/78 Med Informatics L05

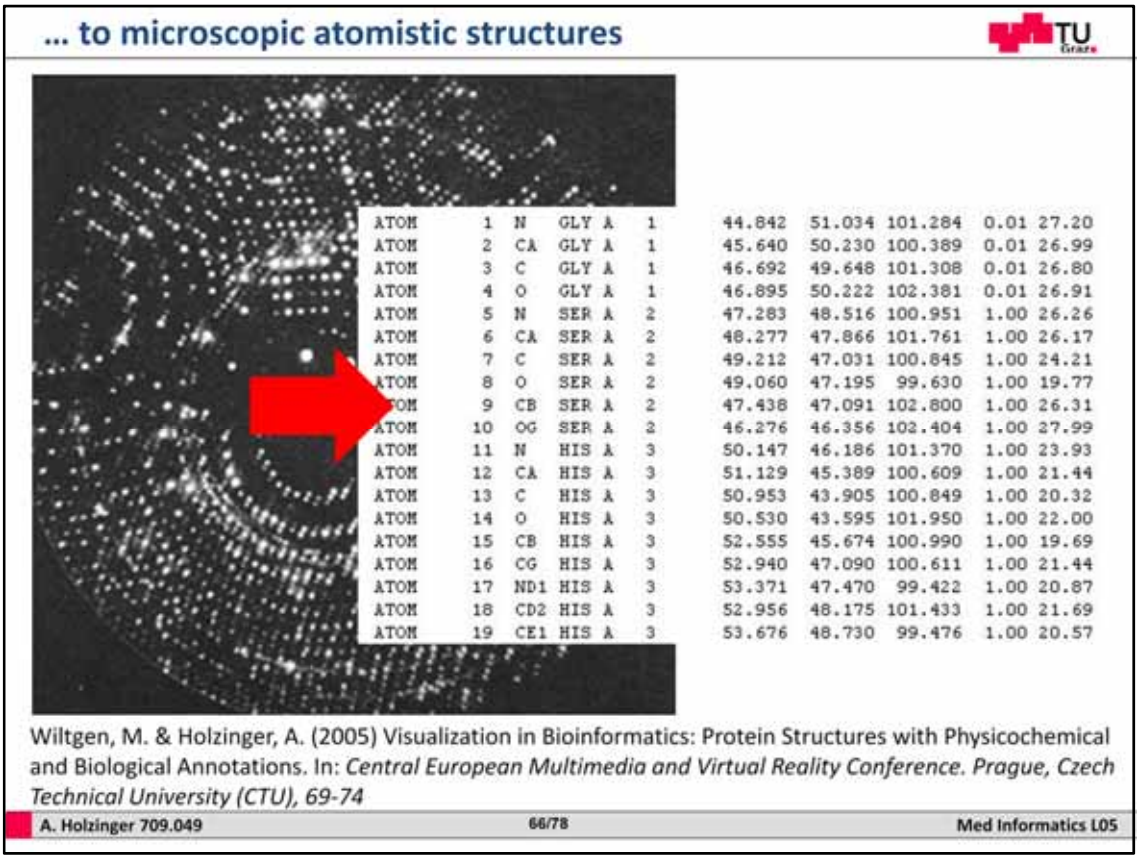
Nesting = recursion, subroutines, information hiding,



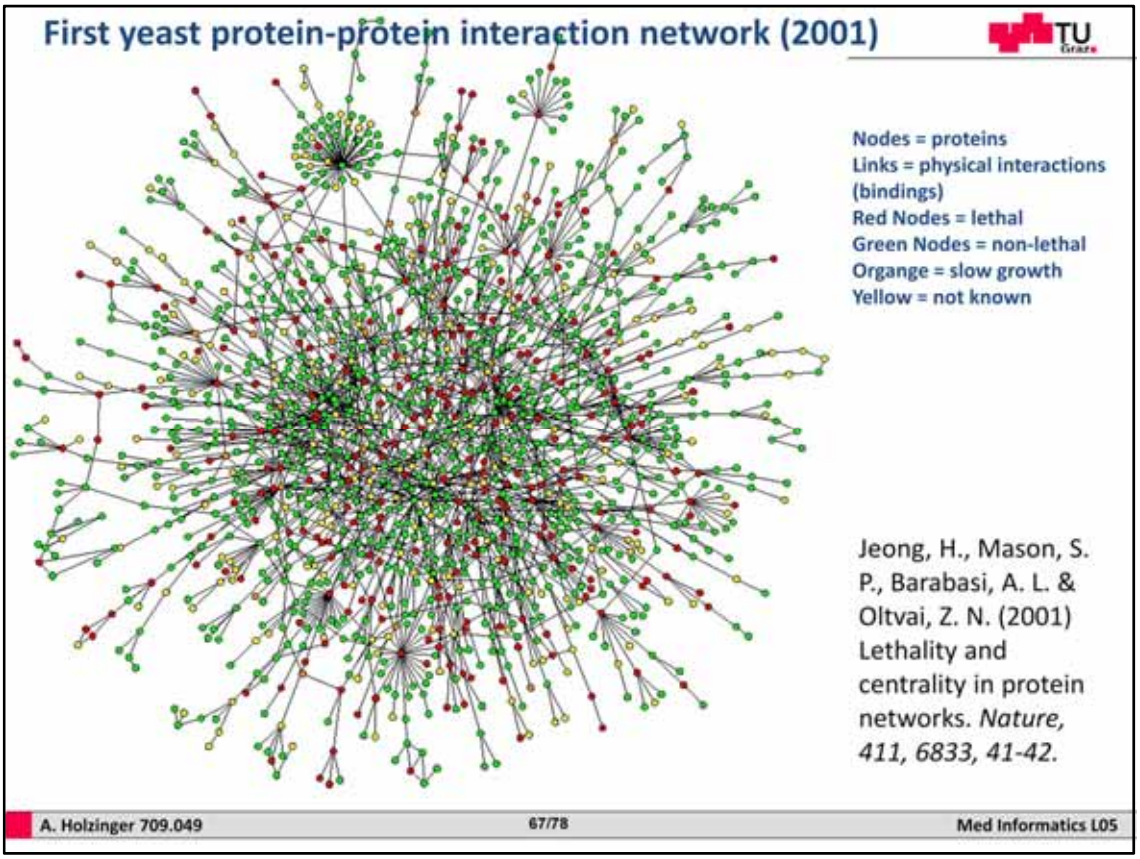

On top in Figure 39 we see a sample XML describing genes involved in long-term memory of a sample specimen Drosophila melanogaster . Nested within the gene elements, are sub-elements related to the parent. The first gene includes two nucleic acid sequences, a protein product, and a functional annotation. Additional information is provided by attributes, such as the organism. This example illustrates the difficulty of modeling many-to-many relationships, such as the relationship between genes and functions. Information about functions must be repeated under each gene with that function. If we invert the nesting (i.e., nesting genes inside function elements), then we must repeat information about genes with more than a single function. At the bottom in Figure 39 we see the same information about genes, but using RDF and OWL. Both genes are instances of the class Fly Gene, which has been defined as the set of all Genes for the organism D. melanogaster. The functional information is represented using a hierarchical taxonomy, in which Long-Term Memory is a subclass of Memory (Louie et al., 2007).



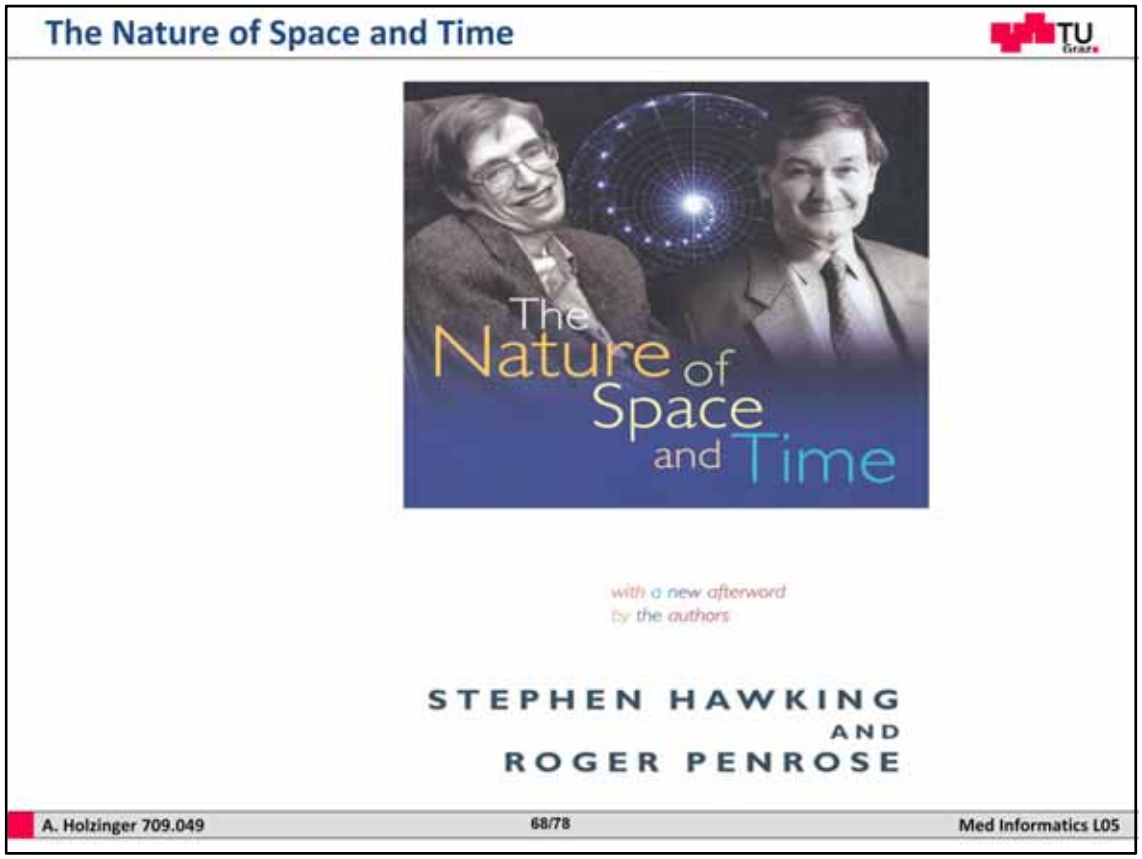

This is star cluster structure M30 Let us look into the macroscopic area first and let us look for some similarities …
This is star globular star cluster M30 (NGC 7099), including some 100.000 stars a diameter of about 100 light-years, approx. 40,000 light-years away from earth – look at the structure – look at the similarity – and consider the time, if our eyes see this structure they might be vanished (Darwin Channel)
Macroscopic structure




From this large macroscopic structures to tiny microscopic structure
Here a X-ray crystallography, which is a standard method to analyse the arrangement of objects (atoms, molecules) within a crystal structure. This data contains the mean positions of the entities within the substance, their chemical relationship, and various others … and the data is stored, for example – if having a protein structure – in a Protein Data Base (PDB). This database contains vast amounts of data. If a medical professional looks at the data, he or she sees only lengthy tables of numbers …

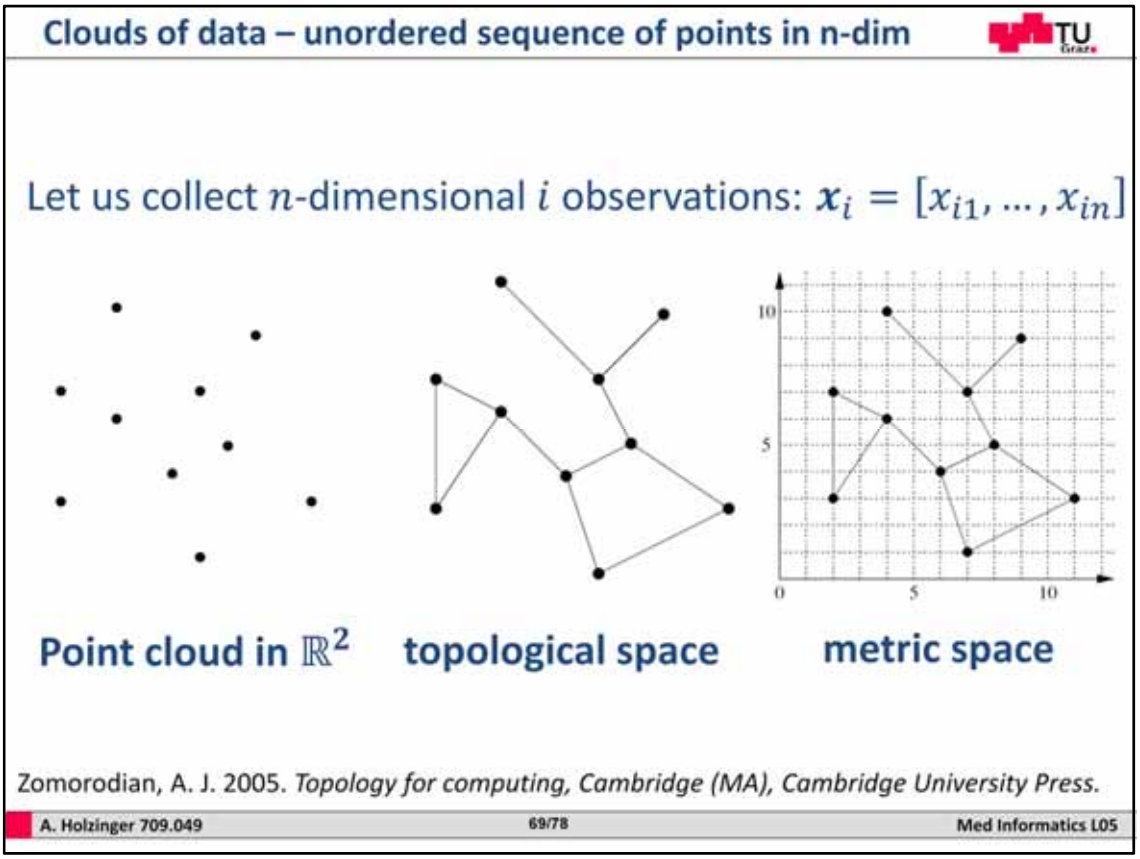



Structures! This is now our keyword. When we talk about structures, we will see some really interesting aspects of structures. A good example for a data intensive and highly complex microscopic structure is a yeast protein network.
Note: Yeasts (Hefe) are eukaryotic micro-organisms (fungi) with 1,500 known species currently, estimated to be only 1% of all yeast species. Yeasts are unicellular, typically measuring 4 µm in diameter.
In this picture you can see the first protein interaction network (published by Jeong et. al, 2001). The nodes are the proteins. The links are the physical interactions (bindings). The red nodes are lethal to the organism, the green ones are non-lethal and the yellow ones are not yet known (still unknown).
You may ask whether this structure is useful? Well, what we get out by this yeast is something which some of us may really like: Prost!
The problem with such structures is that they are very big and that there are so many! Knowledge Management can help to discover such unknown structures amongst the enormous set of uncharacterized data. We will come back to such structural homologism later. Now let us make a closer look on what Knowledge Management can do for us.




When thinking about data, we should always keep two fundamental physical aspects in mind: time related aspects (e.g. entropy of data) and space related aspects (e.g. topology of data).

http://www.youtube.com/watch?v=oBkOYQ02chs
TedxWarwick 2010 Roger Penrose in Space-Time Geometry.
http://www.youtube.com/watch?v=aSz5BjExs9o
Visualizing Eleven Dimensions

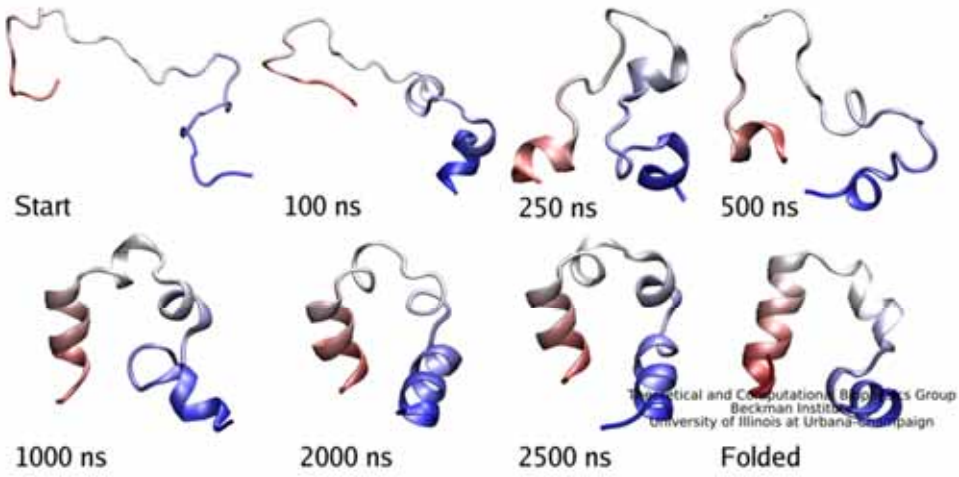



Clouds of data. Very often, data is represented as an unordered sequence of points in a Euclidean n-dimensional space En. Data coming from an array of sensor readings in an engineering testbed, from questionnaire responses in a psychology experiment, or from population sizes in a complex ecosystem all reside in a space of potentially high dimension. The global 'shape' of the data may often provide important information about the underlying phenomena which the data represents.

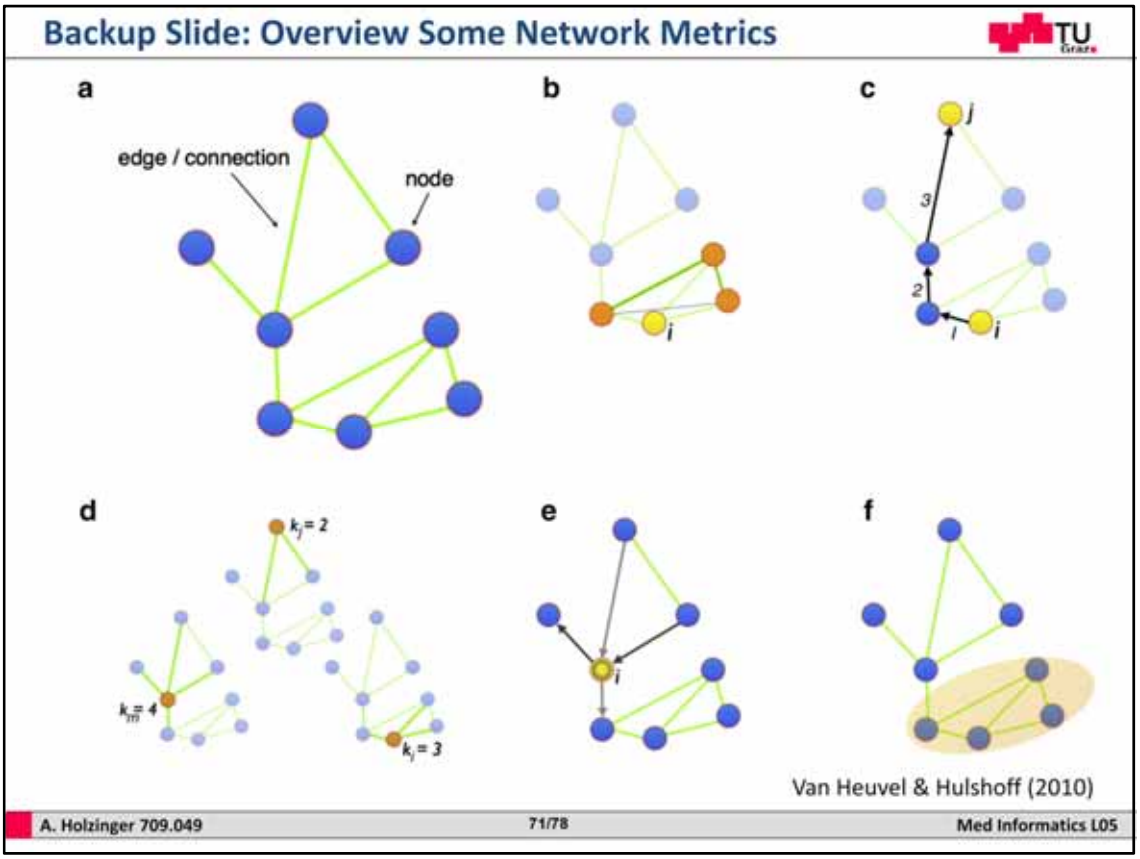
One type of data set for which global features are present and significant is the so-called point cloud data coming from physical objects in 3-d. Touch probes, point lasers, or line lasers sweep a suspended body and sample the surface, record-

ing coordinates of anchor points on the surface of the body. The cloud of such points can be quickly obtained and used in a computer representation of the object. A temporal version of this situation is to be found in motion-capture data, where geometric points are recorded as time series. In both of these settings, it is important to identify and recognize global features: where is the index finger, the keyhole, the fracture?

## Example: To predict the folding of a protein



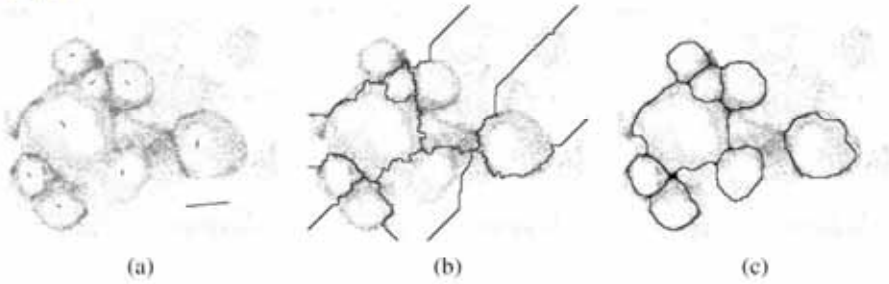

Source: Theoretical and computational Biophysics Group: http://www.ks.uiuc.edu/




a = order
b = clustering coefficient
c = path length
d = centrality
e = nodal degree
F = modularity
Network metrics

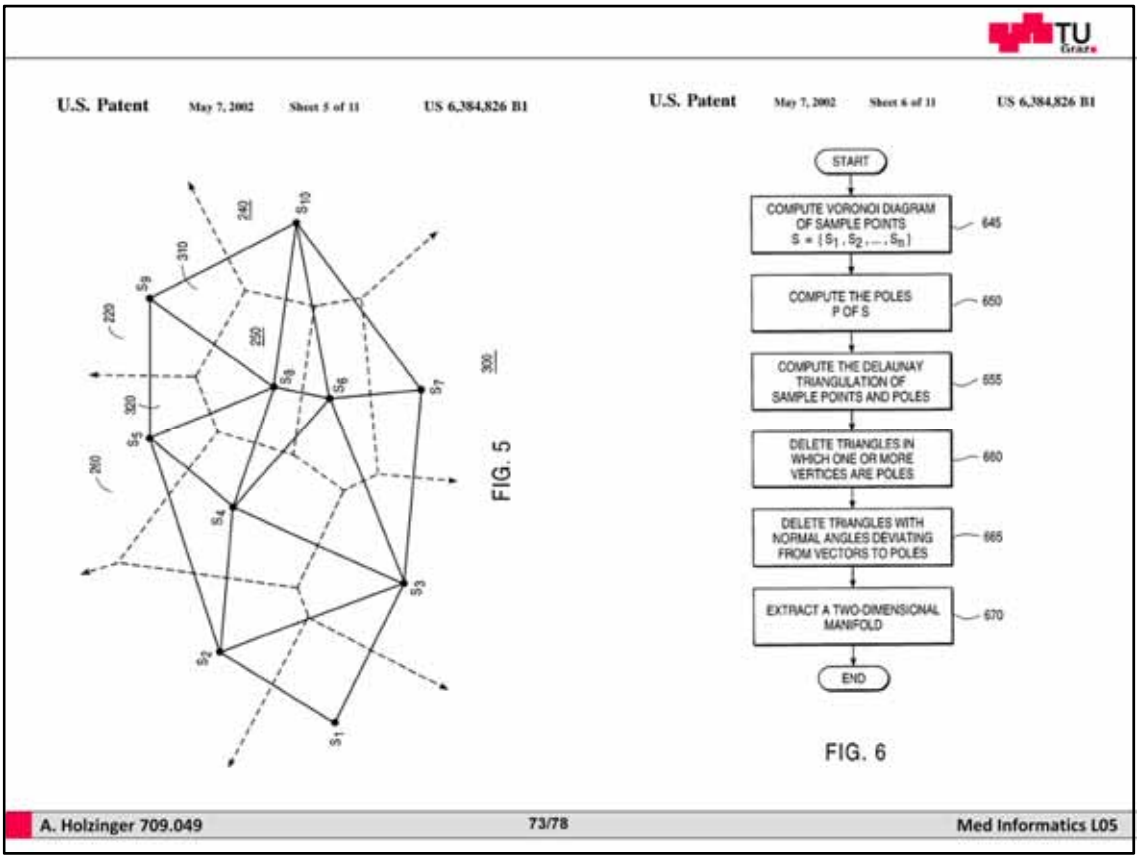
## Slide 5-19: Watershed Principle

- Catchment basins:
  - treating an image as a height field or landscape, regions where the rain would flow into the same lake



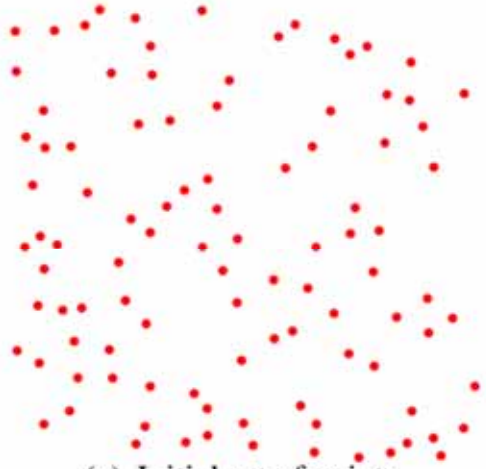
(a) (b) (c)

- Start flooding from local minima, and label ridges wherever differently evolving components meet

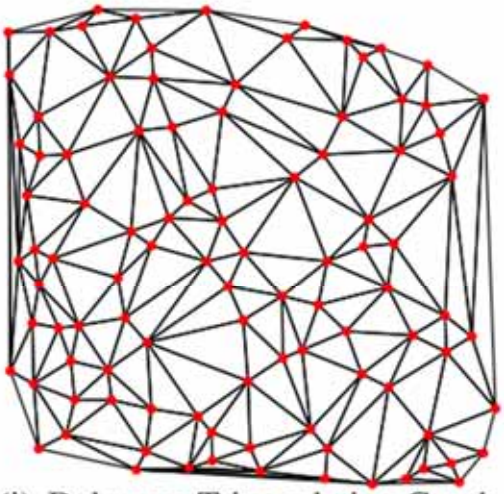



http://www.google.com/patents/US6384826

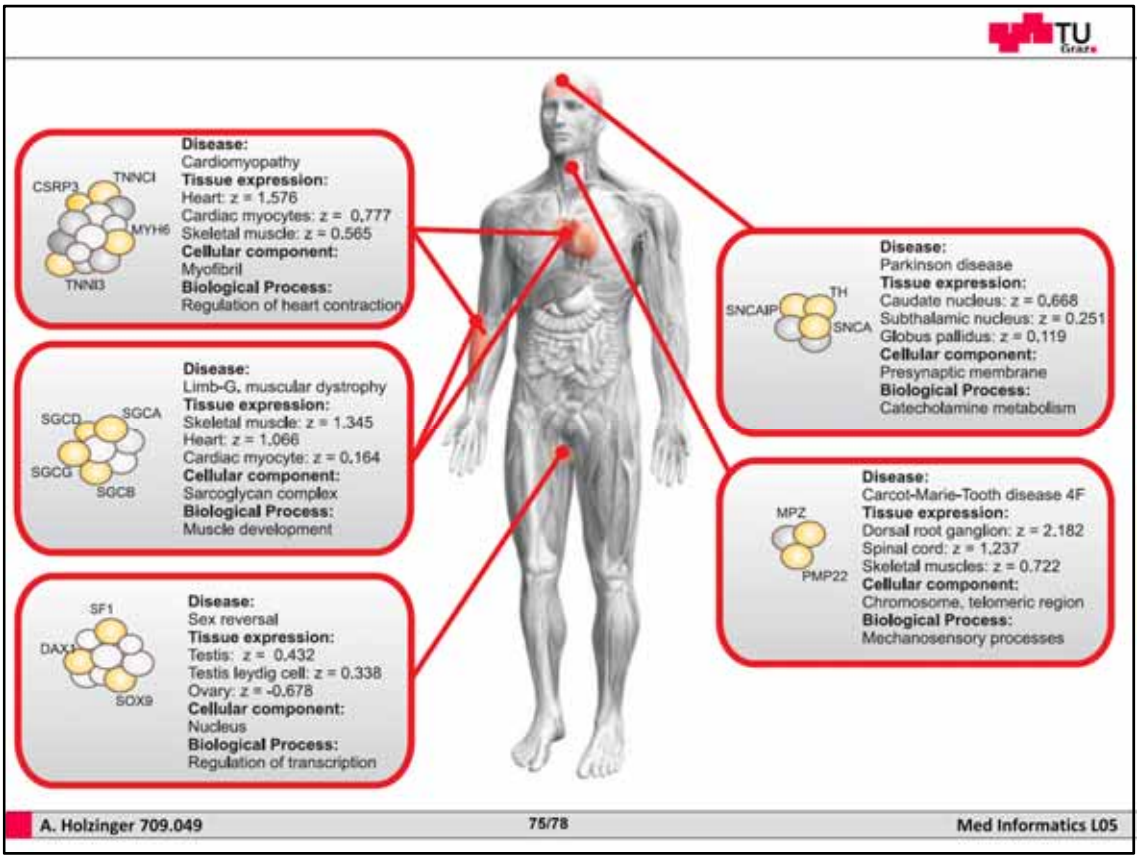
## Slide 5-15 Graphs from Images: Voronoi <> Delauney



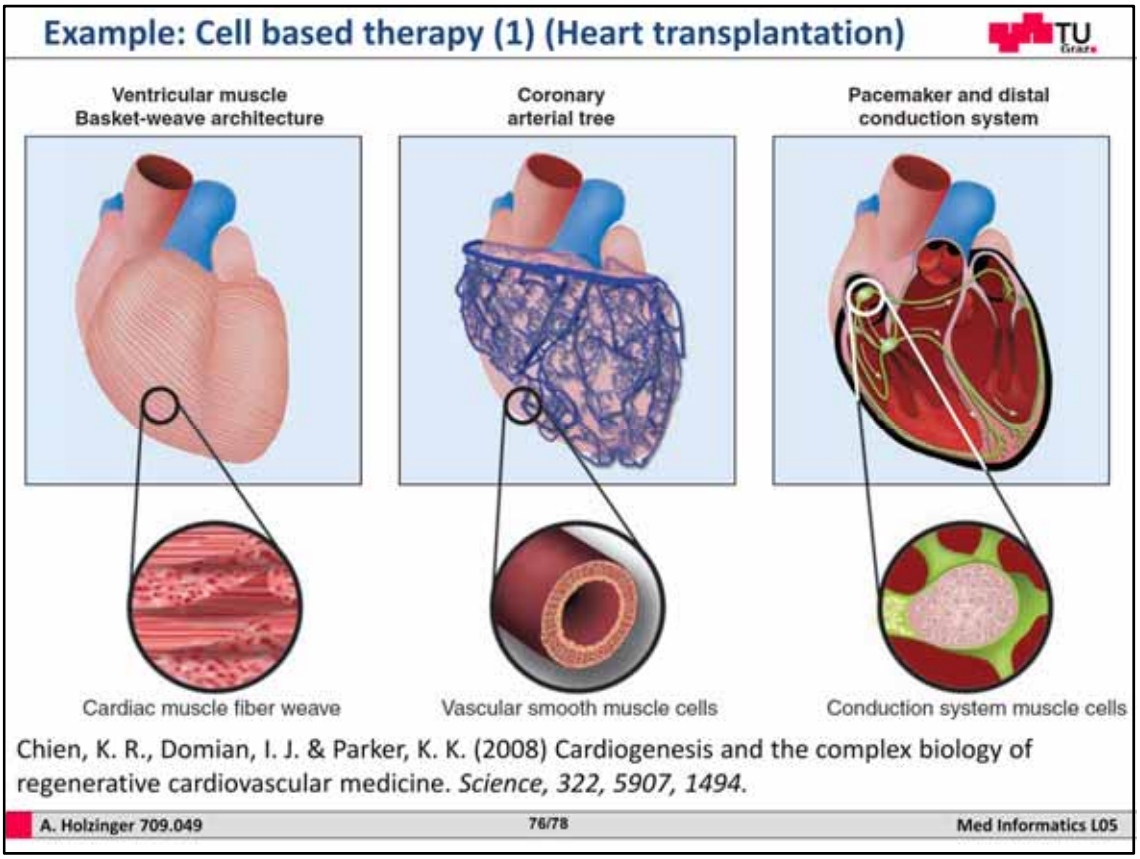
(a) Initial set of points.



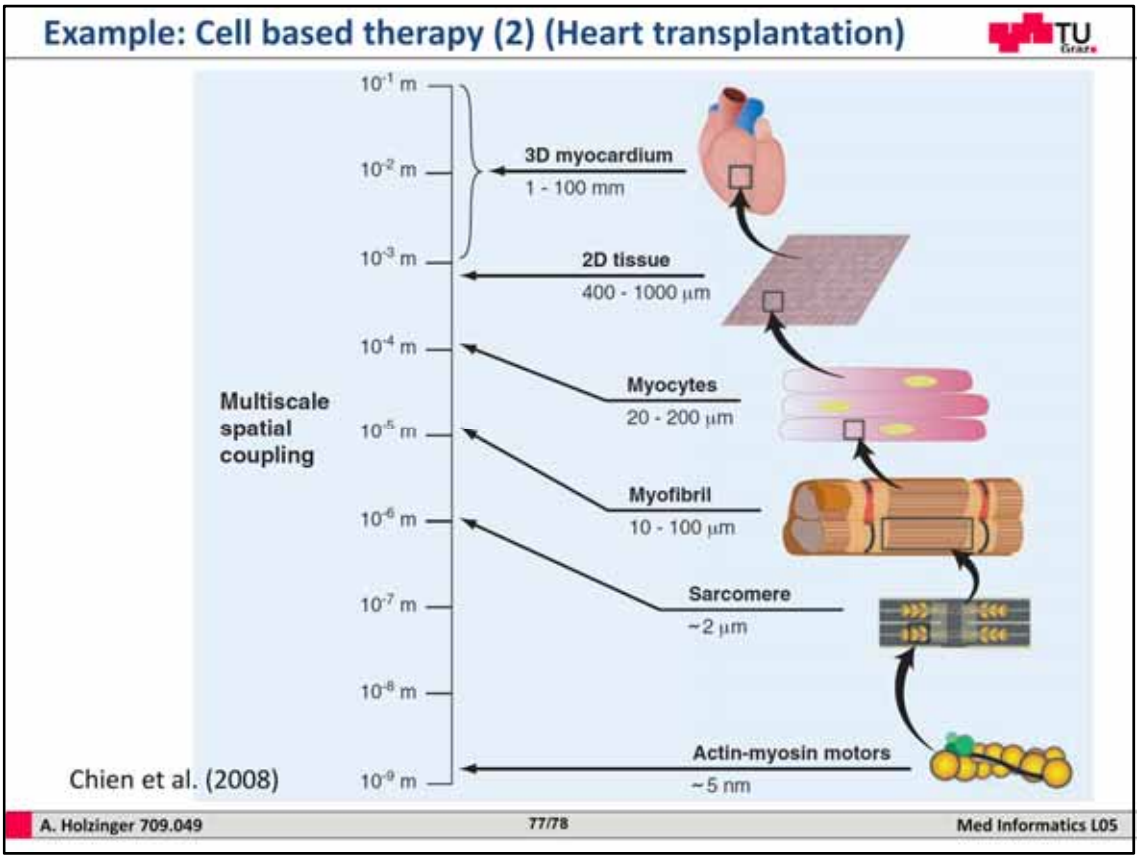
(i) Delaunay Triangulation Graph.




Representative examples of disease complexes are displayed. Diseases are associated with tissues by using our disease–tissue matrix, and expression data are from
the GNF dataset. The expression levels of complexes are shown as z scores. If a disease is associated with more than 3 tissues, only the 3 most associated tissues are shown for
clarity. In a given complex, proteins relevant to the disease in question are yellow. The figure shows the general tendency of overexpression of the complexes in the tissues
in which they are involved in pathology compared with their expression level in other tissues. All members of the complexes can be seen in

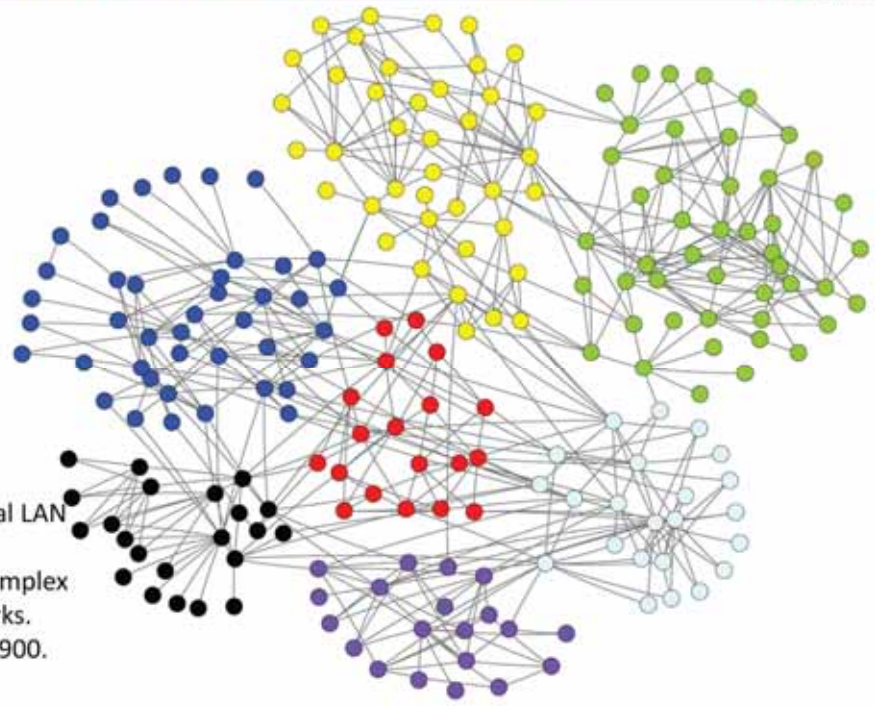



Three-dimensional structure of ventricular muscle basket weave, coronary arterial tree, and pacemaker
and conduction system. One of the central challenges of cell-based therapy for regenerating specific heart
components is guiding transplanted cells into a functional syncytium with the existing three-dimensional
architecture. Transplanted cells must make functional connections with neighboring specialized heart cells to
result in a net gain of global function. Transplanted myogenic progenitors, for example, must align with and
integrate into the existing ventricular muscle basket weave to allow synchronous contraction and relaxation of
graft and host myocardium. Integration of pacemaker and conduction system progenitors into the appropriate
tissue type is necessary to generate a biological pacemaker and avoid cardiac arrhythmia. For example, having a
transplanted heart muscle progenitor integrate into the conduction system might have arrythmogenic consequences,
as would the introduction of cells with independent pacemaker potential in the heart. Similarly, cell-based
therapies to promote coronary collateral formation or neo-arteriogenesis require functional integration of transplanted
cells with the host coronary arterial tree.




Three-dimensional structure of ventricular muscle basket weave, coronary arterial tree, and pacemaker
and conduction system. One of the central challenges of cell-based therapy for regenerating specific heart
components is guiding transplanted cells into a functional syncytium with the existing three-dimensional
architecture. Transplanted cells must make functional connections with neighboring specialized heart cells to
result in a net gain of global function. Transplanted myogenic progenitors, for example, must align with and
integrate into the existing ventricular muscle basket weave to allow synchronous contraction and relaxation of
graft and host myocardium. Integration of pacemaker and conduction system progenitors into the appropriate
tissue type is necessary to generate a biological pacemaker and avoid cardiac arrhythmia. For example, having a
transplanted heart muscle progenitor integrate into the conduction system might have arrythmogenic consequences,
as would the introduction of cells with independent pacemaker potential in the heart. Similarly, cell-based
therapies to promote coronary collateral formation or neo-arteriogenesis require functional integration of transplanted
cells with the host coronary arterial tree.

## Example: Network Generated by Gene Duplication



High Modularity (Modularity = 0.6717, Scaled Modularity = 29); Different colors represent different modules identified by Guimera and Amaral's algorithm [28].

Guimera R, Amaral LAN (2005) Functional cartography of complex metabolic networks. Nature 433: 895–900.

Wang & Zhang (2007)

A. Holzinger 709.049 78/78 Med Informatics L05